STEMFORMATICS

The Stemformatics Virtual Lab

More than genomic data visualisation in the cloud

19 October 2017 eResearch Australasia

Overview

Introduction

Data visualisation

Collaboration

Trusted Data

Data Mining

Future of Stemformatics

Stemformatics is...

- A web based pocket dictionary
- An Atlas to benchmark against 350+ public datasets, manually curated, high quality
- A host of private datasets
- A resource focused on helping researchers
- Rely on NeCTAR for all of our websites

Human Gene Expression

Landscape of the bioinformatics ecosystem by Rowland Mosbergen

Biologists

- Want something easy to use
- Minimal bioinformatics knowledge
- Want to benchmark their work
- Don't want to manipulate data
- Happy with default settings

Command line bioinformaticians

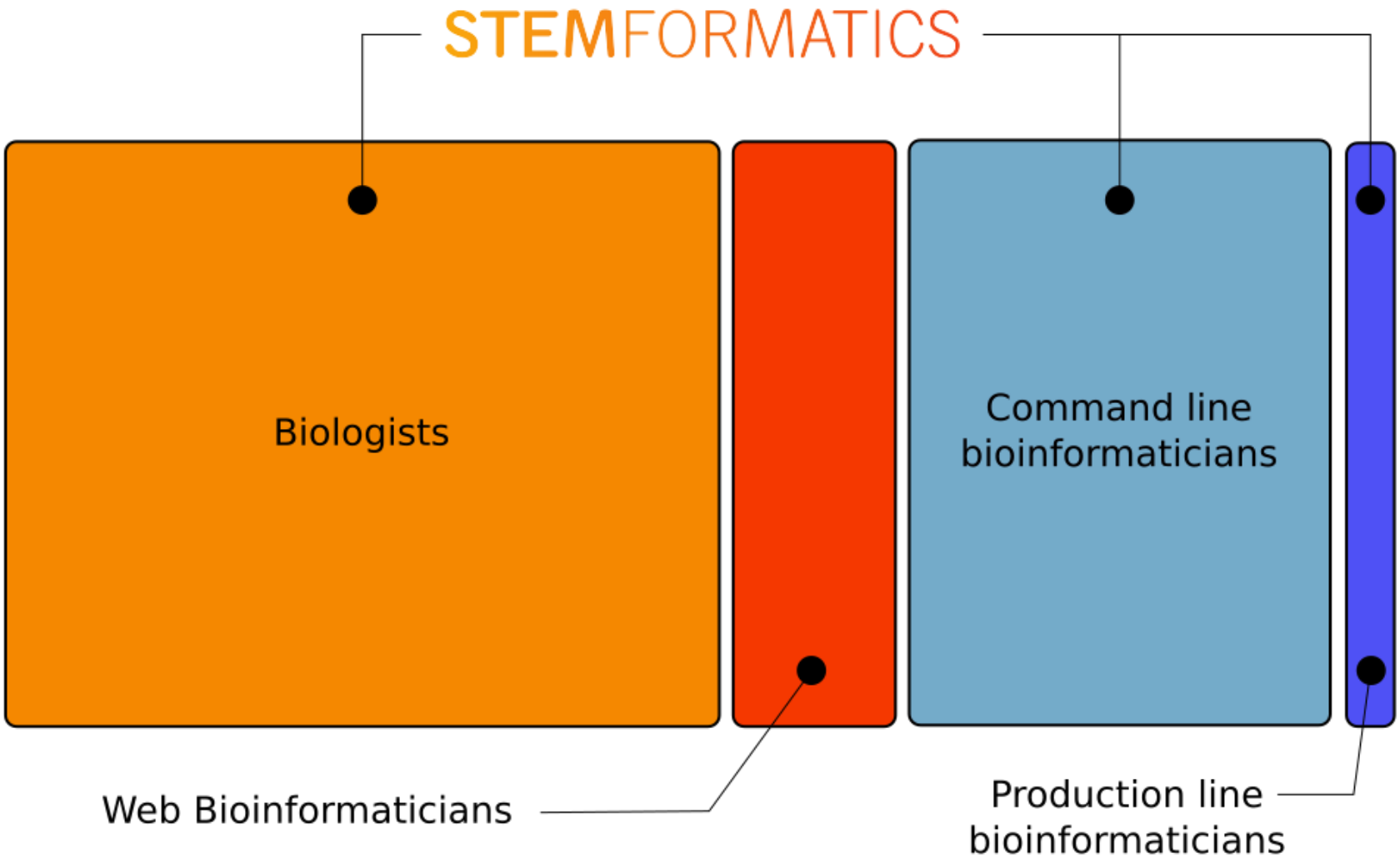
- Want to choose methods
- Want cutting edge methods
- Want access to full settings
- Can manipulate data
- Runs on laptop or HPC
- Includes biostatisticians

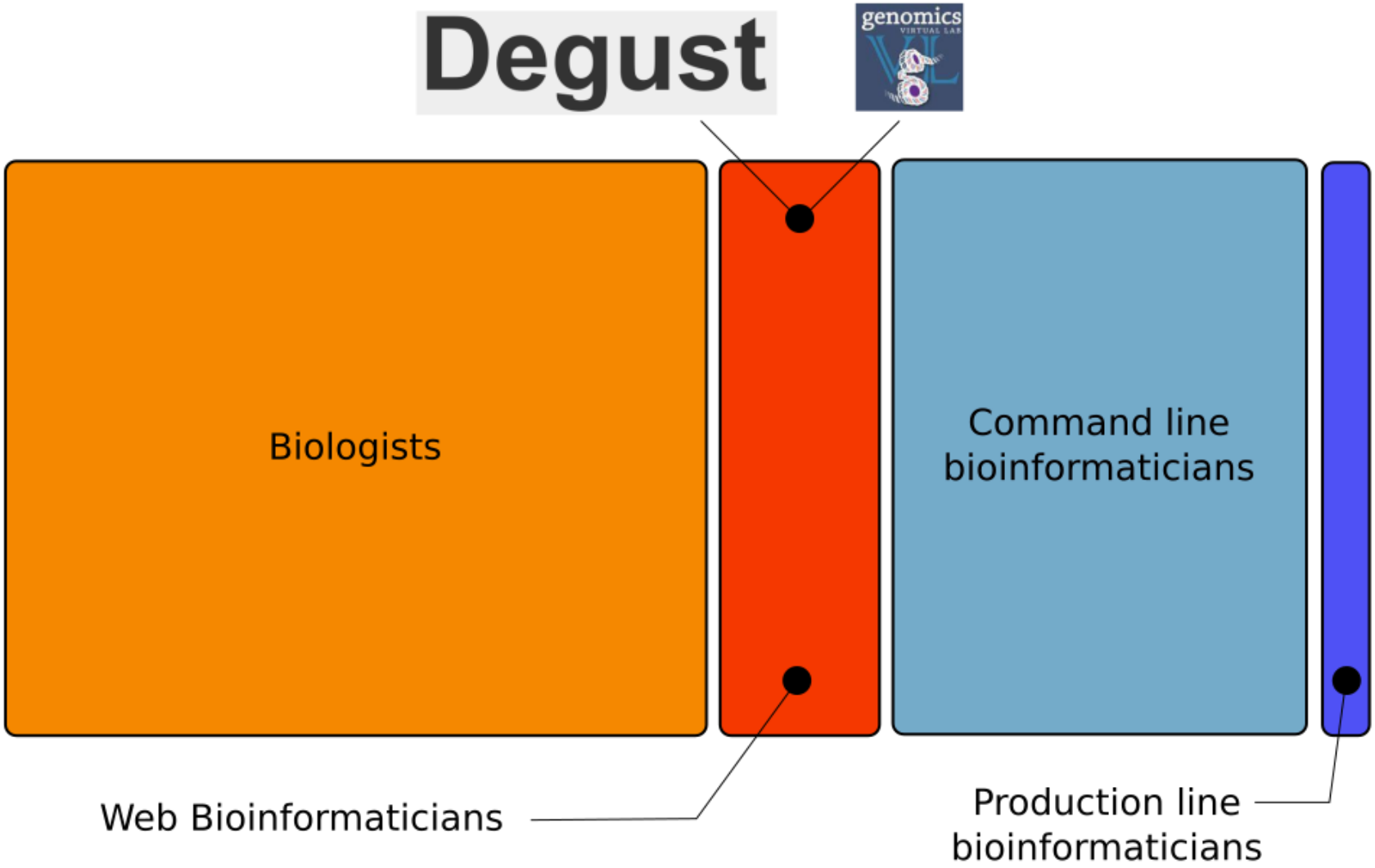
Web Bioinformaticians

- Happy with default methods
- Happy with non-cutting edge methods
- Want to change settings
- OK to manipulate data

Production line bioinformaticians

- Pump out multiple datasets per day
- Constant pipeline, very little variance
- Particular focus on goal/service
- Runs on Amazon or HPC


STEMFORMATICS
Biologists
Command line bioinformaticians
Web Bioinformaticians
Production line bioinformaticians


Degust
genomics
VIRTUAL LAB
Biologists
Command line bioinformaticians
Web Bioinformaticians
Production line bioinformaticians

Data Visualisation

STEMFORMATICS

Login
Register
Forgot pass phrase

GENES > | DATASETS > | GRAPHS > | ANALYSES > | MY JOBS > | ABOUT US >

Search Gene in Stemformatics

Enter Symbol, Ensembl, Entrez, HGNC, MGI or RefSeq IDs for more precise results. It will provide suggestions via an autocomplete after four characters.

mincle | SEARCH

CLEC4E	Homo sapiens	CLECSF9 MINCLE	C-type lectin domain family 4, member E
Clec4e	Mus musculus	C86253 Clecsf9 Mincle	C-type lectin domain family 4, member e

Analyses

My Gene Lists

Browse my gene lists

Kegg Pathways

Browse Kegg pathways

Public Gene Lists

Browse Public gene lists

Feature Search

Search miRNA

STEMFORMATICS

Login
Register
Forgot pass phrase

GENES > | DATASETS > | GRAPHS > | ANALYSES > | MY JOBS > | ABOUT US >

Gene Expression Graph

One gene, one dataset

MultiGene Graph

Multi gene, one dataset

Multiview

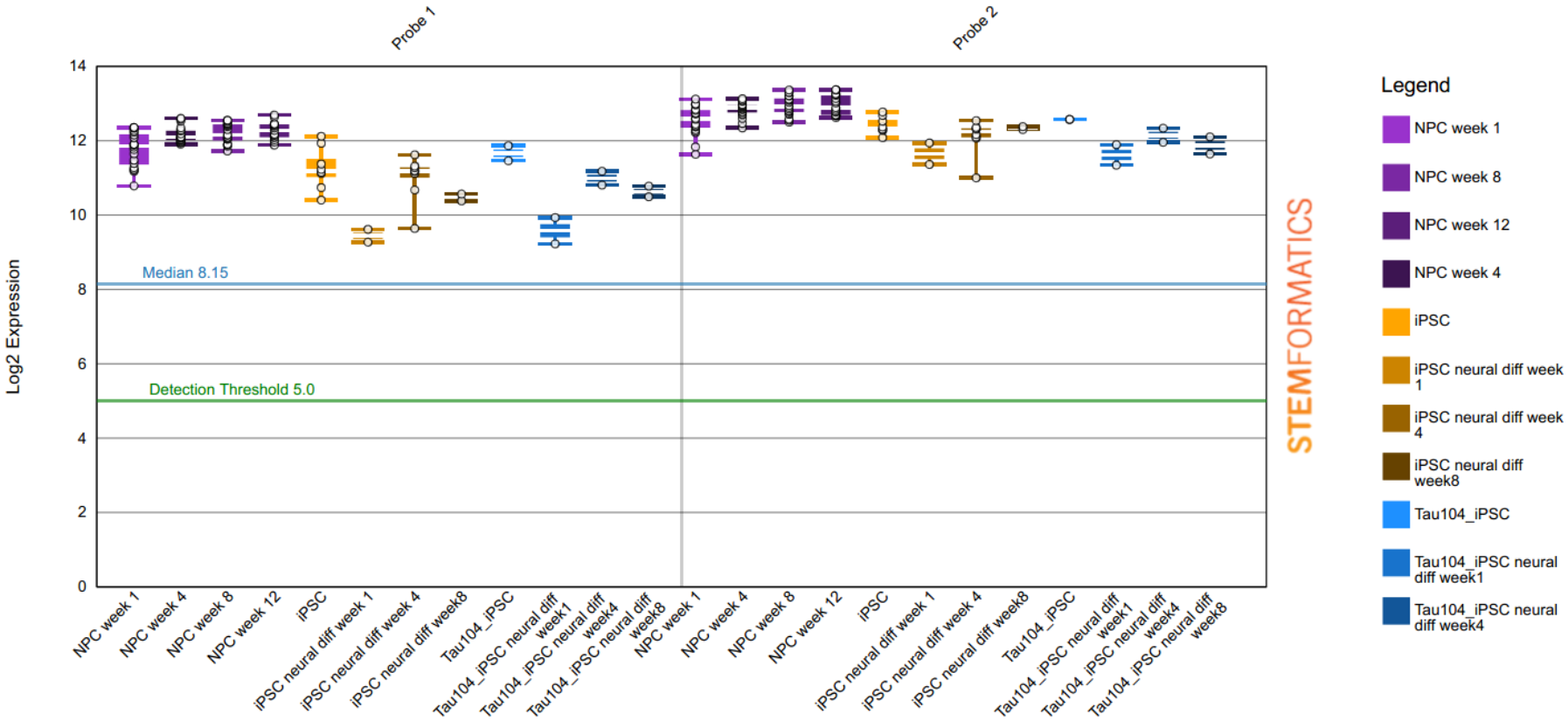
One gene, up to four datasets

YuGene

One gene, all datasets

Gene Expression Graph for Gene SOX2 grouped by Sample Type

A quantitative framework to evaluate modeling of cortical development by neural stem cells



STEMFORMATICS

Login
Register
Forgot pass phrase

GENES > | DATASETS > | GRAPHS > | ANALYSES > | MY JOBS > | ABOUT US >

Hierarchical Cluster

Clustering

Gene Expression Profile

Find similar profiles

Rohart MSC Test

View MSC and non-MSC status

Multiple Dataset Downloader

Search and download datasets

Fold Change Viewer

x2

Avg fold change

UCSC Browser

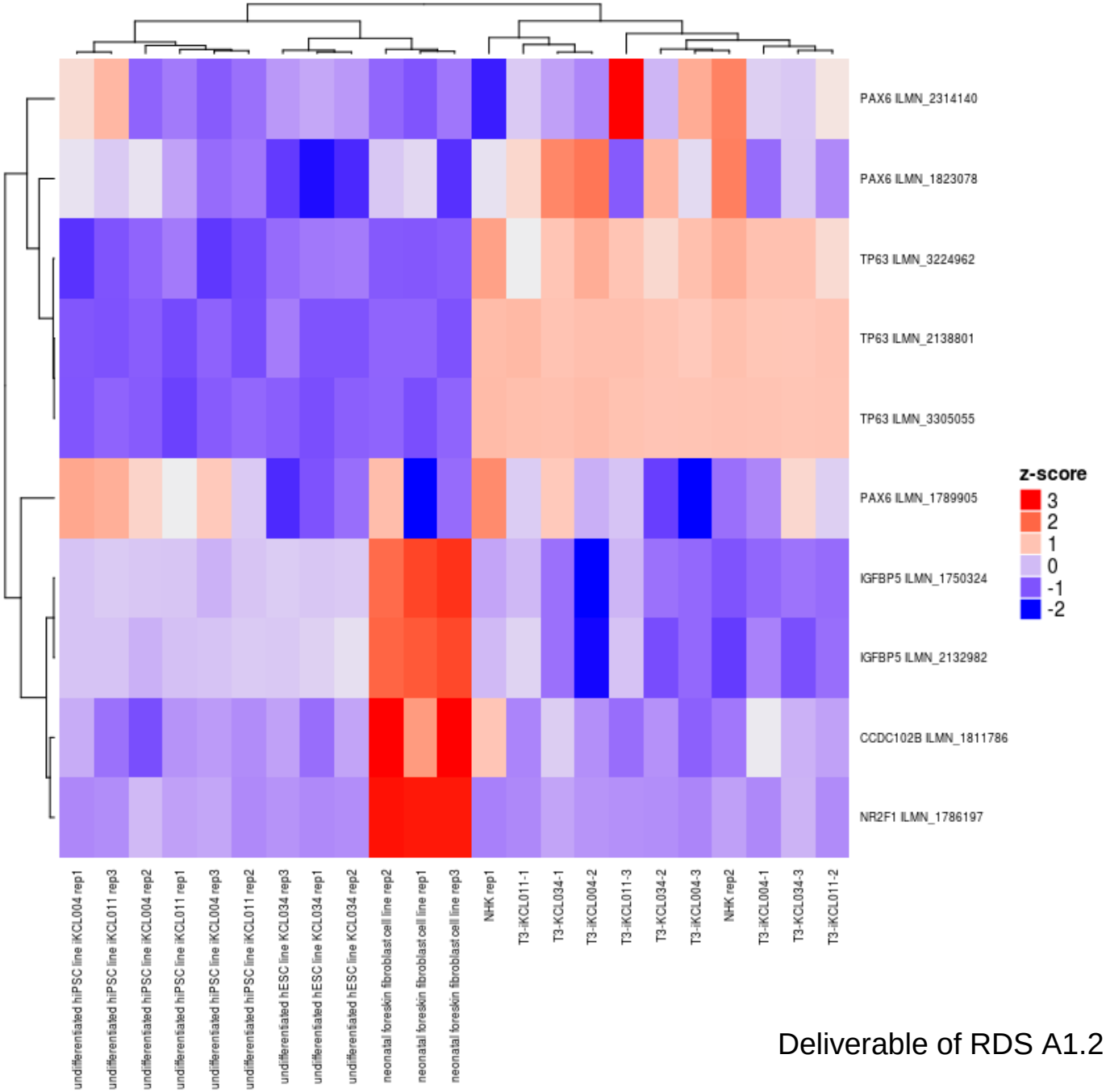
Browsing UCSC

Gene List Annotation

Find Kegg Pathways

My Jobs

Pending and finished jobs



Deliverable of RDS A1.2 Project

Collaboration

Collaboration Platform

- Project Grandiose is a good example
- 2 Nature and 3 Nature Communication papers
- 13 datasets over multiple labs and continents
- Most accessed datasets in Stemformatics
- Stemformatics Landing page

Google

project grandiose

All | Images | News | Shopping | Videos | More | Search tools

About 568,000 results (0.32 seconds)

Stemformatics - Project Grandiose
https://www.stemformatics.org/**project_grandiose**
Project Grandiose defines two reprogramming trajectories, which arrive at distinct pluripotent states: the "F-class" and embryonic stem cell (ESC)-like iPSCs.

Stemformatics - Find expression data from leading stem cell ...
www.stemformatics.org/
Log in to run and save your own analyses. START EXPLORING. 1234. prevnext. **Project Grandiose**. See **Project Grandiose's** roadmap of cell reprogramming.

Stem cells: The black box of reprogramming : Nature News ...
www.nature.com/.../stem-cells-the-black-box-of-reprogramming-1.1652...
Dec 10, 2014 - This week, the biggest such project — an international collaboration audaciously called **Project Grandiose** — unveiled its results. The scientists ...

All Images News Shopping Videos More Settings Tools

About 15,000,000 results (0.37 seconds)

Stemformatics - Project Grandiose

https://www.stemformatics.org/project_grandiose

Project Grandiose defines two reprogramming trajectories, which arrive at distinct pluripotent states: the "F-class" and embryonic stem cell (ESC)-like iPSCs.

[PDF] Project Grandiose - Lunenfeld-Tanenbaum - The Lunenfeld ...

research.lunenfeld.ca/rss/files/file/PG_journalist_summary_w_Figures_MP_v3.pdf

Dec 11, 2014 - **Project Grandiose**: High definition characterization of the process of ... Summary of **Project Grandiose** Nature/ Nature Communications cluster ...

Project Grandiose - Stem Cells Australia

www.stemcellsaustralia.edu.au › News & Events › News

Dec 11, 2014 - At the heart of **Project Grandiose** is the Stemformatics team from The University of Queensland's Australian Institute for Bioengineering and ...

Introduction

Epigenome

Transcriptome

Proteome

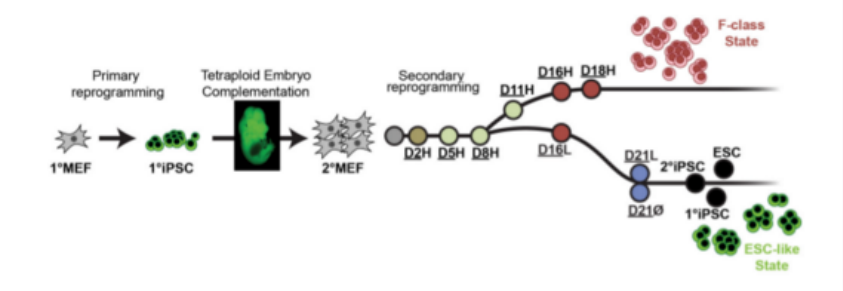
Data Visualisation

Project Grandiose

Differentiated cells are presumed to have a molecular network that is 'hard-wired' to dictate the restricted phenotype of that cell. The discovery that the over-expression of four transcription factors (reprogramming factors), Oct4, Sox2, Klf4 and c-Myc (OSKM) could disrupt this hard-wiring, and revert cells to stem-cell-like phenotypes (induced pluripotent stem cells or iPSCs) challenged many of our assumptions about cell fate.

Project Grandiose defines two reprogramming trajectories, which arrive at distinct pluripotent states: the "F-class" and embryonic stem cell (ESC)-like iPSCs. The F-class state represents reprogramming factor dependent cells, whilst the (ESC)-like iPSCs state represent reprogramming factor independent cells.

A highly efficient mouse secondary reprogramming system was used to characterise the molecular trajectories leading to these two distinct pluripotent cell types at multiple omic levels. NGS was used to profile the transcriptome (miRNA, lncRNA, mRNA), genome wide CpG methylation and chromatin marks (H3K4me3, H3K27me3 and H3K36me3) in addition to quantitative mass spectrometry profiling of the global and cell surface proteome. These analyses were coordinated to be performed in parallel on the same cell collections at the same time points and cell states indicated in Figure 1.



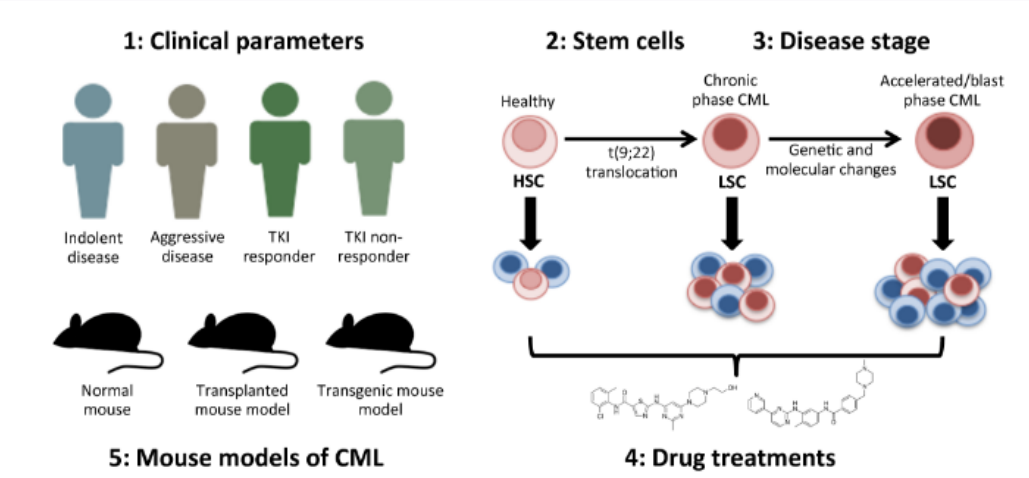
Welcome

Data

LEUKomics

As we learn more about the molecular changes that drive cancer, so our power to combat it grows. Researchers of many cancer types aspire to build our understanding of these drivers and use this knowledge to create targeted therapies. High-throughput technologies, which are becoming ever more powerful, have the potential to lead us to a truer picture of the molecular features of cancer cells.

CML researchers led the way in the search for targeted cancer therapies through discovery of the BCR-ABL1 tyrosine kinase inhibitors, leading to huge improvements in patient care. However, challenges still remain in CML. At LEUKomics we believe that the groundbreaking CML research happening across the world could be enhanced by use of a growing collection of high throughput data. Our goal is to bring these datasets together and create a tool to facilitate in depth analysis of the wealth of information that resides within them.

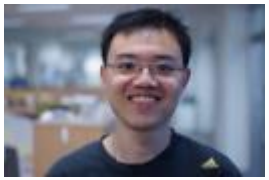


Collaboration: What users ask for

- More related data for their particular research area (eg. leukaemia, pain)
- More multi-omics and single cell RNASeq datasets
- Easier to share interesting graphs with collaborators
- More tools to interact with the data
- Landing pages to attract attention

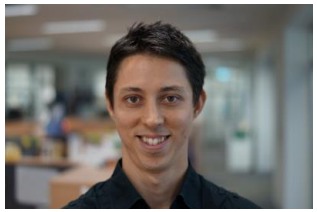
Trusted Data

Overall failure rate (19/09/17)



Cannot identify samples	2.51%
Failed QC Experiment Design	9.92%
Failed QC Normalisation	8.1%
No Raw Data	8.66%
Total datasets failed	**29.19%**

Dataset Types

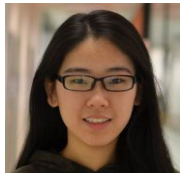


Pipeline	**Hosted**
Microarray	CAGE
RNA-Seq	Protein
ChIP-Seq	Single Cell qPCR
small RNA (miRNA)	Methylation
ATAC-Seq	

Trusted Data: What users ask for

- More related data for their particular research area (eg. leukaemia, pain) that they can trust
- More multi-omics and single cell RNASeq datasets that they can trust
- Bioinformaticians and Biostatisticians want to use this trusted data too!
- This is why we created the Data Portal with student interns Sadia and Huan

STEMFORMATICS



Enter search terms: mincle **Search**

All Categories

Species

- Mus musculus (2)

Tissue Types

- NULL (1)

Filter Clear Reset

Export selected dataset metadata | Export selected sample metadata | Export download script for selected datasets

2 datasets found.

Search:

Show 100 **entries**

Click to toggle	ID	Handle	Title	Species	Samples/Total	Actions
☑	6498	Tanaka_2014_25236782 [Microarray]	Macrophage-inducible C-type lectin underlies obesity-induced adipose tissue fibrosis.	Mus musculus	5/11	Actions...
☑	6731	Arumugam_2015_S4M-6731 [Microarray]	The role of Clec4e (Mincle) in microglia response to transient ischemic injury	Mus musculus	0/16	Actions...

Showing 1 to 2 of 2 entries

Previous 1 Next

https://portal.stemformatics.org/

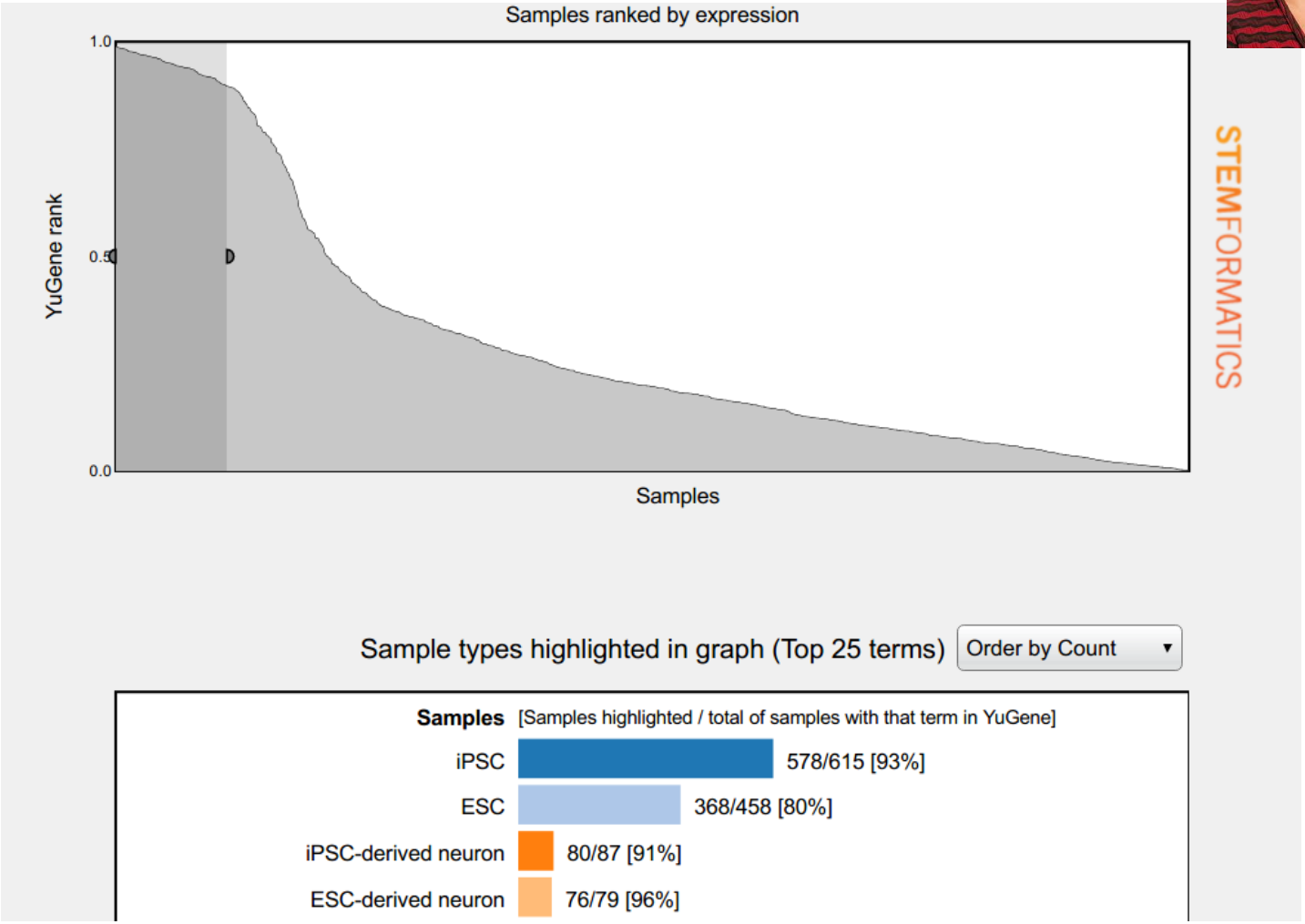
Bursary from Linux Australia

Data Mining

YuGene

- Working with biostatistician Kim-Anh Lê Cao
- Need to reduce the dimensionality of data
- ie. remove batch effects of dataset or platform
- YuGene helps to reduce this
- Microarray only
- Used for comparing all samples of one gene in Stemformatics
- I call this “Manual data mining”

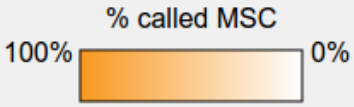
YuGene Graph SOX2 (Human)



Rohart MSC Test

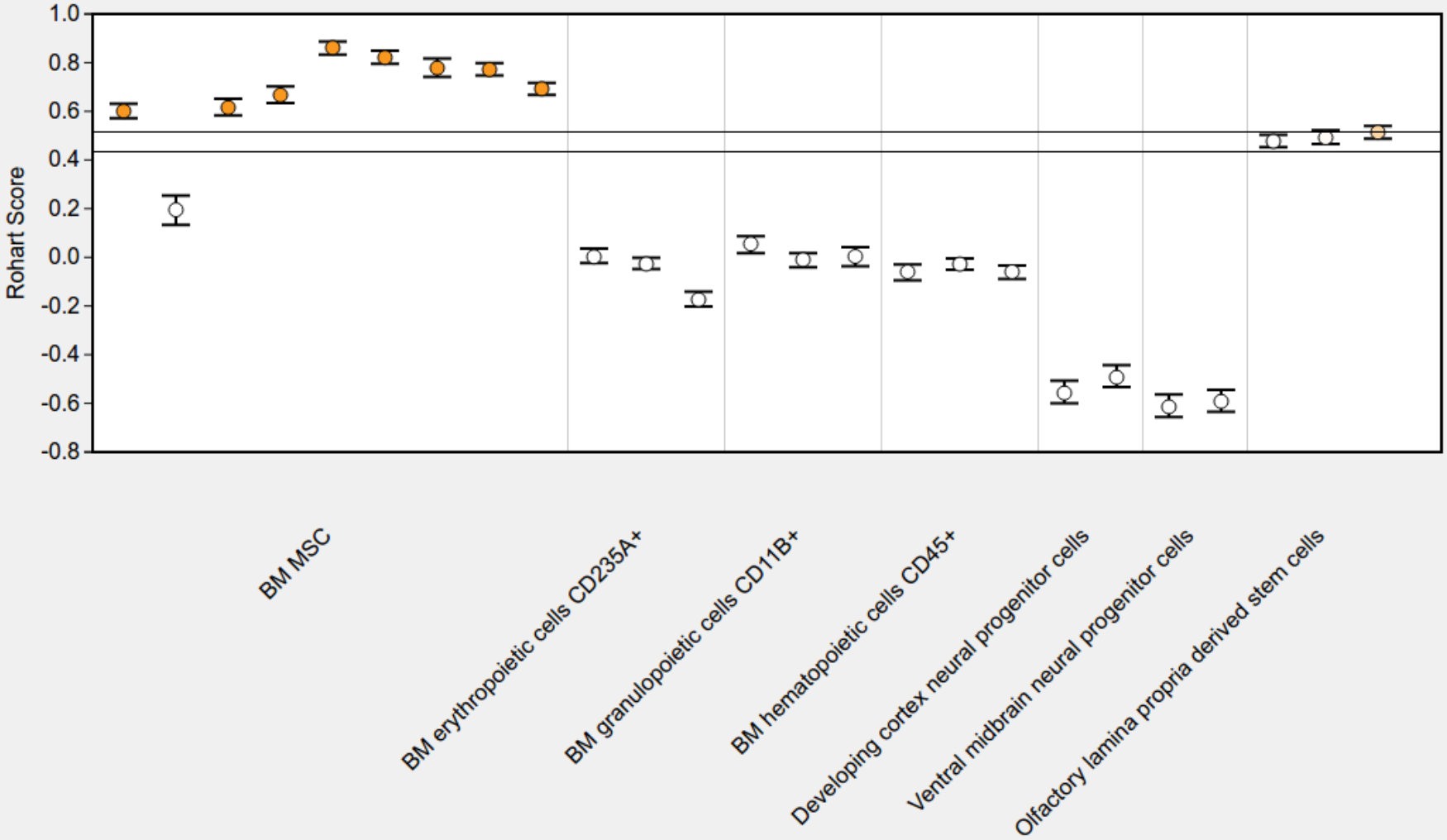
- Working with biostatistician Kim-Anh and Florian Rohart
- Curating metadata is hard!
- Data mining across multiple datasets
- Used YuGene data to create signature with machine learning
- Ability to identify MSC with 97%+ accuracy

Rohart MSC Test for Dataset ID 6037



The human nose harbours a niche of olfactory ecto-mesenchymal stem cells displaying neurogenic and osteogenic properties

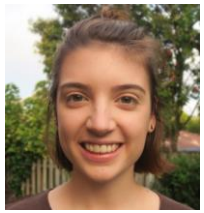
Bone marrow derived mesenchymal stem cells (MSC), hematopoietic cells, neural progenitors, olfactory-derived stem cells



Data Mining: What users ask for

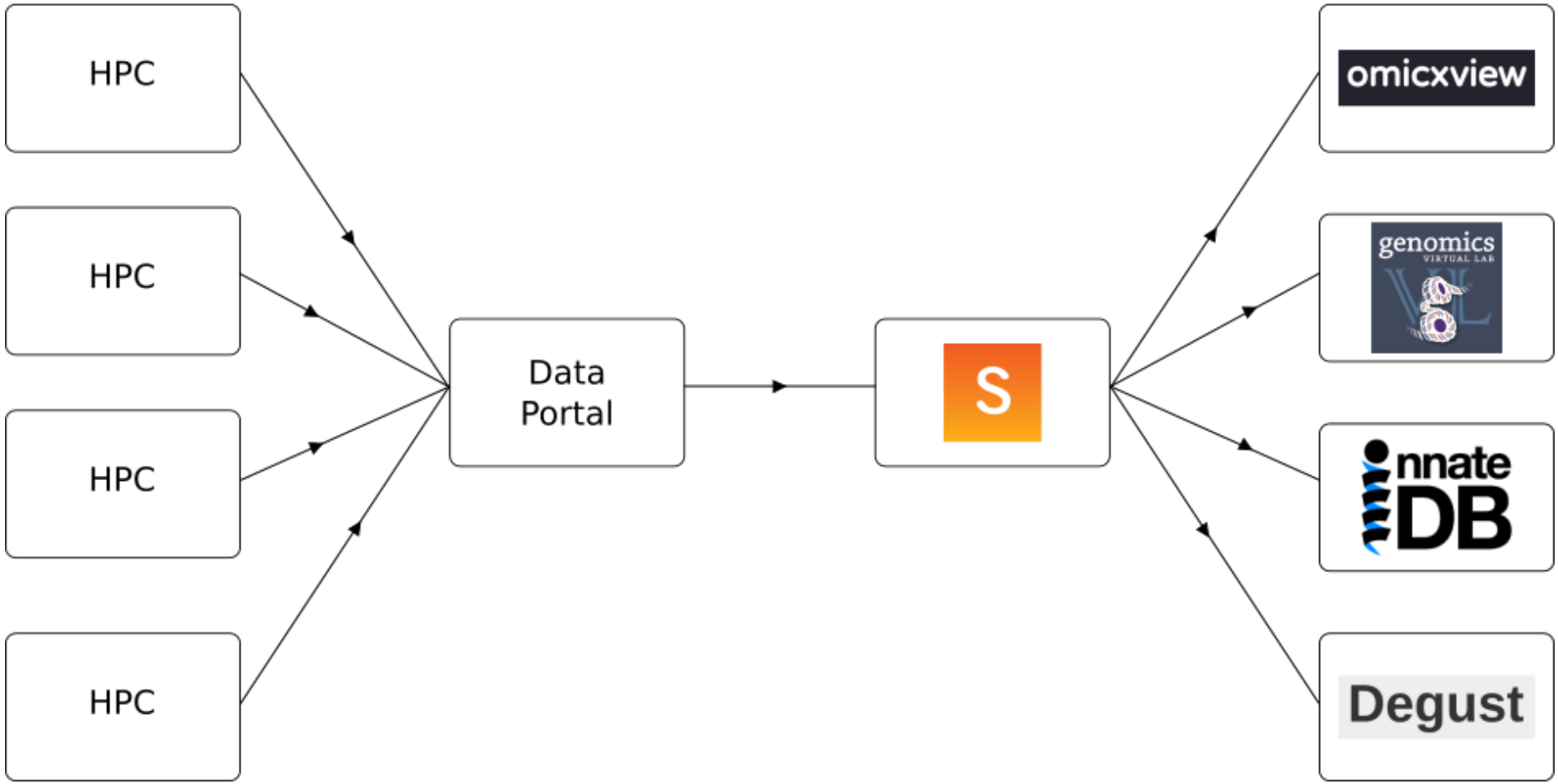
- Can you run the Rohart MSC Test on my dataset?
- Can you make it easier to show my cell types of interest in the YuGene Graph?
- Can you run YuGene analyses across genes?
- Consistent sample annotations!

Future of Stemformatics



- Ecosystem of tools (eg. Omicxview, Data Portal)
- Easy to setup pipelines & tools via Ansible
- More online communities (eg. LEUKomics)
- More data mining opportunities
- National and International engagement
- Long term sustainability both in technical and financial terms

Future Ecosystem Flow of Data



Acknowledgements

UoM

Christine Wells

Rowland Mosbergen

Tyrone Chen

Isha Nagpal

Sadia Waleem

Huan Wang

Jarny Choi

Elizabeth Mason

Chris Pacheco Rivera

UoM (Systems Genomics)

Kim-Anh Lê Cao

AIBN (UQ)

Othmar Korn

Ariane Mora

Steve Englart

Travelling

Florian Rohart

Command Line Domain Training