Summary of Upcoming Work in the Rubin-Gerstein Collaboration

Lucas Lochovsky

gTech subgroup

Stardate 2011.179

Outline

- LL's work: Regulatory Network Disruptions in Cancer
- R01 Grant
- Indel work

Outline

- LL's work: Regulatory Network Disruptions in Cancer
- R01 Grant
- Indel work

Previously on LL's work

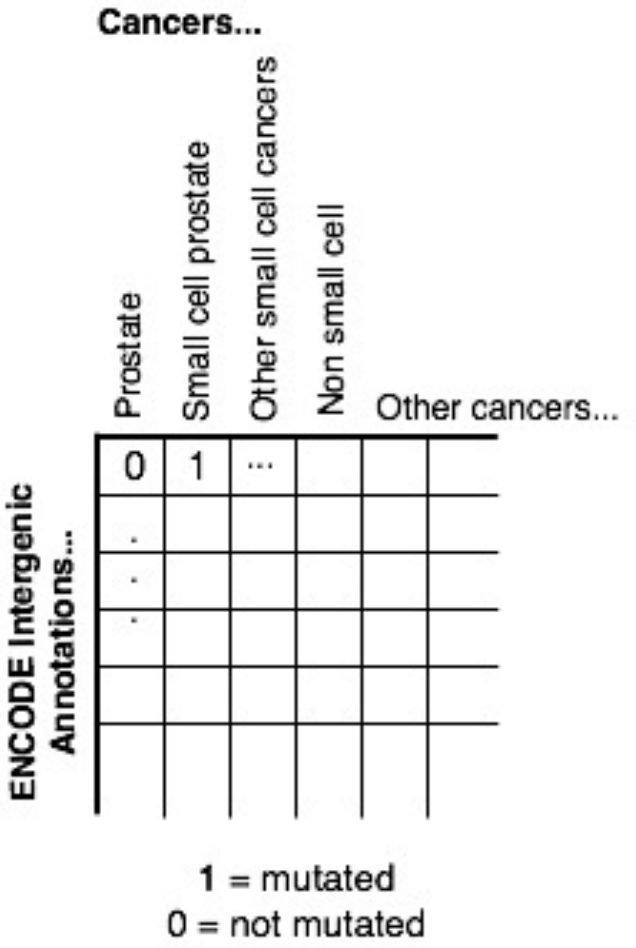
- Finding mutated hubs and bottlenecks in HPRD, and later the ENCODE TF network
- **Mutation data sources**
  - COSMIC: Dumping ground for a large number of experiments over the past 20 years
  - ICGC simple somatic mutation data (Better organized than COSMIC)
- **On the last episode:** Constructed a matrix of ENCODE TF inhubs vs. ICGC cancers, and indicated which inhubs had a simple somatic mutation in which cancer
  - Ditto outhubs and bottlenecks

Current Plan

Start with:

1. ENCODE Intergenic Annotations
   - Enhancers
   - Promoters
   - ncRNA
2. Cancer variant data
   - Prostate data from the Rubin lab
   - Other cancers from ICGC
   - Must be certain of what types of analyses produced the data, and how many samples were involved

Current Plan



Current Plan

- For genes whose regulatory sites have been mutated, find genes that are downstream in the ENCODE TF network
- See what effect this has on expression of gene with the mutation, and on downstream genes' expression
  - Requires RNA-seq data for the genes involved (TCGA and ICGC)
  - Do we see an expression change between cancer and normal samples in the gene with the mutation, and is there a corresponding fold change in downstream genes' expression?

Other Questions

- Do we see significant enrichment/depletion of cancer variants in ENCODE intergenic annotations compared to the whole genome?
- Is there a significant enrichment/depletion of prostate cancer variants in certain regions compared to all cancer variants?
- Visualize intergenic region mutations in the ENCODE TF network as disrupted edges
  - Find which TFs tend to have its sites knocked out most often across cancer types
  - Which TFs have the most knocked out edges?
- Investigate ratio of (number of mutated inhubs/outhubs/bottlenecks in top 100):(total number of mutated genes) for each cancer
  - If we expand the inhub, outhub, and bottleneck matrices to the top 200 on the ENCODE TF network, does this ratio stay the same?

ASIDE: Birney file

- Used ICGC sample database to determine how many samples are part of each dataset
  - Mix of tumor tissue, xenografts, and cell line samples
- US datasets only display mutations in gene coding regions and UTRs
  - All other datasets have mutations labelled "upstream", "downstream", "ncRNA", and "intergenic"
- Downloaded all simple mutation data from ICGC data portal and did Unix diff on Birney file
  - The two files are completely identical

Outline

- LL's work: Regulatory Network Disruptions in Cancer
- R01 Grant
- Indel work

R01 Grant

- Starting Aug 1, will start resequencing prostate cancer and benign prostate
- Sequencing targeted at regions that can serve as biomarkers for prostate cancer
  - Novel TARs
  - Other mutations

Outline

- LL's work: Regulatory Network Disruptions in Cancer
- R01 Grant
- Indel work

Indel Plans

- **Good News:** No one's studied indels like we have
- **Bad News:** The indel data we do have is weird
  - **Operational definition of "weird" here:** Many indels are in repetitive regions, hence we're not completely confident in our indel calls
- **Solution:** Targeted indel calling on a calibration set to estimate the error rates of indel callers

Indel Plans

- Mark Rubin would like to prioritize characterization of important genes
- LH to find genes to prioritize for this characterization
  - Regions that affect expression (promoters, enhancers, etc.)
  - Pathways (PTEN, PI3K)
  - Tyrosine kinases (RTKs)—mostly upstream activators
  - DNA binding domains
  - Knocked out genes in prostate cancer
  - Tumor suppressors
  - Pseudo~~jeans~~ genes

Indel Plans

- Rubin lab responsibilities
  - # of indels
  - # of samples
  - Comparison with small cell cancers
  - Validation assay throughput
- Gerstein lab responsibilities
  - Find out what was used in 1KG for experimental validation of indels (is it Sequenom?)

Other Questions

- Conduct validation in cell lines?
- Look for statistically significant enrichment/ depletion of indels in certain genes

Acknowledgements

- Lukas Habegger (Lukas 1.0)
- Andrea Sboner
- Mark Gerstein
- Mark Rubin + lab
- UCs