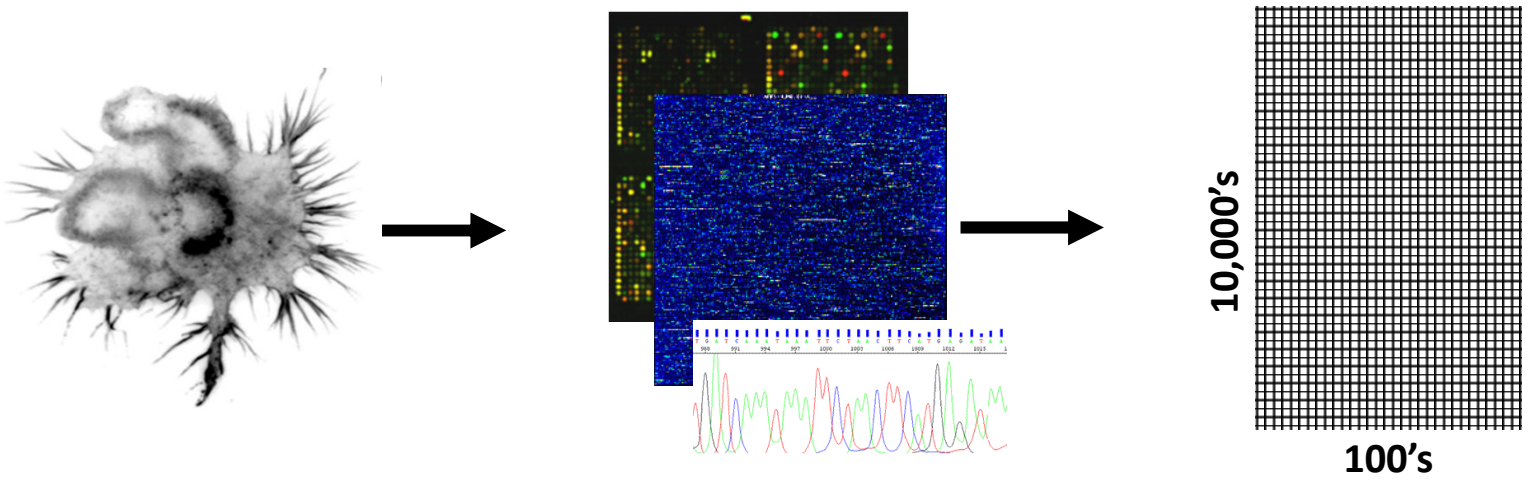
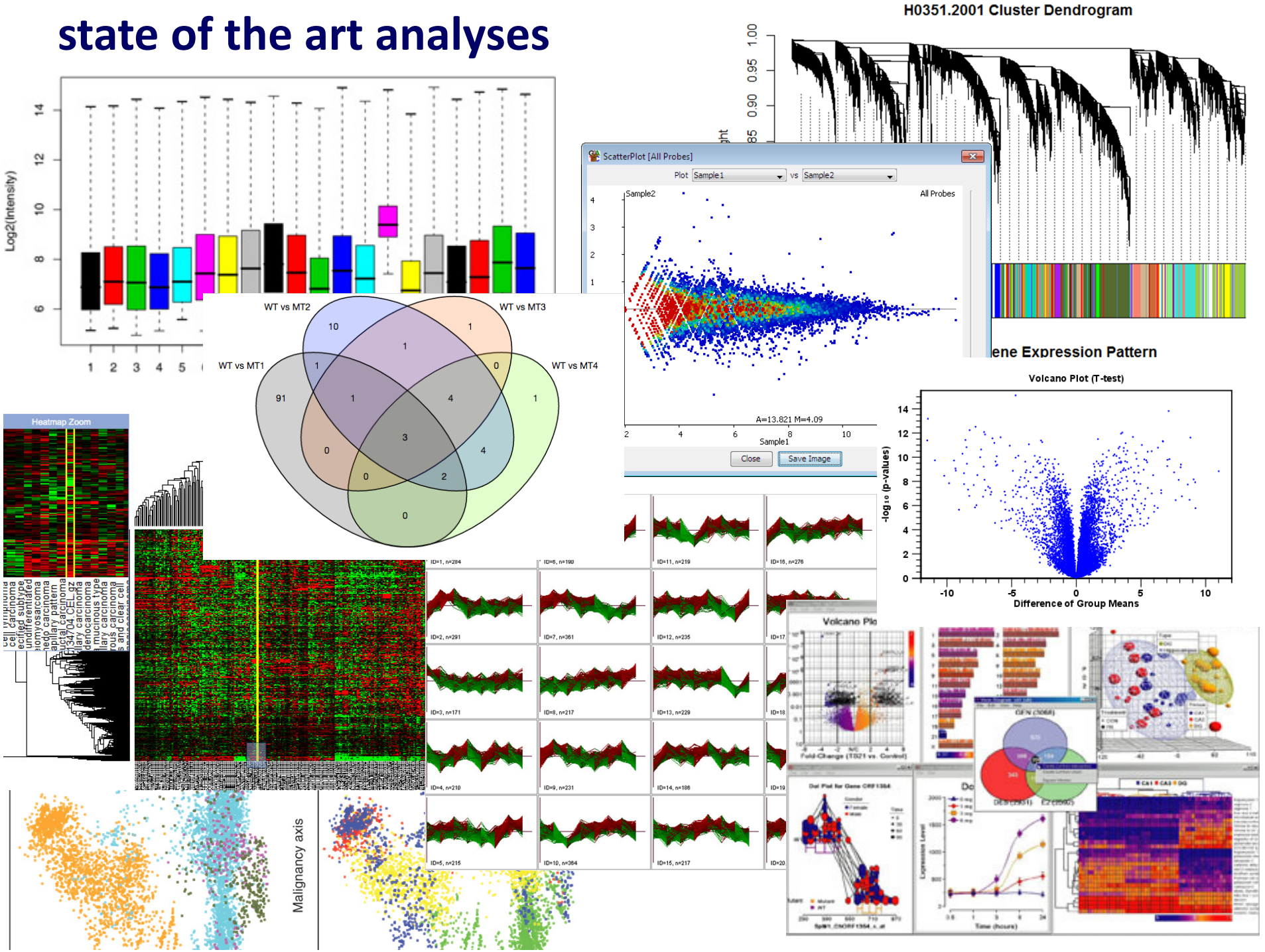
network analysis of biological data: visualising the matrix

Prof Tom Freeman
Systems Immunology Group,
The Roslin Institute,
University of Edinburgh

biological analyses



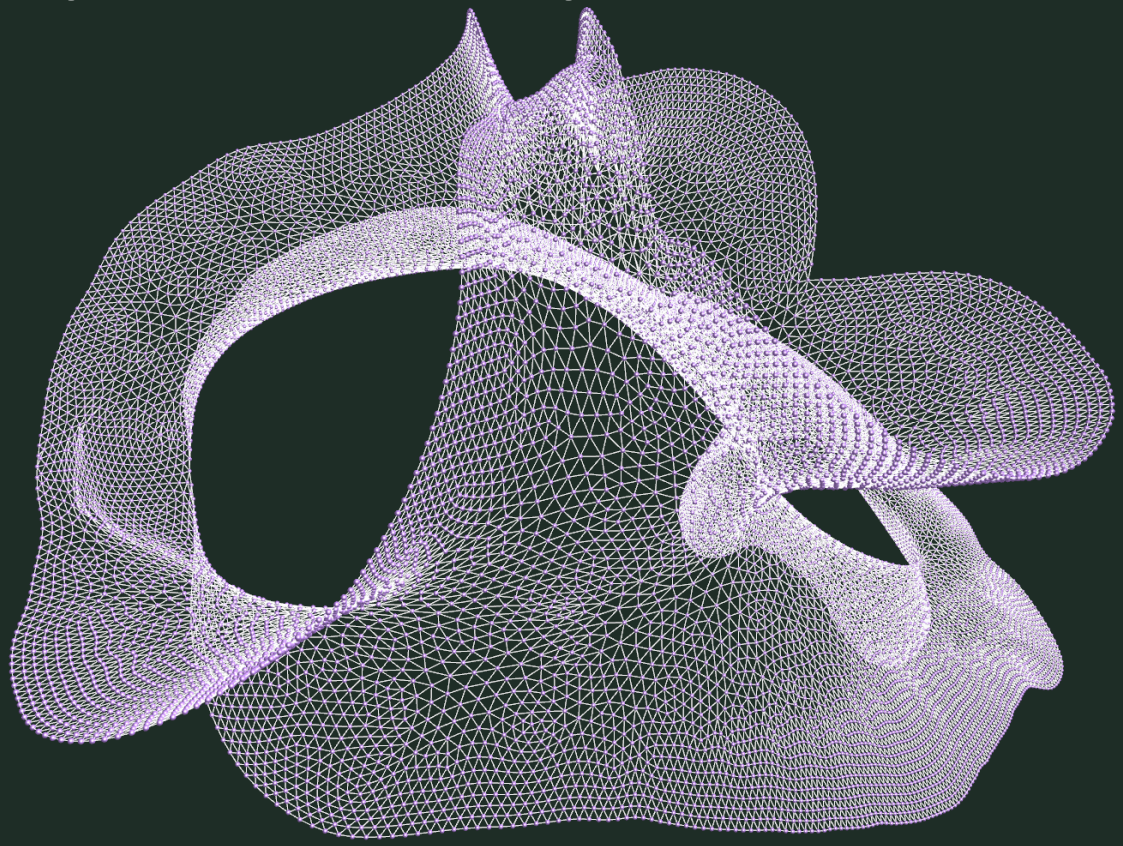
state of the art analyses



networks

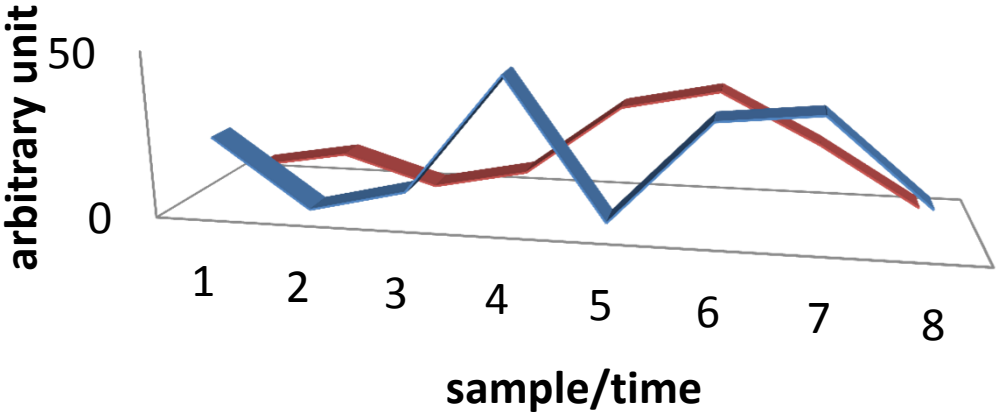
composed of nodes and edges

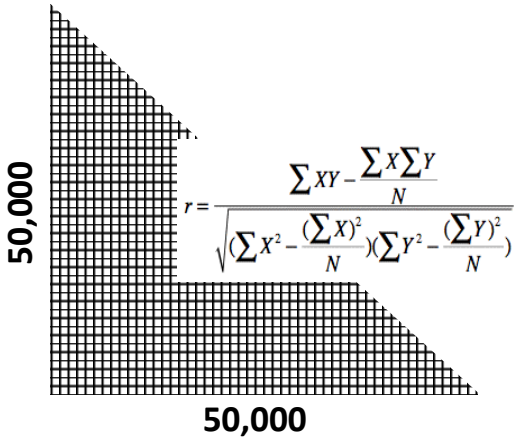
- ***nodes usually represent entities or things***
- ***edges represent relationships between those entities***



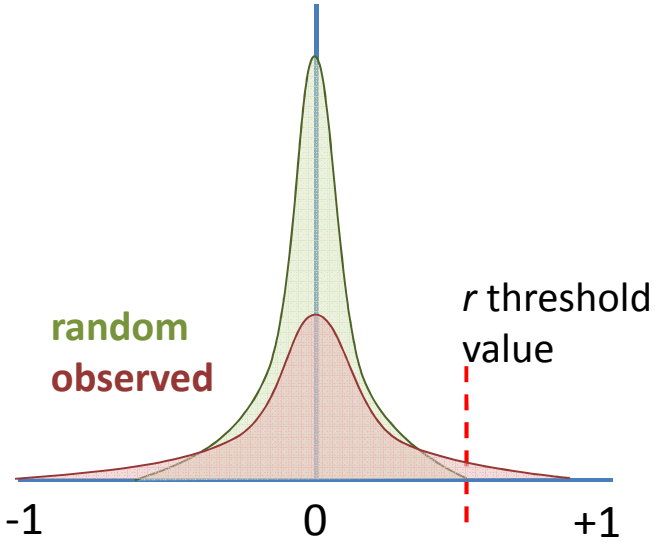
network analyses can be applied to many different types of data and systems

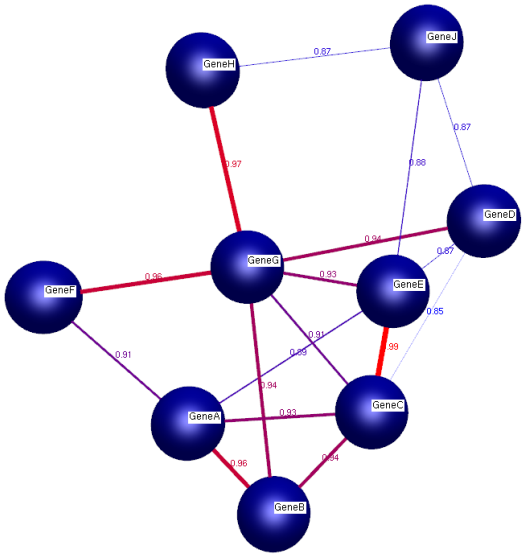
plotting the statistically improbable: principles of correlation (co-expression) networks





1.25 billion calculations



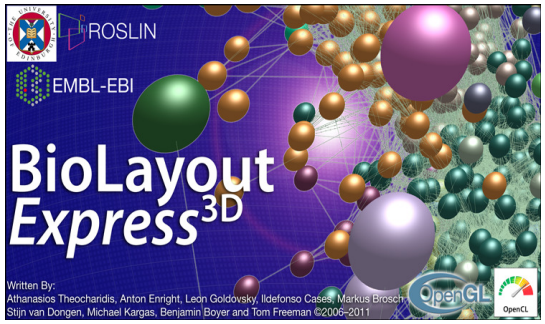


weighted, non-directional graph

BioLayout *Express*^{3D}: one tool many applications

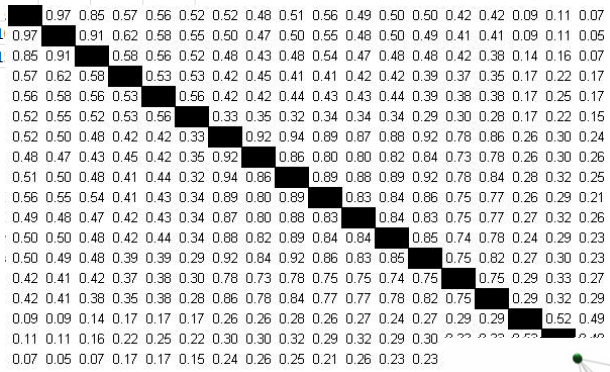
Technology

- developed over 10 year period with by scientists in Cambridge and Edinburgh
- platform independent (Mac, Windows, Linux)
- supports numerous data import formats (txt, gml, graphml, biopax, expression, matrix)
- incorporates a number of highly optimised algorithms
  - correlation (Pearson/Spearman)
  - graph layout (F-R, FMMM)
  - MCL clustering
  - Petri net stochastic flow
- many options for user interaction and querying of graphs and underlying data
- advanced visualization capabilities for interactive graph rendering in 2D and 3D
- scalable and fast

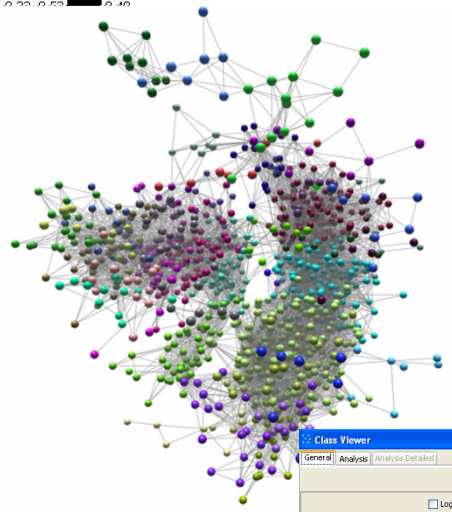


Symbol	0-cont	24-con	0 h	1 h	2 h	4 h	6 h	8 h
OR4F17	14.1	13.7	13.4	13.4	12.8	12.9	12.8	13.1
OR4F16	21.8	21.7	20.7	22.2	20.8	22.5	20.9	19.7
LOC64383	58.1	41.1	37.8	52.7	69.8	112.2	90.3	95.8
SAMD11	159.3	140.5	129.4	103.5	123.3	127.0	134.6	139.6
KLHL17	151.2	132.2	119.9	113.0	92.9	105.5	112.3	117.2
PLEKHN1	95.5	84.4	74.2	69.9	58.4	68.6	71.0	71.7
ISG15	114.0	111.7	97.1	78.2	74.0	85.6	75.8	75.1
AGRN	212.4	201.6	209.8	226.9	175.5	162.0	136.0	143.1
MIR200B	81.3	104.6	85.7	84.9	66.6	65.9	74.0	77.6
MIR200A	16.3	15.4	15.3	15.4	14.8	16.4	15.9	18.0
MIR429	39.2	37.5	42.5					
TTLL10	136.0	131.6	123.0	1				
B3GALT6	197.0	181.7	173.7	1				

high dimensional data
100's columns, thousands rows



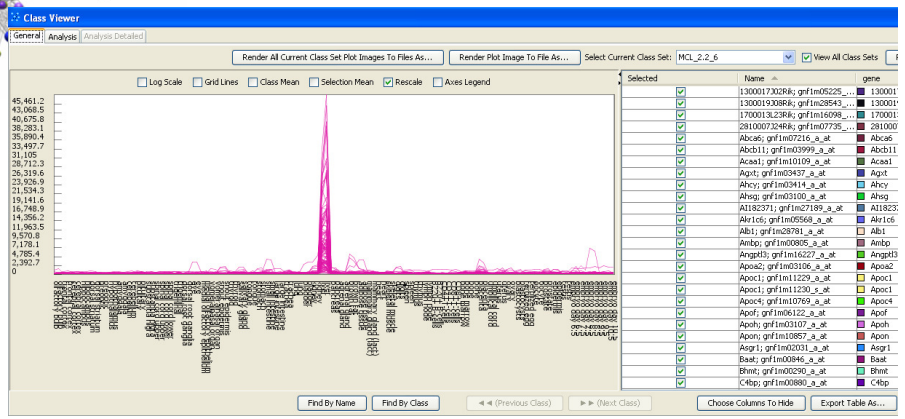
calculate correlation matrix



layout, dynamic visualisation and clustering of graphs (10's thousands nodes, millions of edges) in 2D or 3D

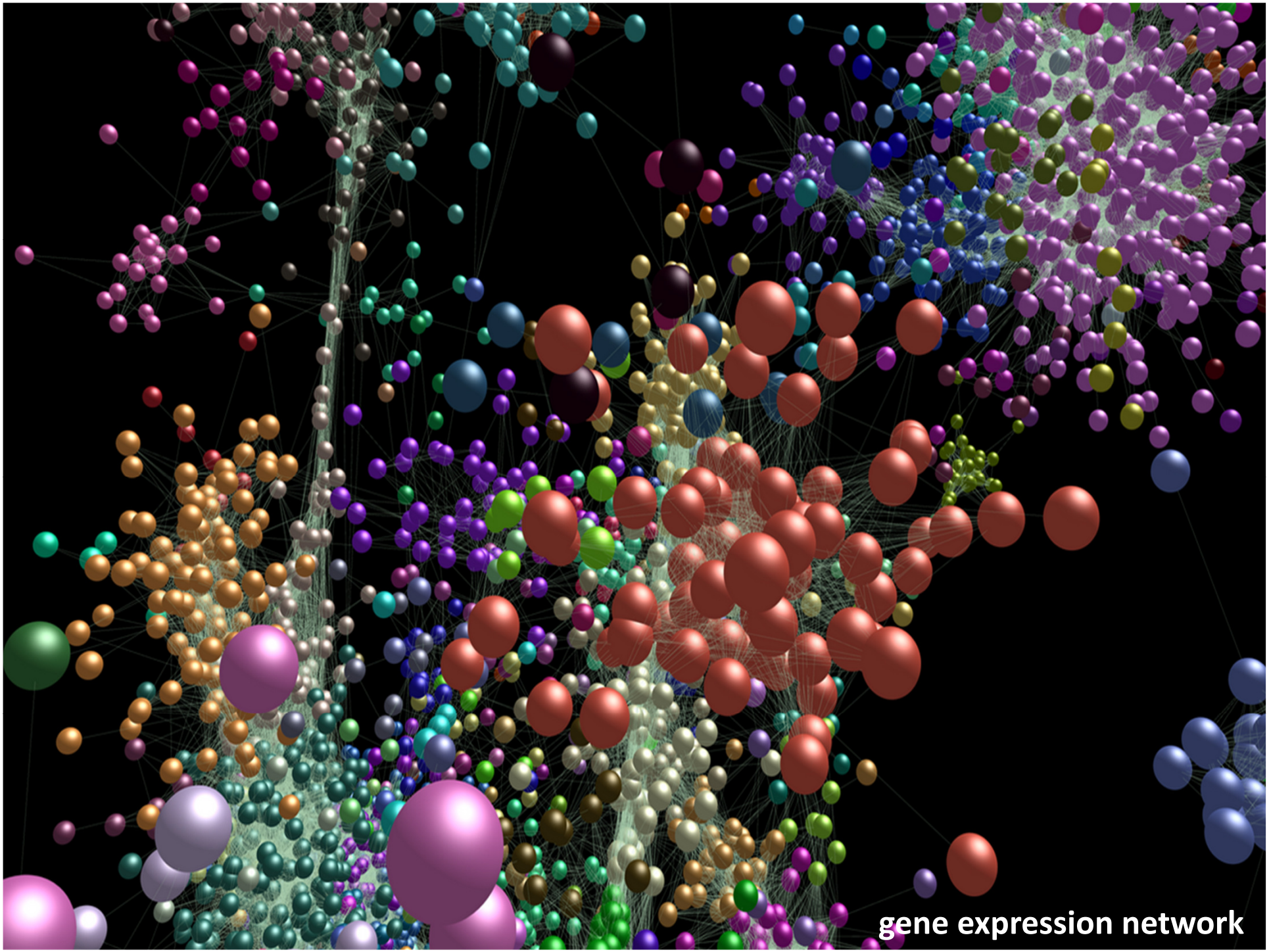
minutes

data integration and exploration



BioLayout *Express*^{3D}
analysis pipeline for high dimensional data

www.biolayout.org


gene expression network

Human Conjunctival Transcriptome Analysis Reveals the Prominence of Innate Defense in *Chlamydia trachomatis* Infection[▽]†

Angels Natividad,[1]‡ Tom C. Freeman,[2] David Jeffries,[1,3] Matthew J. Burton,[1] David C. W. Mabey,[1] Robin L. Bailey,[1] and Martin J. Holland[1,3]*

Department of Clinical Research, London School of Hygiene and Tropical Medicine, University of London, London WC1E 7HT, United Kingdom[1]; The Roslin Institute, R(D)SVS, University of Edinburgh, Roslin Biocentre, Midlothian EH25 9PS, United Kingdom[2]; and Medical Research Council Laboratories, Viral Diseases Programme, Fajara, Banjul, the Gambia[3]

Received 3 August 2010/Returned for modification 17 August 2010/Accepted 24 August 2010

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Functional clustering and lineage markers: Insights into cellular differentiation and gene function from large-scale microarray studies of purified primary cell populations

David A. Hume *, Kim M. Summers, Sobia Raza, J. Kenneth Baillie, Thomas C. Freeman

The Roslin Institute, Roslin Biocentre, Roslin EH25 9PS, Midlothian, UK
Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Roslin Biocentre, Roslin EH25 9PS, Midlothian, UK

Journal of Investigative Dermatology advance online publication 31 October 2013; doi:10.1038/jid.2013.375

Distinct Molecular Signature of Human Skin Langerhans Cells Denotes Critical Differences in Cutaneous Dendritic Cell Immune Regulation

Marta E. Polak[1], Stephen M. Thirdborough[2], Chuin Y. Ung[1], Tim Elliott[2], Eugene Healy[1], Tom C. Freeman[3] and Michael R. Ardern-Jones[1]

Freeman *et al. BMC Biology* 2012, **10**:90
http://www.biomedcentral.com/1741-7007/10/90

BMC Biology
incorporating Journal of Biology

RESEARCH ARTICLE Open Access

A gene expression atlas of the domestic pig

Tom C Freeman[1*], Alasdair Ivens[2,6], J Kenneth Baillie[1], Dario Beraldi[1,7], Mark W Barnett[1], David Dorward[1], Alison Downing[1], Lynsey Fairbairn[1], Ronan Kapetanovic[1], Sobia Raza[1], Andru Tomoiu[1], Ramiro Alberio[3], Chunlei Wu[4], Andrew I Su[4], Kim M Summers[1], Christopher K Tuggle[5], Alan L Archibald[1*] and David A Hume[1*]

JLB

Article

Analysis of the transcriptional networks underpinning the activation of murine macrophages by inflammatory mediators

Sobia Raza, Mark W. Barnett,* Zohar Barnett-Itzhaki,† Ido Amit,† David A. Hume,* and Tom C. Freeman*,1*

*The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Scotland, United Kingdom; and

Contents lists available at ScienceDirect

Immunobiology

journal homepage: www.elsevier.de/imbio

Meta-analysis of lineage-specific gene expression signatures in mouse leukocyte populations

Neil A. Mabbott *, J. Kenneth Baillie, David A. Hume *,1, Tom C. Freeman 1

The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Roslin Biocentre, Roslin EH25 9PS, Midlothian, UK

Mabbott *et al. BMC Genomics* 2013, **14**:632
http://www.biomedcentral.com/1471-2164/14/632

BMC Genomics

RESEARCH ARTICLE Open Access

An expression atlas of human primary cells: inference of gene function from coexpression networks

Neil A Mabbott*†, J Kenneth Baillie, Helen Brown, Tom C Freeman*† and David A Hume*†

Doig *et al. BMC Genomics* 2013, **14**:469
http://www.biomedcentral.com/1471-2164/14/469

BMC Genomics

RESEARCH ARTICLE Open Access

Coexpression analysis of large cancer datasets provides insight into the cellular phenotypes of the tumour microenvironment

Tamasin N Doig[1,3], David A Hume[3], Thanasis Theocharidis[3], John R Goodlad[2], Christopher D Gregory[1] and Tom C Freeman[3*]

ARTICLE

doi:10.1038/nature13182

A promoter-level mammalian expression atlas

The FANTOM Consortium and the RIKEN PMI and CLST (DGT)*

nature

sample-to-sample correlation graphs

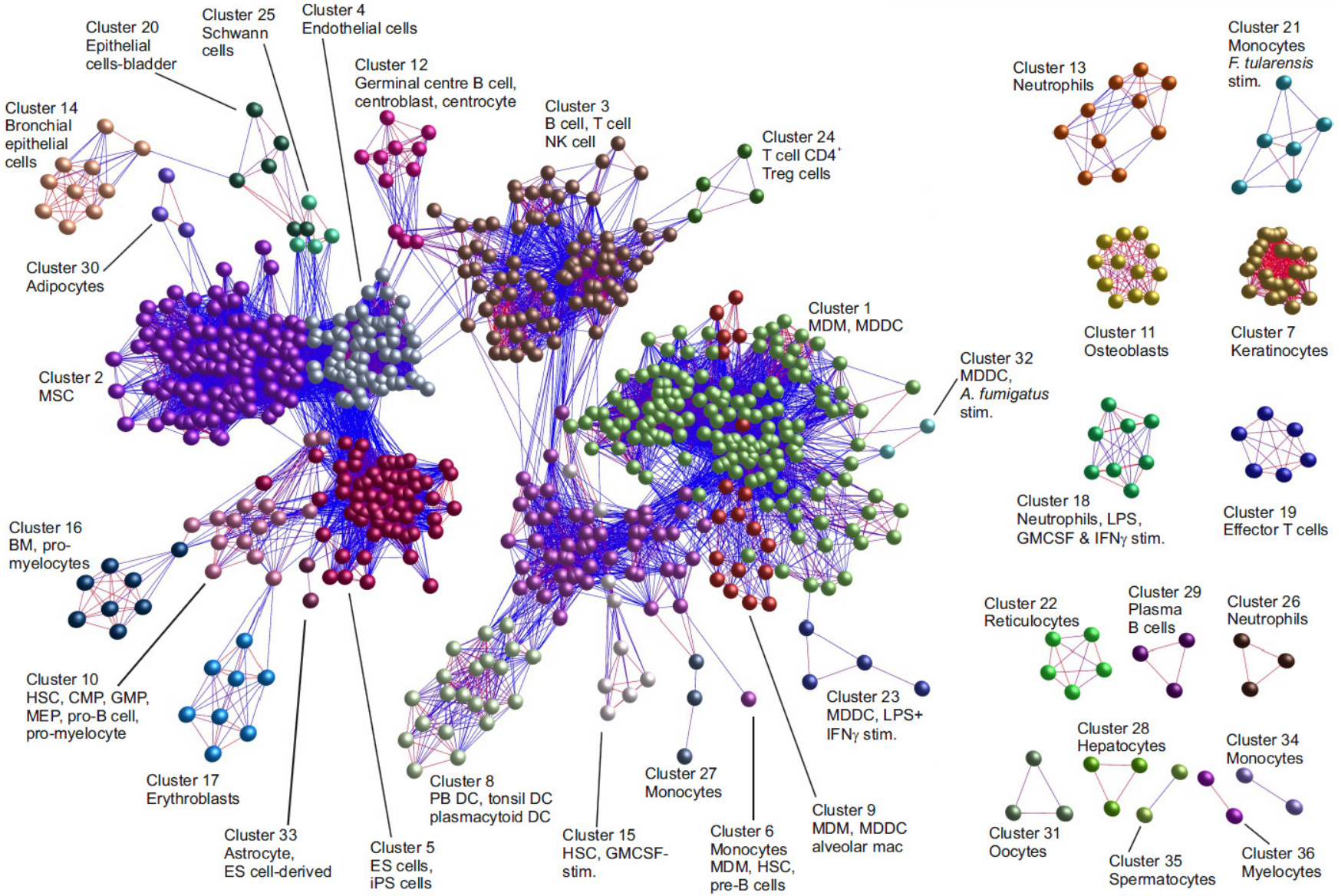
Mabbott et al. BMC Genomics 2013, 14:632
http://www.biomedcentral.com/1471-2164/14/632

BMC Genomics

RESEARCH ARTICLE Open Access

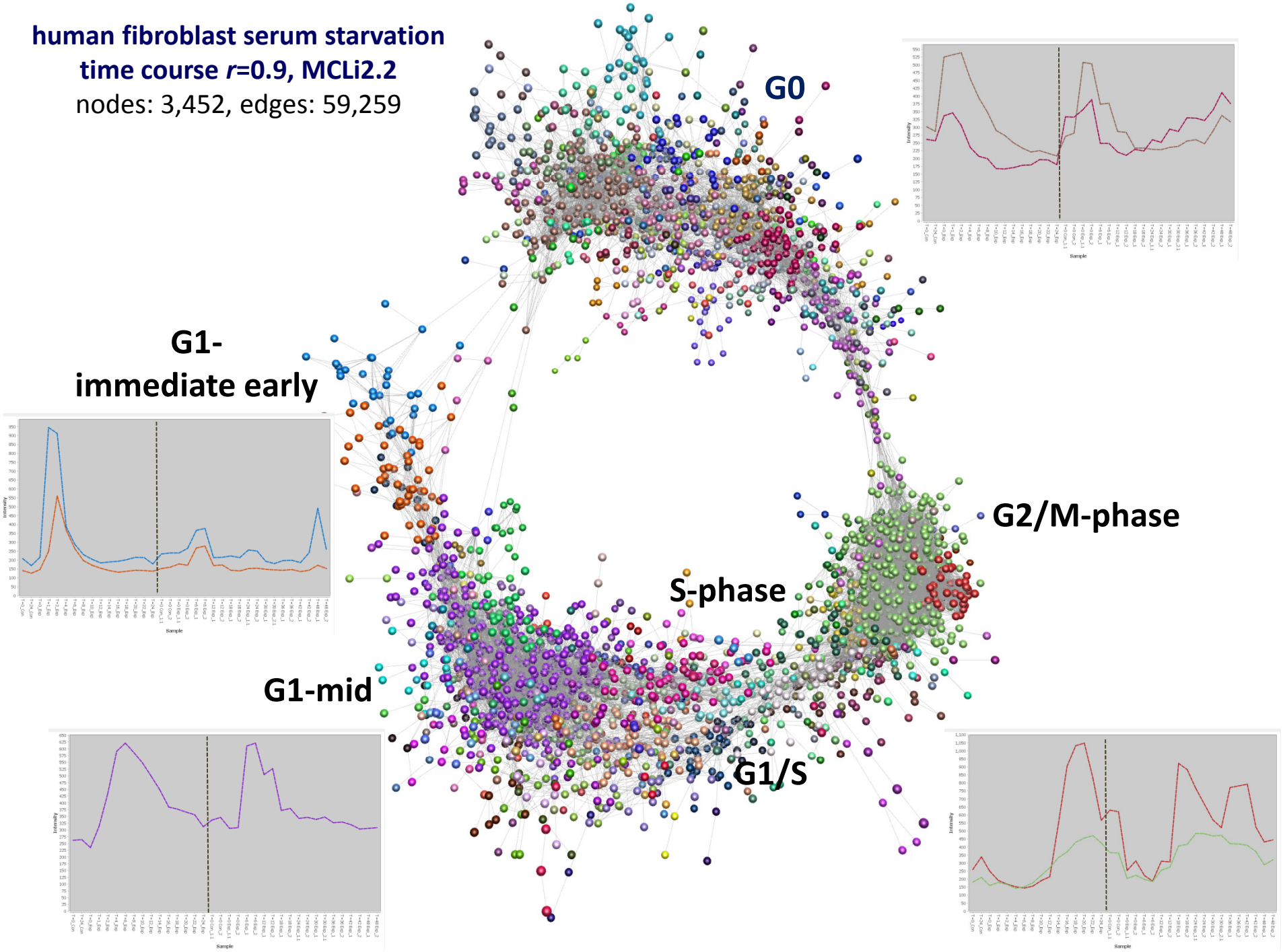
An expression atlas of human primary cells: inference of gene function from coexpression networks

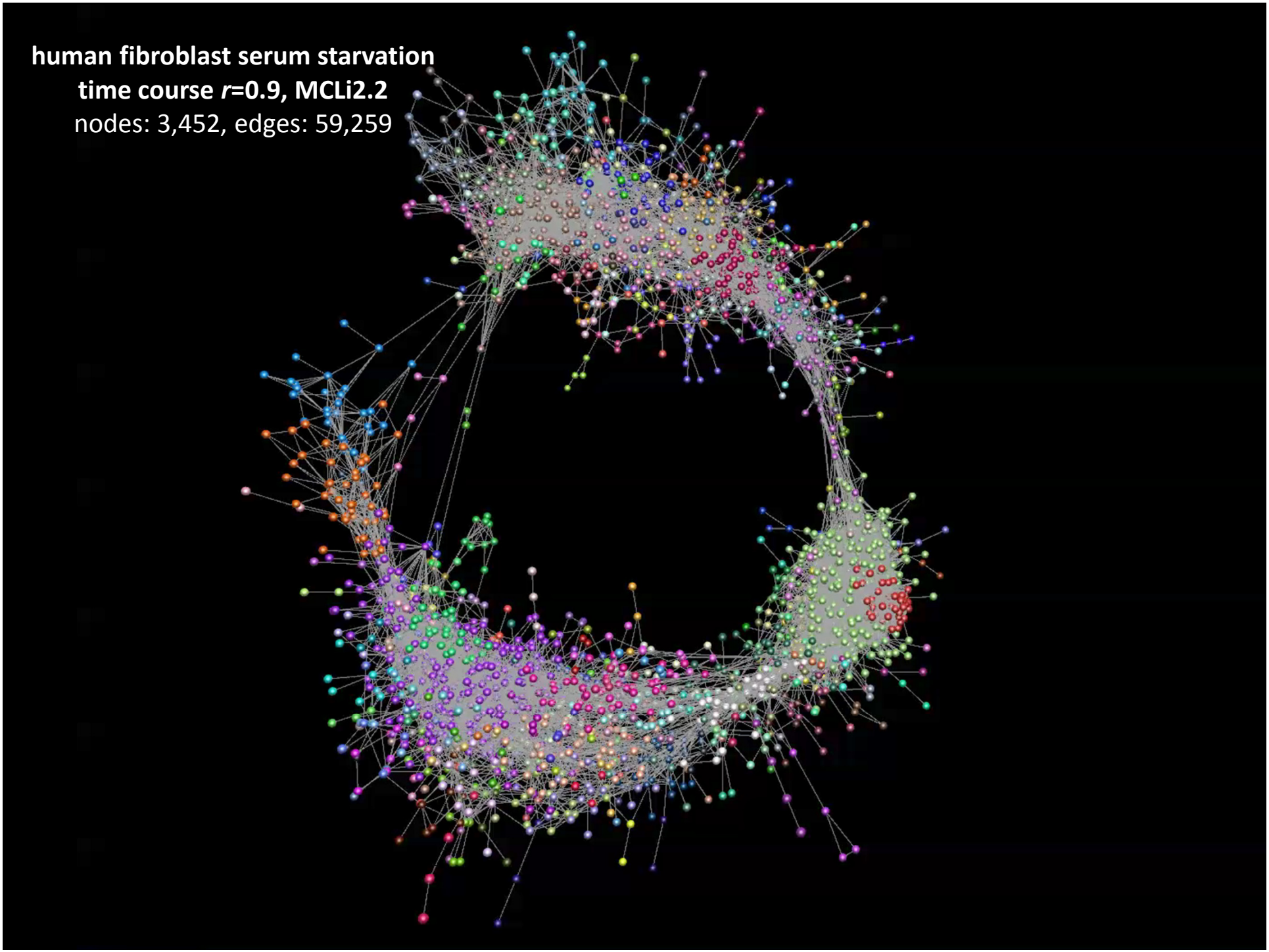
Neil A Mabbott*†, J Kenneth Baillie, Helen Brown, Tom C Freeman*† and David A Hume*†



human fibroblast serum starvation time course *r*=0.9, MCLi2.2

nodes: 3,452, edges: 59,259




human fibroblast serum starvation
time course *r*=0.9, MCLi2.2
nodes: 3,452, edges: 59,259

network graph diffuse large B-cell lymphoma (DLBCL) data

194 patient samples, comprised of ABC, GCB, PMBL and unclassified tumours

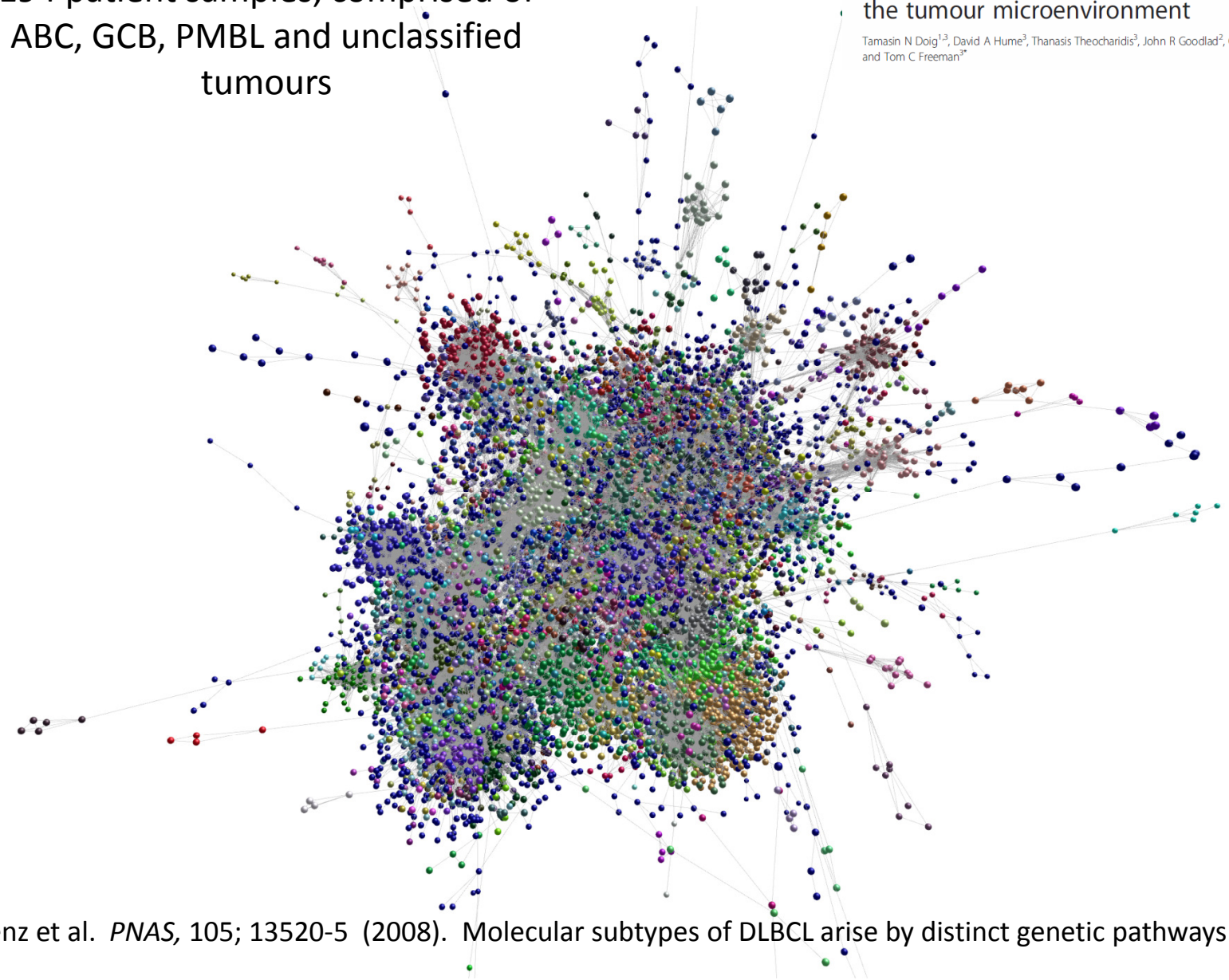
Doig et al. BMC Genomics 2013, 14:469
http://www.biomedcentral.com/1471-2164/14/469

BMC Genomics

RESEARCH ARTICLE Open Access

Coexpression analysis of large cancer datasets provides insight into the cellular phenotypes of the tumour microenvironment

Tamasin N Doig[1,3], David A Hume[3], Thanasis Theocharidis[3], John R Goodlad[2], Christopher D Gregory[1] and Tom C Freeman[3*]



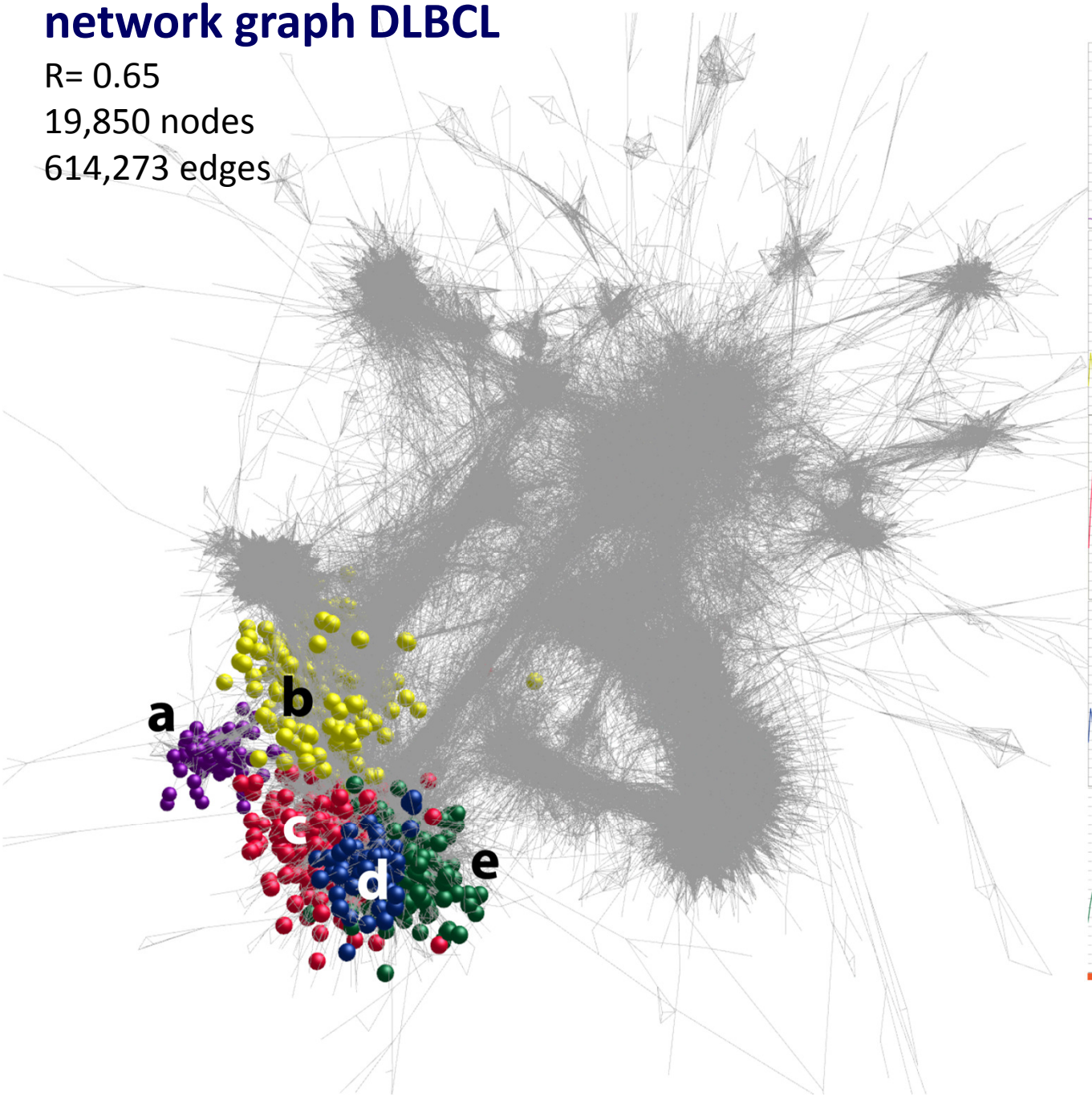
Lenz et al. *PNAS,* 105; 13520-5 (2008). Molecular subtypes of DLBCL arise by distinct genetic pathways.

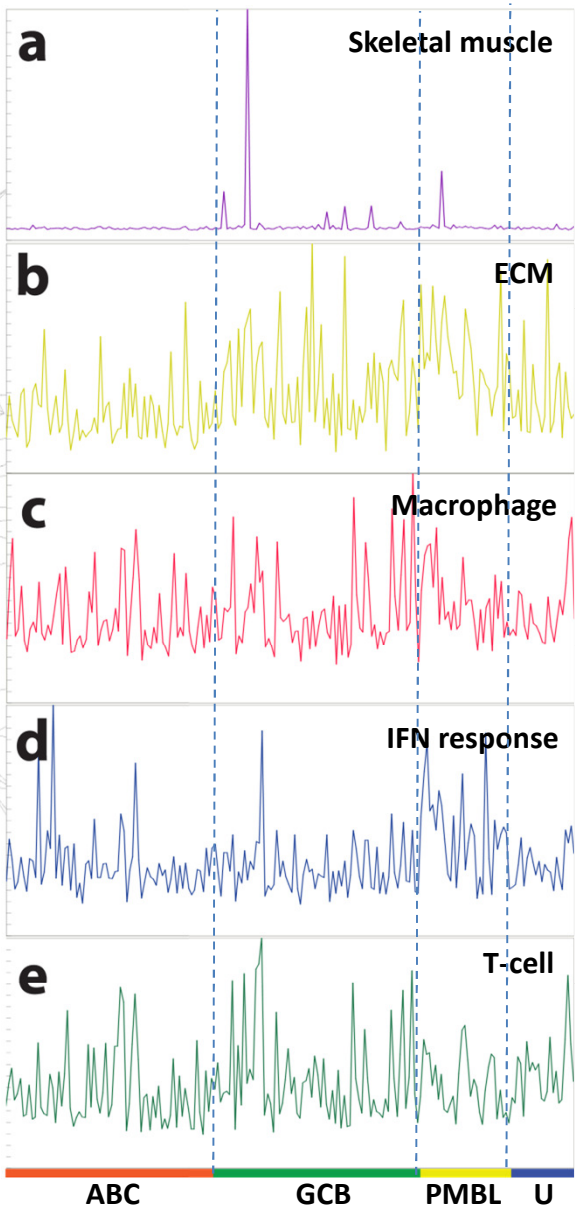
network graph DLBCL

R= 0.65

19,850 nodes

614,273 edges

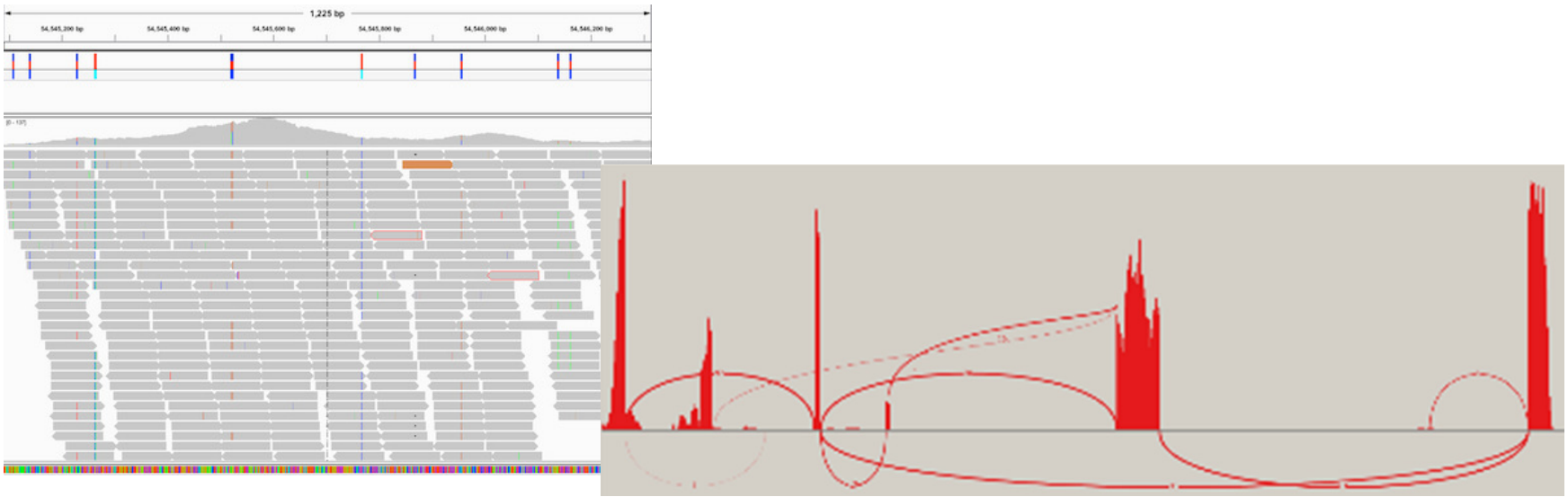




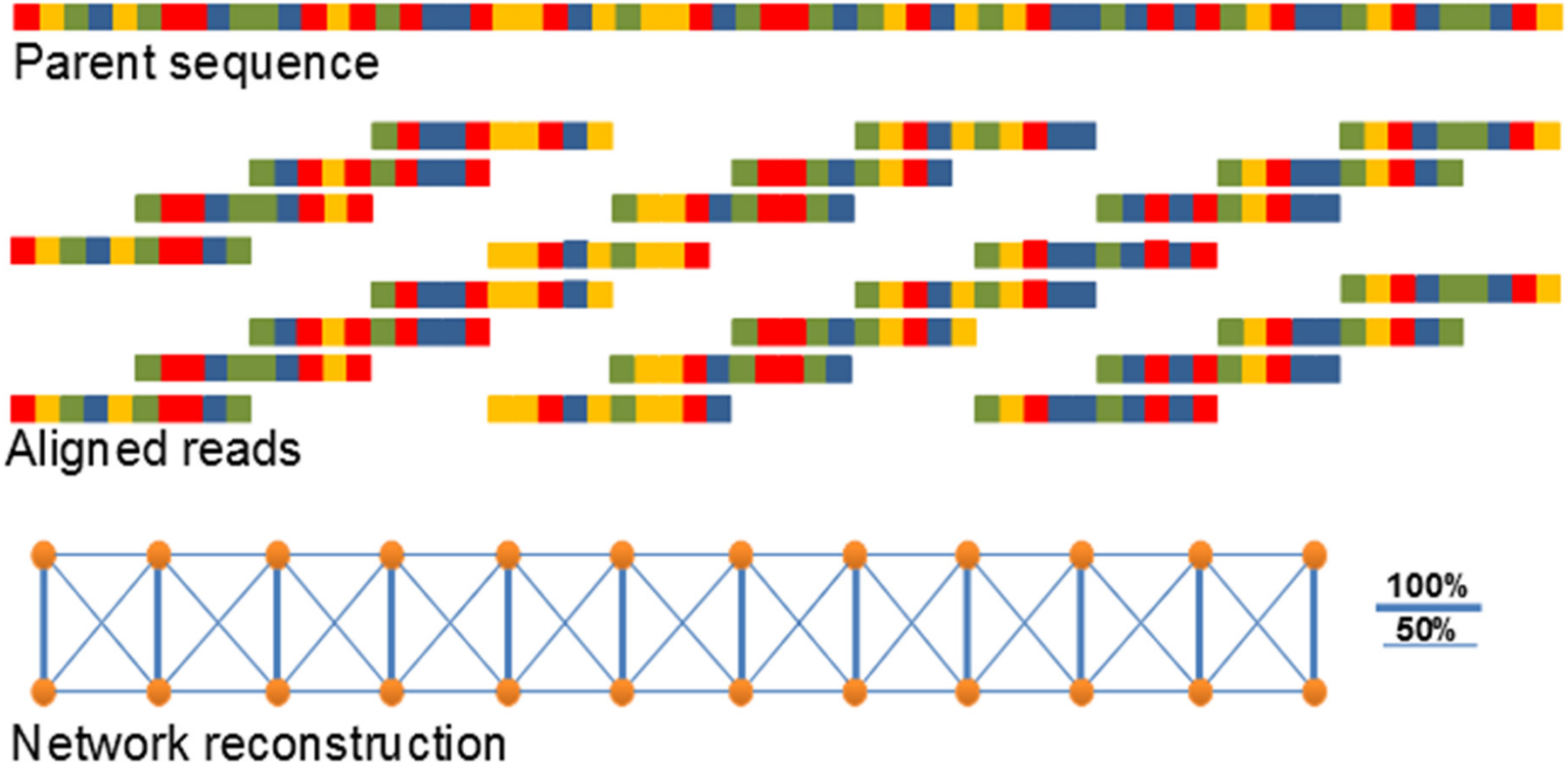
part 2
visualisation of DNA assemblies

Objectives

- to explore the use of graph-based approaches to analyse RNA-seq data
- to better understand the RNA to DNA alignments
- to better identify the splice-variants
- to develop a rapid analysis pipeline for such analyses

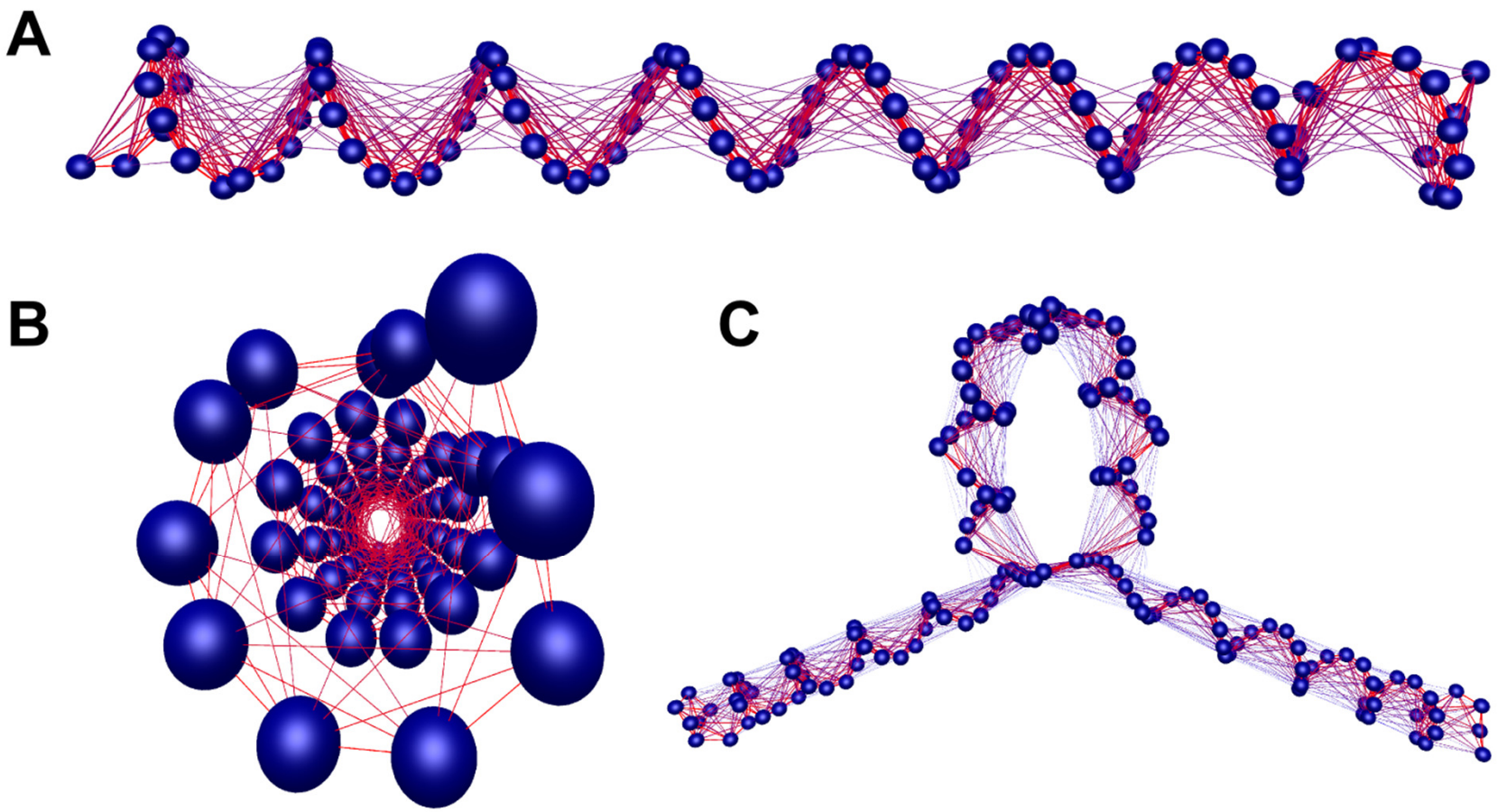


graph paradigm for DNA-seq data

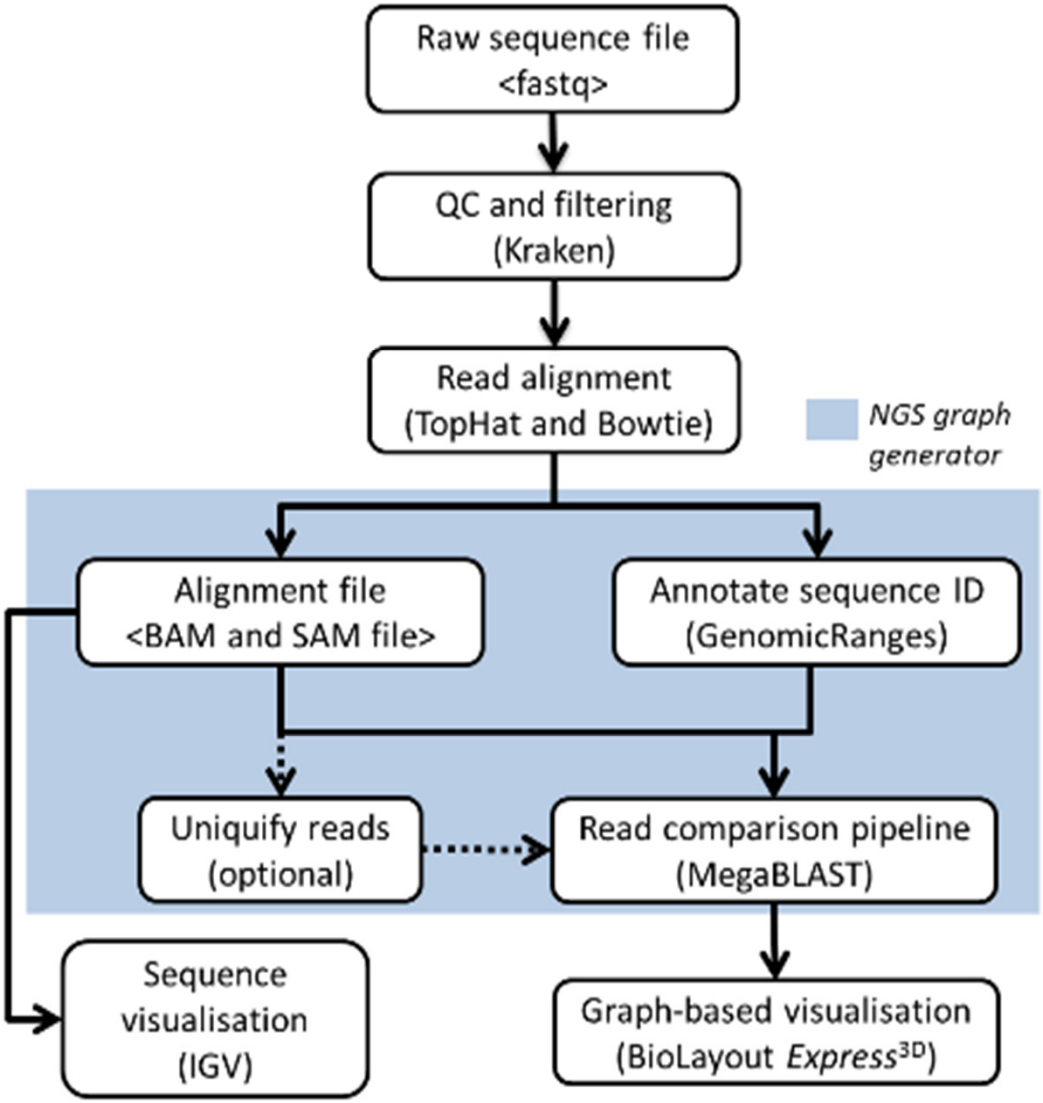


nodes represent reads, edges the similarity score between them as defined by megablast

visualisation of overlap graphs



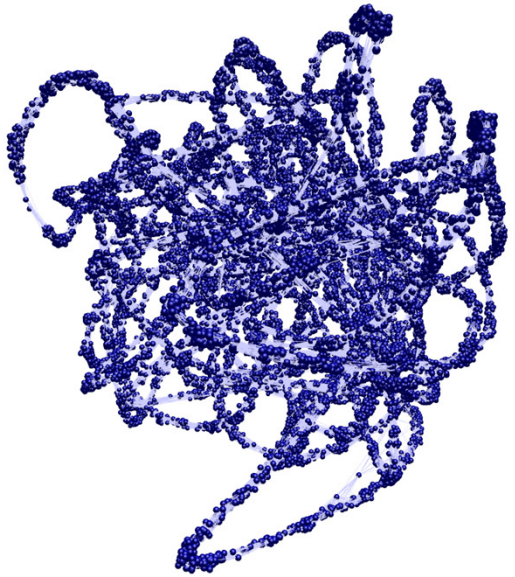
NGS graph-based analysis pipeline



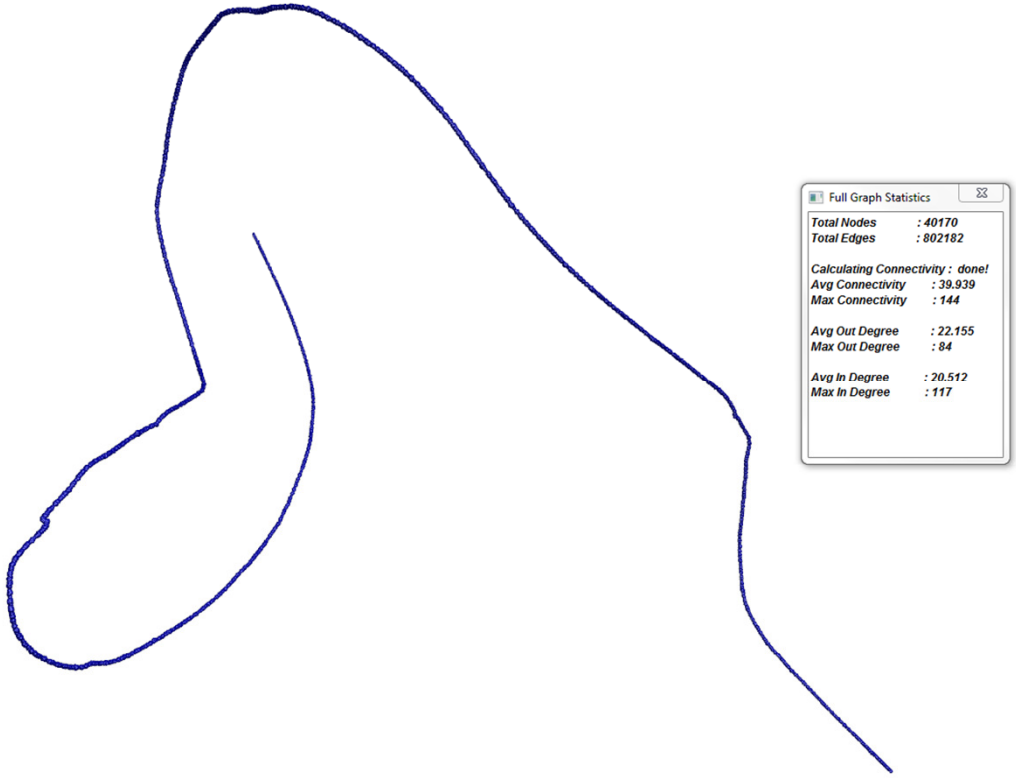
application to RNA-seq data

COL5A1

- one of the most abundantly expressed fibroblast genes
- 8.1kb, 66 exons
- >40,000 reads



graph layout
Fruchterman & Rheingold algorithm



fast multipole multilevel method (FMMM) layout

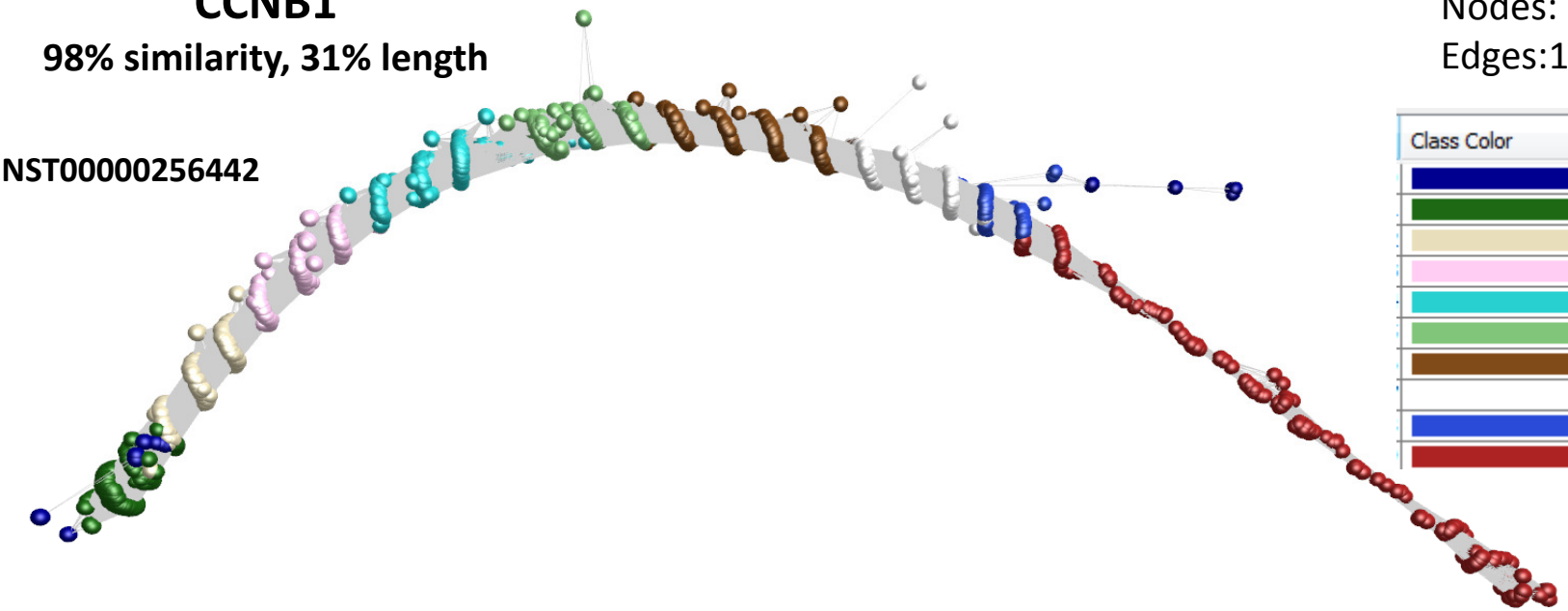
OGDF

CCNB1

98% similarity, 31% length

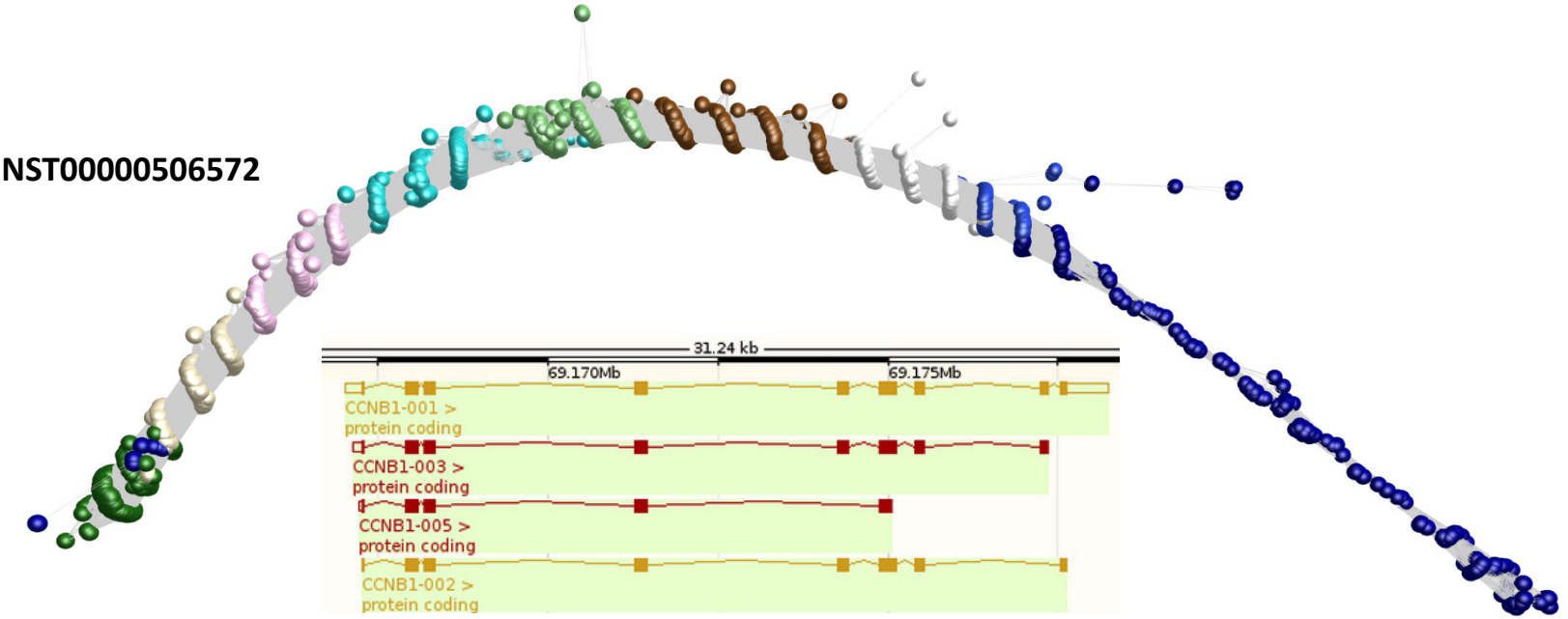
Nodes: 5,996
Edges:1,641,662

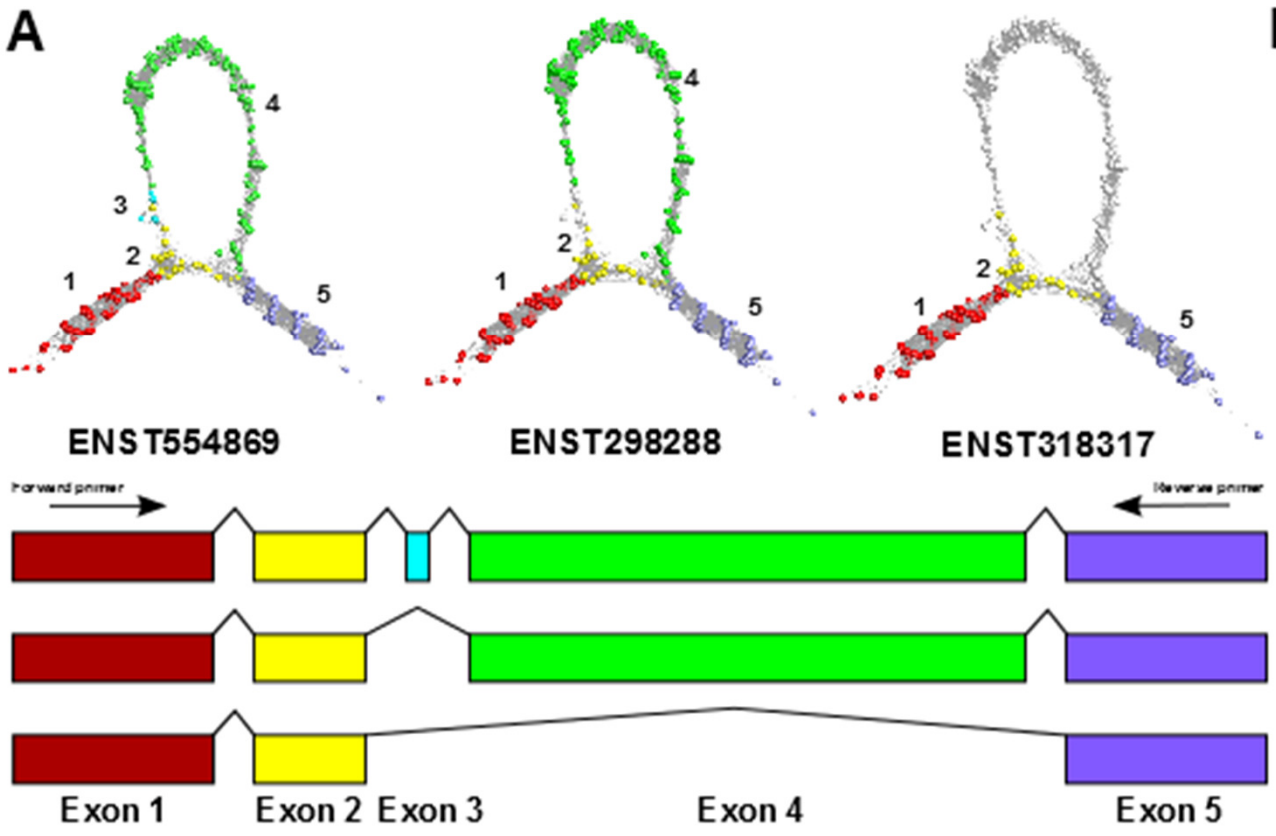
ENST00000256442

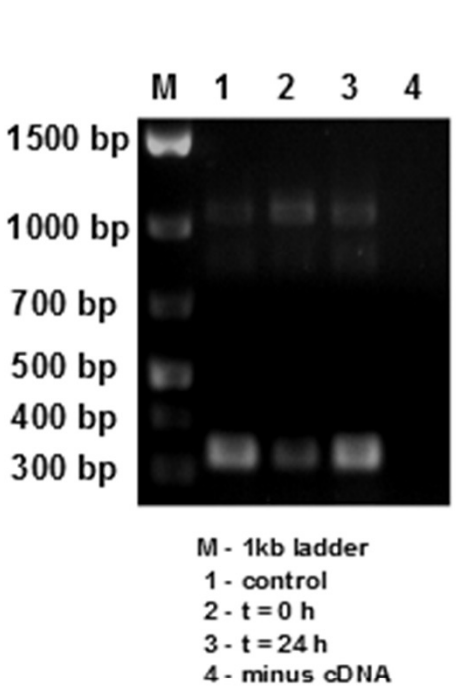


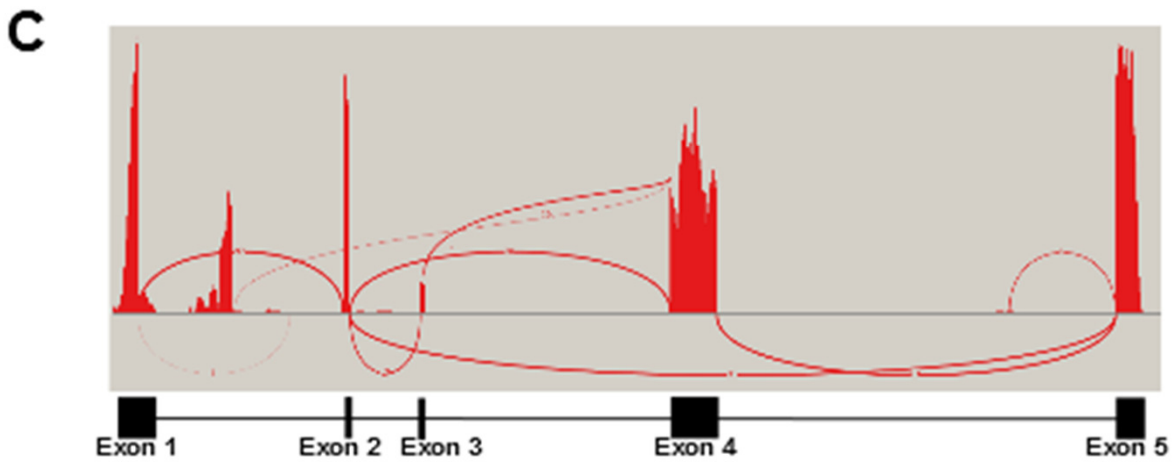
Class Color	Class Description
	No Class
	Exon 1
	Exon 2
	Exon 3
	Exon 4
	Exon 5
	Exon 6
	Exon 7
	Exon 8
	Exon 9

ENST00000506572



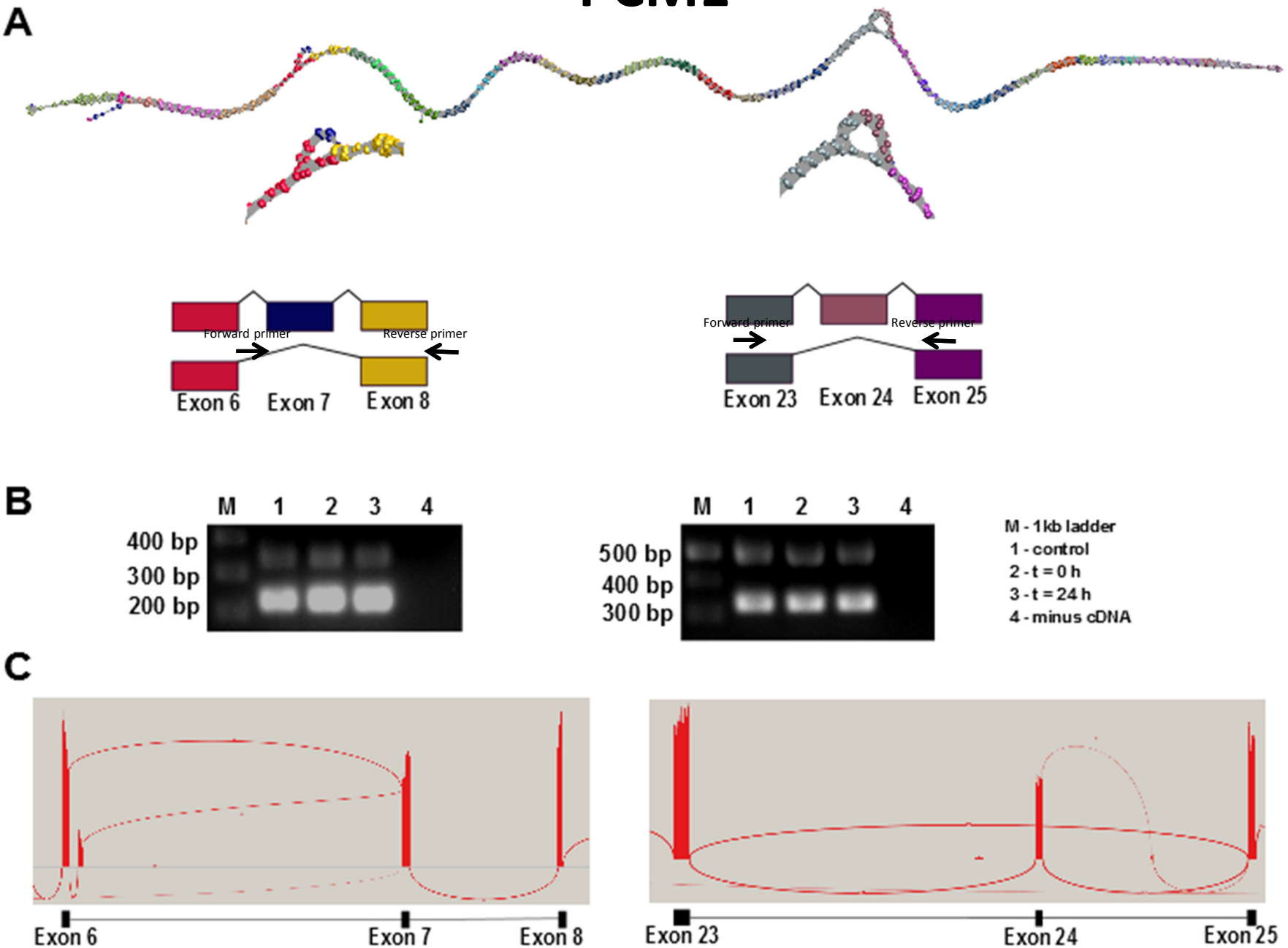

A
ENST554869
ENST298288
ENST318317
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5


B
M 1 2 3 4
1500 bp
1000 bp
700 bp
500 bp
400 bp
300 bp
M - 1kb ladder
1 - control
2 - t = 0 h
3 - t = 24 h
4 - minus cDNA


C
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5

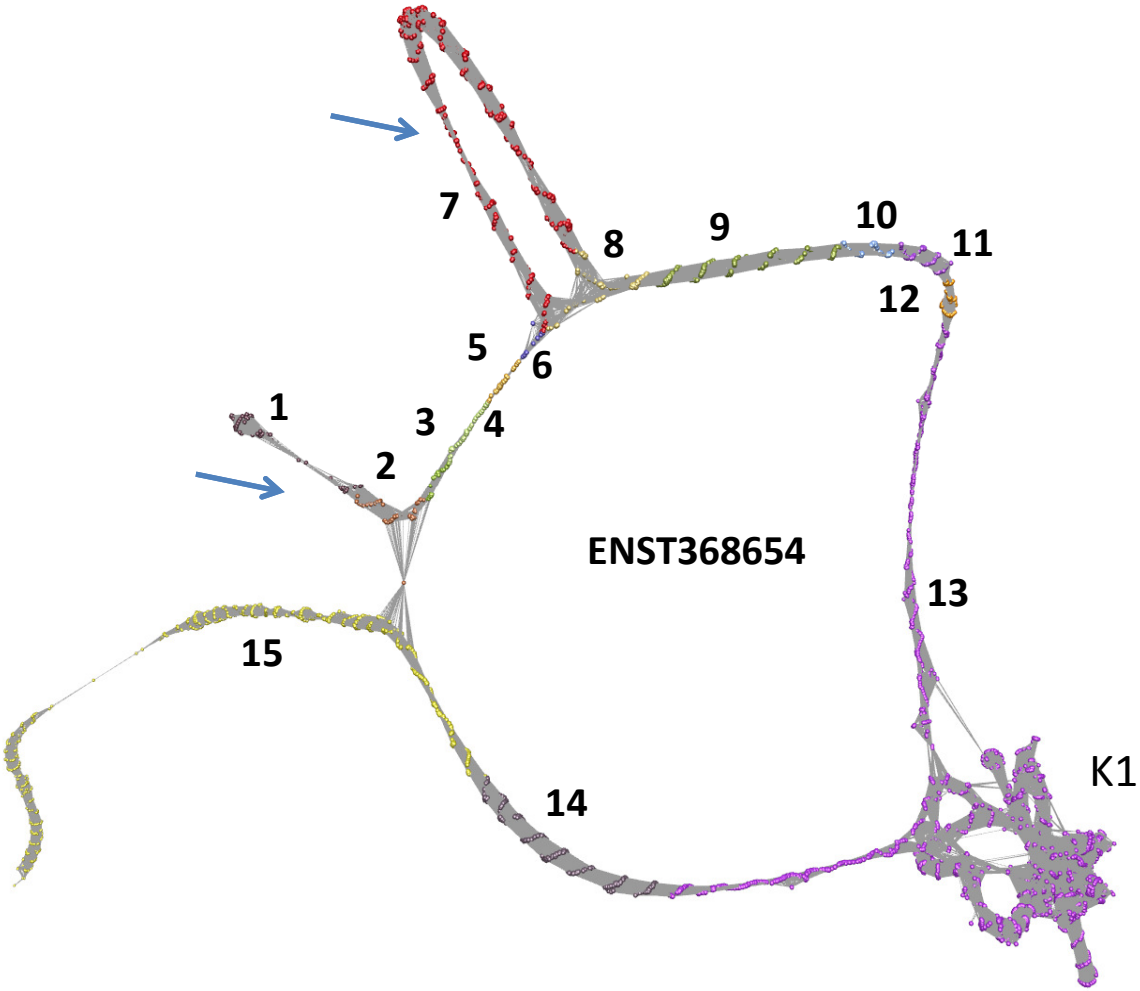
LRR1

PCM1

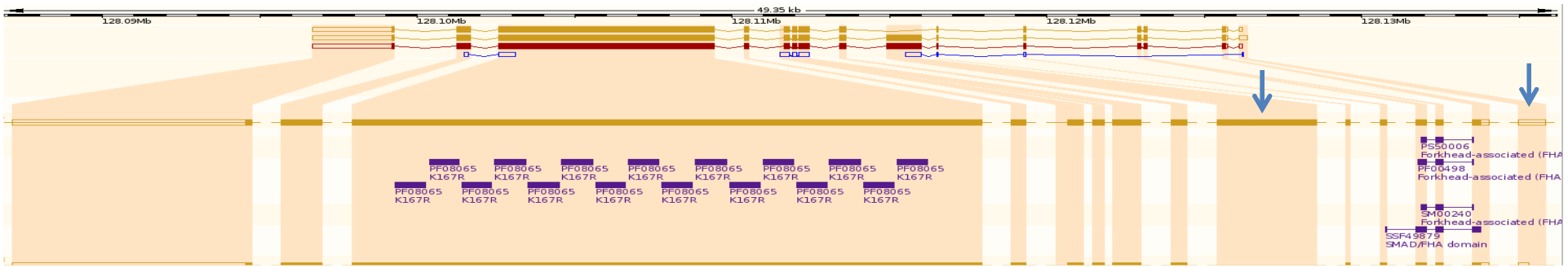

A
Forward primer
Reverse primer
Exon 6
Exon 7
Exon 8
Forward primer
Reverse primer
Exon 23
Exon 24
Exon 25
B
M 1 2 3 4
400 bp
300 bp
200 bp
M 1 2 3 4
500 bp
400 bp
300 bp
M - 1kb ladder
1 - control
2 - t = 0 h
3 - t = 24 h
4 - minus cDNA
C
Exon 6
Exon 7
Exon 8
Exon 23
Exon 24
Exon 25

MKI67

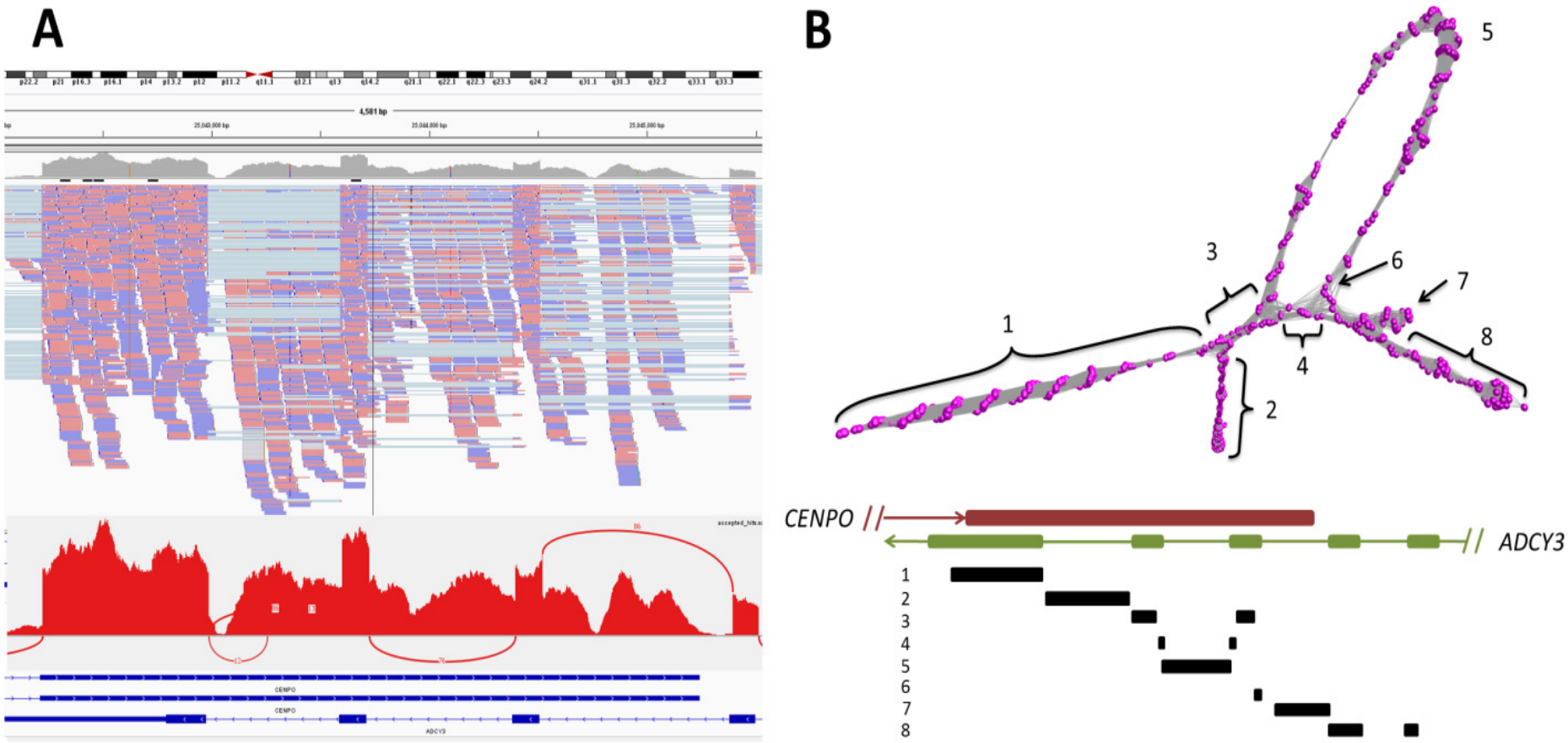
Nodes: 12,062,
Edges: 1,054,685



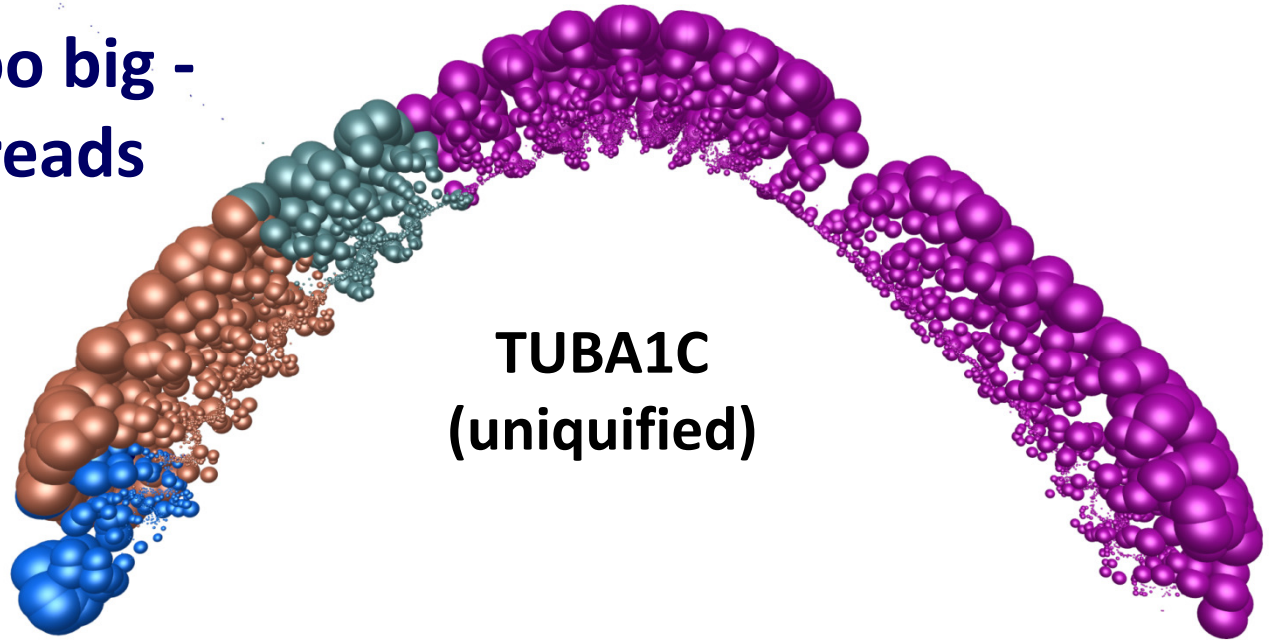
K167/Chmadrin repeats



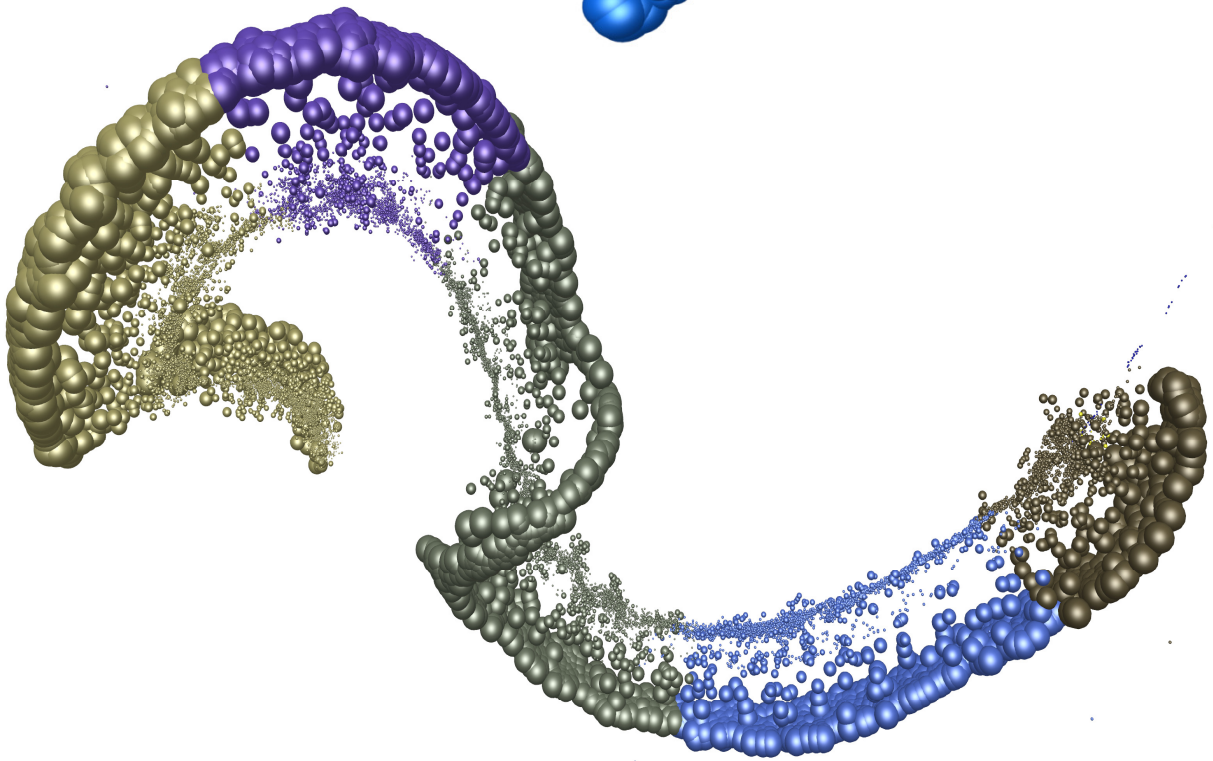
CENPO



When graphs get too big - uniquification of reads

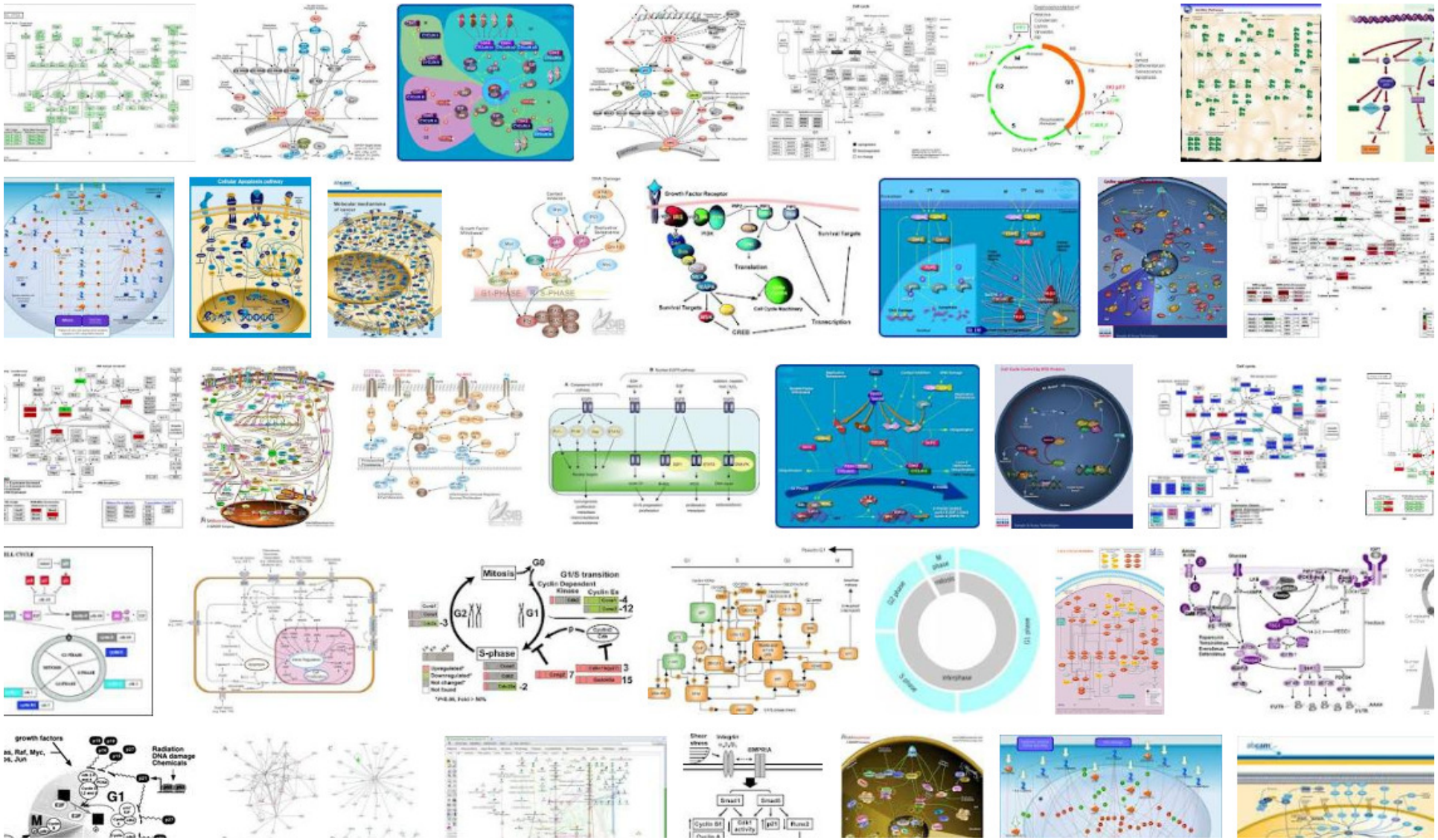


TUBA1C
(uniquified)



ACTB
(uniquified)

part 3
modelling of pathway knowledge

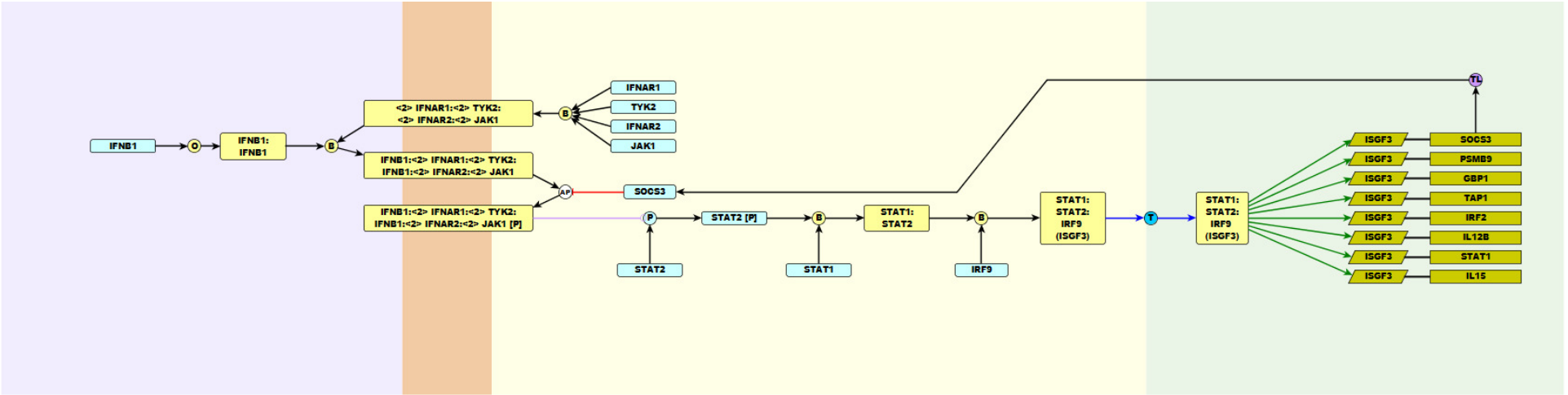


how can we better contemplate the complexity of biological systems? – build models

why?

- **to capture what you think you understand**
- **to map genes to proteins**
- **to place proteins in their site or context of action**
- **to capture a systems level appreciation of the pathway**
- **as a memory aid**

- ***to build a resource for the mapping of data, latest ideas, hypothesis generation, computation modelling***



mEPN Pathway Modelling Language (2104)

PLACES

ENTITY NODES

- PROTEIN COMPLEX
- PEPTIDE, PROTEIN
- GENE
- DNA
- DNA REGION (PROMOTER ELEMENT)
- RNA (Type defined by label e.g. [mRNA] etc)
- RNA REGION (PROMOTER ELEMENT)
- SIMPLE BIOCHEMICAL
- ION/SIMPLE MOLECULE
- GENERIC ENTITY
- DRUG
- ENERGY TRANSFER (CO-REACTION)

BOOLEAN LOGIC OPERATOR

- OR (OPERATRES AS PLACE)

SPN-SPECIFIC NODES

- SPACER NODE

COMPONENT ANNOTATION

PROTEIN 1
[Mod]
(ALIAS)

PROTEIN 1:
<n> PROTEIN 2 [Mod]
(ALIAS)

Protein/Complex State

[A] ACTIVE
[I] INACTIVE
<n> NUMBER OF DEFINED ENTITY

Protein Modifications

[P] PHOSPHORYLATED
[Ub] UBIQUITINATED
[Su] SUMOLAYTED
[Ac] ACETYLATED
[Pr] PRENYLATED
[H] PROTONATED
[Pe] PEGYLATED
[Ox] OXIDISED
[Gy] GLYCOSYLATED
[Me] METHYLATED
[Pa] PALMITOYLATED
[OH] HYDROXYLATED
[S] SULPHATED
[My] MYRISTOYLATED
[Se] SELENYLATED
[t] TRUNCATED

TRANSITIONS

PROCESS NODES

BINDING	B	O	OLIGERMISATION
DISSOCIATION	D	TL	TRANSCRIPTION/TRANSLATION
TRANSLOCATION	T	A	ACTIVATION
MUTATION	M	S	SECRETION
UNKNOWN	?	/	SINK (PROTERSOMAL DEGRADATION)
CLEAVAGE	X	AX	AUTO-CLEAVAGE
CATALYSIS	C	AC	AUTO-CATALYSIS
PHOSPHORYLATION	P	-P	DEPHOSPHORYLATION
AUTO-PHOSPHORYLATION	AP	PT	PHOSPHO-TRANSFER
UBIQUITISATION	Ub	Su	SUMOYLATION
SELENYLATION	Se	Gy	GLYCOSYLATION
PRENYLATION	Pr	Me	METHYLATION
ACETYLATION	Ac	Pa	PALMITOYLATION
PROTONATION	H+	Su	SULPHATION
PEGYLATION	Pe	My	MYRISTOYLATION
OXIDATION	Ox	OH	HYDROXYLATION

PATHWAY MODULE/OUTPUT

BOOLEAN LOGIC OPERATOR

AND (OPERATES AS TRANSITION) &

SPN-SPECIFIC TRANSITIONS

TOKEN INPUT

DISTRIBUTION NODE (20, 20)

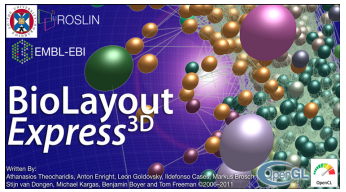
SPACER NODE/TOKEN OUTPUT (10, 10)

EDGES

EDGE ANNOTATION

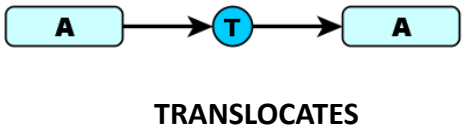
- INTERACTION
- CATALYSIS
- ACTION POTENTIAL
- NON COVALENT INTERACTION
- INHIBITION
- INHIBITION (NON-COMPETITIVE)
- INHIBITION (COMPETITIVE)

yED yWorks

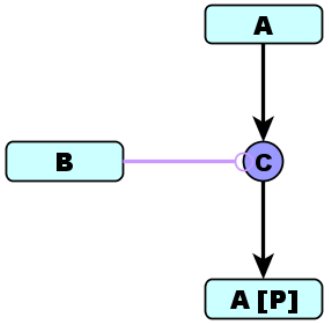


Freeman et al., 2010

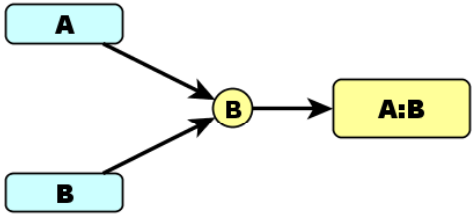
basics of mEPN usage



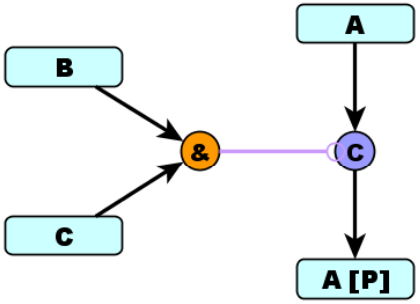
TRANSLOCATES



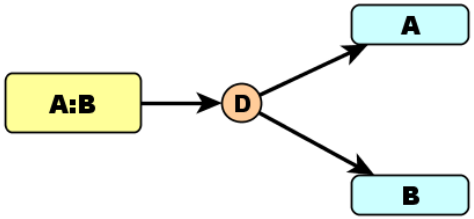
B CATALYSES THE PHOSPORYLATION OF A



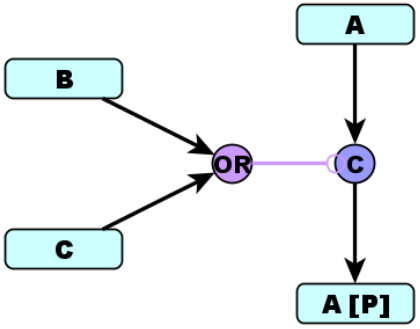
BINDS



B AND C CATALYSE THE PHOSPORYLATION OF A



DISSOCIATES



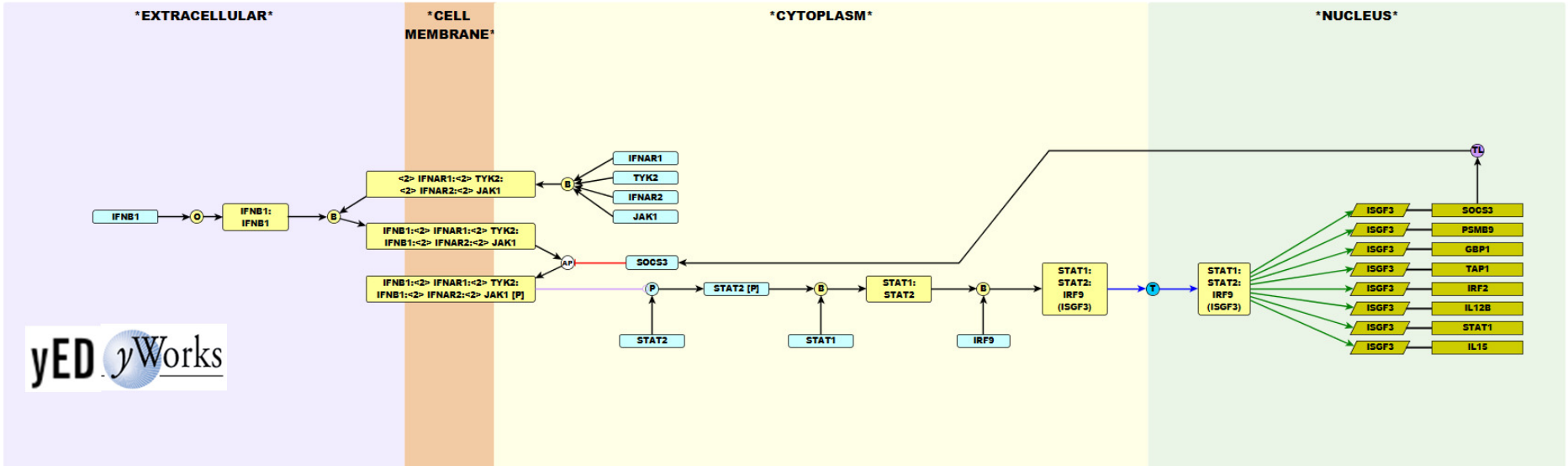
B OR C CATALYSE THE PHOSPORYLATION OF A

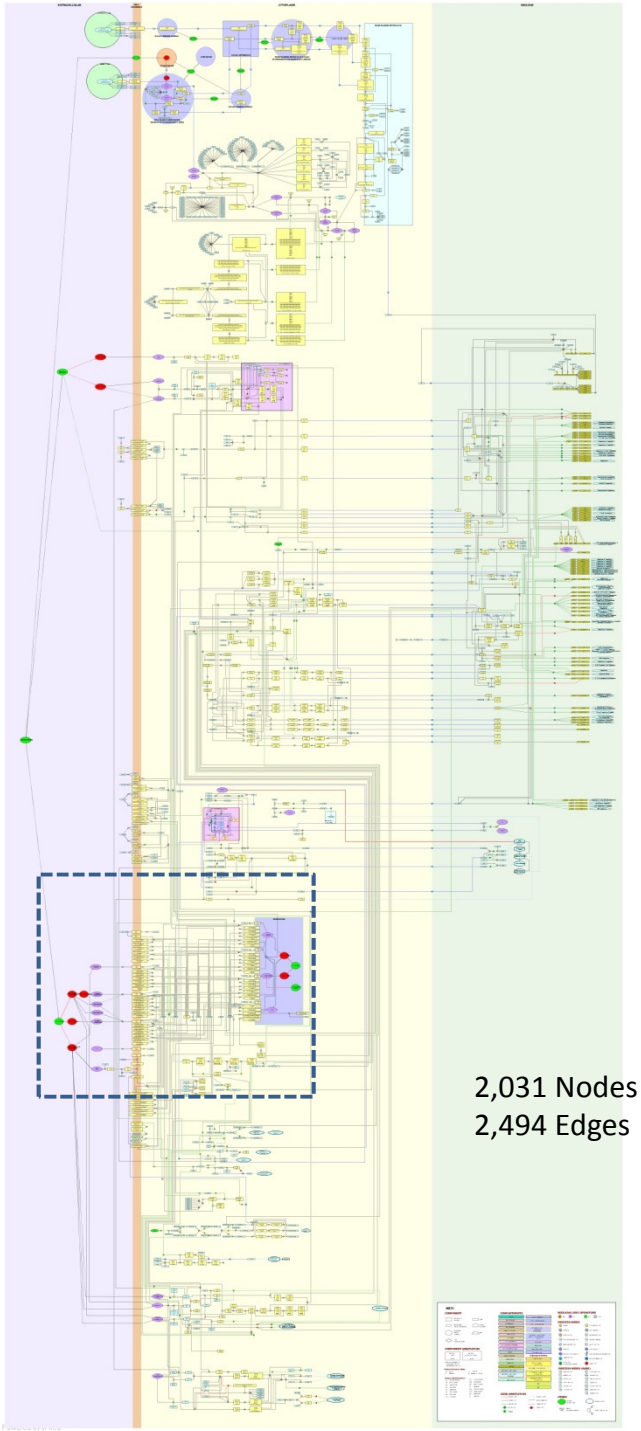
turning text into models

Pathway Description (**Components** – **Interactions** – **Compartments**)

Interferon beta (**IFNB1**) **homodimerises** and **binds** to its **cell surface** receptor complex, composed of the **transmembrane** proteins **IFNAR1** and **IFNAR2** and the **intracellular** kinases **TYK2** and **JAK1**. The complex is composed of 2 of each of these proteins. **Binding** causes a conformational change in the complex, resulting in the **autophosphorylation** of **JAK1**. Once activated, the complex catalyses the **phosphorylation** of **STAT2**, which forms a heterodimer with **STAT1**. This complex then **binds** interferon regulatory factor 9 (**IRF9**) forming the complex often referred to as **ISGF3** and **translocates** to the **nucleus**. Here it **activates the transcription** of a number of genes including **IRF2**, **IL12B**, **STAT1**, **IL15**, **TAP1**, **GBP1**, **PSMB9**, and **SOCS3**. In turn, **SOCS3** **inhibits** the **autophosphorylation** of the receptor, thereby preventing further activation.

graphical representation

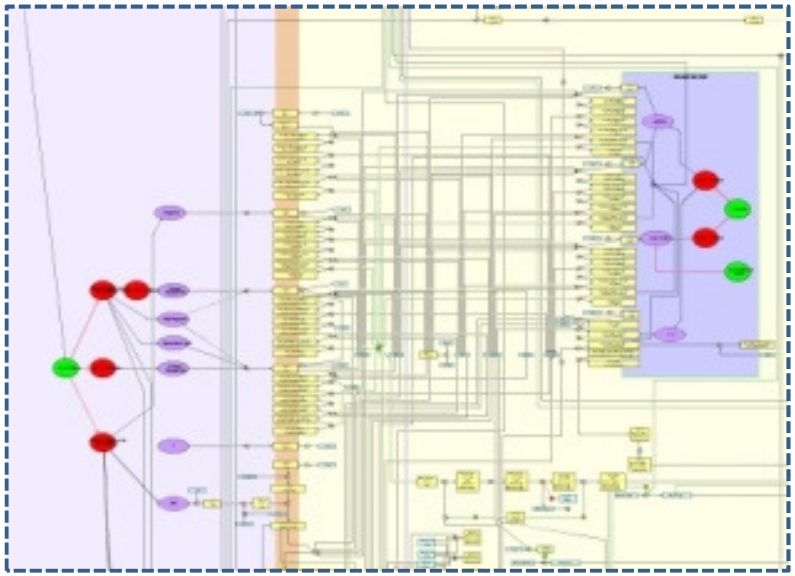


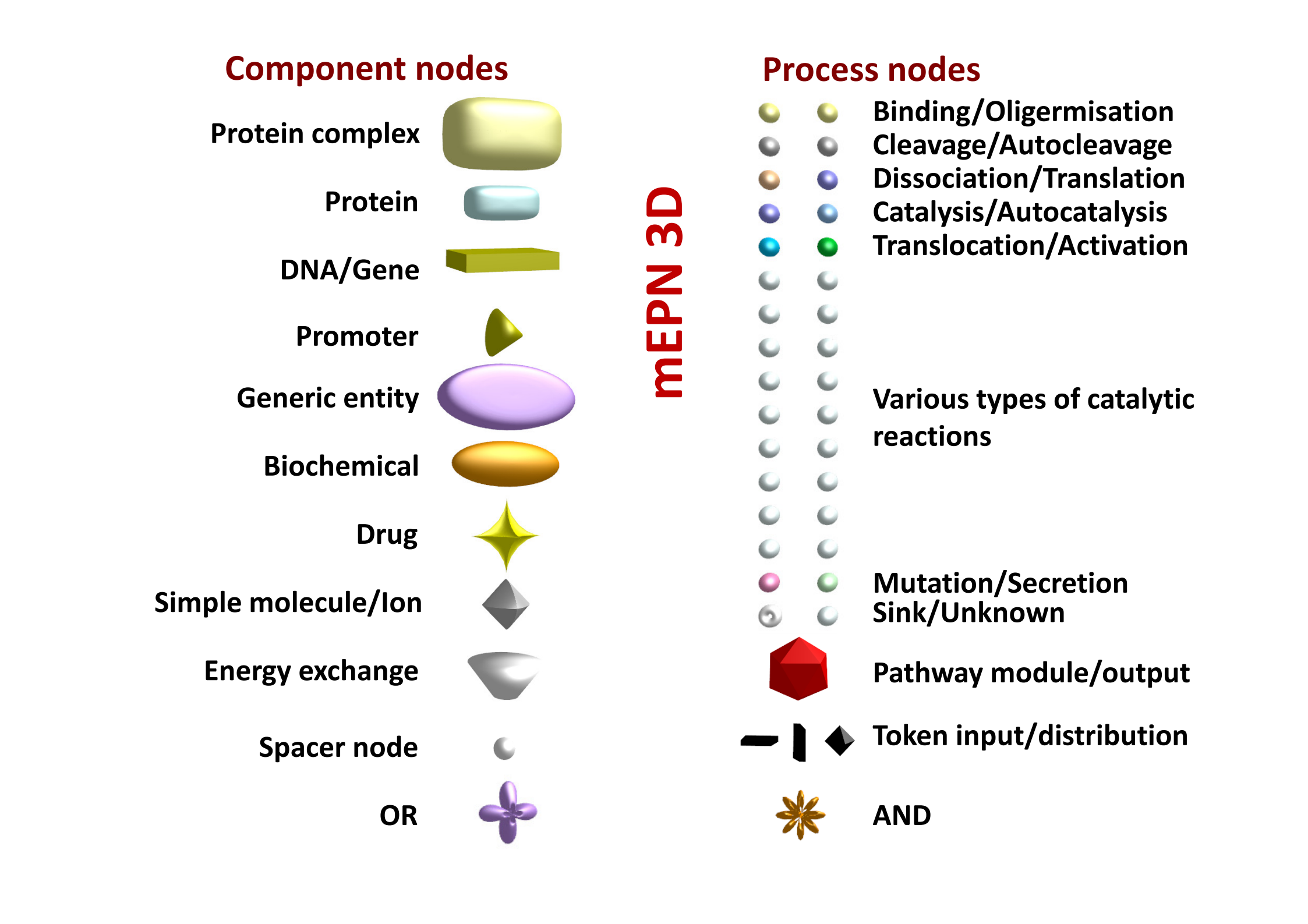


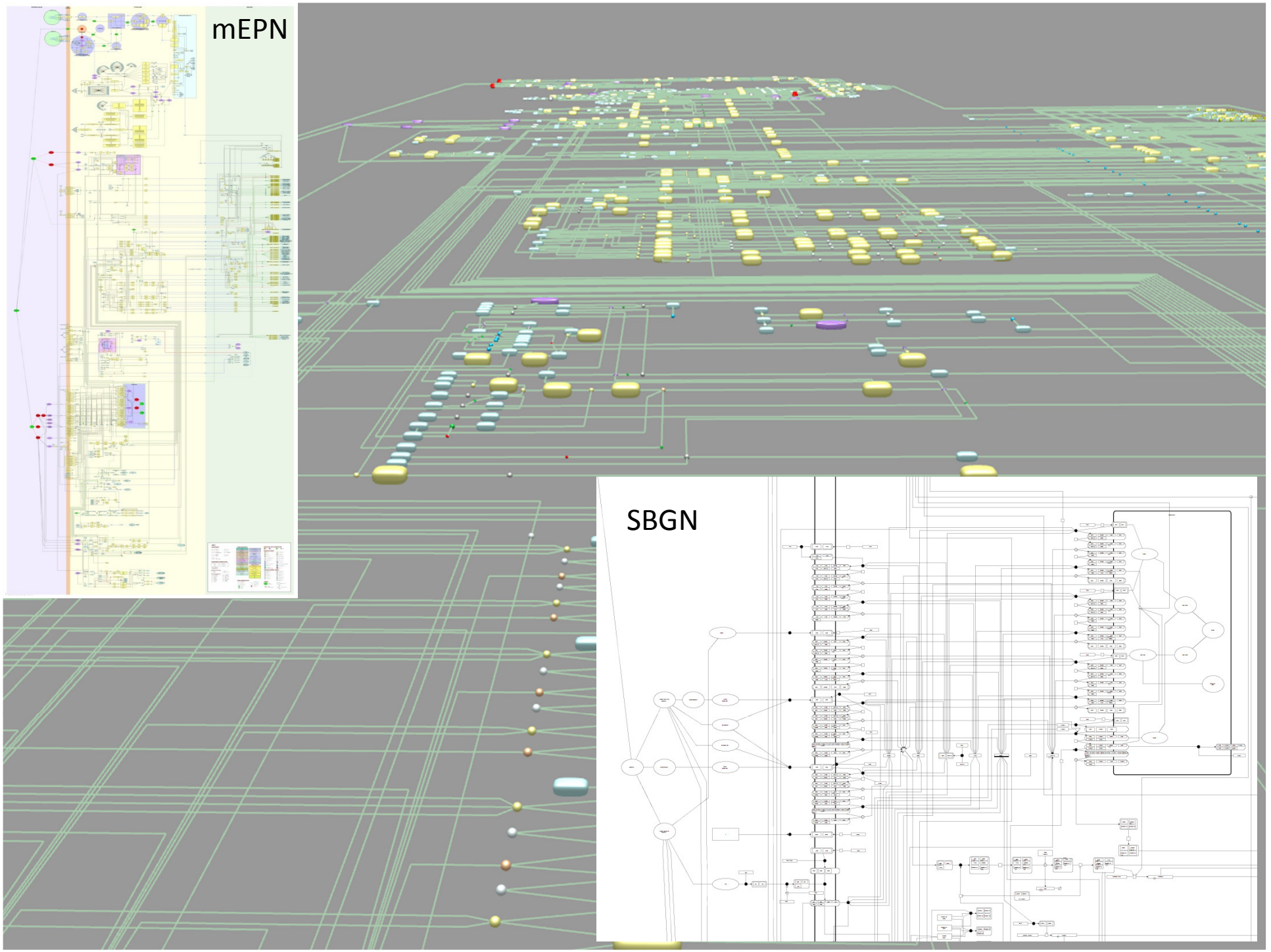
2,031 Nodes
2,494 Edges

integrated view of macrophage activation and effector pathways

Raza et al., BMC Systems Biology 4:63 2010
(available from www.macrophages.com)




Component nodes
Protein complex
Protein
DNA/Gene
Promoter
Generic entity
Biochemical
Drug
Simple molecule/Ion
Energy exchange
Spacer node
OR
mEPN 3D
Process nodes
Binding/Oligermisation
Cleavage/Autocleavage
Dissociation/Translation
Catalysis/Autocatalysis
Translocation/Activation
Various types of catalytic reactions
Mutation/Secretion
Sink/Unknown
Pathway module/output
Token input/distribution
AND


mEPN
SBGN

Signaling Petri Net:
stochastic flow simulation

- a Signaling Petri Net (SPN) is a Petri net-related method proposed recently for simulations of biological pathways

  See:

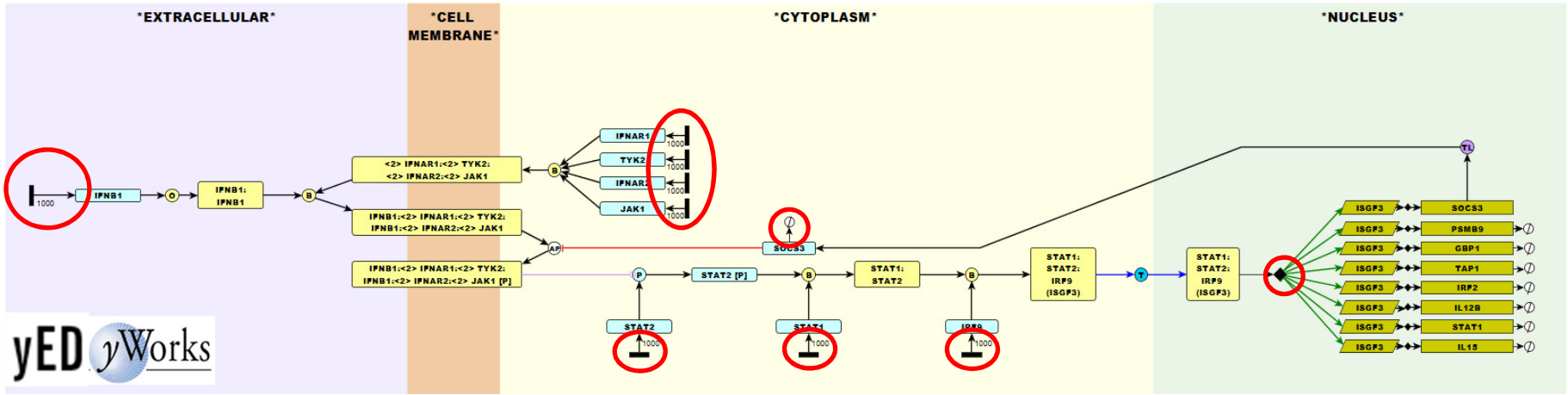
  Ruths *et al.* BMC Systems Biology 2:76 (2008); PLoS Comp Biol. 4:76 (2008)

- it is an alternate version of Petri net and the algorithm models the stochastic flow of a *variable* number of tokens

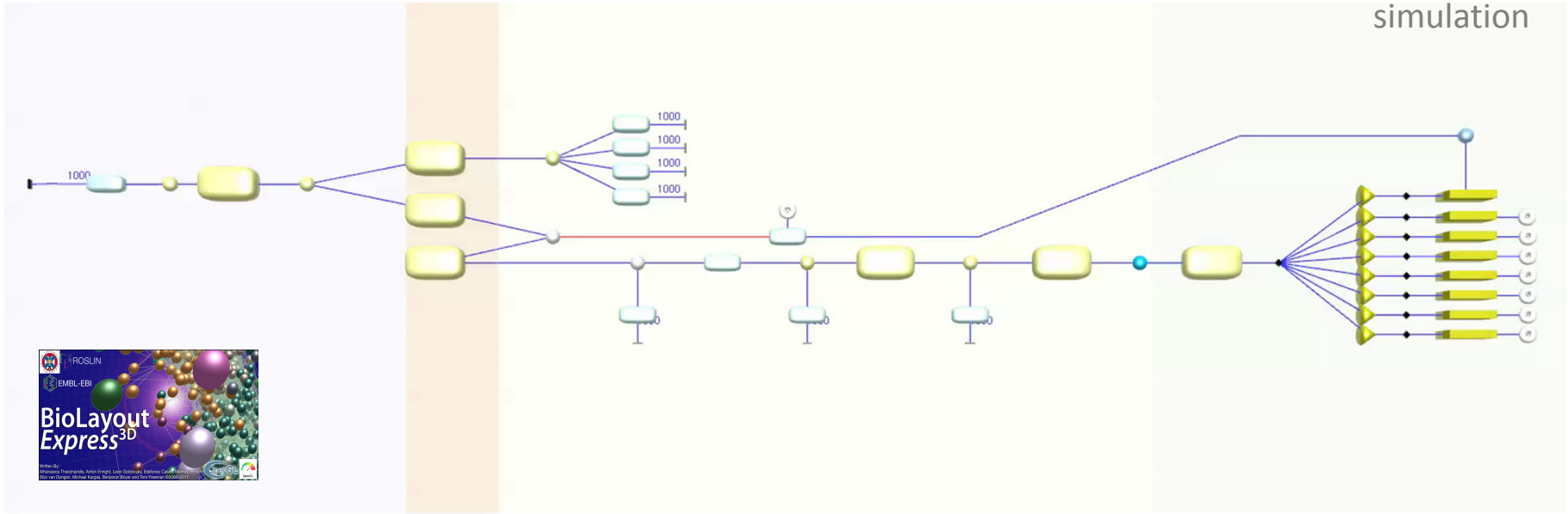
- it doesn't need kinetic parameters of reactions/transitions

- it’s fast and intuative!

parameterisation



simulation



summary

- if you have complex data or dealing with complex biological systems – think networks

- BioLayout *Express*^{3D} provides fast and intuitive platform to explore high dimensional data and examine its 'structure' in a hypothesis-free manner – designed to be biologist friendly i.e. GUI driven but takes time to learn

- graphical modeling of pathway systems very useful for recording known biology of a system – *although not easy and time consuming*

- useful for teaching, data interpretation, hypothesis generation and computation modeling

- now developing it for analysis of other data types i.e. sequencing data and live cell recording and applications outside of biology

Acknowledgements

Systems Immunology Group, The Roslin Institute

Sobia Raza
Fahmi Wan Mohamad Nazarie
Bruno Giotti
Mark Barnett
David Chen

BioLayout *Express*^{3D}

Thanasis Theocharidis, Tim Angus, Derek Wright, The Roslin Institute

Anton Enright, Stijn van Dongen, Matloob Qureshi, EBI

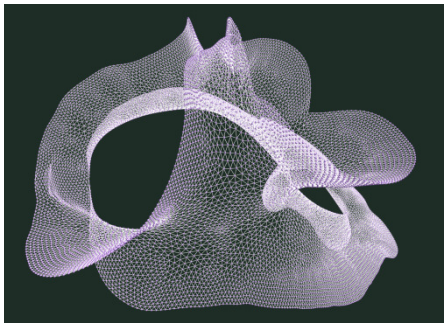
Info Viz Res Group, Sydney

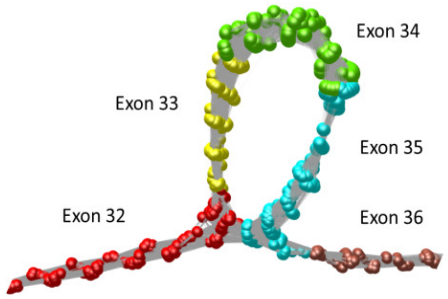
Karsten Klein

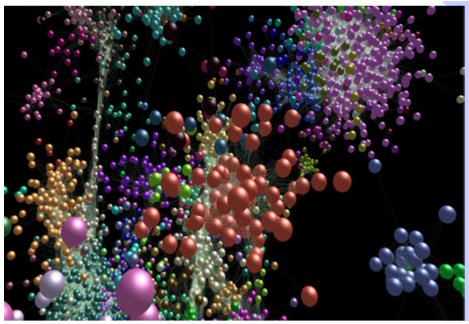
www.biolayout.org

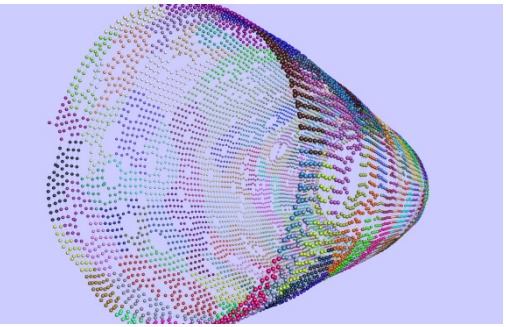
www.macrophages.com

www.virtuallyimmune.org









tom.freeman@roslin.ed.ac.uk