An R/Bioconductor package for tabular omics data analysis using a supra-hexagonal map

Hai Fang*, Julian Gough

Computational Genomics Group
Department of Computer Science

*hfang@cs.bris.ac.uk

University of BRISTOL

Some facts on omics data

- **table-like matrix**:
  digitise bioactivities (eg. expression level) of genomic regions (eg. genes) across samples

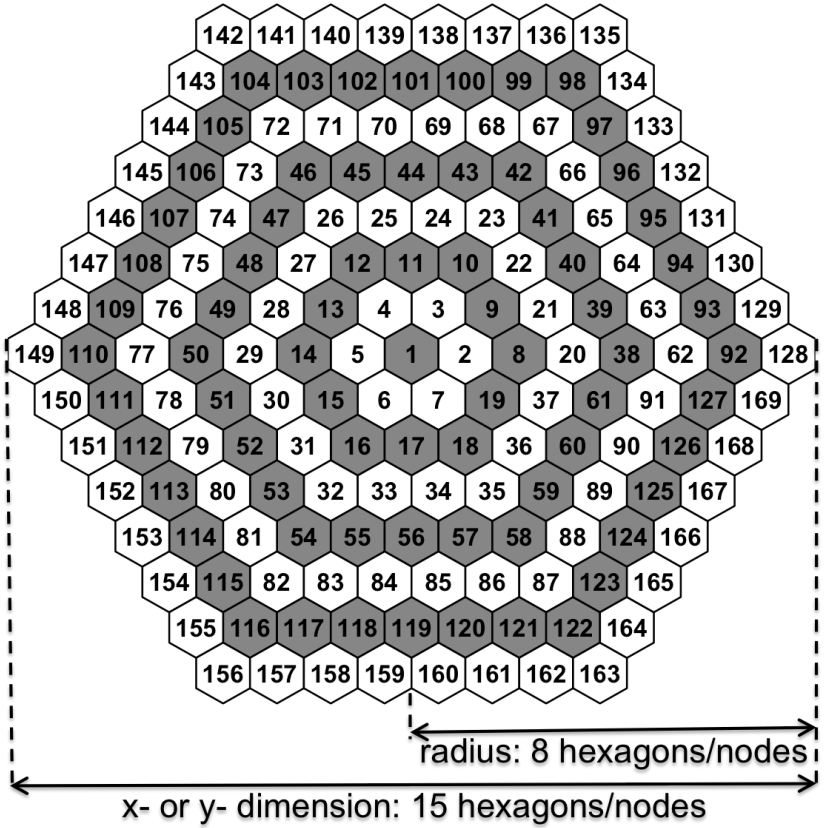
- **dimension curse**:
  number of genes/features **>>** number of samples

- **rational behind data normalisation**:
  no changes in most genes

- **existence of a center with radial symmetry**:
  most of no-changed genes are mapped onto this center

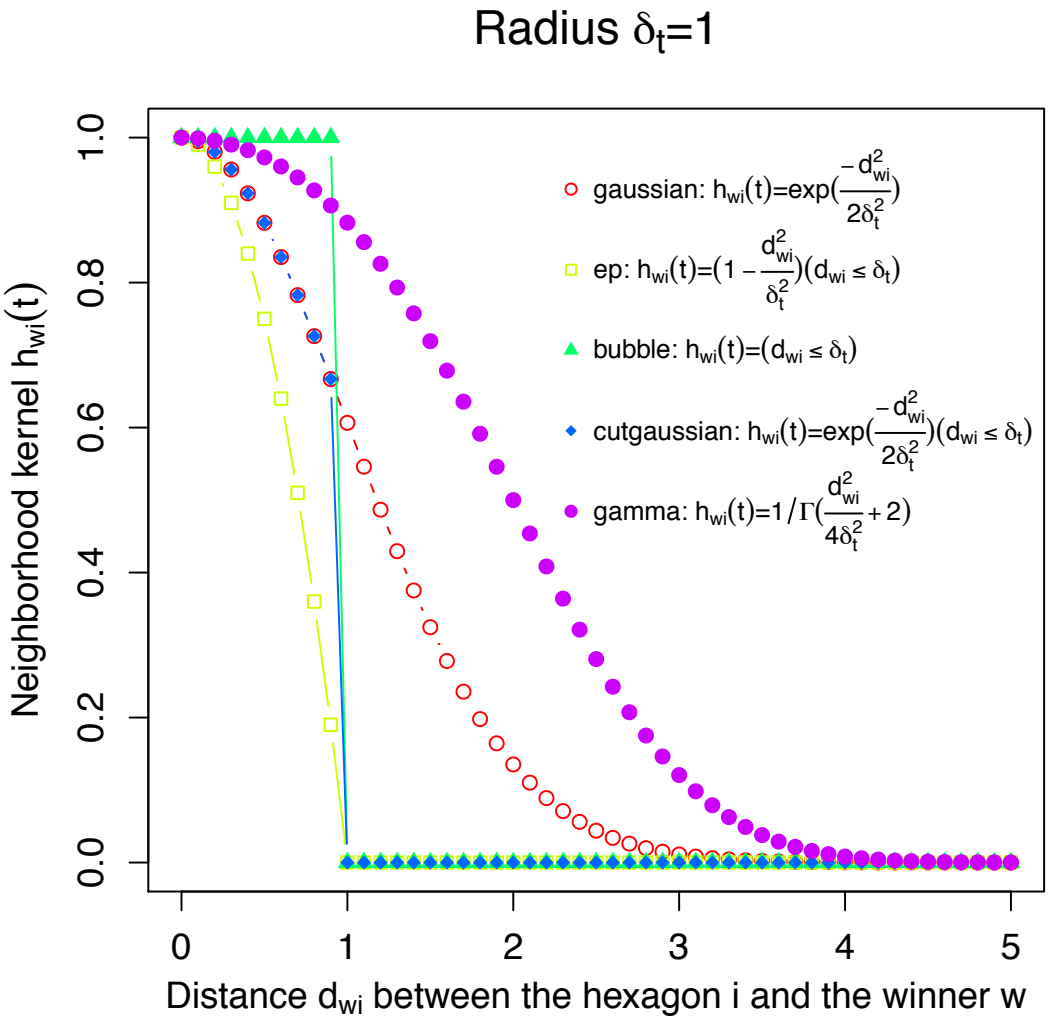
Architectural design of a supra-hexagonal map

- **inspired** by omics data structure
- **node indexing** from the center, radiating circularly outwards
- **determinant**: radius only
- **neighbors**:
  - 6 for inside nodes
  - 3 for six corner nodes
  - 4 for border nodes
- **two coordinate systems**:
  - 2D output space
  - High-dimensional input space

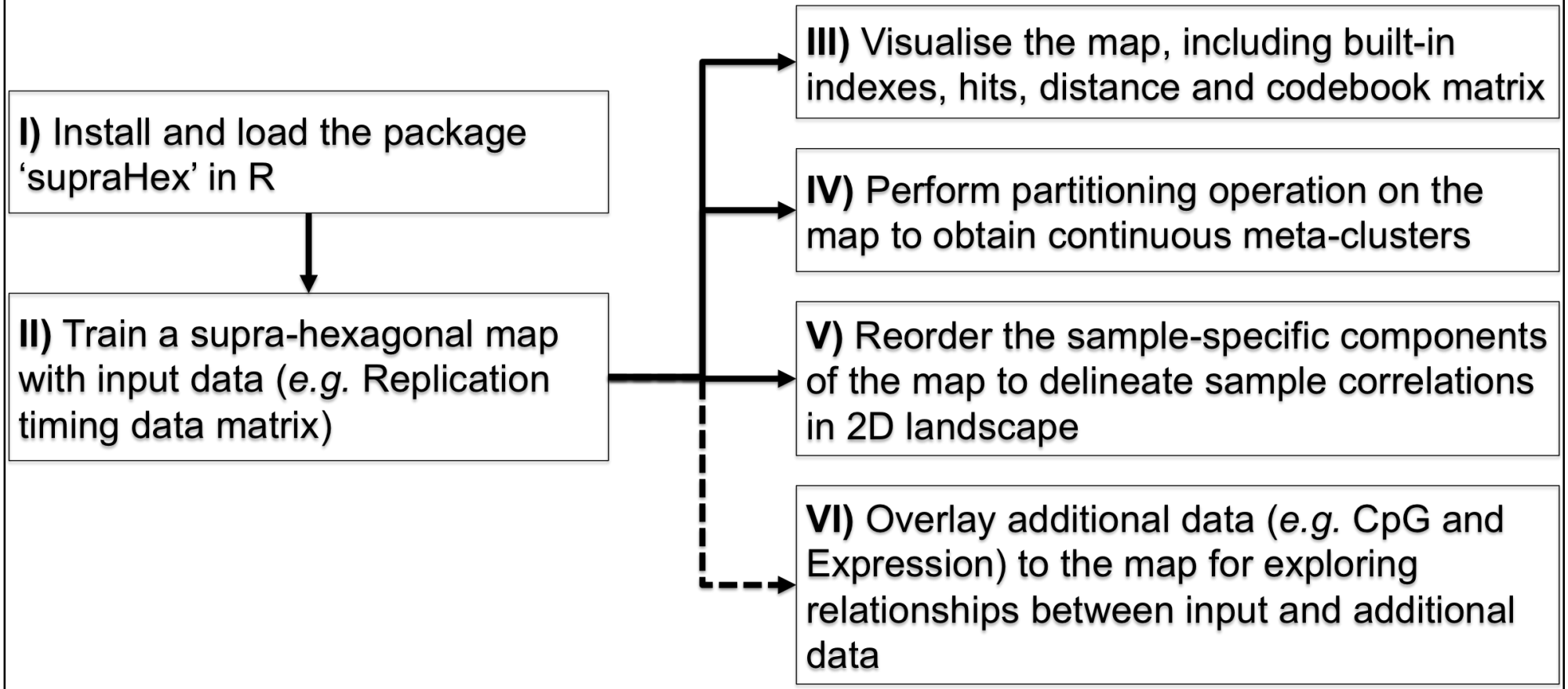


The supra-hexagonal map trained via self-organising learning algorithm

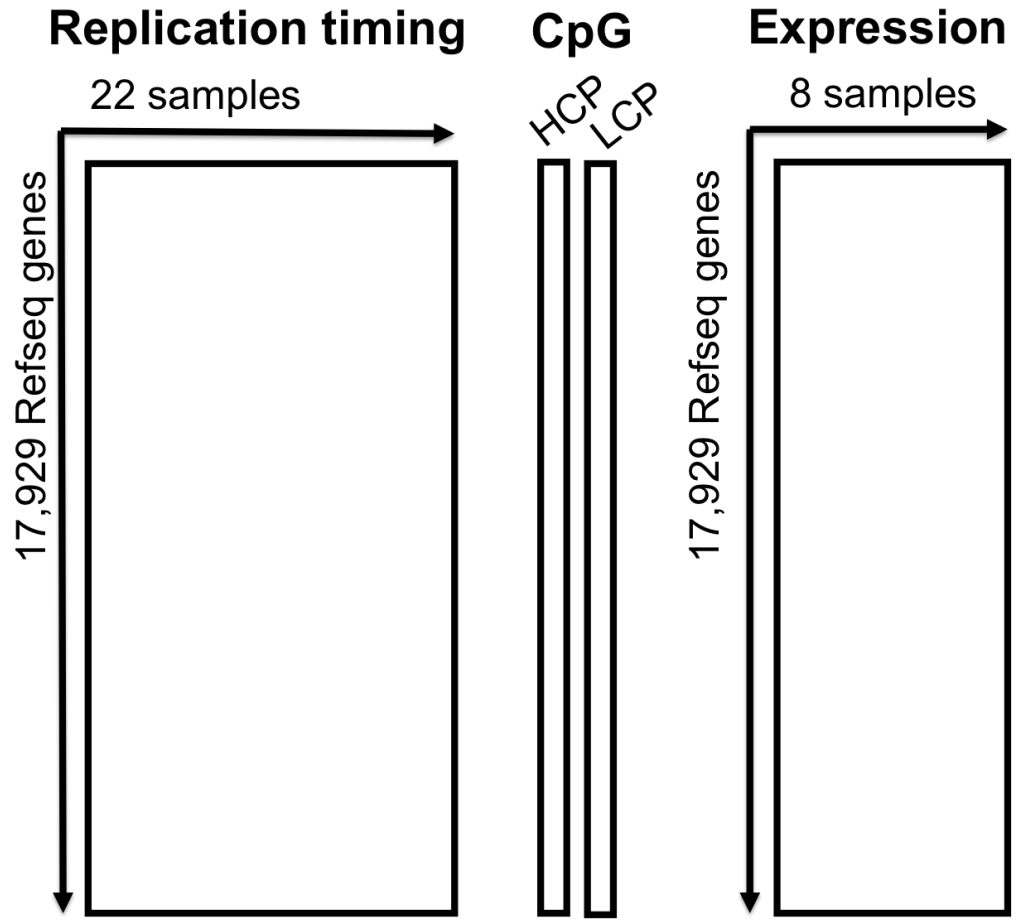
- **in essence**: converting input gene-sample matrix into the output codebook matrix associated with the map
- **outcome**: mapping of similar input data onto neighboring regions
- **neighborhood kernel**: dictate the topology of the trained map



Workflow of the supraHex package



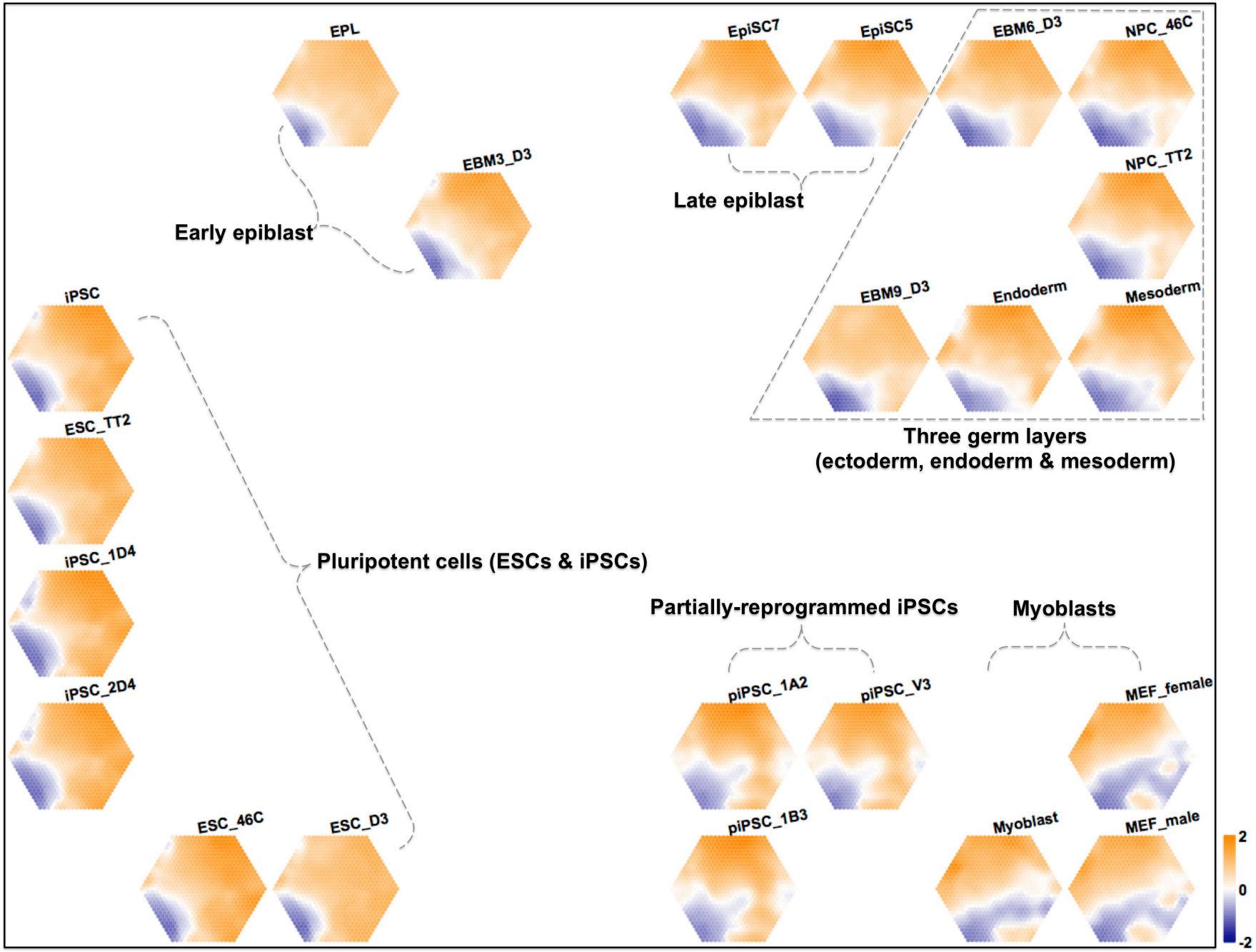
Case in analysing DNA replication timing, CpG and expression



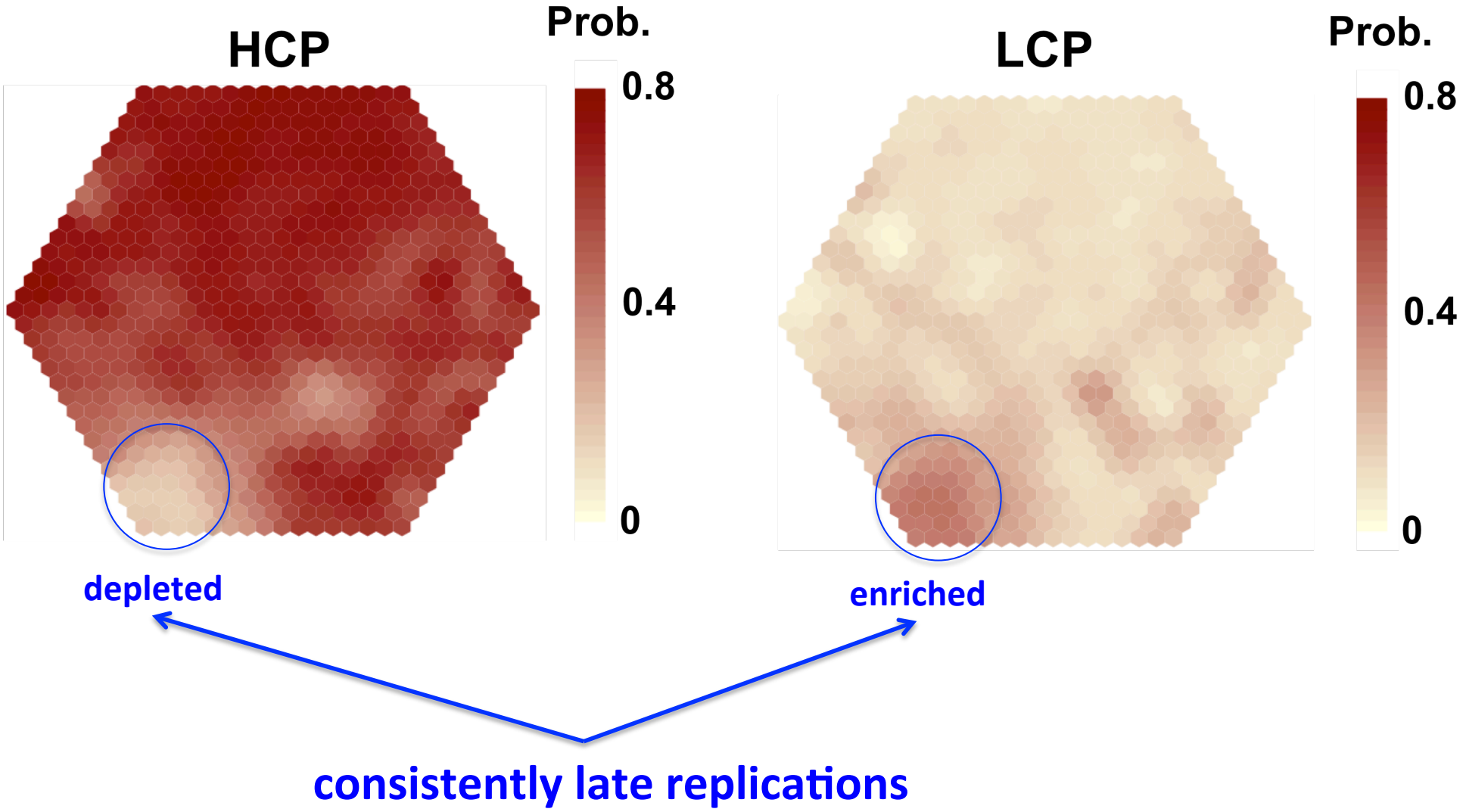
Input data for training **Additional data for overlaying**

*1. Hiratani, et al., **Genome Res**. 20 (2010) 155–169; 2. Mikkelsen, et al., **Nature**. 448 (2007) 553–60.*

Replication timing map

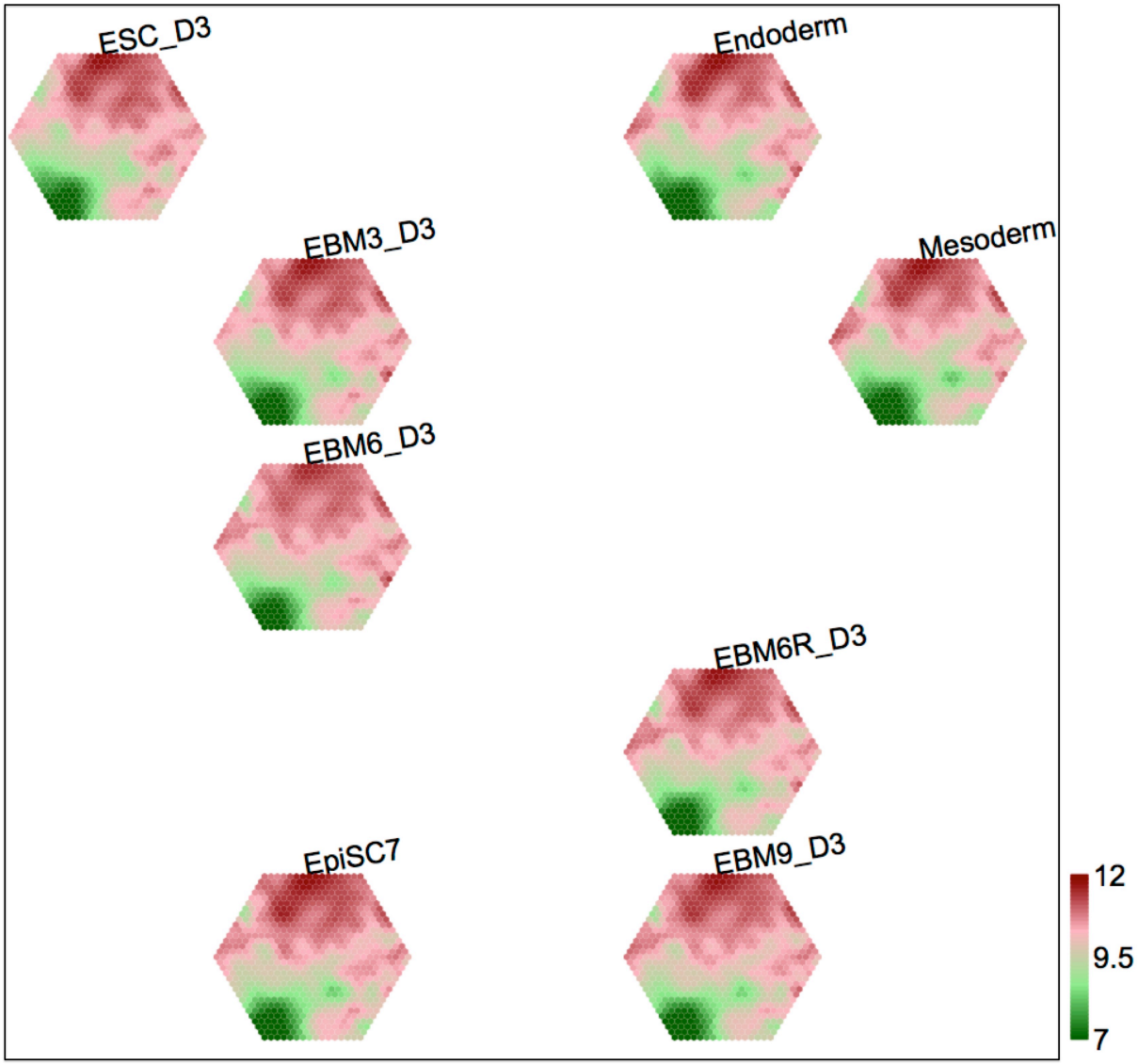


CpG data overlaid onto replication-timing map



Expression data onto replication-timing map

- **Expression map:** purely according to replication timing
- **Inherent relationship:** late replication of genes with low expression
- **Trajectory**: neural differentiation from embryonic stem cell



Comparisons to other tools

Features	supraHex	somtoolbox	kohonen	Cluster3.0
Programming language	R	Matlab[a]	R	C
Map shape	supra-hexagon	sheet[b]	sheet	sheet
Visually friendly	Yes	Yes	Yes	No
With neighbour kernels	Yes	Yes	No	No
Meta-clustering	Yes	Yes	No	No
Sample reordering	Yes	Yes	No	No
Overlaying with additional data	Yes	No	No	No
Bioconductor project	Yes	No	No	No

[a]Needs commercial license
[b]Also supports the cylinder and toroid shapes but are less popular

- visual novelty;
- simultaneous analysis of genes and samples;
- multilayer omics data comparisons;
- self-explanatory and reproducible results.

Summary: cover *2* things

supraHex Concept

- architectural design of a supra-hexagonal map
- self-organising learning algorithm

supraHex Functionality

- intuitive visualisations
- gene clustering and meta-clustering
- sample correlations/landscape
- additional data overlaying for multilayer omics data comparisons
- available at http://supfam.org/supraHex or Bioconductor website (http://bioconductor.org)

Hai Fang and Julian Gough, ***Biochem and Biophys Res Commun****. 2013 (DOI:*10.1016/j.bbrc.2013.11.103*)*