RIKEN CAGE Symposium
September 13, 2016

Transcription start site centered transcriptome profiling strategy

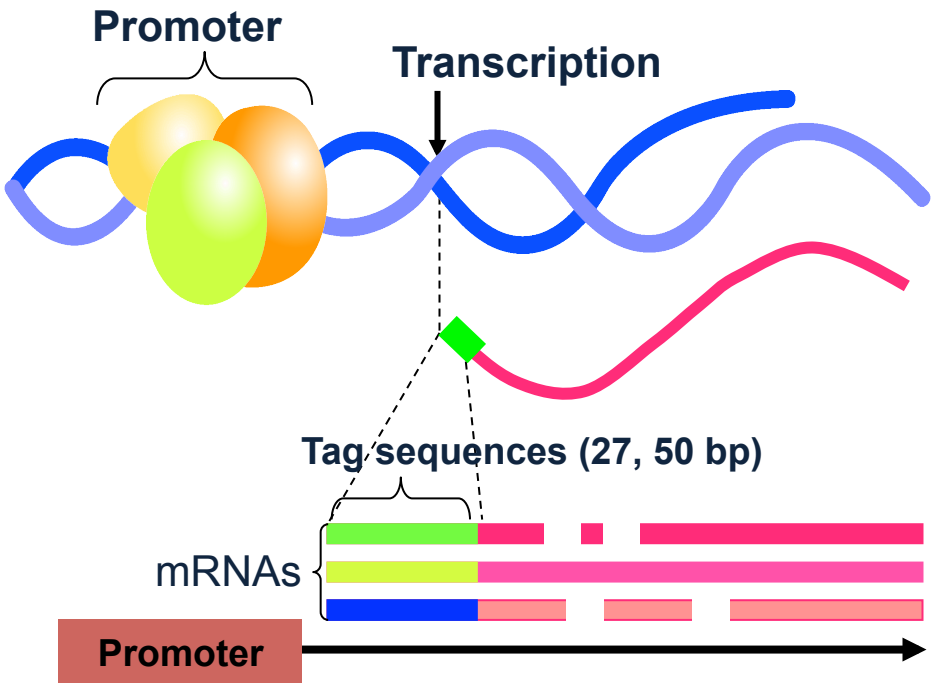
Piero Carninci

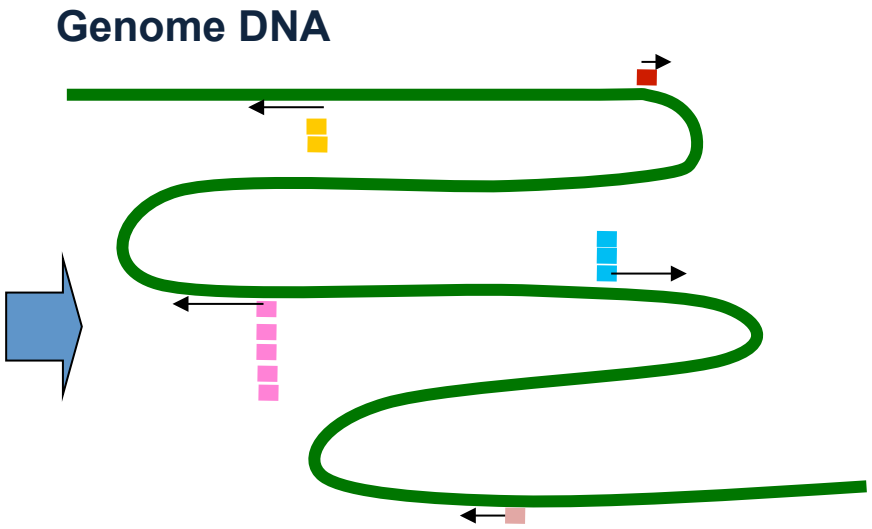
RIKEN Center for Life Science Technologies (CLST), Deputy director
Division of Genomic Technologies (DGT), Director

CAGE (Cap Analysis of Gene Expression)

RIKEN original technology to reveal “GENE REGULATION”

CAGE analyzes 5’-end of the capped transcripts by DNA sequencing.
1) Precise transcriptional starting sites (TSSs) are clarified.
2) Expression profile at each promoter (not gene) could be analyzed.





CAGE (Cap Analysis of Gene Expression)

RNA extraction

CAGE library preparation

1. CAP trapper
2. Trehalose extension method
3. CAGE library

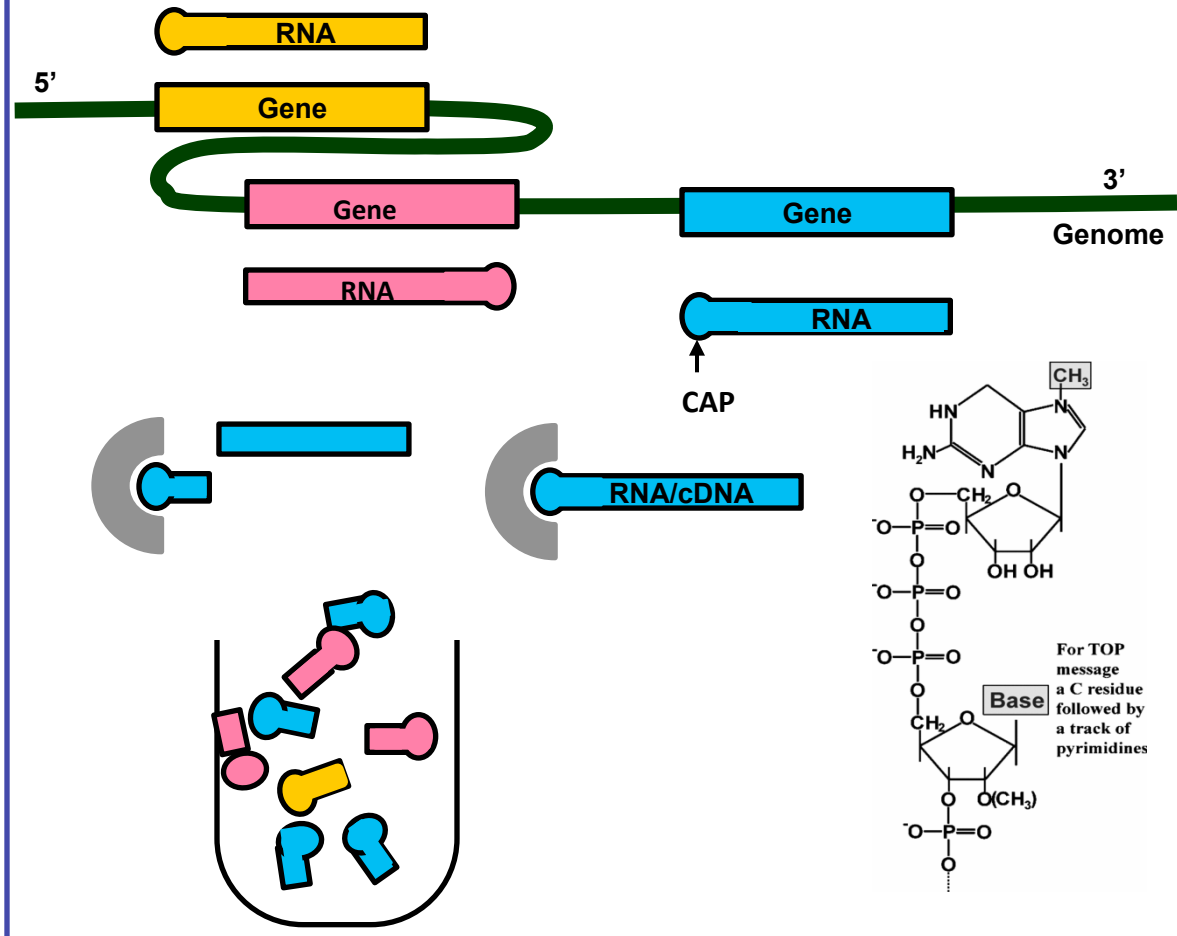
Sequence

Tag sequencing with the next-generation sequencer

1. Illumina HiSeq 2500
2. Helicos HeliScope

Data processing

Quality control
Statistical variation of the obtained sequence
Extraction of tag sequences
Clustering
Mapping
Statistical variation of the mapping result
Visualization with “genome browser”
Statistical analysis



Shiraki *et al*. *Proc Natl Acad Sci U S A* 100, 15776 (2003)
Kodzius *et al*. *Nat Methods* 3, 211 (2006)
Takahashi *et al*. *Nat Protoc* 7, 542 (2012)

CAGE (Cap Analysis of Gene Expression)

RNA extraction

CAGE library preparation

1. **CAP trapper**
2. **Trehalose extension method**
3. **CAGE library**

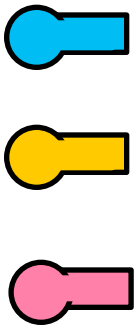
Sequence

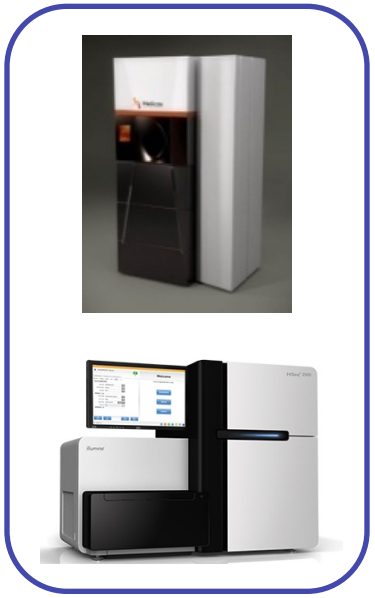
Tag sequencing with the next-generation sequencer

1. **Illumina HiSeq 2500**
2. **Helicos HeliScope**

Data processing

Quality control
Statistical variation of the obtained sequence
Extraction of tag sequences
Clustering
Mapping
Statistical variation of the mapping result
Visualization with “genome browser”
Statistical analysis

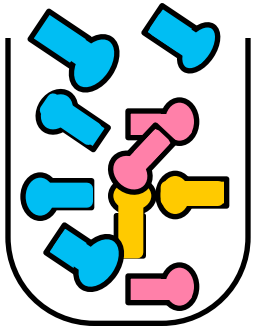




CGCATGGTCGATAGACTTG

GTGCGCGTCGAATATCGAT

CGAATATCGATAGACTTG



CAGE (Cap Analysis of Gene Expression)

RNA extraction

CAGE library preparation

1. **CAP trapper**
2. **Trehalose extension method**
3. **CAGE library**

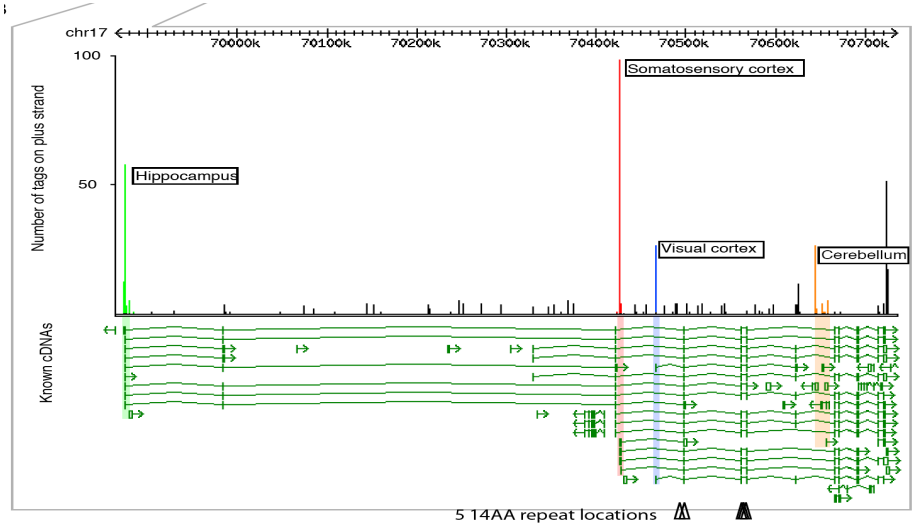
Sequence

Tag sequencing with the next-generation sequencer

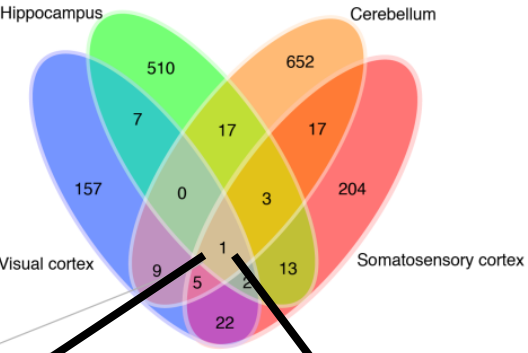
1. **Illumina HiSeq 2500**
2. **Helicos HeliScope**

Data processing

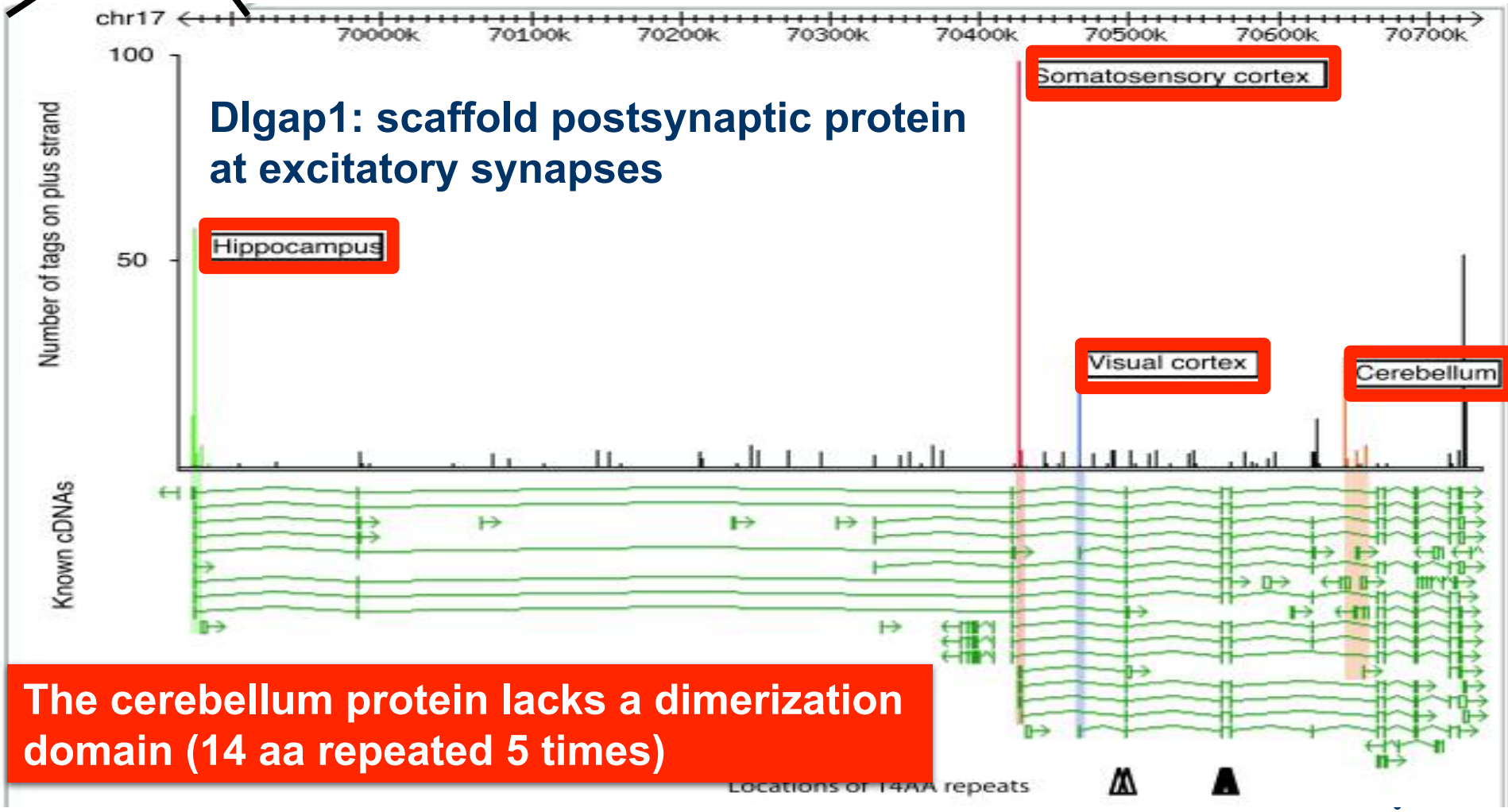
Quality control
Statistical variation of the obtained sequence
Extraction of tag sequences
Clustering
Mapping
Statistical variation of the mapping result
Visualization with “genome browser”
Statistical analysis



Preferentially expressed promoters (PEPs) drive functional variability of the proteome



Dlgap1: scaffold postsynaptic protein at excitatory synapses



The cerebellum protein lacks a dimerization domain (14 aa repeated 5 times)

Valen E. *et al. Genome Research* 19, 255-65 (2009)

Advantages of CAGE

- ✓ Expression profile at TSSs
  - Promoter discovery and activity
  - Enhancer discovery and activity
  - Transcription factor (TF) binding site motifs
- ✓ Quantitative
- ✓ High resolution
- ✓ Genome-wide analysis
- ✓ Sequencing base
- ✓ Wide dynamic range

Examples

Application of CAGE

TSSs identification and activity analysis

Enhancers
- Identification
- Activity
- Mapping
- Cell specific activity

Promoters
- Identification
- Activity
- Mapping
- Cell specific activity

lncRNAs
- Identification
- Cell specific expression
- Function

TF binding sites
- Motif activity
- Transcriptional regulatory NW
- Cell specific NW signatures

Transcription starting sites complexity ⇨ Gene regulation

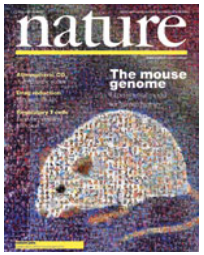
- ✓ Discovery of biomarkers
- ✓ Discovery of drug targets
- ✓ Test drug/cosmetics efficacy/toxicity at gene regulation level
- ✓ Regulatory networks reconstruction (cell conversion)
- ✓ Quality control of iPS cells

FANTOM activities

Functional Annotation of the Mammalian Genome

Annotation of 60,770 full-length mouse cDNAs

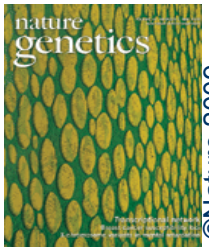
Okazaki *et al. Nature* 420, 563 (2002)
Standardize full-length mammalian cDNAs.



©Nature 2002

Transcriptional regulatory network

Suzuki *et al. Nat. Gen.* 41, 553 (2009)
The first report of detail transcriptional regulatory network in differentiation.



©Nature 2009

Phase1: The atlas of promoter & enhancer in diverse cell types

Forrest *et al, Nature* 507, 455 (2014), Andersson *et al*, *Nature* 507, 462 (2014)

©Nature 2014

FANTOM 1
Mouse full-length cDNA

FANTOM 2
cDNA Clone bank Libraries

FANTOM 3
Promoter Analysis Database

FANTOM 4
Basin Network analysis

FANTOM 5
Cellular diversity

FANTOM 6
lncRNAs Function
Ongoing!

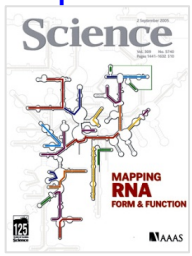
Developed functional gene annotation

Kawai *et al*. *Nature* 409, 685 (2001)
Collection of 20,000 fl-cDNA.
First annotation of transcriptome.

©Nature 2001

Discovery of new transcriptional landscape.

Carninci *et al. Science* 309, 1559 (2005)
- >70% of the genome is transcribed.
- >50% of transcripts are ncRNA.



©AAAS 2005

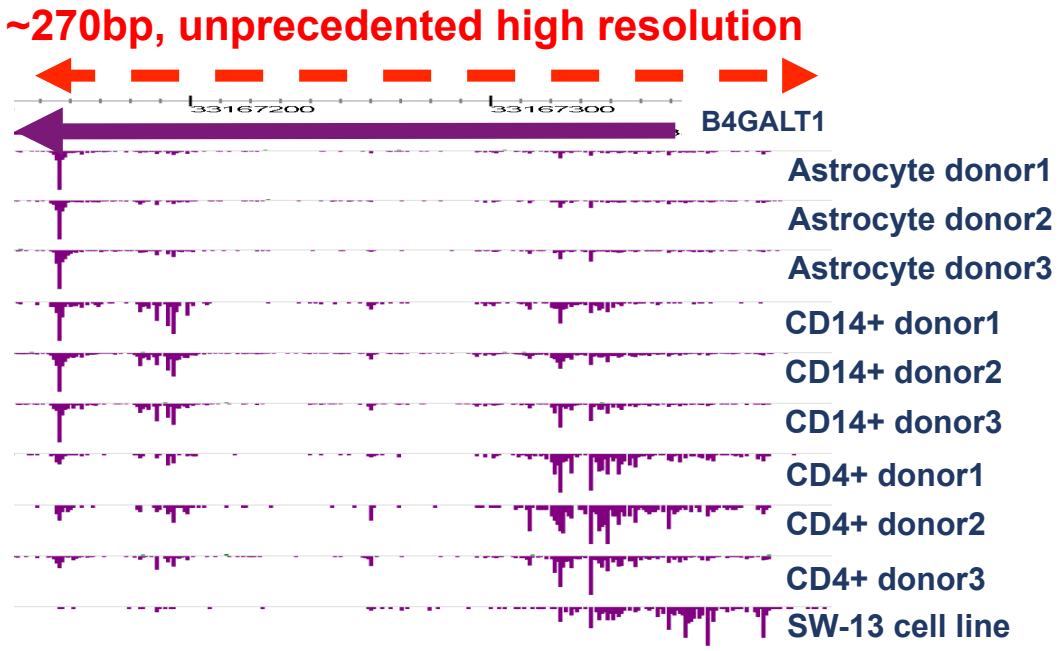
Phase 2: Enhancers dynamics

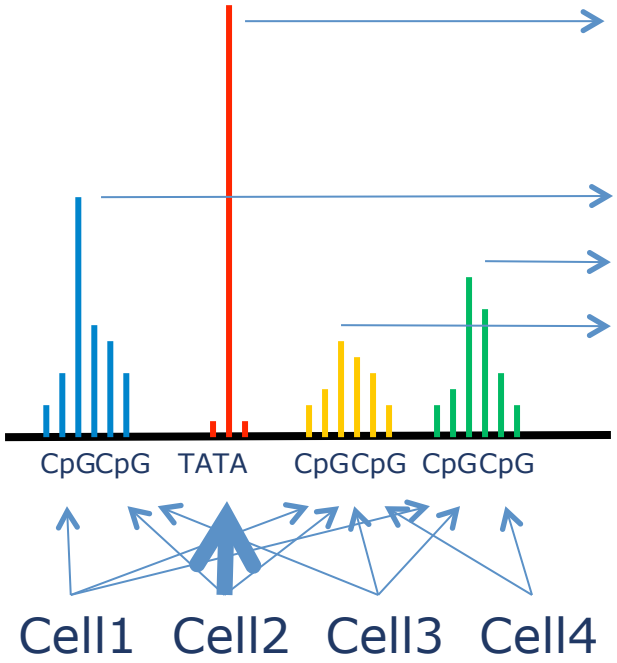
Arner *et al, Science* 347, 1010 (2015)
Enhancer broadly initiates and coordinates transcription.

©AAAS 2015

Higher tissue and tag coverage:

understanding composite promoter architectures and its mixed modes of regulation





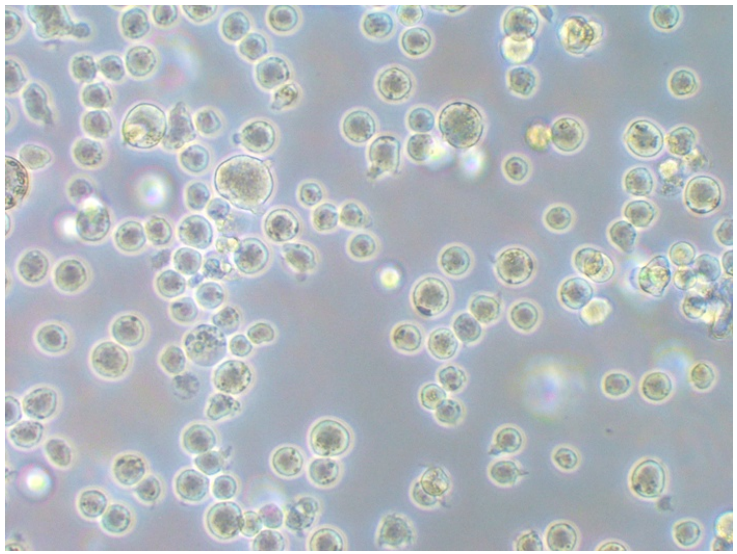
TSS preferences:

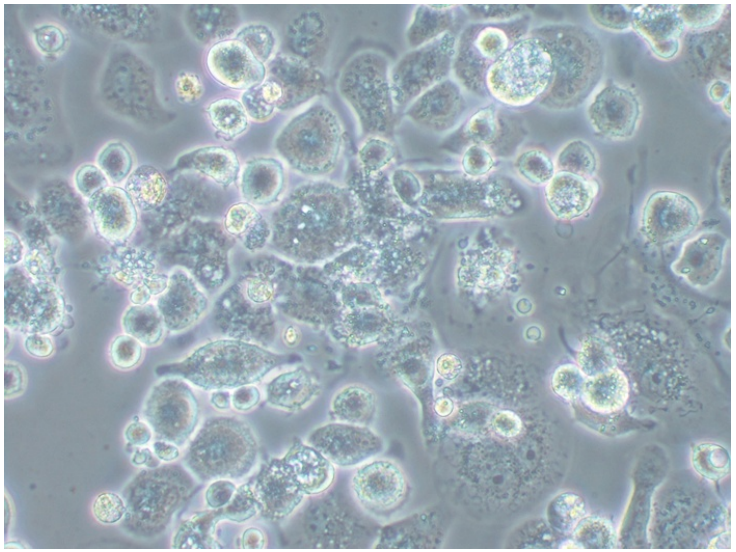
- **B4GALT1 core promoter**
- **Primary Astrocytes**
- **CD14+ monocytes**
- **CD4+ T-cells**

- **Blue, yellow and green: Broadly used**
- **Red: Cell2 specific and highly expressed**

223,428 in human and 162,264 in mouse of reference TSS

FANTOM4 samples analysis stream



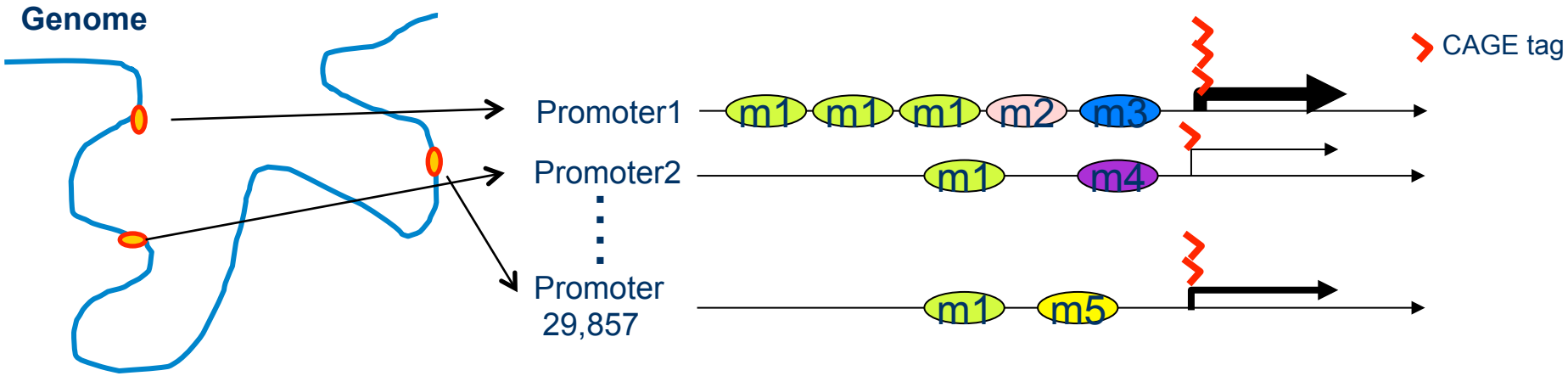


THP-1 cells are a monoblastic leukemia cell line which upon PMA treatment can differentiate into an adherent monocyte like cell ($CD14^+$, $CSF1R^+$)

Suzuki *et al*. *Nature Genetics* 41, 553 (2009)

Motif activity concept

29,857 promoters were identified. Using TFBS (motif) information and linear expression model, we calculated each motif activity.



Number of CAGE tags that mapped on the same site

$$e_{ps} = \sum_m R_{pm} A_{ms}$$

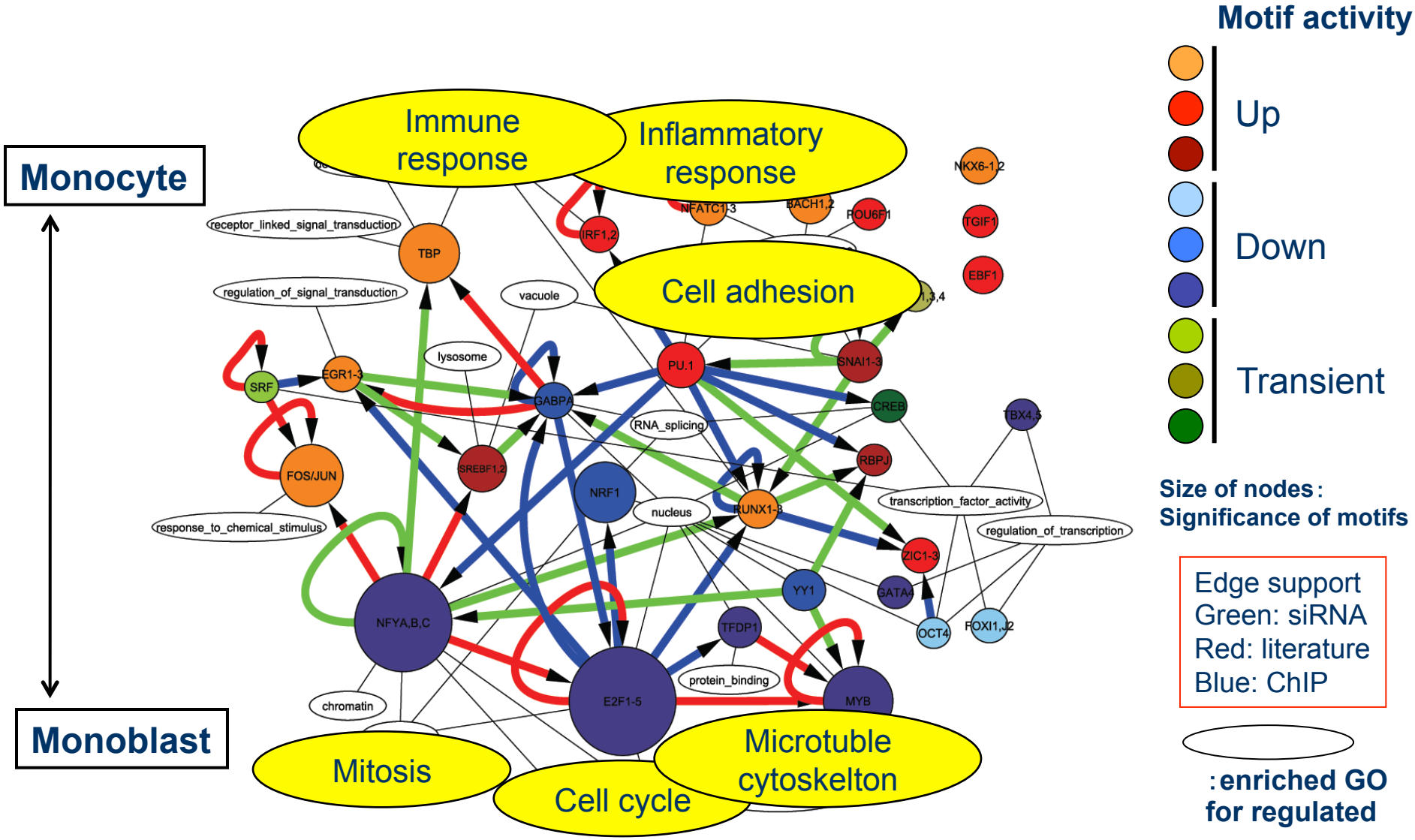
Reaction efficiency

- Number of possible binding sites
- Degree of conservation of the motif
- Chromatin status

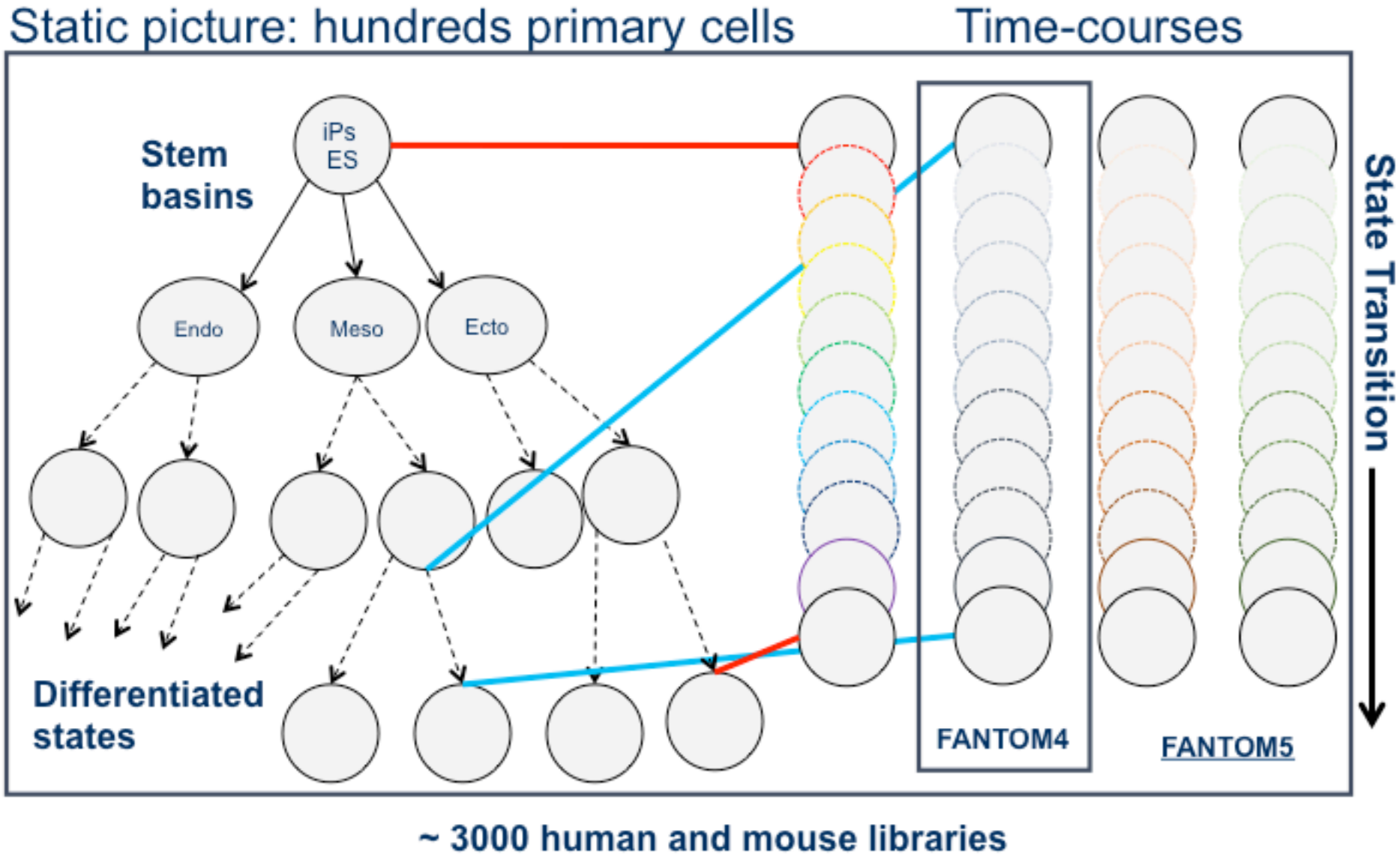
Activity of motif m

Transcriptional regulatory network consisting of 30 core motifs

55 out of 86 edges were supported by experiments/in the literature.

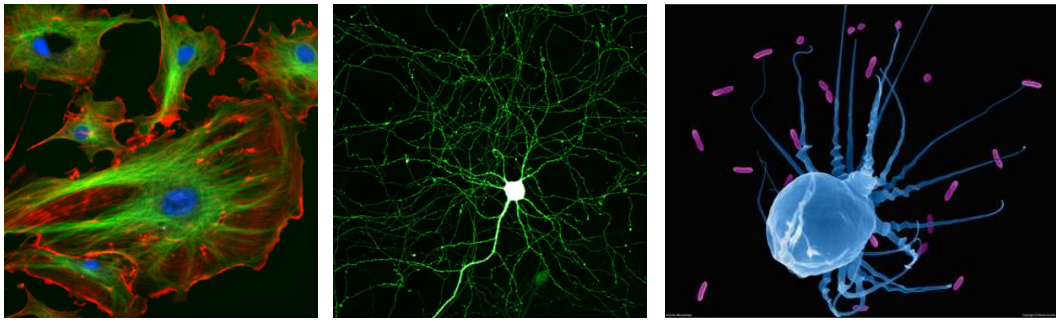


FANTOM5



Forrest AR *et al., Nature* 507, 462-70 (2014)

1000 human samples types (500 mice) + time courses perturbations



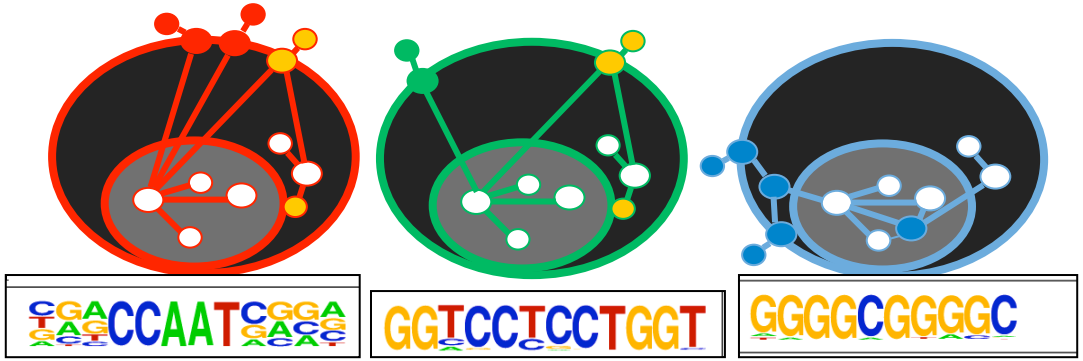
Primary cell compendium

Integrated transcript sequencing

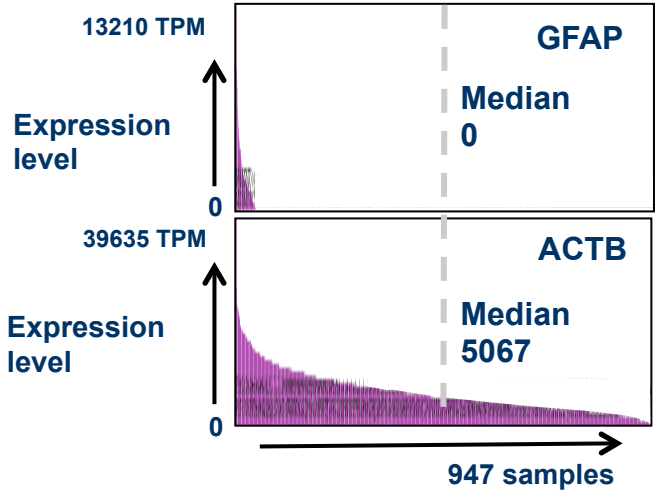
- CAGE promoter map
- RNA-seq transcript map
- Short RNA processing map

Transcript discovery

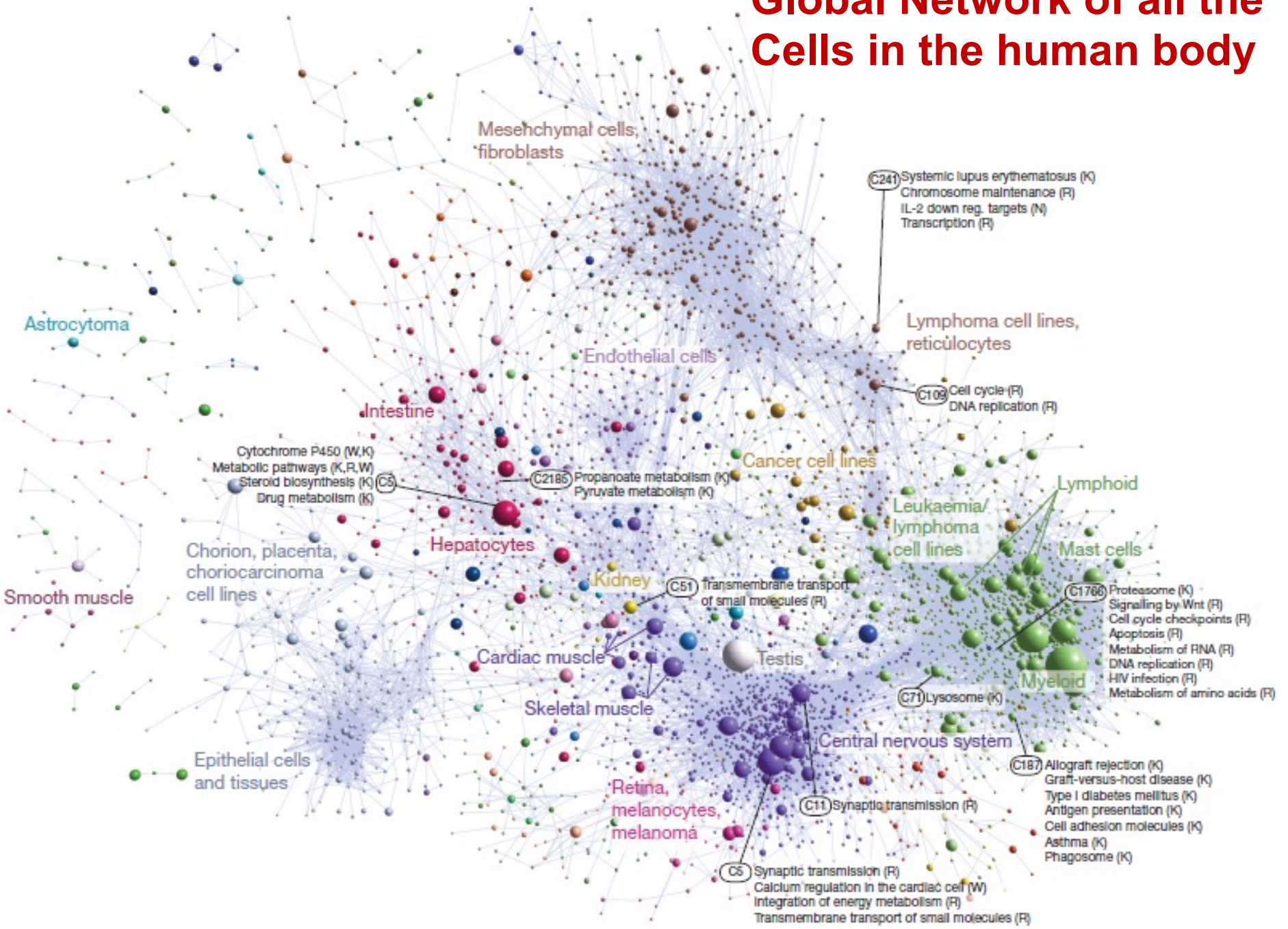
- Better gene models
- New lncRNAs
- New insights on processing
- Promoter-centered expression map



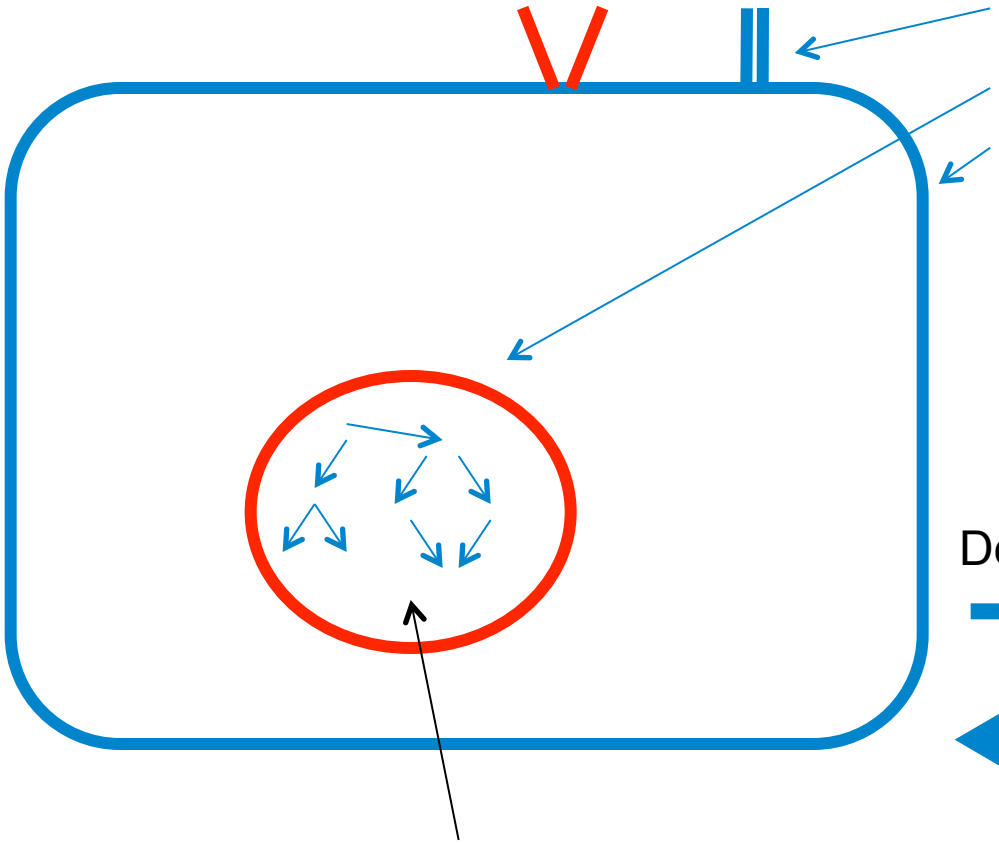
Cell specific network models
Key transcription factors, Key motifs



Global Network of all the Cells in the human body



How are cells characterized?



- Surface markers
- Morphology (shape, volume, polarity)
- Single or multinucleated, enucleated
- Ploidy
- Motility (adherent, resident, migratory)
- Differentiation potential
- Self renewal potential
- Developmental/lineage history
- Tissue of origin
- Developmental age (doublings?)
- Doubling time

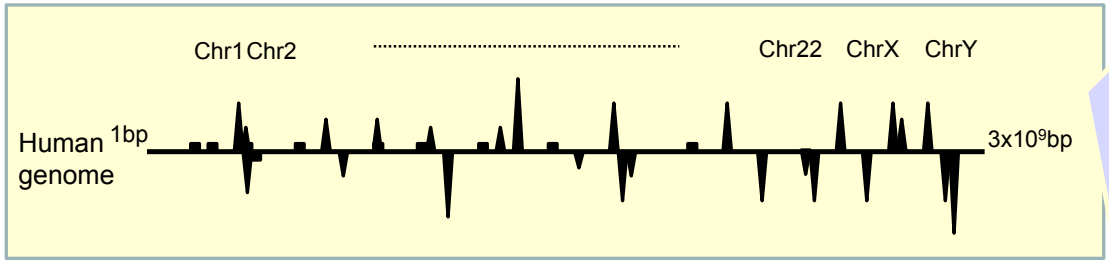
Defined outputs (eg growth factors)

Response to inputs

Self reinforcing stable internal network

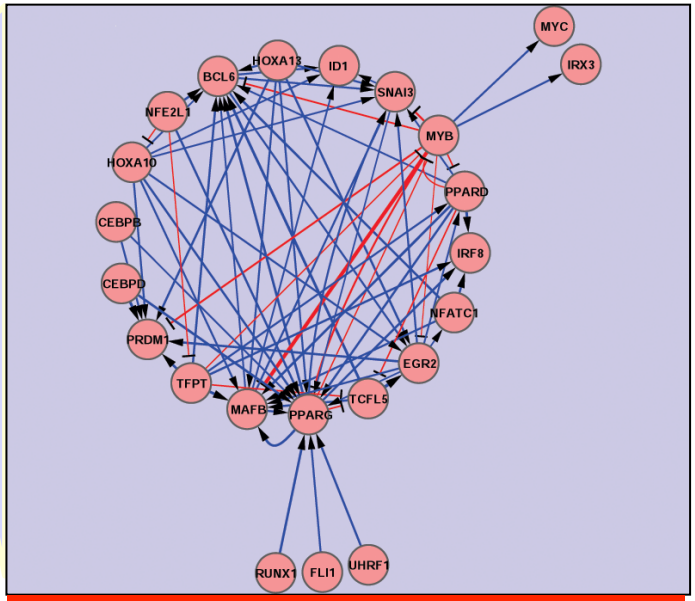
How are cells characterized?

Definition of cell	Objective?
Surface markers	Not sufficient
Morphology (shape, volume, polarity)	Definition is very difficult for non professionals
Ploidy, single or multinucleated, enucleated	Not informative
Motility (adherent, resident, migratory)	Not informative



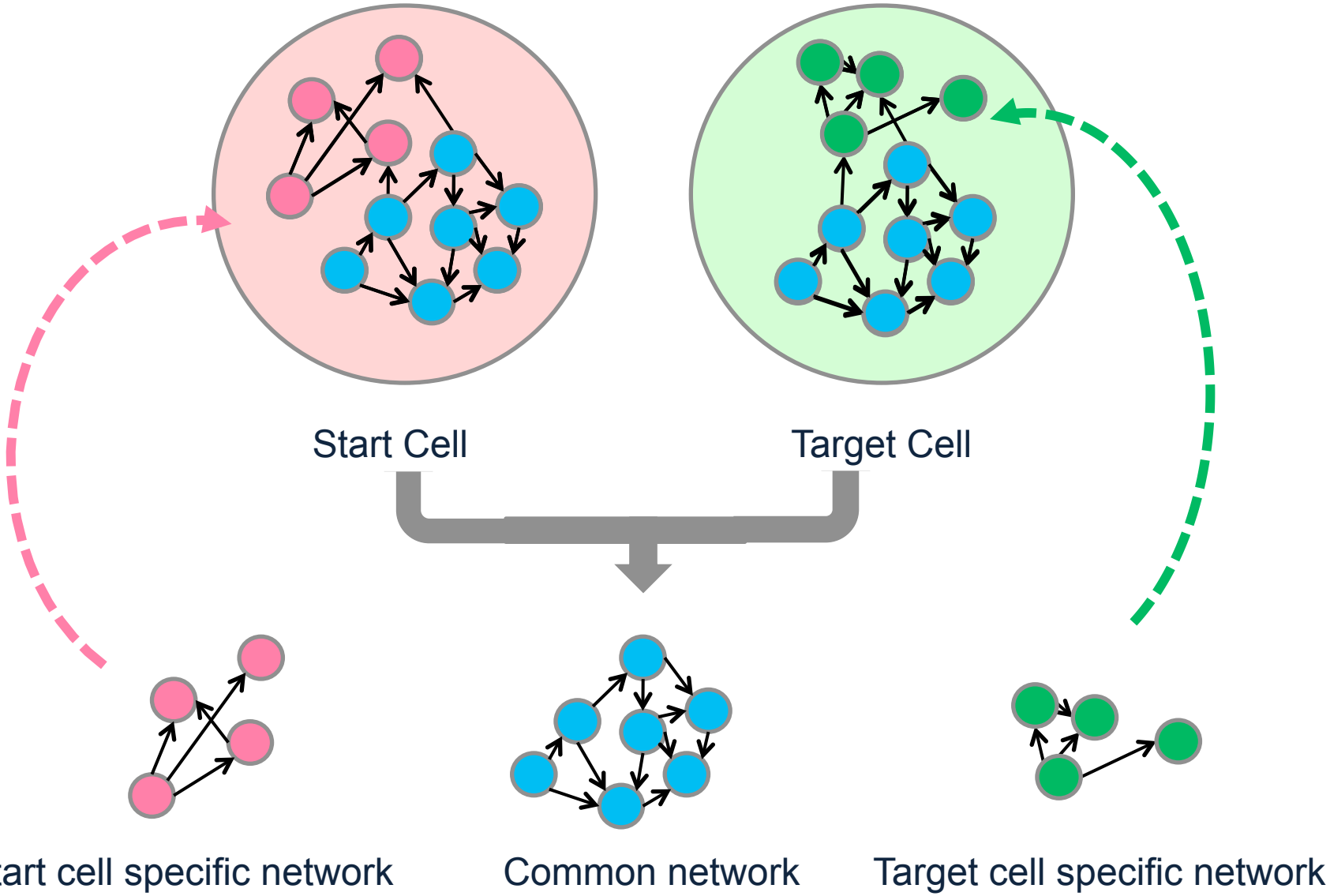
All promoters on human genome will be revealed.

The most objective definition of the cell !!



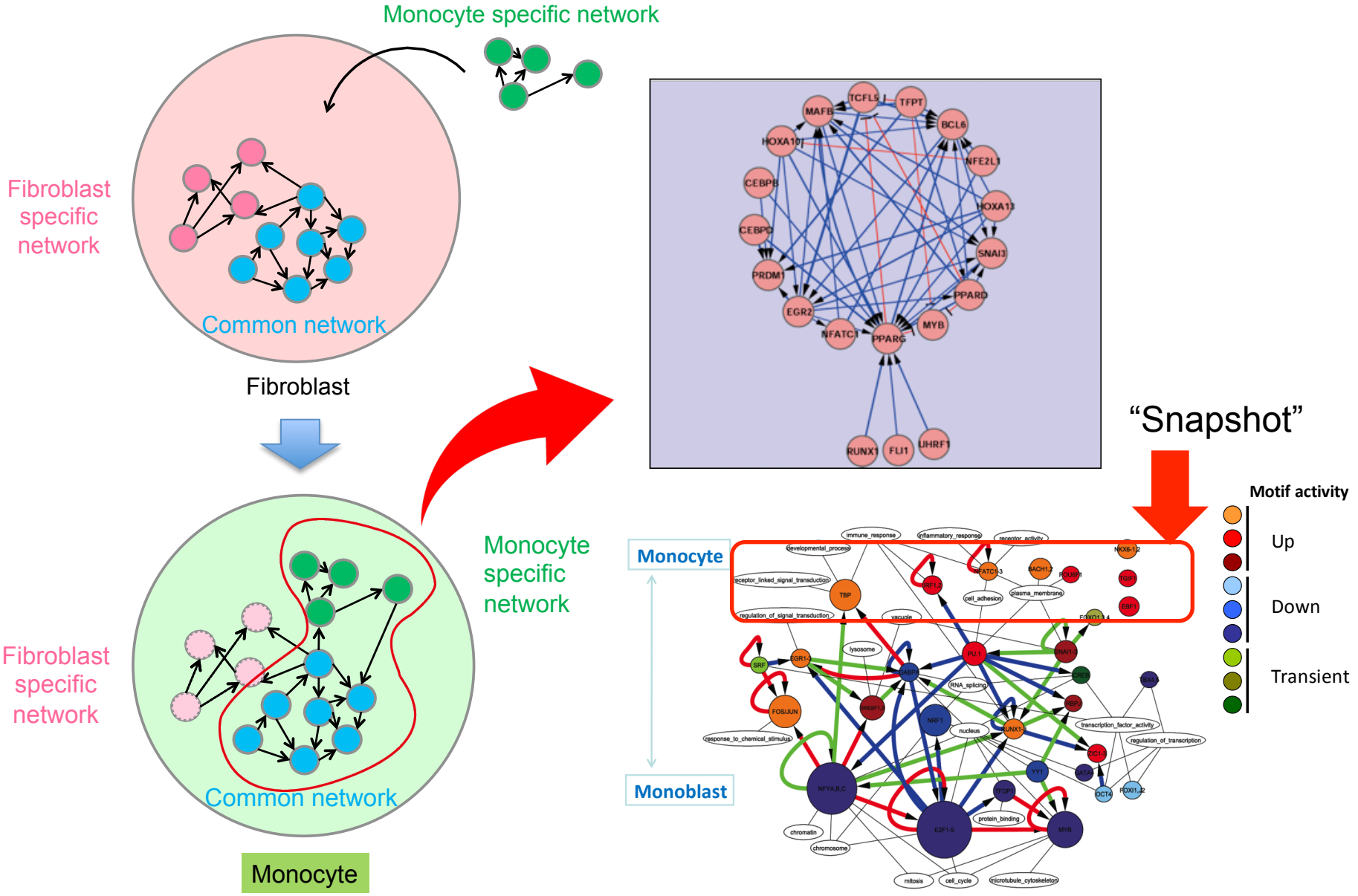
Transcriptional regulatory NW

The Application of Basin Network Analysis (Cell conversion)

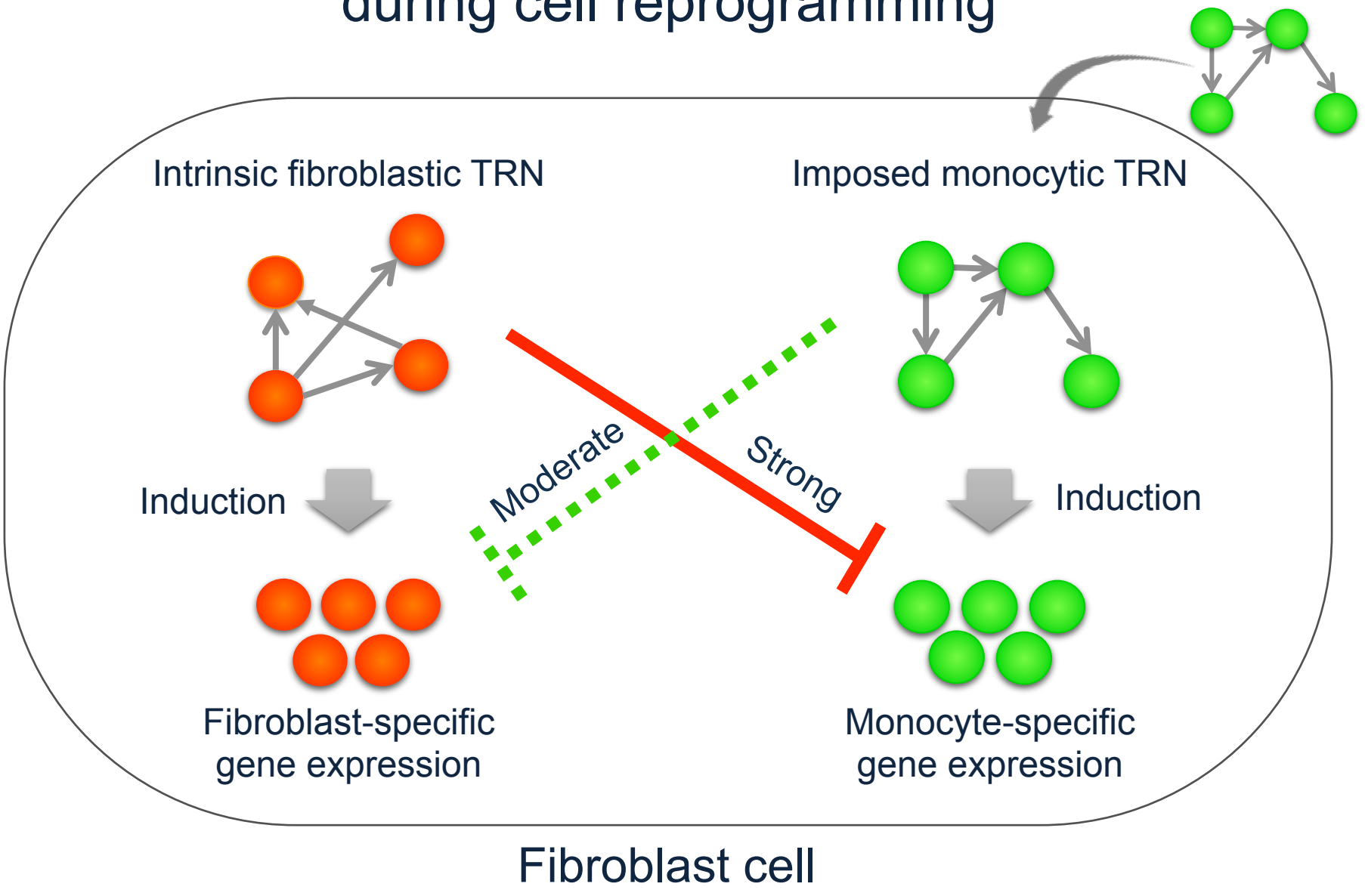


The Application of Basin Network Analysis (Cell conversion)

In the case of monocyte

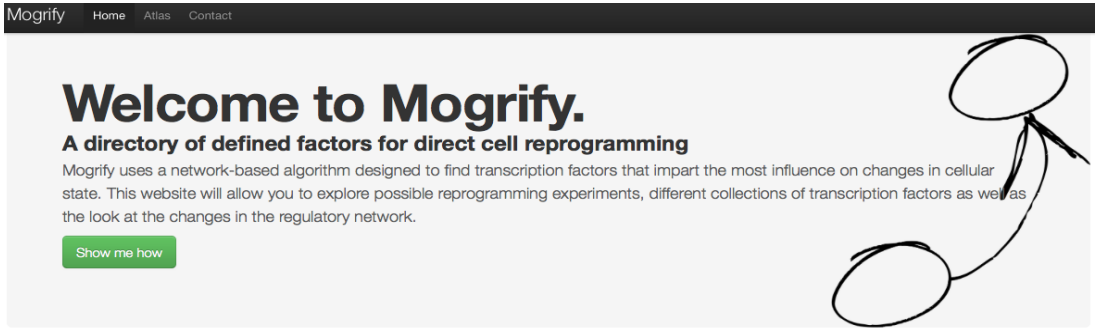


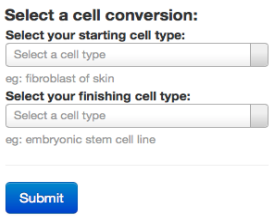
Asymmetric regulation by two networks during cell reprogramming

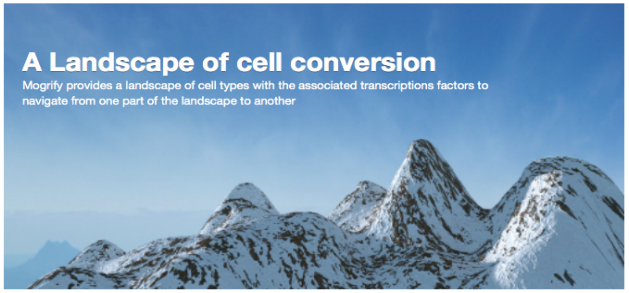


Li JR *et al. PLoS One* 11, e0160459 (2016)

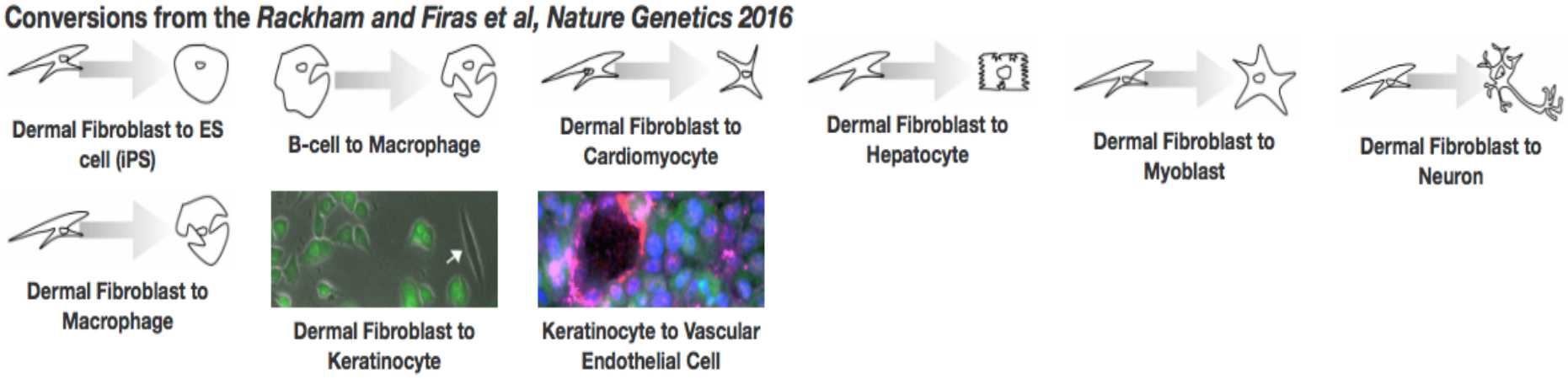
Mogrify: Predictive system to find TFs to induce direct cell reprogramming





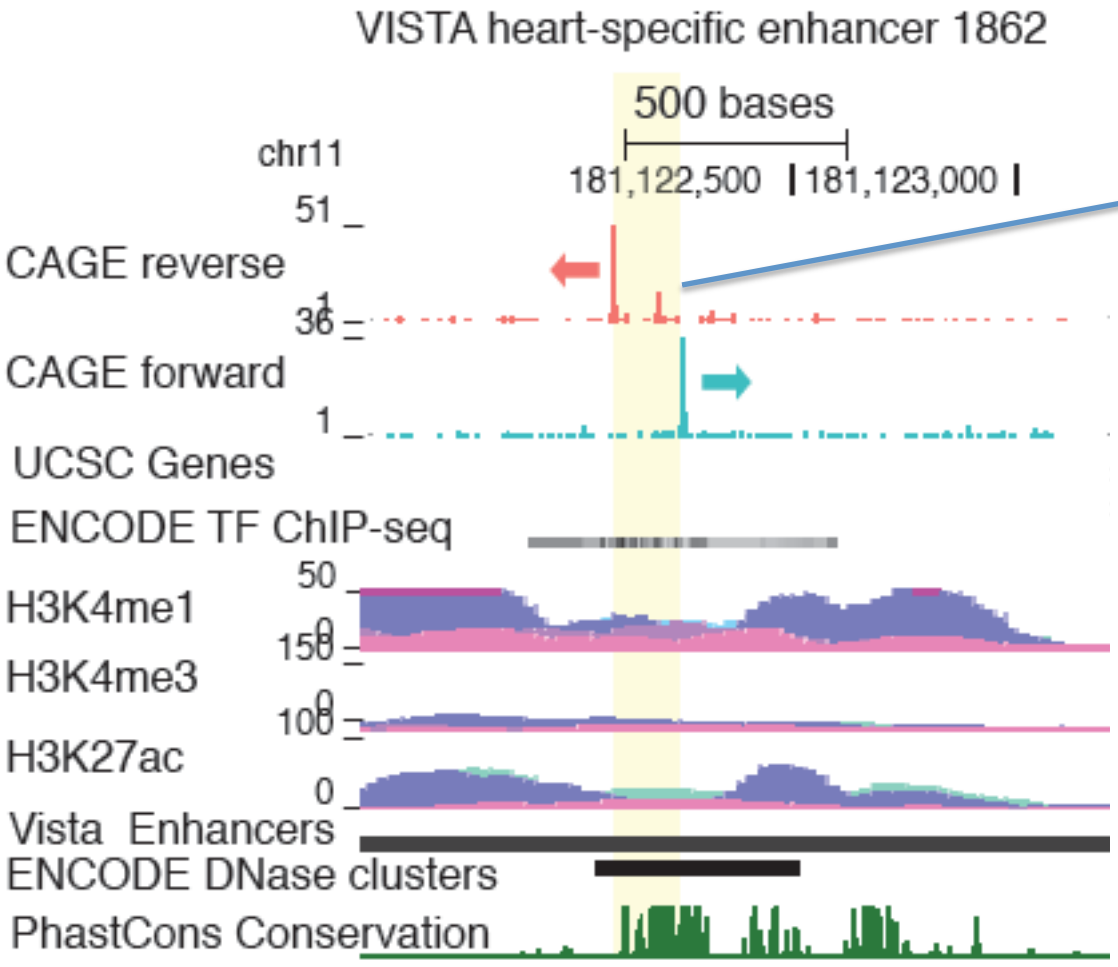


- Based on
  - Gene expression data
  - Regulatory NW info.
- Publicly available
  http://www.mogrify.net



Rackham OJ *et al. Nature Genetics* 48, 331 (2016)

CAGE locates known enhancers *in vivo*



Enhancers have bidirectional CAGE transcription. Bidirectional transcription identifies the nucleosome boundary.

Based on this, we make a rule to locate novel transcribed enhancers over the whole FANTOM collection.

Analysis led by Andersson and Sandelin, Univ. of Copenhagen

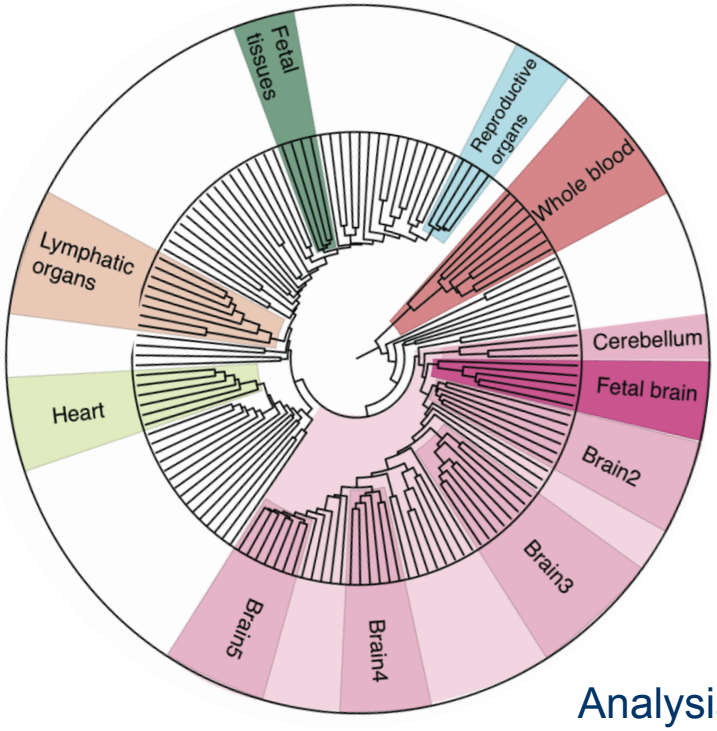
CAGE locates known enhancers *in vivo*

ARTICLE

doi:10.1038/nature12787

An atlas of active enhancers across human cell types and tissues

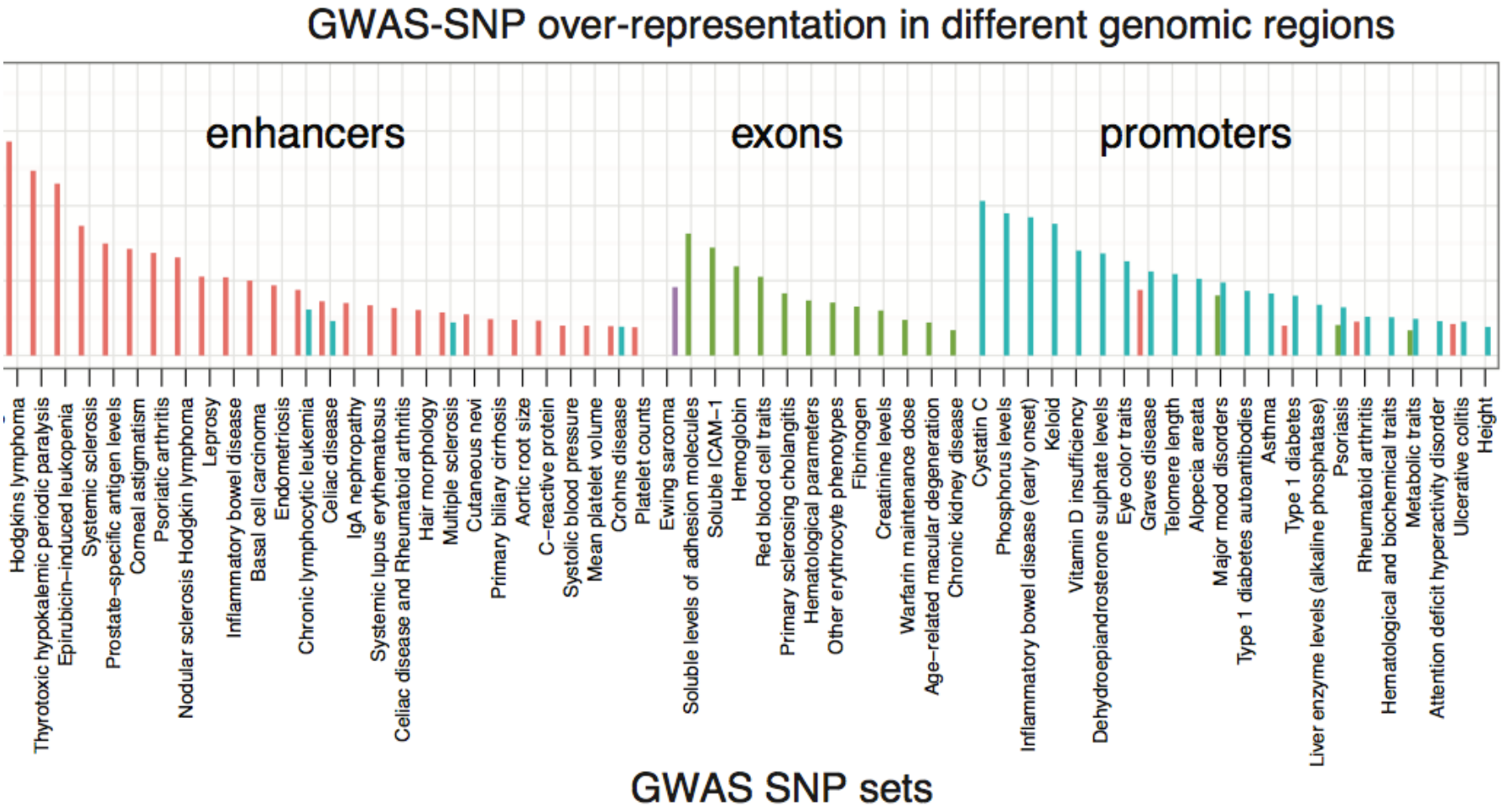
Robin Andersson[1]*, Claudia Gebhard[2,3]*, Irene Miguel-Escalada[4], Ilka Hoof[1], Jette Bornholdt[1], Mette Boyd[1], Yun Chen[1], Xiaobei Zhao[1,5], Christian Schmidl[2], Takahiro Suzuki[6,7], Evgenia Ntini[8], Erik Arner[6,7], Eivind Valen[1,9], Kang Li[1], Lucia Schwarzfischer[2], Dagmar Glatz[2], Johanna Raithel[2], Berit Lilje[1], Nicolas Rapin[1,10], Frederik Otzen Bagger[1,10], Mette Jørgensen[1], Peter Refsing Andersen[8], Nicolas Bertin[6,7], Owen Rackham[6,7], A. Maxwell Burroughs[6,7], J. Kenneth Baillie[11], Yuri Ishizu[6,7], Yuri Shimizu[6,7], Erina Furuhata[6,7], Shiori Maeda[6,7], Yutaka Negishi[6,7], Christopher J. Mungall[12], Terrence F. Meehan[13], Timo Lassmann[6,7], Masayoshi Itoh[6,7,14], Hideya Kawaji[6,14], Naoto Kondo[6,14], Jun Kawai[6,14], Andreas Lennartsson[15], Carsten O. Daub[6,7,15], Peter Heutink[16], David A. Hume[11], Torben Heick Jensen[8], Harukazu Suzuki[6,7], Yoshihide Hayashizaki[6,14], Ferenc Müller[4], The FANTOM Consortium†, Alistair R. R. Forrest[6,7], Piero Carninci[6,7], Michael Rehli[2,3] & Albin Sandelin[1]



We identified 65,423 and 44,459 enhancers in human and mouse. 60% are over-represented in one cell/tissue group

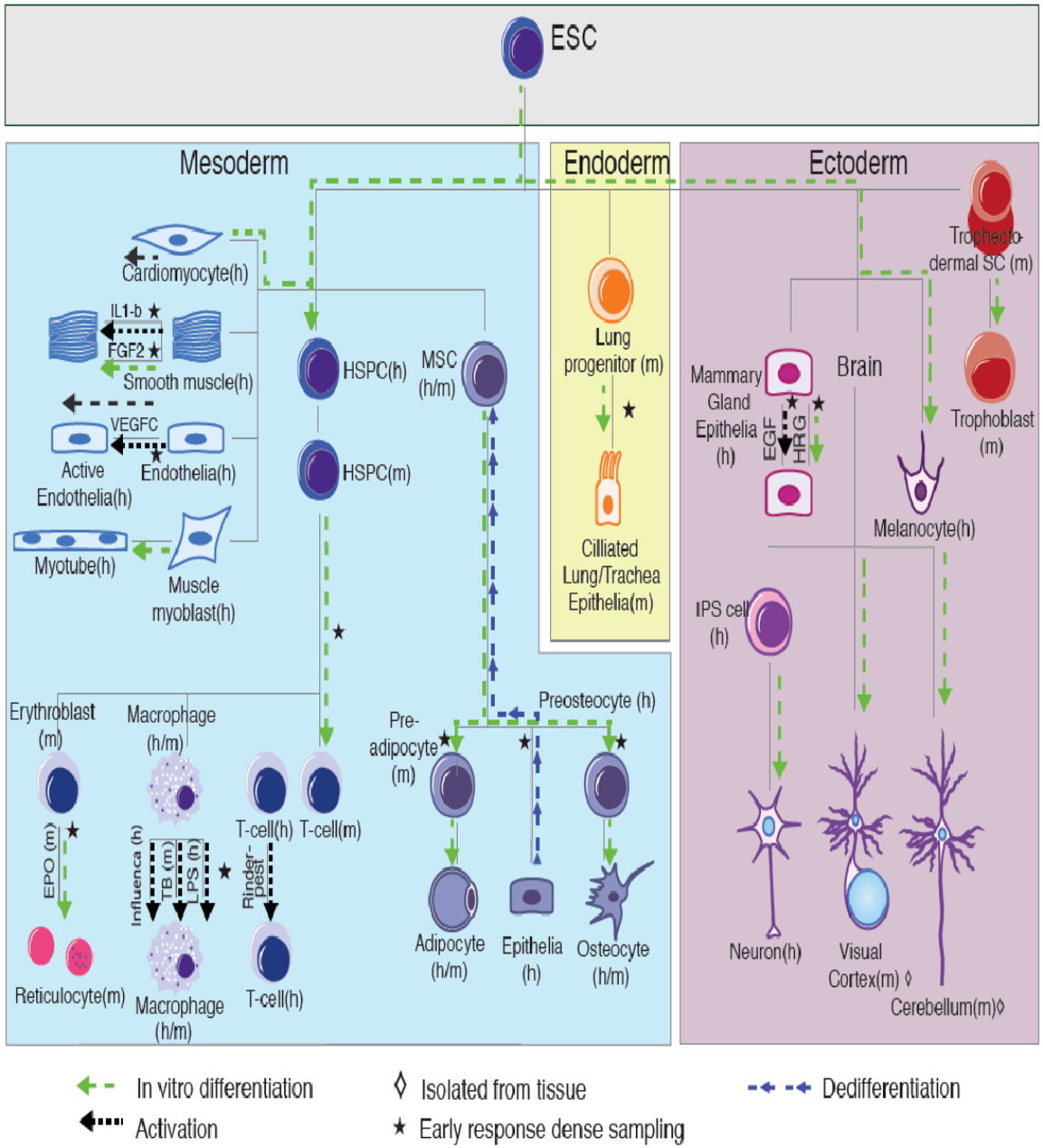
Analysis led by Andersson and Sandelin, Univ. of Copenhagen

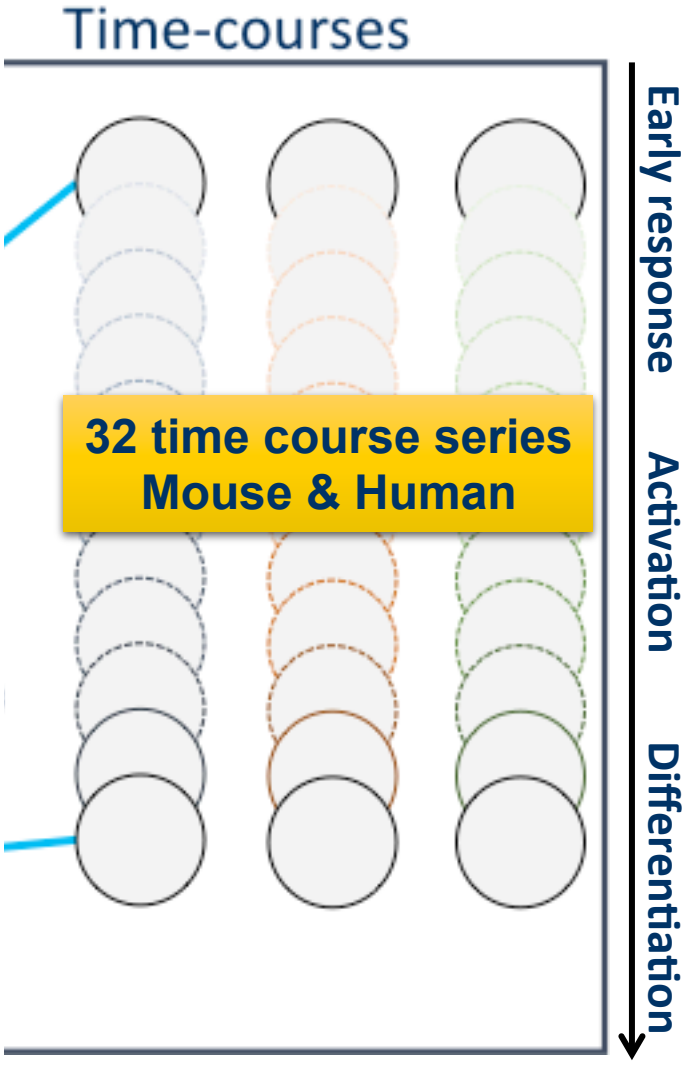
Disease-associated SNPs are enriched in enhancers…



Contributed by Baillie K. *et al*. The Roslin Institute

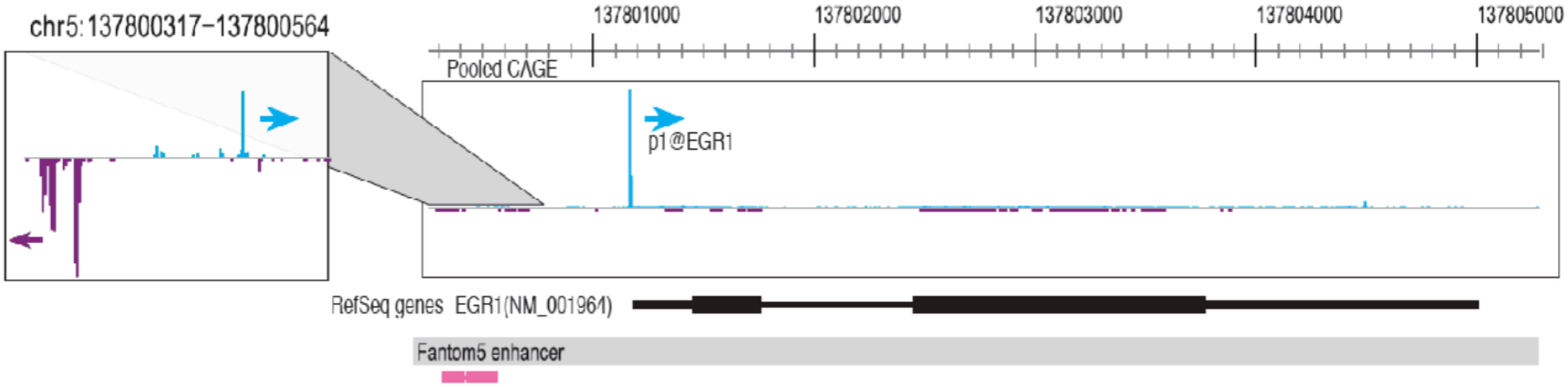
FANTOM5

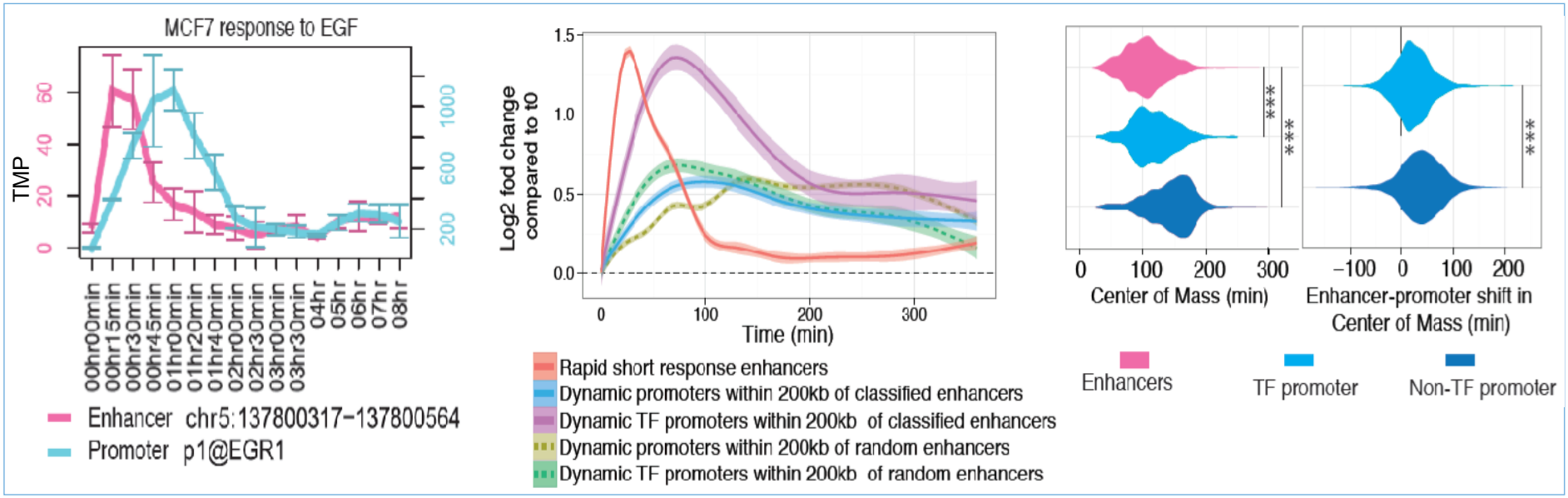




Erik Arner *et al. Science* 347, 1010 (2015)

New paradigm: Enhancer/promoter activation shift





Erik Arner *et al. Science* 347, 1010 (2015)

Enhancers are in control !!!

- They broadly initiate and coordinate biological responses
- Promoters of Transcription Factors follow
- Everything else follows

Sciencexpress

Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells

Erik Arner,*† Carsten O Daub,*† Kristoffer Vitting-Seerup,*† Robin Andersson,*† Berit Lilje, Finn Drablos, Andreas Lennartsson, Michelle Rönnerblad, Olga Hrydziuszko, Morana Vitezic, Tom C Freeman, Ahmad Alhendi, Peter Arner, Richard Axton, J Kenneth Baillie, Anthony Beckhouse, Beatrice Bodega, James Briggs, Frank Brombacher, Margaret Davis, Michael Detmar, Anna Ehrlund, Mitsuhiro Endoh, Afsaneh Eslami, Michela Fagiolini, Lynsey Fairbairn, Geoffrey J Faulkner, Carmelo Ferrai, Malcolm E Fisher, Lesley Forrester, Daniel Goldowitz, Reto Guler, Thomas Ha, Mitsuko Hara, Meenhard Herlyn, Tomokatsu Ikawa, Chieko Kai, Hiroshi Kawamoto, Levon Khachigian, Peter S Klinken, Soichi Kojima, Haruhiko Koseki, Sarah Klein, Niklas Mejhert, Ken Miyaguchi, Yosuke Mizuno, Mitsuru Morimoto, Kelly J Morris, Christine Mummery, Yutaka Nakachi, Soichi Ogishima, Mariko Okada-Hatakeyama, Yasushi Okazaki, Valerio Orlando, Dmitry Ovchinnikov, Robert Passier, Margaret Patrikakis, Ana Pombo, Xian-Yang Qin, Sugata Roy, Hiroki Sato, Suzana Savvi, Alka Saxena, Anita Schwegmann, Daisuke Sugiyama, Rolf Swoboda, Hiroshi Tanaka, Andru Tomoiu, Louise N Winteringham, Ernst Wolvetang, Chiyo Yanagi-Mizuochi, Misako Yoneda, Susan Zabierowski, Peter Zhang, Imad Abugessaisa, Nicolas Bertin, Alexander D. Diehl, Shiro Fukuda, Masaki Furuno, Jayson Harshbarger, Akira Hasegawa, Fumi Hori, Sachi Ishikawa-Kato, Yuri Ishizu, Masayoshi Itoh, Tsugumi Kawashima, Miki Kojima, Naoto Kondo, Marina Lizio, Terrence F. Meehan, Christopher J Mungall, Mitsuyoshi Murata, Hiromi Nishiyori-Sueki, Serkan Sahin, Sayako Sato-Nagao, Jessica Severin, Michiel JL de Hoon, Jun Kawai, Takeya Kasukawa, Timo Lassmann, Harukazu Suzuki, Hideya Kawaji,† Kim M Summers,† Christine Wells,† FANTOM Consortium, David A Hume,†‡ Alistair RR Forrest,†‡ Albin Sandelin,†‡ Piero Carninci,†‡ Yoshihide Hayashizaki†‡

Erik Arner *et al. Science* 347, 1010 (2015)

FANTOM

FUNCTIONAL ANNOTATION OF THE MAMMALIAN GENOME

http://fantom.gsc.riken.jp/

日本語

Home | Data | Views | Protocols | Software | Papers | FAQ

Zenbu

A new data integration, data processing, and expression enhanced visualization system with secured data upload and sharing designed for big data genomics projects (FANTOM5, FANTOM4, FANTOM3, ENCODE).

SSTAR

(Semantic catalogue of Samples, Transcription initiation, And Regulations) An interface to explore the cell type, co-expression cluster specific annotations, motifs and transcription factors. Summary of

Files

Selected datafiles for FANTOM5. Alternatively you may view the data file server directly at http://fantom.gsc.riken.jp/5/datafiles/

Number of page view a year

Total: 9,217,440

ZENBU: 2,927,780 (FY2015)

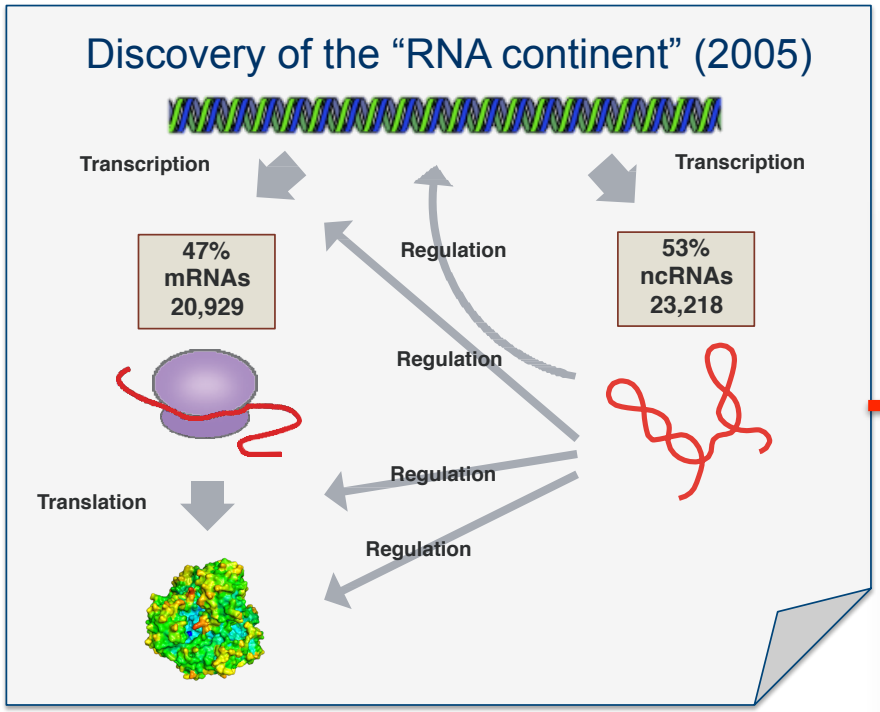
Databases by Hideya Kawaji, Takeya Kasukawa *et al*.

ZENBU: Jessica Severin *et al*. *Nature Biotechnology* 32, 217 (2014)

SSTAR: Imad Abugessaisa *et al*. *Database* pii: baw105 (2015)

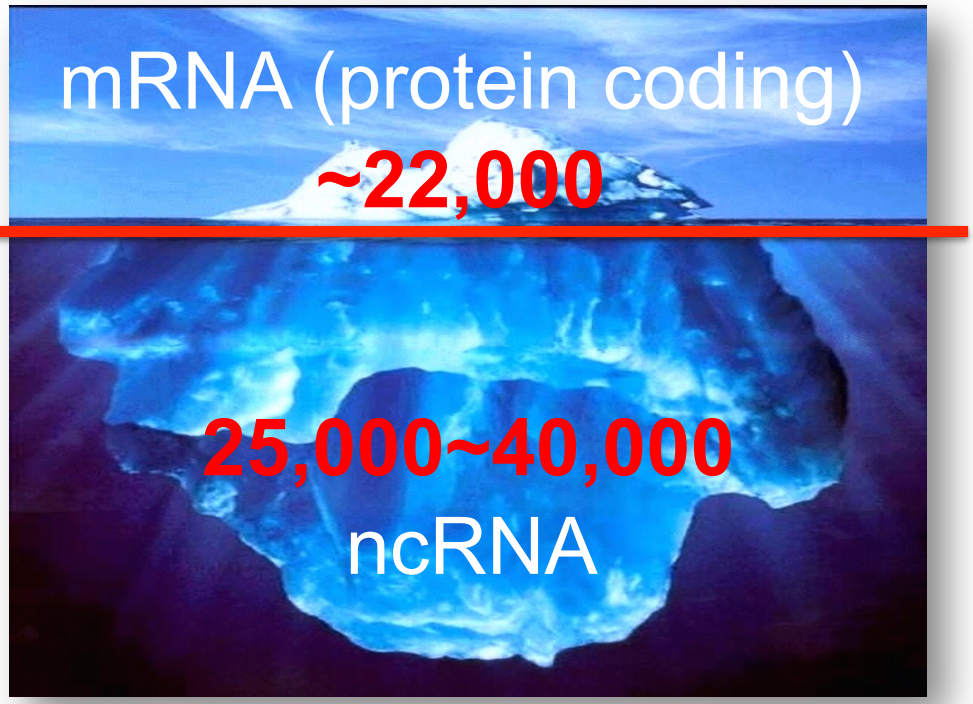
New frontier of ncRNA

We know very little yet



>67.0%

Multiple references in PubMed



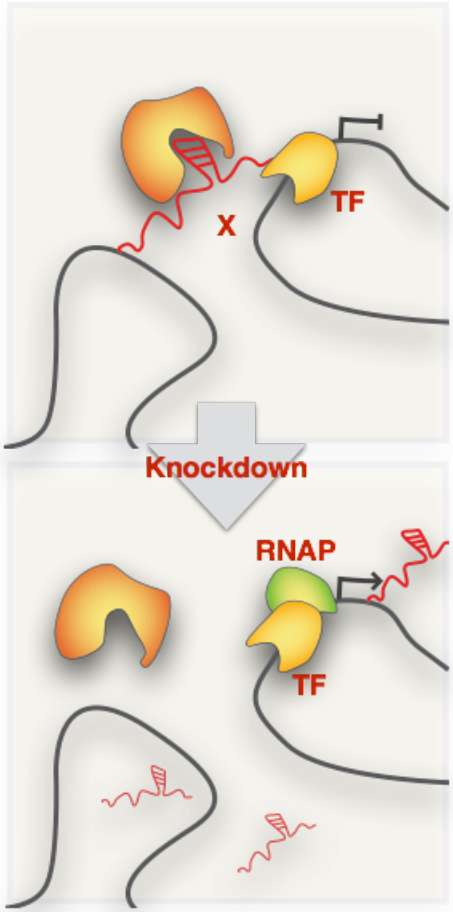
98.4%

No reference in PubMed

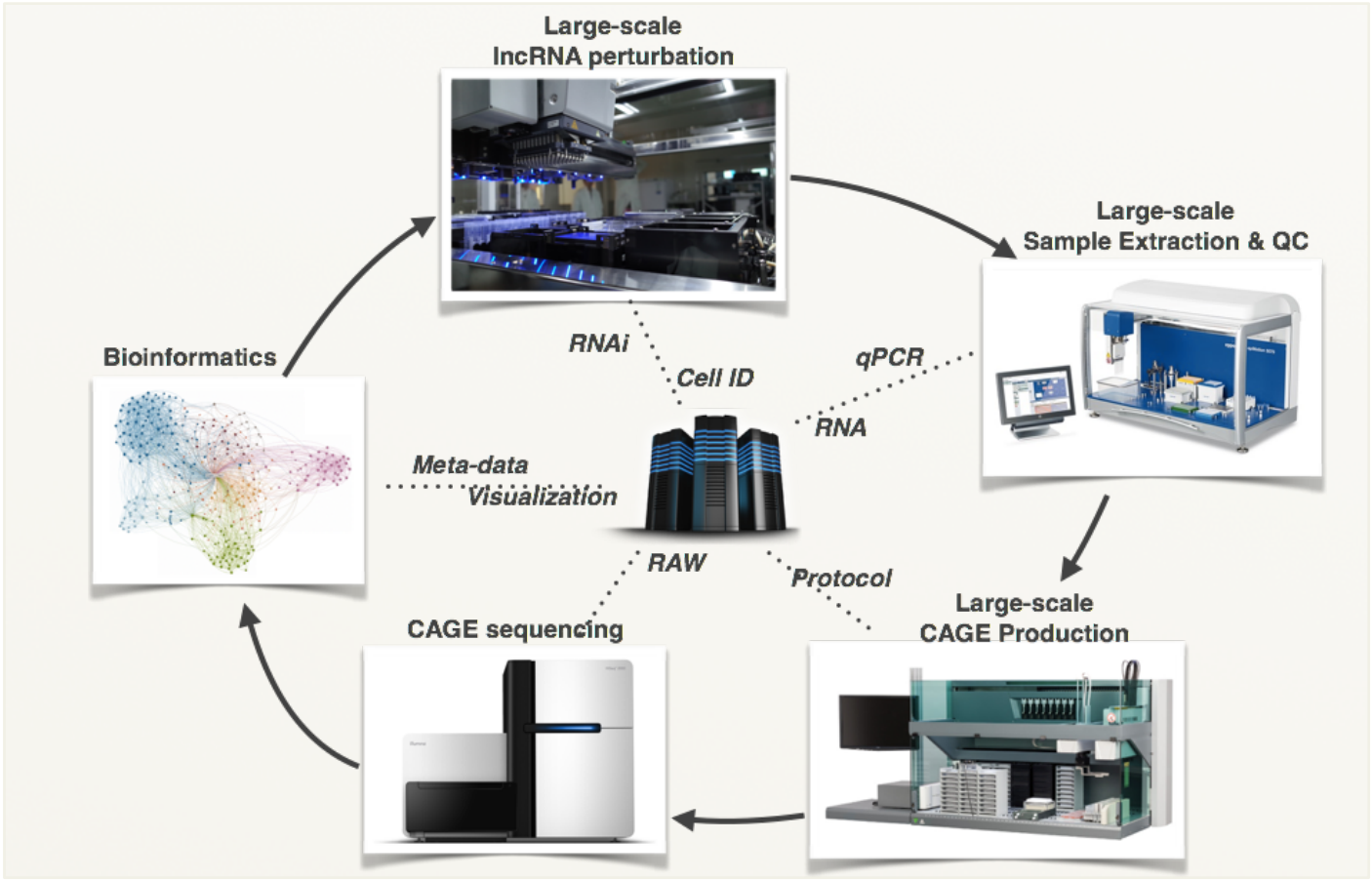
de Hoon *et al., Mammalian Genome* 26, 391 (2015)

FANTOM6 Strategy

Principle



Systematic pipeline



What can technologies (cap-trap, CAGE) contribute to:

GENOME Projects

- annotation using cDNAs, CAGE
  - human, mouse genome (transcriptome data to find genes)
- Nature 2001, 2002 Genome Res 2008 (hypertensive rat)

ENCODE

- Contribution with CAGE, nanoCAGE, etc.
- Nature 2007, two Nature 2012 special issues

ModENCODE Transcriptome

- Nature 2014

Zeprome (Zebra fish Promoterome)

- Nature 2014

Future possibilities with CAGE

- ✓ Comprehensive promoter & enhancer analysis
  - All genes & ncRNAs
  - All developmental stages
  - All species
  - Health & diseases
- ✓ Characterization and quality control of iPS cells
- ✓ Expression quantitative trait loci (eQTL)
- ✓ Drug response analysis

Special thanks to:

- RIKEN
  - Erik Arner
  - Jessica Severin
  - Imad Abugessaisa
  - Hideya Kawaji
  - Masayoshi Itoh
  - Li Jing-Ru
  - Michiel de Hoon
  - Jay Shin
  - Charles Plessy
  - Takeya Kasukawa
  - Harukazu Suzuki
  - Yoshihide Hayashizaki
  - GENAS
- Harry Perkins Institute of Medical Research
  - Alistair Forrest
- Telethon Kids Institute
  - Timo Lassmann
- Karolinska Institutet
  - Carsten Daub
- The Roslin Institute
  - David Hume
  - Kenneth Baillie
  - Tom Freeman
  - Kim Summers
- Univ. Edinburgh
  - Martin Taylor
- Univ. Copenhagen
  - Albin Sandelin
  - Robin Andersson
  - Kristoffer Vitting-Seerup
- Birmingham Univ.
  - Ferenc Mueller
- Imperial College London
  - Boris Lenhard
- Univ. Hosp Regensburg
  - Michael Rehli
- DZNE
  - Peter Heutink
- Univ. Sheffield
  - Winston Hide
- Vavilov Institute of General Genetics
  - Vsevolod Makeev
- KAUST
  - Vladimir Bajic
- Univ. Melbourne
  - Christine Wells
- Univ. Bristol
  - Julian Gough
  - Owen J L Rackham
- Monash University
  - Jose M Polo

Funding

- Research Grants for RIKEN Omics Science Center, RIKEN Preventive Medicine and Diagnosis Innovation Program and RIKEN Centre for Life Science Technologies, Division of Genomic Technologies from MEXT, Japan
- A grant of the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT, Japan

And to many more FANTOM Consortium members, Lab members, ENCODE members, Students and historical collaborators…

FANTOM Collaborators: Thanks!