UMASS CCTS Seminar Series
The Bioinformatics Core
and
The Garber Lab

Manuel Garber
Nov 29th 2012
Plummeting cost of DNA sequencing

- Capillary
- Automated capillary
- 454
- Illumina GA
- Illumina GAII
- Illumina GAIIx
- Illumina HiSeq

Redefined the use of sequencing

Estimated cost per Mb in large scale centers

Cost per Q20 megabase ($) vs. Year

- Moore's Law
- $ / MQ20

1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010

$20,000 / Mb

$0.2 / Mb

~100,000X
Illumina Length, quality, yield continue to improve

**Base position**

**Base quality**

2007

2008

2009

HiSeq

Passing filter (%)

36b, GAII, Jan. 2007

51b, GAII, Jan. 2008

76b, GAII, Jan. 2009

101b, GAIIx, Jan. 2010

101b, HiSeq, April 2010
Two step approach – sequence is expensive
• Sequence once, a reference
• Build arrays to explore samples

Toolkits
- Affy/Agilent expression arrays
- Affy genotyping arrays
- Conservation databases

Achievements
• Reference human and other species genomes
• Reference transcriptome
• Reference variation map (HapMap)

Which has drastically change informatics, from
To DNA sequence being general purpose tool

- Normal human variation and association studies
- Human genetics and gene discovery
- Cancer genomics
  - Map translocations, CNVs, structural changes
  - Profile somatic mutations
- Genome assembly
  - Virus
  - Bacteria/fungi
  - Mammals
- Transcriptomics
  - Comprehensive genome annotation
  - Expression dynamics (DGE)
  - micro- and small RNAs
  - Immunogenomics
- Epigenomics
  - Map histone modifications
  - Map DNA methylation
- Polymorphism/mutation discovery (SNPs and structural)
  - Bacteria
  - Genome dynamics/directed evolution
  - Exon (and other target) sequencing
  - Disease gene sequencing
- Ancient DNA (Neanderthal)
- Pathogen discovery
- Metagenomics
  - Human microbiome
“The result is that the ability to determine DNA sequences is starting to outrun the ability of researchers to store, transmit and especially to analyze the data.”

“Data handling is now the bottleneck,” said David Haussler, director of the center for biomolecular science and engineering at the University of California, Santa Cruz. “It costs more to analyze a genome than to sequence a genome.”
Challenges of big data

- Large datasets are “easy”, “fast”, and “cheap” to generate. But they are time consuming and expensive to analyze.
- Looking at data is **crucial** to data analysis.
- Thinking about **how to analyze** the data is crucial.
- *Data analysis involves many similar steps with only variations on the approach*

My Goal is to have a core focused on enabling sequence data analysis.
Steps for analyzing NGS data

*Biological insight*

**Finding signal**, what are my differentially expressed genes, which peaks are in samples vs. controls etc.

**Data processing**
Sequences → Mapping, assembly, peak calling, transcript quantification → tables, browser tracks

**Data setup**: format and make data accessible to bioinformatics programs
Most of the technical, computationally intensive is generic and “core”

Biological insight

Finding signal, what are my differentially expressed genes, which peaks are in samples vs controls, etc

Data processing
Sequences → Mapping, assembly, peak calling, transcript quantification → tables, browser tracks

Data set up accessible to bioinformatics programs
Maximum impact

Biological Insight → Results, grants, papers

Provides standard analysis options and supports analysis developed at UMASS

Implements best of breed, UMASS specific data processing pipelines

Eases data access and manipulation
An informatics community around the core

Education:
- How to interpret processed data
- Statistics
- Visualization

Training sessions:
- Hands on training
- Consultation
- Q&A

Development
- New tools
- New pipelines

Community
- Regular presentations
- User discussion forum
Current Progress

• Bioinformatics seminar (We’ve had 2 so far, all are welcomed)
  – Occurs the second Friday of every month at 11:00 am
  – Two short talks
    • A computational talk:
      – Algorithm
      – Pipeline
      – Method
    • A data centric analysis:
      – An integrative analysis
      – A preliminary analysis that looks for feedback
      – Data from novel techniques

• Redesigned Website
  – Dynamic website with documentation on pipelines
  – Hot-ticketing system when users experience problems with core-supported pipelines

• One Hire and one more being recruited

• Listserv for online discussions: bioinfo@list.umassmed.edu
Short Term Goals

- Integrate user Galaxy and similar tools with the HPC cluster
- Implement standard pipelines for
  - RNA sequencing analysis
    - including small RNA
  - ChIP sequencing analysis
  - Variant calling from deep sequence or exome data
- Make this pipelines available through Galaxy so that most users can take advantage of them
A typical pipeline (e.g. RNA-Seq)

- Upload your sequence data (fastq)
- Align to the ribosome (Bowtie)
- Output ribosomal contamination metrics report
- Align remaining reads to genome (TopHat)
- Produce RNA-Seq report % aligned, % intergenic, % exonic, % UTR
- Produce IGV and UCSC friendly files
- Quantify transcriptome
- Produce a table with normalized expression values
- Call differentially expressed genes (if multiple samples)
- Report pairwise significant genes that are differentially expressed
Pipelines will be available in the HPCC cluster

- For those unafraid of UNIX, pipelines will be available to execute from the command line:
  - Write a script

```
#!/bin/bash
# Now set up some environment stuff

export PATH=/share/apps/bin:$PATH

tophat --num-threads 4 --GTF /seq/lincRNA/data/mm9.mrna.10.31.gtf --prefilter-multihits --output-dir ACTAAG.th1.4.1.g15 --segment-length 20 --max-multihits 15 /seq/lincRNA/data/mm9.nonrandom.bowtie /seq/dcchip/mouse/DC/rnaSeq/DGE/hiseq_1-20-12/split_fqs/ACTAAG.fq

- And submit to the server farm
```

Of course not everyone is comfortable with UNIX and scripting!
And there is a solution for this ...
Goal: abstract the technical complexity, let labs leverage their intuition

Repeat analysis using different parameter settings
Specific analysis incorporating biological insight
Custom analysis
Custom figures
Writing my paper

Provides standard analysis options and supports analysis developed at UMASS

Implements best of breed, UMASS specific data processing pipelines

Eases data access and manipulation
Help for the last step

• Discussion:
  – Monthly meeting, 2\textsuperscript{nd} Friday of every month at 11:00 am all are welcomed.
    • Next meeting: methods to annotate and quantify piRNAs
  – Mailing list: bioinfo@list.umassmed.edu

• Training:
  – Invited speakers:
    • Genome Space team (February)
    • R workshop (possibly in March)
    • Planning an RNA-Seq analysis workshop
  – Training on supported tools and methods

Questions on the bioninformatics core?
WHAT DOES MY LAB DO?
The genome encodes many different elements

• What is out there?
• Non-coding genes
  – Finding them
  – Characterizing function
  – Mechanism and evolution
• Regulatory elements
Each element set has a unique code

- Genetic Code
- Cis-regulatory code
- RNAi and miRNA codes
- RNA Code
- Histone Code
Most of which we do not understand

✓ Genetic Code

? Cis-regulatory code

? RNAi and miRNA codes

? RNA Code

? Histone Code
Most of which we do not understand

✓ Genetic Code
?
?
?
?

RNAi and miRNA codes
?

RNA Code
?

Histone Code
Our work

Estimate the “functional genome” by finding what is under selection

- Develop informatics tools for new methods
- Develop models of transcriptional regulation
- Develop models of epigenetic interactions
- Evolution of large non-coding RNAs
Project: Transcription regulation in DCs

Input

- Bacteria
- Viruses

Output

- Anti-bacterial program (inflammation)
- Anti-viral program (interferon)

How is this response controlled?

1800 genes

Amit, Garber et al, Science 2010
Strategy: Genetic + physical mapping

What are the direct targets of transcription factors?
Only possible with High throughput ChIP-Sequencing

Crosslink → Sonicate → Precipitate → Purify DNA → Indexed library → Sequence

Magnetic beads

Robot-automated

24 libraries/lane

Ido Amit
Systematic mapping of the DC LPS-response network

- RNA-Seq, Pol-II ChIP (time-course of DC response to LPS)
- Find the ~200 top expressed TF; identify key time-points
- Curated available reagents
- Transcriptional analysis

- QC
- 29 targets for ChIP-Seq
- ~300 Antibodies
- HT-ChIP
Dataset: temporal view of expression and state

85% of high scoring TF peaks fall within annotated cis-regulatory regions
An example: Stat1, a late induced gene

Stat1 expression is a combination of pre-binding and dynamic binding
Transcription factors control specific pathways

What are the differences between sub-clusters?
Specific factors control amplitude of expression

Binding of stat1/2 controls inductions levels
Immediate early genes are highly bound

What are the differences between sub-clusters?
Immediate early gene programs

Immediate Early 1 (125 genes)

Immediate Early 2 (73 genes)
Current models under consideration

Late induced
Stat regulated

Immediate early
highly induced

Two forms of regulation?
Regulatory modes are established hierarchically.

PU.1 coincides with or precedes Stat1 binding.

Out degree
100%

≥75%

≤50%

Bound early
Bound late
Conclusions and considerations

• A large fraction of binding exist prior to stimulus
• Immediate vs. late regulation is quite distinct:
  – Early induced genes regulators are more redundant
  – Late induced regulators are less redundant
  – Are the early inflammation pathways evolutionary more malleable?
• Factors act in layers, consistent with previous reports
• Genomic approaches like this are applicable to many systems
  – Protocols can handle smaller input material (Alon Goren, Oren Ram)
• Test models using a genome wide genetic screen
• Map TFs with no available antibodies
• Currently building maps of another 20 factors for which antibodies became available
Next steps: Perturbing each factor

TF binding map

Loss of function screen
An expensive proposition…

\(~100\) genes KD \(* \) replicates = \(\text{LARGE NUMBER of samples}\)

\(\checkmark\) high cost
\(\checkmark\) limited starting material
An expensive proposition…

~100 genes KD * replicates = LARGE NUMBER of samples

✗ high cost
✗ limited starting material

Previous solution
- qPCR
- Fluidigm
- Luminex
- NanoString
- microarray

Problem: need to choose your genes in advance and limited genes assayed

Cost/sample

#Genes assayed

10s 100s 1000s

10000s

microarray

fluidigm

nanostring

qPCR

Sabah Kadri
An expensive proposition…

~100 genes KD * replicates = LARGE NUMBER of samples
  ✗ high cost
  ✗ limited starting material

Previous solution
- qPCR
- Fluidigm
- Luminex
- NanoString
- microarray

Problem: need to choose your genes in advance and limited #genes assayed
Goal: Cheap RNA-Seq for quantification

~100 genes KD * replicates = LARGE NUMBER of samples
  ✗ high cost
  ✗ limited starting material

Previous solution
  - qPCR
  - Fluidigm
  - Luminex
  - NanoString
  - microarray

Problem: need to choose your genes in advance and limited #genes assayed

<table>
<thead>
<tr>
<th>Cost/sample</th>
<th>#Genes assayed</th>
</tr>
</thead>
<tbody>
<tr>
<td>10s</td>
<td>RNASeq</td>
</tr>
<tr>
<td>100s</td>
<td>microarray</td>
</tr>
<tr>
<td>1000s</td>
<td>fluidigm, nanostring, qPCR</td>
</tr>
</tbody>
</table>
Full length RNA-Sequencing

Motivation

Protocol

Pipeline

Quantification

Diff Expr

Annotation

Sabah Kadri
Full length RNA-Sequencing

1. (AAA)$_n$
2. Fragmentation
3. (AAA)$_n$
4. Adapter
5. Library construction

Motivation
Protocol
Pipeline
Quantification
Diff Expr
Annotation

Sabah Kadri, PhD
Full length RNA-Seqencing

Motivation

Protocol

Pipeline

Quantification

Diff Expr

Annotation

Sabah Kadri, PhD
End RNA-Seq

\[(TTT)_{25} (AAA)_n\]
End RNA-Seq

5' End RNA-Seq

- Fragmentation
- 5' exonuclease

3' End RNA-Seq

- Heat Fragmentation
- PolyA selection

Motivation | Protocol | Pipeline | Quantification | Diff Expr | Annotation
Current work: Generating timecourses of KDs
Acknowledgements

Jim Robinson, Helga Thorvaldsdottir, Bang Wong (IGV)
New Visualization for time series ChIP data

Raktima Raychowdhury and Anne Thielke
Automation, Library preparation, cell culture

Brian Minie, Dennis Friedrich, Jim Meldrim, Andrew Barry, Chad Nusbaum (GSAP)
Automation, High Throughput protocols

Oren Ram, Alon Goren
ChIP String, Low input, many interactions

Jim Bochicchio
Christine Cheng
Nir Hacohen
Brad Bernstein
Aviv Regev

I do Amit

Nir Yosef
A non-canonical binder: Runx1

Mechanism unknown, different partner? Different complex?

- 3' binding is stronger
- Higher 3' bound
- Gene expression
- 3' motifs are worse
Understanding the cis-regulome

Loss of function screen

TF binding map
Early inflammatory genes are smaller, have larger enhancers and are farther away from other genes.
Figure S1 (part-III)

<table>
<thead>
<tr>
<th></th>
<th>DC Cebpb 0hr</th>
<th>DC PU1 0hr</th>
<th>Mac Cebpb</th>
<th>Mac PU1</th>
<th>Bcell PU1</th>
<th>Mac Cebpa 2hr</th>
<th>DC PU1 2hr</th>
<th>DC Cebpb 2hr</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC Cebpb 0hr</td>
<td>34467</td>
<td>14574</td>
<td>18060</td>
<td>11883</td>
<td>4601</td>
<td>15196</td>
<td>13360</td>
<td>24328</td>
</tr>
<tr>
<td>DC PU1 0hr</td>
<td>14448</td>
<td>47806</td>
<td>9846</td>
<td>23321</td>
<td>11644</td>
<td>8340</td>
<td>36487</td>
<td>14132</td>
</tr>
<tr>
<td>Mac Cebpb</td>
<td>18060</td>
<td>9884</td>
<td>22733</td>
<td>10718</td>
<td>3465</td>
<td>16895</td>
<td>9175</td>
<td>17457</td>
</tr>
<tr>
<td>Mac PU1</td>
<td>11859</td>
<td>23516</td>
<td>10734</td>
<td>30612</td>
<td>10604</td>
<td>9143</td>
<td>22283</td>
<td>12037</td>
</tr>
<tr>
<td>Bcell PU1</td>
<td>4534</td>
<td>11606</td>
<td>3437</td>
<td>10504</td>
<td>17328</td>
<td>2668</td>
<td>11094</td>
<td>4805</td>
</tr>
<tr>
<td>Mac Cebpa</td>
<td>15203</td>
<td>8361</td>
<td>16917</td>
<td>9139</td>
<td>2675</td>
<td>18201</td>
<td>7747</td>
<td>14781</td>
</tr>
<tr>
<td>DC PU1 2hr</td>
<td>13242</td>
<td>36532</td>
<td>9149</td>
<td>22116</td>
<td>11143</td>
<td>7731</td>
<td>41982</td>
<td>13874</td>
</tr>
<tr>
<td>DC Cebpb 2hr</td>
<td>24059</td>
<td>14198</td>
<td>17335</td>
<td>11993</td>
<td>4892</td>
<td>14680</td>
<td>13955</td>
<td>34643</td>
</tr>
</tbody>
</table>

G

H

I