Stem Cell Research (2013) 10, 387–395

Available online at www.sciencedirect.com

SciVerse ScienceDirect

www.elsevier.com/locate/scr

METHODS AND REAGENTS

Stemformatics: Visualisation and sharing of stem cell gene expression

Christine A. Wells [a,*], Rowland Mosbergen [a], Othmar Korn [b], Jarny Choi [c], Nick Seidenman [c], Nicholas A. Matigian [d], Alejandra M. Vitale [a,1], Jill Shepherd [a]

[a] The Australian Institute for Bioengineering and Nanotechnology, University of Queensland, Brisbane, 4072 Australia
[b] Scholarly Information & Research, Division of Information Services, Griffith University, Nathan, QLD 4111, Australia
[c] The Walter & Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3050, Australia
[d] Australian Centre for Adult Stem Cell Research, Eskitis Institute for Cell and Molecular Therapies, Griffith University, Nathan QLD 4111, Australia

Received 25 September 2012; received in revised form 28 November 2012; accepted 4 December 2012
Available online 20 December 2012

Abstract Genome-scale technologies are increasingly adopted by the stem cell research community, because of the potential to uncover the molecular events most informative about a stem cell state. These technologies also present enormous challenges around the sharing and visualisation of data derived from different laboratories or under different experimental conditions. Stemformatics is an easy to use, publicly accessible portal that hosts a large collection of exemplar stem cell data. It provides fast visualisation of gene expression across a range of mouse and human datasets, with transparent links back to the original studies. One difficulty in the analysis of stem cell signatures is the paucity of public pathways/gene lists relevant to stem cell or developmental biology. Stemformatics provides a simple mechanism to create, share and analyse gene sets, providing a repository of community-annotated stem cell gene lists that are informative about pathways, lineage commitment, and common technical artefacts. Stemformatics can be accessed at stemformatics.org.

Introduction

Gene expression signatures have yielded a wealth of information for benchmarking stem cell phenotypes (for example, (Muller et al., 2008, 2011; Pelekanos et al., 2012)), identifying core signalling networks (Novershtern et al., 2011) and mining novel gene products with roles in the maintenance and differentiation of various stem cell lines. It is increasingly common for researchers to include gene expression profiling as a metric of cell quality, particularly in the derivation of iPSC from various tissues or disease cohorts (Bock et al., 2011; Nayler et al., 2012). As a result, a large number of gene expression datasets generated on various stem cell models, using microarray or sequencing platforms, are publicly available, in gene expression repositories such as ArrayExpress (Parkinson et al., 2011) or GEO (Barrett et al., 2011). However, assessment of gene expression patterns across different stem cell datasets remains difficult, and as such, there is a demand for tools that allow researchers to

[1] Current address: Instituto de Biologia y Medicina Experimental, CONICET, Buenos Aires, Argentina.

1873-5061/$ - see front matter.
http://dx.doi.org/10.1016/j.scr.2012.12.003

determine the reproducibility of genes and gene signatures that correlate with particular stem cell phenotypes.

The identification of new stem cell markers characteristic of a phenotypic subset or experimental state remains a major community research goal. Researchers may wish to assess the uniqueness of an expression profile before embarking on experiments that rely on a reporter gene or antibody to select for a particular stem cell subset. Genes that are highly novel generally have few publications, and while data on these genes is likely to be collected in large-scale genomics datasets, finding and assessing this information can be challenging. Web-based tools such as BioGPS (Wu et al., 2009) or TiGER (Liu et al., 2008) do exist for rapid querying of single gene profiles across a tissue atlas; these tap into a common desire of researchers to quickly assess the expression of a single gene or a small gene set, but they lack relevant stem cell samples. A handful of databases, such as the mouse embryonic stem cell database FunGenEs (Schulz et al., 2009) and the haematopoiesis database SCDb (Hackney et al., 2002), focus on specialist stem cell datasets designed for consortiums around a specific area of stem cell biology. StemBase (Sandie et al., 2009) provides the most comprehensive mouse and human microarray collection focussed on stem cells, but the search terms are dataset-centric rather than gene-centric, and it can be difficult to use without explicit guidance or training. StemCellDB (Mallon et al., 2013) is a recently published expression database focussed on iPSC and ESC, hosting exemplary in-house generated data on highly curated stem cell lines. While meeting the community requirements for exemplary expression datasets on pluripotent stem cells, it lacks the breadth of experimental data available in the public domain and has limited visualisation functionality.

Indeed, user interfaces for most existing gene expression data repositories are designed for bioinformaticians, not biologists (Pavelin et al., 2012). Many focus on data storage, and bioinformatics expertise is required to pull out relevant datasets systematically and analyse them appropriately. The task of identification, downloading, normalisation and analysis of the relevant stem cell datasets offers a significant barrier to this kind of query. Even in databases which offer 'tissue-signatures' that may include stem cell experiments, such as The Gene Expression Barcode (McCall et al., 2011), or the Human Gene Expression Atlas at EBI (Lukk et al., 2010), unwieldy search terms can make datasets of interest difficult to identify. This difficulty in filtering the large number of datasets for relevance and quality acts as a barrier for many stem cell biologists, thus the available expression resources are effectively underutilised.

Stem cell datasets generated on painstakingly prepared cell models, but analysed on older technologies, are often overlooked because of the difficulties associated with comparing data from different platforms. Even qualitative assessment of gene profiles generated across different technologies requires the data to be handled in a systematic, robust and standardised manner. For stem cell groups lacking bioinformatics expertise, such analyses, however worthwhile may not seem possible. The lack of suitable tools for visualisation of data across different 'omics' platforms is another barrier for groups that don't have access to a team of bioinformaticians. Even qualitative assessment of gene profiles generated across different technologies requires the data to be handled in a systematic, robust and standardised manner. The Stemformatics portal was generated to future-proof existing stem cell datasets, by providing straightforward tools for easy visualisation and comparison of gene expression generated across platforms, laboratories and cell models.

Stemformatics.org is an online data portal and a collection of visualisation tools, designed to help stem cell biologists identify and assess relevant datasets, gene sets and pathways. It addresses many of the problems identified above by hosting a growing collection of manually-curated, high-quality public datasets and providing an intuitive biology-centric workflow to assist researchers access gene profiles quickly. The Stemformatics target audience is a stem cell biologist with minimal bioinformatics background and focuses on easy to interpret views of the data using interactive graphs and heat maps. The site provides all data in downloadable formats which can be readily opened by most common desktop spreadsheet programs, as well as a translation feature to assist users who wish to run more sophisticated analyses using external software such as GenePattern (Reich et al., 2006; Kuehn et al., 2008) or MeV (Howe et al., 2011). Stemformatics supports some basic analysis features, including sample comparison and identification of correlated gene patterns. Flexible gene annotation features include the ability to create, manipulate and analyse private gene lists, and an integrated gene annotation function to help predict cell-surface proteins or membership in relevant pathways. Its biology-centric philosophy means that external tools and resources can be accessed quickly using common queries as the starting point, with the application automatically transforming the data into the required formats as needed. The resources described in this manuscript are available at www.stemformatics.org.

Methods

Dataset processing and inclusion criteria

Datasets

The number of GEO datasets at the start of the Stemformatics project which profiled human stem cells was 270. Of these, 118 experiments met the criteria of examining whole-genome expression profiling in adult or embryonic stem cell biology and 50 experiments met the quality control standards set for experimental design and replication. Datasets were excluded if they did not include sufficient biological replication, if the metadata was not sufficiently explicit or if the data itself did not meet expected quality control (QC) standards. Detailed methodology on normalisation and QC is available in Supplement File 1.

In short, our QC standards require that raw data is available and that the experiments are well replicated, with $n>=3$ per experimental group. 12 stem cell experiments, prioritised by the local stem cell community, were included in the phase 1 release. Currently there are 44 public datasets, 34 human and 10 mouse (see also Supplement File 2), and an active curation team means this number is growing with each new release. New datasets are included at the request of our users. Links to the original publications

(via PubMed) and GEO or ArrayExpress accession numbers are provided for all datasets, and dataset contacts are replicated from the information provided by the data depositor.

Platforms

Probe sequences for each platform have been independently mapped to Ensembl annotated genome (coding and non-coding sequences). This process allowed us to maintain consistency in relation to mapping parameters, retain probe-transcript-gene annotations, and identify probe sets with ambiguous mappings or annotations.

Datasets undergo several curation processes prior to inclusion into Stemformatics. In the first instance, specific datasets are chosen to represent significant stem cell populations with well characterised phenotypes available for each of the profiled groups. The Stemformatics team considers all datasets requested by users to ensure that a range of relevant stem cell models is available. All of the datasets are manually curated from the relevant original publication to make sense of the sample metadata in a biological context, rather than simply relying on machine-driven definitions from the public microarray repositories. In some circumstances, where insufficient metadata was publicly available for datasets considered important for inclusion, we contacted the authors to obtain the relevant metadata.

Microarray datasets considered for inclusion in Stemformatics are subjected to both pre- and post-normalisation quality control checkpoints requiring validation by experienced bioinformaticians. Quality control steps include: a check for completeness of raw data; technical and biological validation of sample quality using quantitative relative log expression (RLE) analysis of housekeeping genes; sample hierarchical clustering and principal component analysis (PCA) before and after normalisation and dataset expression density check for well formed (bi-modal or uni-modal, as appropriate) probe expression distributions. Based on the overall QC results and evidence provided by these multiple measures, each dataset included in the Stemformatics resource meets strict criteria for technical robustness.

An in-house web tool tracks the status of datasets in our processing queue as they move through the QC pipeline, through to experiment and sample metadata annotation, post-upload validation into our Beta server, and finally release to our public Stemformatics website. Manual metadata annotation by stem cell scientists ensures adequate capturing of relevant biological and experimental factors of interest using controlled vocabularies for core fields such as cell and sample type annotations. Where possible, we also maintain, in parallel, the author's nomenclature for sample naming to ensure that samples are always easily traceable to the relevant publication. Metadata is drawn from GEO, Array Express and the original publication. An in-house annotation wizard includes an automated 'validator' to ensure the integrity of the metadata prior to merging with a normalised dataset for publishing in Stemformatics. By focussing on high quality, human readable metadata, we provide an essential framework for future meta-analyses across disparate data sources. All the metadata for the samples in the system can be viewed through the application.

Results

Features

Gene query

A motivation for developing the Stemformatics resource was the desire of our collaborators to quickly interrogate expression profiles of their gene-of-interest across different stem cell experiments. Such investigations can provide information about a gene's expression pattern across various conditions of interest to the researcher, and allows users to quickly make qualitative comparisons between their own results and those of other studies. As summarised in Fig. 1, Stemformatics provides a simple search interface to find genes of interest; users do not need to look up accession numbers, probe IDs or gene symbols as these are suggested to match user input terms. For example, POU5F1 can be identified using POU5 as the search term, or Oct3, or OCT4 or variations on these themes, as well as via the common NCBI and Ensembl accession numbers, and via platform identifiers such as the probe number. To be sure that the correct species-specific gene symbol is returned, the search workflow automatically prompts species selection (currently mouse or human) and provides synonyms and descriptions for the selected gene. A summary of each gene is given, including links to external sources such as Entrez Gene and Pubmed. To view gene expression profiles, users are prompted to select the dataset of interest, and a free-text search box allows rapid filtering of the list of available experiments. Datasets can be filtered on author, sample/cell type or platform attributes.

The graph view (Fig. 1C) summarises the expression data for each gene by displaying every probe (in the case of microarray data) or Ensembl transcript (in the case of sequence data) that are annotated to that single gene. Two lines intersect the graph — the top (green) line provides the median expression of all data points in a particular experimental series, and the lower (blue) line provides the detection threshold for the dataset. Together this information allows users to judge "high", "medium" or "low" expression of the gene-of-interest relative to all genes in that experimental series. The style of graph can be changed; the scatter plot shows every data point, whereas the bar graph or boxplot provides summarised information. The data may be further summarised according to different manually curated experimental parameters, such as sample, cell type, time point, or disease state. The gene search area of Stemformatics can be used without logging in, making this a quick gene-centric look-up tool for the research community.

Most expression platforms take multiple measurements across a gene using different oligonucleotide probes (for microarrays), or transcripts (for RNASeq). Many downstream analysis methods are designed to force a 1:1 relationship between gene and expression measure, so some databases will show only summary expression for a gene, rather than describe the behaviour of each probe. Others (like BioGPS) will only allow examination of one probe at a time. Common methods to summarise this data include averaging across all probes mapping to a gene, taking the first (numerical/alphabetical) probe or selecting the probe that has the highest expression value. However understanding potential

ambiguities in the data at the probe level is essential before undertaking downstream validation of gene-level behaviour. Stemformatics displays all of the measurements collected for a single gene on one graph, allowing rapid assessment by the user of any potential inconsistencies. As shown in Fig. 1, and Supplementary File 1, probes whose design permits cross-hybridisation across multiple genes are highlighted in red and hyperlinked from the graph to a table that lists all possible matches. POU5F1, a critically important gene in many stem cell analyses is an example of a gene measured using multiple probes on either Affymetrix or Illumina microarray platforms, where multiple probes may cross-hybridise with other members of the POU homeodomain family.

Multiview

Qualitative assessment of expression of a single gene across different datasets is provided via the multi-experiment viewer. Up to four different datasets can be visualised simultaneously, with gene expression profiles from each dataset graphed separately. This tool allows users to visually assess concordance (or discordance) of gene expression patterns across different samples, or across different 'omics' measurements for a dataset series. This is useful when comparing between microarray or RNAseq profiles, or when miRNA, Methylseq or ChIP-seq data is available on the same experimental series. Because the site needs to save session information in order to display multiple datasets, users must be 'signed in' to the Stemformatics workbench to use the multiview feature. This permits the dynamic updating of all four windows with the interactive features of the graph.

Case study

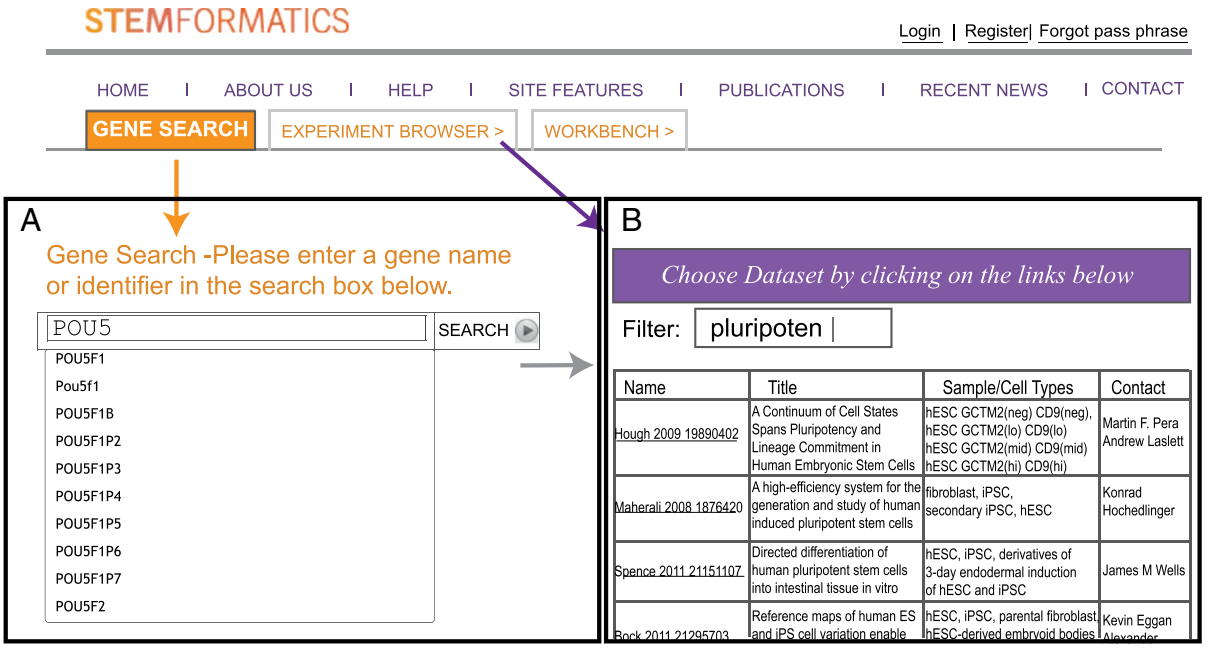
CD9 and GCTM2 are cell surface markers which have been used to discriminate the layers of cells in a human embryonic stem cell (hESC) colony with varying capacity for self-renewal and differentiation. The 2009 dataset of Hough and colleagues used Illumina microarrays to profile gene expression across four fractions of hESC: $CD9^{hi}/GCTM2^{hi}$ was designated P7 and had the highest self-renewal capacity, P6 $CD9^{mid}/GCTM2^{mid}$, P5 $CD9^{lo}/GCTM2^{lo}$ and P4 $CD9^{neg}/GCTM2^{neg}$ with the lowest self-renewal capacity (Hough et al., 2009). The CD9 transcript was expressed in the same pattern as the cell surface protein, low in P4 with step-wise increasing expression across the fractions. Fig. 2 demonstrates the ease of comparing CD9 expression with additional, independently derived datasets in the 'Multiview' window, allowing a quick confirmation that CD9 is indeed highly expressed across different human ESC and iPSC lines (Evseenko et al., 2010; Bock et al., 2011; Vassena et al., 2011), is down-regulated in mesodermal progenitor cells derived from hESC (Evseenko et al., 2010) and is expressed abundantly in fertilised oocytes, very early (2–8 cell) human embryos, with highest expression in the morula and with equivalent levels in human blastocyst and hESC (Vassena et al., 2011).

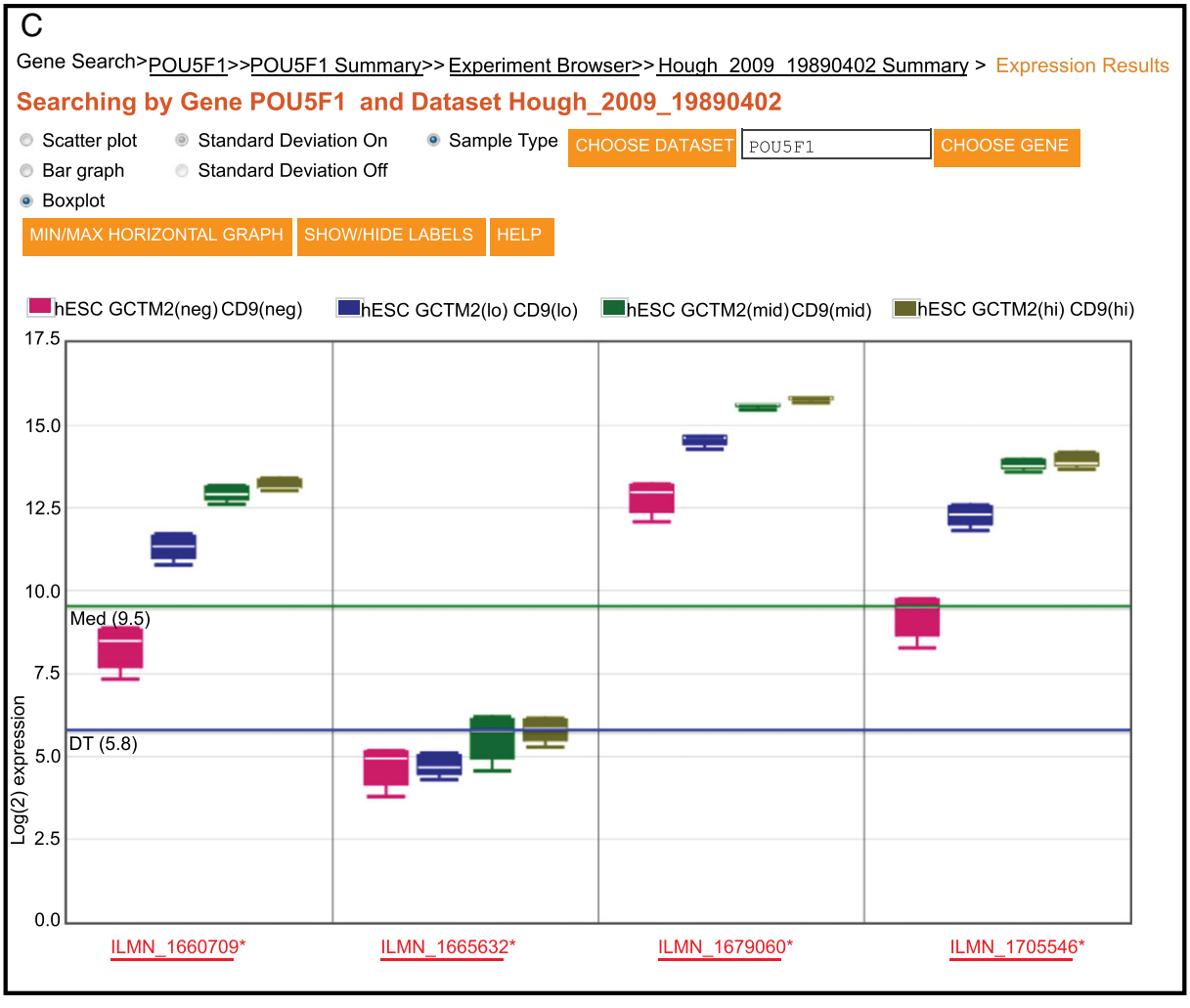
In order to find genes in the Hough dataset with closely correlated patterns of expression, the 'Gene Neighbourhood' function on the gene-view page was used to initiate a Pearson correlation and for this analysis returned 623 matches ($r<=0.8$). The list contains many genes previously associated with self-renewal and pluripotency, including members of the Oct3/4 transcription factor network. The gene list was then viewed in the Stemformatics workbench (includes Illumina probe ID, gene symbol and p-value), and can be downloaded into a standard spreadsheet. Saving the list as a gene set in the Stemformatics workbench provides access for future analyses. As shown in Fig. 2, clustering the gene list using the hierarchical clustering tool provides visual confirmation of the step-wise expression of genes in this list from high in the P7 fraction to low or absent in the P4 cells.

The gene set is now available for analysis against other datasets. For example, one might wish to know how generalizable the expression of this set of genes is in additional human ESC lines, or in very early human embryos. When the gene set generated in the example above is compared with the human early embryo series of Vassena and colleagues, which was generated on the Affymetrix HuGene-1_0-ST microarray platform (Vassena et al., 2011), we can quickly see that the pattern of expression is split into two major clusters. The first cluster is highly expressed in the Barcelona hESC, as well as the 8dpc, morula and blastocyst stages. The second cluster shows highest expression in very early embryos, including the metaphase oocyte, 2dpc, 4dpc and 6dpc embryos. The images and the ranked gene list can be exported or shared via email from this page.

Annotation of the transcripts in this gene set through Stemformatics Workbench identifies the subset which contains a transmembrane domain, or signal peptide, and whose products are therefore likely to be present on the cell surface, or secreted into the local environment (Fig. 2). Additionally, genes that are annotated to KEGG pathways or other public gene sets are identified, and a simple over-representation test ranks pathways according to the degree of overlap. In this case study, several of the genes with the highest expression in the early embryo cluster were members of metabolic pathways, whereas the genes highest in the >8dpc/hESC cluster included members of protein synthesis and cell remodelling pathways. Approximately one third of each gene list contained gene products predicted to contain a transmembrane domain, making them possible targets for the development of novel hESC surface markers. More potentially secreted products were identified in the hESC cluster than the early embryo cluster. Further study of the potentially secreted products across the four hESC

Figure 1 An overview of the Gene search, experiment browser, and interactive graphs. Panel A demonstrates the free-text search and auto-suggest feature; Panel B demonstrates free-text filtering of experiments; Panel C shows the features of the Gene View boxplot. The interactive features include summarising data using different experimental parameters, such as grouping cell types or disease states. The lines intersecting the graph indicate the detection threshold and median expression line for each experiment. Stemformatics highlights platform ambiguities such as multiple measurements per gene, or probes that cross-hybridise with other genes, in this example POU5F1 has four probes on the Illumina microarray platform, and each probe cross-hybridises with other members of the POU family.


STEMFORMATICS
Login | Register| Forgot pass phrase
HOME | ABOUT US | HELP | SITE FEATURES | PUBLICATIONS | RECENT NEWS | CONTACT
GENE SEARCH
EXPERIMENT BROWSER >
WORKBENCH >
A
Gene Search -Please enter a gene name or identifier in the search box below.
POU5
SEARCH
POU5F1
Pou5f1
POU5F1B
POU5F1P2
POU5F1P3
POU5F1P4
POU5F1P5
POU5F1P6
POU5F1P7
POU5F2
B
Choose Dataset by clicking on the links below
Filter: pluripoten
Name
Title
Sample/Cell Types
Contact
Hough 2009 19890402
A Continuum of Cell States Spans Pluripotency and Lineage Commitment in Human Embryonic Stem Cells
hESC GCTM2(neg) CD9(neg), hESC GCTM2(lo) CD9(lo) hESC GCTM2(mid) CD9(mid) hESC GCTM2(hi) CD9(hi)
Martin F. Pera Andrew Laslett
Maherali 2008 1876420
A high-efficiency system for the generation and study of human induced pluripotent stem cells
fibroblast, iPSC, secondary iPSC, hESC
Konrad Hochedlinger
Spence 2011 21151107
Directed differentiation of human pluripotent stem cells into intestinal tissue in vitro
hESC, iPSC, derivatives of 3-day endodermal induction of hESC and iPSC
James M Wells
Bock 2011 21295703
Reference maps of human ES and iPS cell variation enable
hESC, iPSC, parental fibroblast, hESC-derived embryoid bodies
Kevin Eggan Alexander


C
Gene Search>POU5F1>>POU5F1 Summary>> Experiment Browser>> Hough_2009_19890402 Summary > Expression Results
Searching by Gene POU5F1 and Dataset Hough_2009_19890402
Scatter plot
Bar graph
Boxplot
Standard Deviation On
Standard Deviation Off
Sample Type
CHOOSE DATASET
POU5F1
CHOOSE GENE
MIN/MAX HORIZONTAL GRAPH
SHOW/HIDE LABELS
HELP
hESC GCTM2(neg) CD9(neg)
hESC GCTM2(lo) CD9(lo)
hESC GCTM2(mid) CD9(mid)
hESC GCTM2(hi) CD9(hi)
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Log(2) expression
Med (9.5)
DT (5.8)
ILMN_1660709*
ILMN_1665632*
ILMN_1679060*
ILMN_1705546*

fractions, P4–P7, could provide valuable information about intercellular communication and cellular metabolism within the heterogeneous hESC colony.

The utility of Stemformatics as an exploratory tool for stem cell researchers is highlighted by recent publications of Stemformatics users (Nayler et al., 2012) and (Vitale et al., 2012).

Workbench

Registered users have access to the Stemformatics Workbench area, which provides a private area to upload, save, analyse and share lists of genes. The Stemformatics team regularly engages with the stem cell community in order to assess how to best meet their bioinformatics needs. The Workbench in its present form has distilled the most common bioinformatics queries that the Stemformatics team is asked. The major workbench analysis features are summarised in Table 1, and Fig. 2.

Workbench access requires an account, which is used to remember an individual's job requests and stores the analysis results so that they can be securely accessed at any time within a three month window. Guest users who do not wish to register may access workbench via a public 'Guest account', which is available at the log-in screen. Most of the current workbench functions utilise GenePattern (Reich et al., 2006; Kuehn et al., 2008) as the analysis engine. Leveraging existing and well-established analysis tools such as GenePattern allows us to focus on bridging the user gap between the data and the tool and provides access to a wide range of gold-standard statistical tools.

The queries use short wizards which guide the user through a biology-centric workflow, hiding the computational pipeline in the backend. For example, when the user clicks on hierarchical clustering a short workflow directs the user to selecting the right gene list and experimental dataset to cluster. The user is then informed that the job was submitted and is directed back to the analysis area, or forward to the 'pending jobs' area. In the background, the application generates a GCT file, which is a format that GenePattern's 'HierarchicalClustering' module requires as input. It then submits this job to a local GenePattern server. In this example, the application has used the output of the 'HierarchicalClustering' module as input for the 'HierarchicalClusteringViewer' module of GenePattern. This workflow is automated and does not require any additional input from the user, utilising the pipelining feature of GenePattern. Both the image, and the gene list in clustered order, can be downloaded from Stemformatics.

Hamlet

Hamlet is a web-based tool for clustered genes and samples and generating heat-maps, and can be accessed from the summary page of each dataset. It provides users with an interactive tool to cluster and explore data housed in Stemformatics. The tool provides various clustering algorithms for the users to select from, allows changes in the colour scheme returned for the heat-map, and provides a zoom-in and zoom-out feature to explore gene membership across various clusters. Gene lists can be created by simple point, zoom and highlight commands, and then exported to the Stemformatics Workbench or downloaded to the user's computer.

Gene sets

Stemformatics provides a means to upload and save a list of genes as a gene set. The gene set may be the result of browsing or searches made within Stemformatics, or uploaded into the application through its Bulk Import Manager. All the saved gene sets are private and secure to each user, and it is simple to assess gene expression patterns across the list. Users can generate profiles from any of the Stemformatics datasets, visualising the entire list using the Hierarchical Cluster tool for gene sets of any size, or Histogram tool for gene sets with fewer than 50 members. Gene sets can be shared between collaborators using email notification links, and users may chose to publish their gene sets to the public gene set folder for general use by the stem cell community.

Geneset annotation

The Gene Annotation feature allows for viewing and filtering of a gene set on two main attributes. Using the annotation data available for Ensembl BioMart (Guberman et al., 2011; Zhang et al., 2011), any gene with a transmembrane domain or signal peptide is identified, and as can be seen in Fig. 3, a gene set can be quickly filtered on the type of domain annotation. Secreted proteins are predicted using SignalP, an online tool that predicts classical signal peptides by identifying the presence of recognition sites for signal peptidase I (SPase) enzyme (Kall et al., 2004). Pathway membership is determined using the public KEGG pathways (Kanehisa and Goto, 2000; Wrzodek et al., 2011) hosted by BioMart.

A simple gene set enrichment test provides the ability to examine the overlap between public gene lists, which include public-domain pathways, and the users own gene sets. The resulting gene lists can be ranked on p-value or by pathway membership and exported as a text file.

Figure 2 A case study (Panel A) demonstrating the multiview feature, identifying a single gene profile in four different datasets – in this example the gene CD9 is examined in HES2 fractions ((Hough et al., 2009) Illumina HumanWG-6V2 dataset), a reference map of hESC and iPSC ((Bock et al., 2011) Affymetrix HT-HG-U133A dataset), mapping the first stages of mesoderm commitment during differentiation of hESC ((Evseenko et al., 2010) Affymetrix HG-U133_Plus_2 dataset) and the transcriptional program initiation during human preimplantation embryonic development ((Vassena et al., 2011) Affymetrix HuGene-1_0-ST V1 dataset); (Panel B) Using the workbench tools to cluster and visualise across a heatmap the geneset derived from the GeneNeighbourhood analysis of CD9 to build a gene set of highly correlated genes from the Hough 2009 dataset; (Panel C) visualising the same CD9 correlated gene set using the Hierarchical Clustering tool in an independent dataset (early human embryo development of Vassena 2011); (Panel D) Using the workbench Annotate Geneset tool to identify transmembrane domains, signal peptides and overlapping KEGG pathways in a subset of genes identified from the CD9 correlated gene set.

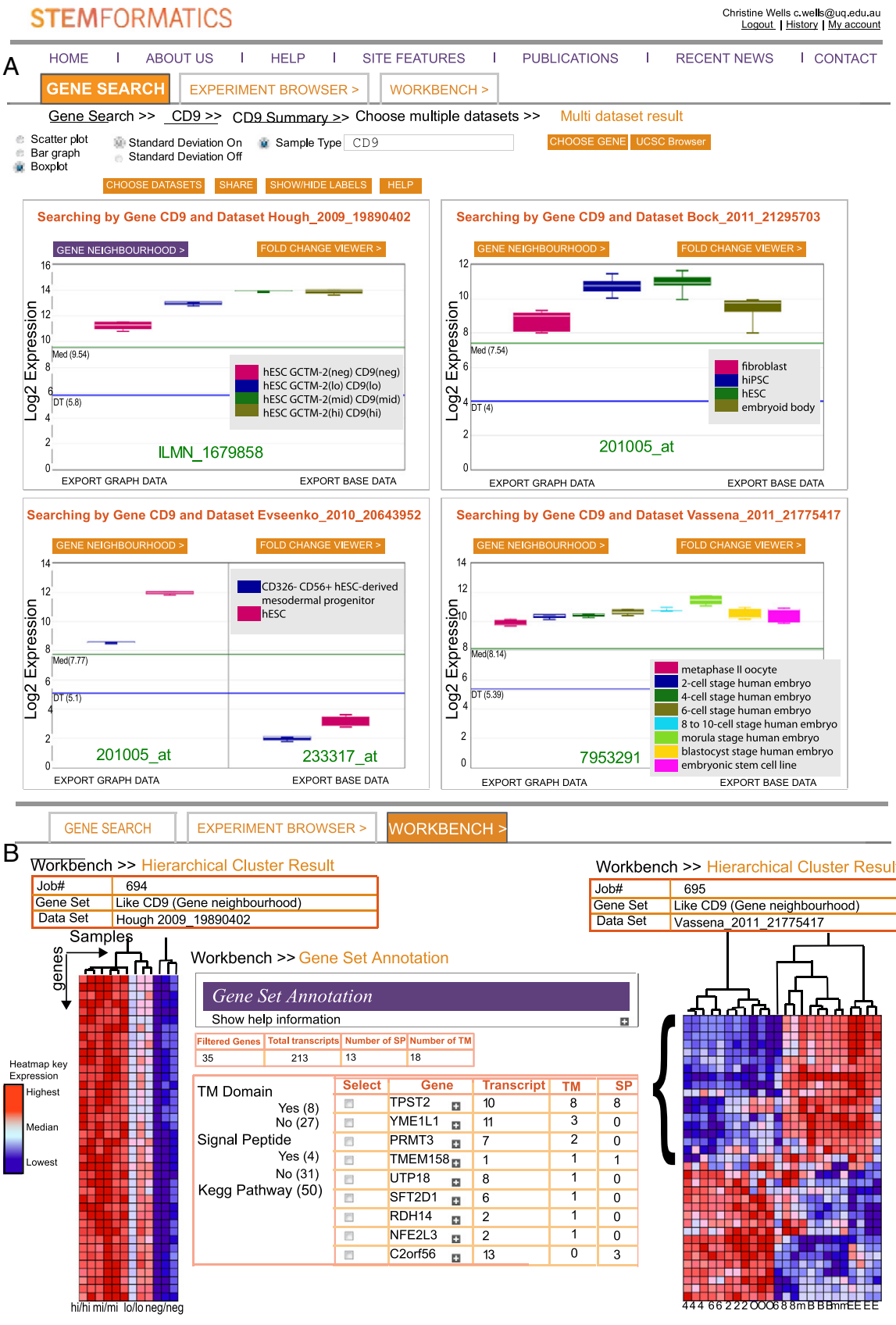

STEMFORMATICS
Christine Wells c.wells@uq.edu.au
Logout | History | My account
A
HOME | ABOUT US | HELP | SITE FEATURES | PUBLICATIONS | RECENT NEWS | CONTACT
GENE SEARCH
EXPERIMENT BROWSER >
WORKBENCH >
Gene Search >> CD9 >> CD9 Summary >> Choose multiple datasets >> Multi dataset result
Scatter plot
Bar graph
Boxplot
Standard Deviation On
Standard Deviation Off
Sample Type
CD9
CHOOSE GENE
UCSC Browser
CHOOSE DATASETS
SHARE
SHOW/HIDE LABELS
HELP
Searching by Gene CD9 and Dataset Hough_2009_19890402
GENE NEIGHBOURHOOD >
FOLD CHANGE VIEWER >
Log2 Expression
Med (9.54)
DT (5.8)
hESC GCTM-2(neg) CD9(neg)
hESC GCTM-2(lo) CD9(lo)
hESC GCTM-2(mid) CD9(mid)
hESC GCTM-2(hi) CD9(hi)
ILMN_1679858
EXPORT GRAPH DATA
EXPORT BASE DATA
Searching by Gene CD9 and Dataset Bock_2011_21295703
Med (7.54)
DT (4)
fibroblast
hiPSC
hESC
embryoid body
201005_at
Searching by Gene CD9 and Dataset Evseenko_2010_20643952
Med(7.77)
DT (5.1)
CD326- CD56+ hESC-derived mesodermal progenitor
hESC
201005_at
233317_at
Searching by Gene CD9 and Dataset Vassena_2011_21775417
Med(8.14)
DT (5.39)
metaphase II oocyte
2-cell stage human embryo
4-cell stage human embryo
6-cell stage human embryo
8 to 10-cell stage human embryo
morula stage human embryo
blastocyst stage human embryo
embryonic stem cell line
7953291
B
GENE SEARCH
EXPERIMENT BROWSER >
WORKBENCH >
Workbench >> Hierarchical Cluster Result
Job# 694
Gene Set Like CD9 (Gene neighbourhood)
Data Set Hough 2009_19890402
Samples
genes
Heatmap key
Expression
Highest
Median
Lowest
hi/hi mi/mi lo/lo neg/neg
Workbench >> Gene Set Annotation
Gene Set Annotation
Show help information
Filtered Genes 35
Total transcripts 213
Number of SP 13
Number of TM 18
TM Domain
Yes (8)
No (27)
Signal Peptide
Yes (4)
No (31)
Kegg Pathway (50)
Select
Gene
Transcript
TM
SP
TPST2 10 8 8
YME1L1 11 3 0
PRMT3 7 2 0
TMEM158 1 1 1
UTP18 8 1 0
SFT2D1 6 1 0
RDH14 2 1 0
NFE2L3 2 1 0
C2orf56 13 0 3
Workbench >> Hierarchical Cluster Result
Job# 695
Gene Set Like CD9 (Gene neighbourhood)
Data Set Vassena_2011_21775417
4 4 4 6 6 2 2 2 O O O 8 8 8 m B B B m m E E E E

Table 1 A summary of the analysis questions available through Stemformatics workbench.

Menu title	Task	Description
Manage my gene sets	Create a new gene set;	Allows users to create and save new gene sets and visualise gene expression across these gene sets in any of the
	Upload a file to create a new gene set	Stemformatics datasets using histograms.
	Manage current gene sets	Sort existing private gene sets, edit, view, share, export and delete.
	Annotate a gene set	Identify the gene products (transcripts and proteins) predicted to contain trans-membrane domains, or be secreted (contain a signal peptide). Identify members of pathways
	Download a gene set's expression profile	Download a text file or GCT GenePattern file for any list, populated with data from any Stemformatics dataset, for use in external analysis tools such as GenePattern or MeV
	View public gene sets	Lists of genes in the public domain, such as pathways. Lists identified from Stem Cell publications by the Stemformatics team, or shared by other Stemformatics users.
What questions can I ask?	Identify patterns shared between samples or within a given study	Hierarchical clustering and heat map visualisation (GenePattern module)
	Find genes differentially expressed between samples	Comparative marker selection (F-statistic, GenePattern module)
	Display my gene list as a graph	Build a histogram across any selected dataset
	Fold change viewer	Calculate the fold expression difference for any gene between two samples in a given dataset
	Current and pending analysis jobs	The list of jobs submitted and completed by a user over the last 3 months

Help features

Short videos showing users how to use the site features are provided on the home page, as well as the help menu of Stemformatics. All search bars 'suggest' the best match for the free-text entry provided by the user, which helps to overcome gene nomenclature changes and user typing error. These suggestions are prompted from matches to gene symbols, synonyms or accession numbers provided by Ensembl and NCBI Entrez gene, or from the platform identifier provided by each manufacturer. Help menus are provided on each page. Hovering the mouse over certain points provides additional information, for example on sample annotations, probe annotations or genomic features shown in the graph.

Links we like

The Stemformatics team has compiled a list of tools and databases relevant to the stem cell community that we find useful. The list includes links to many of the specialist stem cell databases that provide access to datasets across a range of specialist areas, or who provide different sets of analysis perspectives. Stemformatics is not affiliated with any of these links, but they reflect the tools that we use and routinely share with our collaborators.

Translation

Stemformatics is not a substitute for a good bioinformatics collaboration, but it does aim to provide a broad stepping stone between biologists, datasets and computational tools. Researchers wishing to undertake more sophisticated analyses may wish to utilise the workflows available through public analysis tool such as MeV or GenePattern. Stemformatics assists users by providing a format 'translation' feature, which allows users to download public experiments in formats ready for use in these external tools.

Future directions

Stemformatics has begun to integrate multiple 'omics' platforms for collaborators, and future releases will see inclusion of miRNA, proteomic, RNAseq and epigenomic datasets. The team continues to update the website with new datasets, and users of the site are encouraged to submit exemplar datasets for inclusion. Future releases will include improved interactive capacity with graphics, allowing users to hide or reorder samples in graphs. Additional annotations, including improved pathway annotations will be included as these become available.

Conclusion

Stemformatics.org is a resource that was built as a collaboration platform for Australian Stem Cell Science, is free to use and is now meeting an increasing demand from the international community. We provide a much needed interface between large, and often complex gene expression datasets and stem cell researchers who lack bioinformatics training. By placing common queries back in the hands of the stem cell community, we enable access to 'omics' data, future-proofing valuable public datasets that may otherwise languish in the big repositories, and importantly, allow our users to examine the reproducibility of gene profiles across multiple stem cell lines, under different growth conditions, isolation methods or differentiation protocols.

Author contributions

Contributions: Christine Wells: Conception and design, financial support, and manuscript writing. Rowland Mosbergen: Other: Design and implementation of user interface. Othmar

Korn: Other: Development of backend database, QC workflows and bioinformatics. Jarny Choi: Other: Development of backend database and bioinformatics. Nick Seidenman: Other: Development of backend database and bioinformatics. Nicholas Matigian: Assembly of data, Other: Annotation of datasets, and user feedback. Alejandra Vitale: Assembly of data, Other: Annotation of datasets, and user feedback. Jill Shepherd: Assembly of data, Other: development of stem cell metadata workflows, and manuscript writing.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.scr.2012.12.003.

Acknowledgments

Stemformatics is funded by a grant from the Australian Research Council Special Initiative in Stem Cell Science to CAW through Stem Cells Australia. Stemformatics is housed at the Queensland facility for Advanced Bioinformatics (QFAB). We thank Ted Liefield and Michael Reich from the Broad Institute for help with GenePattern integration. We thank Charles Willmore, Willmore designs, Brisbane Australia, for advice on website design and navigation.

References

Barrett, T., Troup, D.B., et al., 2011. NCBI GEO: archive for functional genomics data sets—10 years on. Nucleic Acids Res. 39 (Suppl. 1), D1005–D1010.

Bock, C., Kiskinis, E., et al., 2011. Reference maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. Cell 144 (3), 439–452.

Evseenko, D., Zhu, Y., et al., 2010. Mapping the first stages of mesoderm commitment during differentiation of human embryonic stem cells. Proc. Natl. Acad. Sci. U. S. A. 107 (31), 13742–13747.

Guberman, J.M., Ai, J., et al., 2011. BioMart Central Portal: an open database network for the biological community. Database 2011: bar041.

Hackney, J.A., Charbord, P., et al., 2002. A molecular profile of a hematopoietic stem cell niche. Proc. Natl. Acad. Sci. 99 (20), 13061–13066.

Hough, S.R., Laslett, A.L., et al., 2009. A continuum of cell states spans pluripotency and lineage commitment in human embryonic stem cells. PLoS One 4 (11), e7708.

Howe, E.A., Sinha, R., et al., 2011. RNA-Seq analysis in MeV. Bioinformatics 27 (22), 3209–3210.

Kall, L., Krogh, A., et al., 2004. A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol. 338 (5), 1027–1036.

Kanehisa, M., Goto, S., 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28 (1), 27–30.

Kuehn, H., Liberzon, A., et al., 2008. Using GenePattern for gene expression analysis. Curr. Protoc. Bioinform. 22:7.12.1–22:7.12.39 (Chapter 7: Unit 7).

Liu, X., Yu, X., et al., 2008. TiGER: a database for tissue-specific gene expression and regulation. BMC Bioinforma. 9 (1), 271.

Lukk, M., Kapushesky, M., et al., 2010. A global map of human gene expression. Nat. Biotechnol. 28 (4), 322–324.

Mallon, B.S., Chenoweth, J.G., et al., 2013. StemCellDB: the human pluripotent stem cell database at the National Institutes of Health. Stem Cell Res. 10 (1), 57–66.

McCall, M.N., Uppal, K., et al., 2011. The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. Nucleic Acids Res. 39 (Suppl. 1), D1011–D1015.

Muller, F.J., Laurent, L.C., et al., 2008. Regulatory networks define phenotypic classes of human stem cell lines. Nature 455 (7211), 401–405.

Muller, F.J., Schuldt, B.M., et al., 2011. A bioinformatic assay for pluripotency in human cells. Nat. Methods 8 (4), 315–317.

Nayler, S., Gatei, M., et al., 2012. Induced pluripotent stem cells from ataxia-telangiectasia recapitulate the cellular phenotype. Stem Cells Transl. Med. 1 (7), 523–535.

Novershtern, N., Subramanian, A., et al., 2011. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell 144 (2), 296–309.

Parkinson, H., Sarkans, U., et al., 2011. ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. Nucleic Acids Res. 39 (Suppl. 1), D1002–D1004.

Pavelin, K., Cham, J.A., et al., 2012. Bioinformatics meets user-centred design: a perspective. PLoS Comput. Biol. 8 (7), e1002554.

Pelekanos, R.A., Li, J., et al., 2012. Comprehensive transcriptome and immunophenotype analysis of renal and cardiac MSC-like populations supports strong congruence with bone marrow MSC despite maintenance of distinct identities. Stem Cell Res. 8 (1), 58–73.

Reich, M., Liefeld, T., et al., 2006. GenePattern 2.0. Nat. Genet. 38 (5), 500–501.

Sandie, R., Palidwor, G., et al., 2009. Recent developments in StemBase: a tool to study gene expression in human and murine stem cells. BMC Res. Notes 2 (1), 39.

Schulz, H., Kolde, R., et al., 2009. The FunGenES Database: a genomics resource for mouse embryonic stem cell differentiation. PLoS One 4 (9), e6804.

Vassena, R., Boué, S., et al., 2011. Waves of early transcriptional activation and pluripotency program initiation during human preimplantation development. Development 138 (17), 3699–3709.

Vitale, A.M., Matigian, N.A., et al., 2012. Variability in the generation of induced pluripotent stem cells: importance for disease modeling. Stem Cells Transl. Med. 1 (9), 641–650.

Wrzodek, C., Drager, A., et al., 2011. KEGGtranslator: visualizing and converting the KEGG PATHWAY database to various formats. Bioinformatics 27 (16), 2314–2315.

Wu, C., Orozco, C., et al., 2009. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol. 10 (11), R130.

Zhang, J., Haider, S., et al., 2011. BioMart: a data federation framework for large collaborative projects. Database (Oxford) (bar038).