Running a core without a budget

Michael Poidinger – PI Bioinformatics
SLgN, Singapore
SIgN was launched in January 2008 by Singapore’s A*STAR with the aim of expanding and strengthening immunology in Singapore.

Mission & Vision

- To be a leading international research centre in Human Immunology
- Expand the knowledge base in immunology research, ensuring the translation of research discoveries towards medical applications.
Bioinformatics team and resources

- 5 (soon 9) post-doctoral bioinformatic scientists
  - 2 with pharma industry experience
  - Experience with the cross analysis of heterogeneous data sets
  - Experience in biology and specifically immunoinformatics

- Commercial grade hardware
  - Multi-node 64bit clusters
  - Many Terabytes of storage
  - ASTAR Internal Cloud
Funding/work model

- Each RI has a (generous), ongoing core fund (~70%) with access to various competitive grant schemes and industry collaboration (~30%)
- Staff work in close contact with the 25 PIs and
  - 180 researchers in the institute.
- Every analysis is special
  - Biologists ask us what can we do with the data
- Projects often extended due to findings
Clinical studies

- SIgN is involved with a number of clinical studies
  - Allergies
  - Sepsis
  - Kidney rejection
  - Urinary tract infection
  - HIV
  - Elderly cohort
  - Nasopharyngeal carcinoma
  - H1N1 vaccination
NGS studies

- RNAseq on sorted immune cells from healthy, latent and active TB infection
  - mRNA and miRNA
- Variant detection in yeast mutants
- RNAseq on immune cells from bats
- ChIPseq and methylseq on underweight babies
- Microbiome of ICU patients and prediction of candidiasis
- Tcell and Bcell repertoire in response to Dengue
Heterologous Data Integration

- Many projects comprise data from a range of source
  - Clinical survey
  - Clinical tests (e.g. skin prick for allergy)
  - Cell based assays
  - Usual suspects of HT data (microarray, NGS)
  - Luminex (measures analytes in blood)
  - Flow cytometry
  - Mass cytometry
We don’t...

- Crank the handle on generic protocols
- Work on our own research
  - Unless it derives specifically from an existing project
- Build, install or require systems like Galaxy, that push analysis back to the biologist

...not that there’s anything wrong with that
Software/tools

- Commercial licenses for a range of products
  - Thompson Reuters/ GeneGO
  - Ingenuity Pathway Analysis
  - Oncomine
  - Transfac

- Pipeline Pilot from Accelrys Dessault
- Spotfire from Tibco Perkin Elmer
Reporting via Moin Moin Wiki

SInG Bioinfo wiki

Welcome to the SInG Bioinfo wiki. All experiments and analyses involving the Core and/or Bioinformatics teams will be stored here. Experiments are stored by experiment name and experimental data is restricted according to the directions of the PI.

To gain access to this wiki, use your system username/password (the same one you use to access email). Anyone that had a specific wiki account created before we changed over require access to any of the experimental pages, or would like to utilise this wiki for the management of your own data, please send a request to Michael.

Please note this is an intranet site, access from outside requires VPN.

Information

- System administration information
- Information - general information and knowledge sharing
- Infrastructure
- Functional Genomics Laboratory
- Functional Genomics Laboratory Technical Information

Analysis results

- AlessandraMortillaro
- AnnaMatoFairhurst
- ChengWang
- EeCheeRen
- EvanNewell
- FlorentGinhoux
- FlowCore
Updated analysis

The visualization has been updated to include the latest January cohort data. The updated visualization is available at:


20110923 Initial analysis

The RNA quality was assessed by the Microarray Core group and was made available as an Excel spreadsheet. This was processed and stored into the database prior to being combined with the rest of the cohort data to determine if the samples suitable for microarray had any bias in them.

The results are located in the following Spotfire visualization:


The findings are as follows:

1. Kruskal-Wallis tests were conducted to determine if any of the continuous parameters was significantly different between samples suitable for microarray and not.
   1. The RNA RIN value and yield showed significant results and that is to be expected since they were used to classify whether the samples were suitable for microarray.

2. Chi-square tests were conducted for discrete parameters.
   1. The first 3 parameters are to be expected since they are part of the parameters that determined microarray suitability.
   2. 3 SNPs were also found to be significant but again they are to be expected without any multiple testing correction.
   3. Spearman correlation tests were also conducted against the RNA yield.
      1. Although many of the tests were significant, the $R^2$ is very low for most of the significant results.
Pipeline Pilot

A screenshot of a software interface for pipeline development, showing various components and processes involved in pipeline creation. The interface includes tabs for different sections such as "Profiles", "Components", "Network", and "Repository". The interface displays various elements related to data processing, including file handling, quality control, and alignment tools.
Pipeline Pilot components

- Database and Application Integration
  - Analysis and Statistics
    - Basic Math and Statistics
    - Data Modeling
    - R Statistics
      - Analysis
        - R ANOVA
        - R Correlation Matrix
        - R Factor Analysis
        - R K Nearest Neighbors
        - R K Nearest Neighbors CV
        - R One-Variable Tests
        - R Principal Components Analysis
        - R Probability Distributions
        - R Robust Dose Response Fit
        - R Two-Variables Tests
      - Charts
      - Clustering
  - Charts
  - Clustering

- Next Gen Sequencing
  - Analysis
    - ChIP-Seq
      - Detect Enrichment Peaks (CisGenome)
      - Detect Enrichment Peaks (CisGenome2)
      - Detect Enrichment Peaks (MACS)
    - de novo Assembly
    - Mapping
      - RNA-Seq
        - Compare Transcript Expression (Cuffdiff)
        - Compare Transcript to Reference (Cuffcompare)
        - Cufflinks Transcript Assembler
        - Find Union Transcripts (Cuffcompare)
        - TopHat Spliced Read Mapper
    - Variant Detection
    - Filters
    - Manipulators
    - Readers
      - Mapped Reads
        - BAM Reader
        - Illumina Export Results Reader
        - Paired Illumina Export Results Reader
        - SAM Reader
      - Unmapped Reads
        - FASTQ Directory Reader
        - FASTQ Reader
        - Illumina QSEQ Directory Reader
        - Illumina QSEQ Reader
        - Roche 454 FASTA Directory Reader
        - Roche 454 FASTA Reader
        - Simulated Reads Generator (wgsm)
        - SOLID CSFASTA Directory Reader
        - SOLID CSFASTA Reader
      - Utilities
      - Viewers

- Biology
  - EMBOSS
  - Gene Expression
    - Analysis
      - Clustering
      - Selectors
        - Fold Change
        - Intensity Variation
        - Microarray Statistics
        - Pairwise Differential Expression
        - Selection Consensus
      - Annotators
**RNAseq Workflow**

1. FASTQ Reader → fastQC → ERCC bowtie → View ERCC
2. Picard for insert size → Tophat & Cufflinks
   - cuffmerge → cuffcompare → cuffdiff → visualization
Origins of data

Parameter values are captured
# Spotfire – RNAseq

## Table

<table>
<thead>
<tr>
<th>Transcription ID</th>
<th>Reference Gene</th>
<th>Gene Name</th>
<th>Length</th>
<th>C</th>
<th>H</th>
<th>PRB</th>
<th>IFN</th>
<th>C_HIF</th>
<th>C_PRB</th>
<th>C_IFN</th>
<th>C_IFN sig</th>
<th>C_PRB sig</th>
<th>C_HIF sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD37</td>
<td>CD37A</td>
<td></td>
<td>1582</td>
<td>0.20</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>2.50</td>
<td>28.94</td>
<td>25.60</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>CD38</td>
<td>CD38A</td>
<td></td>
<td>1912</td>
<td>0.30</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>2.80</td>
<td>28.94</td>
<td>25.60</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>CD39</td>
<td>CD39A</td>
<td></td>
<td>1686</td>
<td>0.20</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>2.50</td>
<td>28.94</td>
<td>25.60</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

## Bar Chart

- Y-axis: Number of counts
- X-axis: Gene expression levels
Spotfire – mass cytometry

DIAGRAM:
- DC
- monocytes
- Granulocytes
Spotfire Pairwise correlations
Spotfire – Signature profiles
Spotfire – Frequency distribution