Systems Immunology:

A viewpoint from the Immunological Genome Project
The Immunological Genome Project

- Generate rigorously comparable expression datasets for all populations/states in the mouse immune system (>200 states)

- Computationally work out regulatory connectivity, using genetic and molecular perturbations to validate and refine the network
Members:

- Regev lab
- Hardy lab
- Laidlaw lab
- Wagers lab
- Rossi lab
- Stem Cells
- Goldrath lab
- T cell activation
- Benoist / Mathis lab
- Brenner lab
- aβ T Cells
- Merad lab
- Randolph lab
- Turley lab
- Myeloid Cells
- Kang lab
- γδ T Cells
- Lanier lab
- NK Cells
- UMass
- UCSF

Advisory Board:
L. Berg  
D. Littman  
J. Quackenbush  
D. Raulet  
L. Staudt  
E. Wakeland

ImmGen participants:
Phase-1: all on Affy ST1.0 microarrays
All JAX B6 mice, all sampled @ 9 a.m.
Each lab sorts cell populations according to identical SOP
RNA prepared, labelled, hybridized together
Strict data QC (~8% dropout), normalization

Data are rapidly public upon curation, “Ft Lauderdale principles”
www.immgen.org
Scope:

The full spectrum
B Lymphocytes

Hardy lab, Fox Chase, Philadelphia
Myeloid cells

Turley, Merad, Randolph labs
(Dana Farber, Mt Sinai)

Altogether, aiming to profile
~ 250 different cell types/states
Status?

Phase-1: Microarray compendium
- 651 samples sorted and processed
- 580 samples passing QC
- 162 populations "partial to complete"

Web server active, 2 display formats
~ 70% there
Status?

**Phase-1: Microarray compendium**
- 651 samples sorted and processed
- 580 samples passing QC
- 162 populations “partial to complete”
- Web server active, 2 display formats

**Phase-2:**
- Genetic variation, age
  - Network perturbations: genetic, chemical, RNAi
- RNA-seq (deep discovery, splicing, miRNA)
What makes this project work is that each lab is asking specific questions, while contributing to building the whole project.

- Kang lab: cell populations that express different Vg genes are really different beasts
- Turley/Merad/Randolph labs: location matters a lot to DCs
- Brenner lab: novel distinctions between NKT subsets
- CBDM lab: when does a T cell become a T cell
No other biological system has the breadth of knowledge and tools to tackle lineage variation at this level of resolution
ImmGen is an open resource project:
- All data are public
- Open to suggestions/participation...
Integrating it all:

What is the true landscape of transcription?
The genome is very active

Radu Jianu, Jeff Ericson, David Laidlaw
A few coregulated segments, but not the norm
Integrating it all:

Distribution of transcription factors
Transcription Factors: all…

- ~ 50% are expressed in immune cells
- Uniformity is the majority
The “one factor, two cells, three functions” theme (Tbet, Irf4, Batf)
Integrating it all:

Can we better define populations and relations between them?
How are myeloid and lymphoid cells related?

 Won’t enter into this debate,  
   But relative distances between lymphoid/myeloid populations should be very relevant to the question
Multidimensional scaling
(best approximation in 2D space of distances in N dimensional space)

But the relative disposition of mature T, B and myeloid cells may be the most intriguing

All stem and pro-/pre-T and B cells bunch together

γδT cells are interspersed between αβT cells

NK cells are close to activated T cells

mdscale function, Pearson distance, transcripts with 0.2 quantile highest interpopulation CV
How are myeloid and lymphoid cells related?

Not really…
(Myeloid cells cluster together
No evidence for T-like or B-like myeloid cells)

… but maybe not either…
(Myeloid cells are clearly more B- than T-like)
Integrating it all:

“Reverse-engineering” of the genetic regulatory network
**Regulation as Linear Regression**

The goal is to define modules of co-regulated genes, and the putative regulators that control them.

\[ E_{\text{Targets}} = w_1 x_1 + \ldots + w_N x_N + \epsilon \]

“Elastic Net” regression (evolved from Tibshirani’s Lasso regression)

\[ \minimize w (w_1 x_1 + \ldots + w_N x_N - E_{\text{Targets}})^2 + \sum C |w_i| + \sum D w_i^2) \]

Induces sparsity (reduces large number of negligible / low influences while preserving “true” regulatory influences in the model).

**Particularities in the ImmGen context:**
- Try to use the lineage/population structure in the ImmGen data to support/refine the reconstruction/predicting “true” regulatory in:
- How common are regulatory controls across different lineages?
- Try to work out population relationships when they are unclear (e.g. myeloid cells)

Vladimir Jojic, Daphne Koller
Tal Shay, Aviv Regev
T-cell Selection Module

Enriched for genes activated upon positive selection

Inhibitor of Tgf-beta signaling pathway, responsible for T-cell differentiation and apoptosis
## Context Likelihood of Relatedness

### Similar goals, different mathematics...

<table>
<thead>
<tr>
<th>Condition</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>m</td>
<td></td>
</tr>
</tbody>
</table>

### Mutual Information

\[ Z = \frac{X - \mu}{\sigma} \]

- Select the Regulator/Gene pairs that are the most likely from each distribution
- Also incorporate the modules (from Tal/Vladimir) and GeneOntology classifiers

**James Costello, Jim Collins**
Of regulatory interactions, how many are lineage-specific, how many operate across lineages?
Context Likelihood of Relatedness

James Costello, Jim Collins
Network/cluster analysis

Input:

- All cells
- Compendium dataset

“Perturbed” data

- Genetic perturbation (outbred mouse strains)
- Small molecules
- Genetic switches
- RNAi

Lineage branching, but still a “static” view of regulatory network
How do you get knowledge from data like these?

Designing new tools to display these data and meta-data

Radu Jiany, David Laidlaw (Brown SciVi)
Systems biology:

the strip-tease problem

Too much information.
Participants:
Alexis Battle, Christophe Benoist, Scott Davis, Graeme Doran, Anancia Goldrath, Nir Hacohen, David Hafler, Philip de Jager, Vladimir Jojic, Daphne Koifer, Mark Lee, Diane Mathis, Khadr Kaddus, Aviv Regev, Suzanne Simmons, Barbara Stranger, Michael Wilson, Irene Wood, Xia Zong, ImmGen participants.
600 healthy donors
(200 African, 200 East Asian, 200 European)

Activated DC responses
Hacohen

Immunophenotyping
(PBMC, ~50 markers)
Hafler

Baseline gene expression

Activated T cell responses
cbdm

SNP map
(10^6 dense)
De Jager
600 healthy donors

Immunophenotyping (PBMC, ~ 50 markers)

Baseline gene expression

Activated DC responses

Activated T cell responses

SNP map (10^6 dense)

1. Purified CD4+ naïve T and CD14+ monocytes
2. 2 platform microrray profiling
3. Multiplexed RNAseq for miRNA
4. RNAseq (discovery)

Compare variation in sorted cells, whole PBMC, whole blood

Parallel eQTL in 2 cells
1. Determine subset of genes that can respond in activated T cells
   (TCR +12 different costim, 33 different cytokines, other signaling ligands > ~100 conditions)
2. Compute “representative” proxy signature set
3. Detailed analysis, limited donor set, full network reconstruction
   (time-course, combinatorial stimulation, Ab blockade - Nanostring)
4. Full donor set, 2-3 stimulation conditions (“response eQTL”)
1. Determine subset of genes that can respond in activated DC cells (TLR +12 different costim, 33 different cytokines, other signaling ligands > ~100 conditions)

... 

... 

...
Thank you for your attention