Gene expression databases

Sean Eddy, PhD
Outline

• How to measure gene expression?
  – Microarrays/RNA-seq
• What are gene expression databases?
• Which ones exist?
• How do they differ?
• How can they be used?
• What needs to be taken care of?
• What can I do with specialized databases?
Microarrays: the beginning of high throughput gene expression

- Compartmentalized chips with sequences bound to the surface
- Sample is applied to surface
- Hybridizes to complementary sequence
- Hybridizations are quantified
  - No binding, “no” signal
  - Can only find what is specifically searched for
  - Usually represented as n x m matrix

http://www.nature.com/nrd/journal/v1/n12/images/nrd961-f1.gif
RNA-seq: the future of high throughput gene expression

- Samples (cDNA libraries) are applied to a sequencer
- Millions of sequences are generated and mapped onto a genome
- Sequence reads are quantified
  - No sequence = no expression
  - Can find novel transcripts, splice isoforms, fusion genes, etc.
- Quality of mapping depends largely on library prep and decisions made on how RNA is initially processed.
Gene expression databases

• Repositories for gene expression data
  – Mostly microarray and now RNAseq
  – Primarily for storage
  – Curated or un-curated
  – Access to data on different levels:
    • Datasets
    • Individual levels

• Integrated databases
  – Contain array data and additional data of the samples
  – Array data tends to be more annotated
  – More analytical tools
  – Smaller (more QC and curation needed)
  – Often no direct data access
Why do they exist

• Transparency/reproducibility of publications
  – Journals require data to be available for analysis
  – Nowadays raw data is required
  – Databases offer single resource and standardized access

• Data was generated for a specific purpose, but is not limited to that purpose
  – Can be reanalyzed in a different context
  – Can be combined with other datasets
  – Can be used as independent validation
Gene expression repository examples

• Gene expression omnibus (www.ncbi.nlm.nih/geo/)
  – 1,117,462 samples, 3848 datasets

• Array express (www.ebi.ac.uk/arrayexpress/)

• Princeton University MicroArray database (PUMAdb)
  – 40084 experiments, 6598 made public

• NCBI SRA, ENA and Princeton HTseq for NGS data
What is in a gene expression database?

• Gene expression data in different forms:
  – Resolution:
    • Gene level
    • Transcript level
    • Exon level
    – And / or raw data
  – Comprehensiveness
    • Targeted arrays
    • Whole genome arrays
  – Different platforms (microarrays, RNAseq)

• Generally only gene expression, may have limited sample information
Where does the data come from?

• Expression profiles of
  – Patients
  – Model systems
  – Cell cultures

• Data used for publication
  – Most journals now require raw data submission
  – Very coarse quality control (peer review)
  – QC depends mostly on authors

• Datasets submitted without publication
  – Little or no QC

• Most datasets are tailored towards a specific question
Example: GEO GSE32591

• Go to http://www.ncbi.nlm.nih.gov/geo/
• Enter GSE32591 into search box
• Click on “Analyze with GEO2R”
  – How would you set up the groups for analysis?
  – What do you get?
    • Does that make sense? How can results be verified?
• Go to “value distribution” tab
  – What do you see?
  – What are possible explanations?
Calculate the distribution of value data for the Samples you have selected. Distributions may be viewed graphically as a box plot or exported as a number summary table. The plot is useful for determining if value data are median-centered across Samples, and thus suitable for cross-comparison. 

GSE32591/GPL14663, selected samples
What can be done with GEO?
What can be done with GEO?

• Programmatic access for data download
  – [http://www.ebi.ac.uk/arrayexpress/help/programmatic_access.html](http://www.ebi.ac.uk/arrayexpress/help/programmatic_access.html) (ArrayExpress)

• Pre-computed analyses and on the fly analyses
  – Search by gene across all GEO experiments
  – Search by experiment to retrieve cluster analysis
  – Search by gene sequence for matching expression profiles
    • Described by Barret and Edgar, Methods Mol. Biol. 2006 “Mining Microarray Data at NCBI’s Gene Expression Omnibus (GEO)”
What questions can be answered?
What questions can be answered?

• If you download: anything
  – Only limited by your knowledge, skills, resources

• Pre-computed results
  – Preselected analysis methods/ sample groups
  – Generally within one dataset

• On-the-fly analyses
  – Sets of genes that cluster in under conditions given
  – Sample properties may not be entirely transparent.
What can be answered by doing it yourself?
What can be answered by doing it yourself?

• The quality of the data
  – Is part of the data low quality?
  – Does some of the data not fit into the set (e.g. batch effect, outliers for other reasons)
  – Is it adequately processed?

• What is the relationship between expression data and non-expression variables?
  – How does my gene (of interest) associated with experimental treatments, clinical parameters?

• What are patterns across datasets?
  – Does my finding hold up across similar analyses in independent datasets?
Why do you have to do it yourself?

• Quality control:
  – QC parameters are often glossed over in papers and in microarray submissions
  – For Affymetrix QC modules are available, freely available and widely accepted in the bioinformatic community
  – Other array types have distinct, but also similar properties
    – http://www.nature.com/nbt/focus/maqc/index.html

• Relations to non-expression data variables
  – Data is often not standardized within fields
Why not?

- Analysis across datasets:
  - Because:.... How?
  - Need to find a common standard for identification
  - Values need to be made comparable
    - If absolute expression values used, dynamic range can be a problem
    - Is ratios used, information about expression level lost
  - Non-expression data even worse
Who is the target group for doing it yourself?

- Users with experience in expression data
  - Crucial information (**STUFF**) is missing
Why is this a problem?

- Excludes investigators with good hypotheses but lacking bioinformatic skills

“Must be a clinical fellow.”
How to fix that?
How to fix that?

• Specialized databases
  – Datasets are easier to find
    • Datasets relevant to specific areas are collected in one place
      – NephroSeq for renal disease
      – Oncomine for cancer
  – Datasets are standardized and expertly curated
    • Controlled vocabulary is introduced for non-expression data
    • Curation of expression possible by introducing standardized references and data transformations across datasets
      – Gene IDs/Gene Symbols as references
      – Z-transformation or median centering of log transformed expression data
Developed for the renal research community, Nephroseq is a platform for integrative data mining of genotype/phenotype data, with optimized workflows that lead from search to visualization and from question to answer to next question:

- The expression of a gene is highly correlated with well-known podocyte genes. Is the gene functionally important in glomeruli?
- A gene is significantly differentially expressed in a subset of disease patients. Is the gene associated with a certain phenotype, severity or sub-category of the disease?
- A set of genes is significantly up-regulated in disease patients. Are the disease genes inversely related to the target profile of a compound/drug?

ABOUT NEPHROSEQ

Originally a collaborative effort, Nephroseq is now solely developed and maintained by the Applied Systems Biology Core at the University of Michigan. This resource combines a wealth of publicly available renal gene expression profiles - gathered and curated by an experienced team of data scientists, bioinformaticians, and nephrologists - with a sophisticated analysis engine and powerful web application designed for data mining and visualization of gene expression data.

Nephroseq provides researchers with a rich set of publicly available renal gene expression data, packaged with the tools and interface necessary to analyze it, all aimed at seeking answers to questions and advancing a molecular understanding of kidney disease to ultimately improve clinical outcomes.

In particular, Nephroseq provides unique access to datasets from the Personalized Molecular Nephrology Research Laboratory incorporating clinical data which is often difficult to collect from public sources.
Oncomine (www.oncomine.com www.oncomine.org)

Oncomine™ Research Edition: 715 datasets and 86,733 samples

Design better experiments. Gain more insights. Prepare to publish faster.

With Oncomine™ Research Premium Edition, you can:

Design better experiments...Answer more questions with fewer experiments, select the most promising gene or cell line, and test your hypothesis.

Gain more biological insights...Discover novel targets for therapeutic development, interrogate gene expression profiles, and identify drug and biological interactions.

Prepare to publish faster...Validate your results faster, visualize data easier and make connections to clinical significance.

The Oncomine™ Platform—from web applications to translational bioinformatics services—provides solutions for individual researchers and multinational companies, with robust, peer-reviewed analysis methods and a powerful set of analysis functions that compute gene expression signatures, clusters and gene-set modules, automatically extracting biological insights from the data. It has become an industry-standard tool cited in more than 1,100 peer-reviewed journal articles. The Oncomine Platform has been used as a foundation for ground-breaking discoveries with unique features that include:

- Scalability — with 700+ independent datasets
- High quality — with expertly curated data
- Consistency — with a rich, extensive and controlled ontology of terms
- Standardized analysis — with conventions that assure clear and consistent interpretations of results

Oncomine Research Edition remains free to the academic and nonprofit cancer research communities.
NephroSeq and Oncomine

• Pros:
  – Each focus on one area of interest
  – Clinical data for many individual samples available
  – Advanced analysis using integrated systems biology tools in a pre-defined automated manner
  – Meta analysis possible
  – User friendly, free accessible for academic users
  – **Hypotheses-generating**

• Cons:
  – No raw data download
  – No programmatic access
  – Only predefined analyzes
NephroSeq main Page

26 datasets (2000 samples)

Analysis type:
- Coexpression analysis
- Differential analysis
- Outlier analysis

Welcome to nephroseq:

This application is a web-based analysis engine for molecular biology researchers and clinician scientists who study renal disease and related disorders. Nephroseq gives access to renal genome-wide gene expression datasets generated by the renal research community. This tool is especially powerful because the data are already pre-analyzed and datasets include clinical data. The 3-tiered user interface moves users from left to right within the application to choose data, sort and prioritize analyses, and visualize and export results.

Analyses that are available include:
- Differential Expression: Identify over- or under-expression for a particular gene
- Coexpression: View genes that are coordinately expressed with your gene of interest across a dataset
- Outlier: Identify outlier patterns where a gene is highly over-expressed in a fraction of samples
- Concept Associations: Identify significant overlap between gene sets that represent underlying biology

In addition, users have the ability to upload gene lists to use as filters and export data and visualizations directly to Excel, PowerPoint and SVG.
Two Search Options

• Gene specific search:
  – Gene

• Dataset search:
  – Specific conditions/diseases
Gene Search

Gene summary view

NPHS2: encodes podocin, a podocyte specific protein

Dataset/Disease type

Demographics
Gene Search

Gene summary view

22 out of 33 analysis meet your threshold for NPHS2 in 2 out of 2 datasets
Four Basic Analysis Modes

• Differential expression
• Co-expression analysis
• Outlier analysis
  – Heterogeneity within predefined groups
• Concepts analysis
  – Gene set (Nephromine & third-party sources)
Gene Search

Differential expression (Box graph)
Tubulointerstitial

Glomeruli
Gene Search

Correlation with clinical continuous variation

Gene: VCAN
Analysis type: GFR
Dataset Type: Diabetes

Legend
1. < 15 ml/min/1.73m² (3)
2. 15 - 29 ml/min/1.73m² (4)
3. 30 - 59 ml/min/1.73m² (7)
4. 60 - 89 ml/min/1.73m² (5)
5. > 90 ml/min/1.73m² (3)

P = 7.71E-7
Correlation: -0.832
Gene Search
Outlier analysis

Outlier analysis helps to identify an expression profile where differential pattern is only seen in a fraction of samples of all patients within a disease type.

**Why do we need it:** 25% of patients show over-expression of a gene. This gene may not generate a significant p-value in a t-test comparing DN relative to normal kidney.

**How to do it:** Transform all samples within a dataset, so that genes could be ranked by their expression from high to low. The data transformation is performed at certain percentile bins (75, 90 & 95%), and a line is drawn at the percentile of that analysis to define outliers.

For example, in an outlier analysis at the 75th percentile, the system draws a line at the point at which only the top 25th percentile samples extend above it.
Gene Search
Outlier analysis

Schmid Diabetes (22)

Controls
Diabetic

Legend
1. Cadaveric Donor Control (4) 3. Minimal Change Disease (4)
2. Healthy Living Donor (3) 4. Diabetic Nephropathy (11)
Differential expression – Dataset search

Export
Differential expression – dataset search – compare analysis

• Compare different analyzes
• Data is standardized on upload (centered to 0 and standardized by variance)
• all features are mapped to common identifier (EntrezGeneID)
Meta analysis

• Find out which genes are significantly more expressed in glomeruli compared to tubulointerstitium
• Can you verify that with another dataset?
• Or with more than one other dataset?
• Does it matter if the datasets are different?
• Can you imagine a use of this functionality for an exclusive filter (NOT)
Example
Concepts Analysis

**Concepts** are sets of genes representing some aspect of biology.

Concepts are derived from both *Nephromine gene expression signatures* as well as *third-party sources* such as Gene Ontology, KEGG Pathways, Human Protein Reference Database, etc.

User can upload a self-defined custom concept (a set of genes) to Nephromine to explore it’s association with Nephromine and third-party concepts.
Concepts Analysis

Upload Custom Concept
Manage My Concepts
Change password

Download list from C-tools to the desktop, then upload

The press "validate"
Concepts Analysis
Upload

<table>
<thead>
<tr>
<th>Concept Name:</th>
<th>Podo-50-symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Set (Text File):</td>
<td>podocyte-50_gene_symbol.txt</td>
</tr>
<tr>
<td>Category:</td>
<td>HUGO Gene Symbol</td>
</tr>
</tbody>
</table>

Null Set(s):
- All Entrez Gene IDs
- CodeLink Human Whole Genome Bioarray

Description (Optional):

50 terms were recognized as distinct HUGO gene symbols and will be uploaded.

Concept (Podo-50-symbol) validated successfully.

Then press “Upload”
Concept (Podo-50-symbol) was successfully uploaded and can be viewed in My Concepts.

Select (Podo-50-symbol) as primary concept now.
### Concepts Summary View

**Nephromine Concept Summary**

- **Threshold (fold change):** 2.0
- **Threshold (p-value):** 1.4
- **Data Type:** All

**Associated Concept Summary for "Podo-50-symbol - My Concepts"**

#### Nephromine Concept Summary

<table>
<thead>
<tr>
<th>Concept Type by Dataset Type</th>
<th>Demographics</th>
<th>Donor Type</th>
<th>Group</th>
<th>Pathway Concepts</th>
<th>Regulatory Concepts</th>
<th>Comparative Map V2 Drug Signatures</th>
<th>Literature-defined Concepts</th>
<th>Mutation Concepts</th>
<th>My Concepts</th>
<th>siRNA Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Age</td>
<td>Risk Index</td>
<td>Race</td>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Aging</td>
<td>Diabetes</td>
<td>FSGS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hypertension</td>
<td>IgAN</td>
<td>Pathway Analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lipid</td>
<td>Normal Tissue Enrichment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Other (Non-Nephromine) Concept Summary**

<table>
<thead>
<tr>
<th>Biological Annotation</th>
<th>Pathway Concepts</th>
<th>Regulatory Concepts</th>
<th>Comparative Map V2 Drug Signatures</th>
<th>Literature-defined Concepts</th>
<th>Mutation Concepts</th>
<th>My Concepts</th>
<th>siRNA Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>7</td>
<td>3</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **4 concepts meet your threshold and are associated with the primary concept**
Coll FSGS vs. Normal Kidney
Nephromine Gene Expression Signatures
$P=1.54E-18$, $q=1.15E-14$, Odds=18
Top 5% Under-expressed
Hodgin FSGS
Concepts Analysis

PowerPoint
Publication-quality graphic (SVG)
Excel - Analysis Comparison
Excel - Analysis Gene List
Excel - Dataset Detail
tranSMART
The Translational Challenge: Data Integration & Analysis

Athey and Omenn, 2009
tranSMART Platform: Enabling Translational research

tranSMART – A platform and community

- Open-source and open-data translational biomedical research community
- Biomedical Researchers, Developers, Service Providers
- Clinician Researchers
tranSMART Platform: Academics and industry

2009 Johnson and Johnson

2010 Sage Bionetworks

2010 Thomson Reuters

2012 FDA

2012 One Mind for Research

2012 Pfizer

tranSMART: controlled vocabulary
Subset selection

Can further specify with AND or exclusion
### Summary statistics 1

#### Query Summary for Subset 1

<table>
<thead>
<tr>
<th>Category</th>
<th>Subset 1 (n)</th>
<th>Subset 1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>FEMALE</td>
<td>3</td>
<td>27.3%</td>
</tr>
<tr>
<td>M</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>MALE</td>
<td>8</td>
<td>72.7%</td>
</tr>
<tr>
<td>UNKNOWN</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>null</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>100%</td>
</tr>
</tbody>
</table>

#### Query Summary for Subset 2

<table>
<thead>
<tr>
<th>Category</th>
<th>Subset 2 (n)</th>
<th>Subset 2 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>FEMALE</td>
<td>3</td>
<td>60%</td>
</tr>
<tr>
<td>M</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>MALE</td>
<td>2</td>
<td>40%</td>
</tr>
<tr>
<td>UNKNOWN</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>null</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Total</td>
<td>5</td>
<td>100%</td>
</tr>
</tbody>
</table>

#### Summary Statistics

**Histogram of Age**

**Comparison of Age**

- **Subset 1**
  - Mean: 43.82
  - Median: 41
  - IQR: 13
  - SD: 18.84
  - Data Points: 11

- **Subset 2**
  - Mean: 29.4
  - Median: 20
  - IQR: 8
  - SD: 6.87
  - Data Points: 5

#### Race

**Subset 1**

- MALE: 72.7%
- FEMALE: 27.3%
- Total: 100%

**Subset 2**

- MALE: 40%
- FEMALE: 60%
- Total: 100%
Differentially expressed genes

### Table of top Markers

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Probe ID</th>
<th>Raw p-value</th>
<th>Bonferroni</th>
<th>Holm's</th>
<th>Hochberg</th>
<th>Sidak α</th>
<th>BH</th>
<th>BY</th>
<th>t (permutation)</th>
<th>Raw P (permutation)</th>
<th>Adjusted P (permutation)</th>
<th>Rank</th>
<th>S1 Mean</th>
<th>S2 Mean</th>
<th>S1 SD</th>
<th>S2 SD</th>
<th>Fold Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>CENPE</td>
<td>A_16_P16795940</td>
<td>0.00000</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00205</td>
<td>-5.789798</td>
<td>0.004578755</td>
<td>0.2426740</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>0.2426740</td>
</tr>
<tr>
<td>LTF</td>
<td>16952883</td>
<td>0.00000</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00205</td>
<td>-5.789798</td>
<td>0.004578755</td>
<td>0.2426740</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>0.2426740</td>
</tr>
<tr>
<td>CORO1A</td>
<td>16817824</td>
<td>0.00000</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00205</td>
<td>-5.789798</td>
<td>0.004578755</td>
<td>0.2426740</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>0.2426740</td>
</tr>
<tr>
<td>ERC2</td>
<td>A_16_P16234993</td>
<td>0.00000</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00205</td>
<td>-5.789798</td>
<td>0.004578755</td>
<td>0.2426740</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>0.2426740</td>
</tr>
<tr>
<td>chr5:095910759-095910818</td>
<td>A_16_P17221956</td>
<td>0.00000</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00205</td>
<td>-5.789798</td>
<td>0.004578755</td>
<td>0.2426740</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>0.2426740</td>
</tr>
<tr>
<td>ILV310</td>
<td>16932966</td>
<td>0.00000</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00205</td>
<td>-5.789798</td>
<td>0.004578755</td>
<td>0.2426740</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>0.2426740</td>
</tr>
<tr>
<td>MARK7</td>
<td>A_16_P17727416</td>
<td>0.00000</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00205</td>
<td>-5.789798</td>
<td>0.004578755</td>
<td>0.2426740</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>0.2426740</td>
</tr>
<tr>
<td>RNU5473</td>
<td>A_16_P1668380</td>
<td>0.00000</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00205</td>
<td>-5.789798</td>
<td>0.004578755</td>
<td>0.2426740</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>0.2426740</td>
</tr>
<tr>
<td>AK052155</td>
<td>A_16_P17081075</td>
<td>0.00000</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00205</td>
<td>-5.789798</td>
<td>0.004578755</td>
<td>0.2426740</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>0.2426740</td>
</tr>
</tbody>
</table>

Enlarged:
Comparisons can be saved/ emailed
tranSMART – why do we care?

- Enables data exploration with low hurdles
- Integrates many different data types
- Has interfaces to real analysis tools
- Provides a consistent data set
- Can be run locally/ institutional etc
- Can possibly be “shared” across institutions
  - McMurry et al, PLOS one: *Shrine: enabling nationally scalable Multi-site disease studies*

- Go to: [http://transmartfoundation.org/](http://transmartfoundation.org/)
Acknowledgements

Matthias Kretzler
Felix Eichinger
Wenjun Ju
Sebastian Martini
Viji Nair
Celine Berthier
Laura Mariani
Becky Steck
Colleen Kincaid-Beal
Rachel Dull

Daniel R. Rhodes
Rodney Keteyian
Becky Steck
Colleen Kincaid-Beal
Rachel Dull
Homework for fun

• Connectivity map
  – Use Diabetes vs. control (tubulointerstitium dataset)
  – Select top 1% overexpressed as primary concept
  – Compare to significantly overlapping concepts with Connectivity map
  – Can you find potential drug candidates? Are there any drugs that work for both glom. and tub?
  – What could be optimized? How will you plan further experiments to test your hypothesis?
In most nephrons, the loop of Henle is relatively short and is located in the cortex.

In some nephrons, the loop of Henle is long and plunges into the medulla.

Final urine to ureter