LARVA Analysis of Prostate Cancer Exome Variants from TCGA (Firehose)

Lucas Lochovsky
Variation subgroup
(Because the name always varies)
April 16, 2014

Data

- We're signed up on the TCGA prostate AWG tasklist for finding recurrently mutated pathways
- Found the exome MAFs on the Broad Firehose dashboard
- Run LARVA on this data
- Variant data:
  - 261 samples
  - 19,400 variants
- Annotation data:
  - Genes
  - KEGG
  - HPRD

Methods

- **Recap:** LARVA-SAM simulates variant distribution under neutral mutation processes
  - Compare to observed data to determine significant enrichment or depletion of recurrent variation
- Current version of LARVA-SAM produces *p*-values for:
  - Number of samples mutated *nsamp* (The number of samples with variants that overlap the given pathway)
  - Number of recurrently mutated annotations *nannot* (The number of annotations (exons) with nsamp >= 2)
  - Number of recurrently mutated genes *ngene* (The number of genes with nsamp >= 2)
  - Number of recurrent variants *nvar* (The number of positions where variants from multiple samples overlap exactly)
- Focused on *nsamp* and *ngene*
  - Those seemed to be the most useful *p*-values
- Employed Bonferroni correction on *p*-values for FDR correction

LARVA(TCGA Firehose Prostate Exome, KEGG)

Significant pathways by nsamp, Bonferroni correction

Pathway	nsamp	nannot	nvar	ngene	# samples mutated rand avg	# samples mutated p-value	Enrichment/Depletion
kegg_olfactory_transduction.txt	134	37	24	38	192.70	1.36E-19	Depletion
kegg_thyroid_cancer.txt	51	9	13	9	17.33	6.01E-18	Enrichment
kegg_prostate_cancer.txt	80	17	20	20	41.35	1.57E-12	Enrichment
kegg_pancreatic_cancer.txt	66	11	21	15	33.36	2.22E-11	Enrichment
kegg_bladder_cancer.txt	49	8	17	11	20.89	3.76E-11	Enrichment
kegg_p53_signaling_pathway.txt	65	10	15	12	31.58	4.59E-11	Enrichment
kegg_non_small_cell_lung_cancer.txt	62	9	18	12	30.95	6.01E-10	Enrichment
kegg_chronic_myeloid_leukemia.txt	69	13	23	15	35.88	1.04E-09	Enrichment
kegg_endometrial_cancer.txt	72	15	18	13	39.56	1.05E-09	Enrichment
kegg_neuroactive_ligand_receptor_interaction.txt	129	25	41	44	173.11	1.73E-09	Depletion
kegg_glioma.txt	67	12	20	15	35.58	2.36E-09	Enrichment
kegg_small_cell_lung_cancer.txt	85	10	22	24	51.11	2.60E-08	Enrichment
kegg_apoptosis.txt	65	9	20	11	38.83	5.71E-08	Enrichment
kegg_melanoma.txt	65	16	19	16	38.39	1.35E-07	Enrichment
kegg_starch_and_sucrose_metabolism.txt	33	3	4	10	62.33	6.64E-07	Depletion
kegg_cell_cycle.txt	81	12	23	20	51.33	8.38E-07	Enrichment
kegg_colorectal_cancer.txt	63	12	19	12	38.88	2.17E-06	Enrichment
kegg_hypertrophic_cardiomyopathy_hcm.txt	87	12	21	23	61.20	4.61E-06	Enrichment
kegg_o_glycan_biosynthesis.txt	12	1	5	4	34.30	1.89E-05	Depletion
kegg_basal_cell_carcinoma.txt	62	10	16	13	39.35	3.17E-05	Enrichment
kegg_propanoate_metabolism.txt	10	0	2	3	29.98	1.04E-04	Depletion
kegg_glycerolipid_metabolism.txt	22	1	3	4	40.56	1.80E-04	Depletion
kegg_drug_metabolism_other_enzymes.txt	25	3	6	6	42.15	2.03E-04	Depletion
kegg_amyotrophic_lateral_sclerosis_als.txt	54	7	15	9	34.33	2.05E-04	Enrichment
kegg_n_glycan_biosynthesis.txt	14	1	4	4	31.27	2.94E-04	Depletion

Bonferroni correction = 0.05/186 ≈ 0.0002688172043 = 2.68E-04

LARVA(TCGA Firehose Prostate Exome, KEGG)

Significant pathways by ngene, Bonferroni correction

Pathway	nsamp	nannot	nvar	ngene	# genes mutated rand avg	# genes mutated p-value	Enrichment/Depletion
kegg_olfactory_transduction.txt	134	37	24	38	126.11	9.79E-29	Depletion
kegg_neuroactive_ligand_receptor_interaction.txt	129	25	41	44	85.98	2.38E-19	Depletion
kegg_alpha_linolenic_acid_metabolism.txt	8	0	0	3	0.79	3.13E-08	Enrichment
kegg_starch_and_sucrose_metabolism.txt	33	3	4	10	18.70	3.82E-06	Depletion
kegg_drug_metabolism_cytochrome_p450.txt	35	3	8	7	16.29	6.09E-06	Depletion
kegg_toll_like_receptor_signaling_pathway.txt	33	4	11	5	12.15	3.11E-05	Depletion
kegg_drug_metabolism_other_enzymes.txt	25	3	6	6	11.77	4.05E-05	Depletion
kegg_retinol_metabolism.txt	40	4	7	7	16.56	4.77E-05	Depletion
kegg_ascorbate_and_aldarate_metabolism.txt	15	2	3	3	7.88	8.44E-05	Depletion
kegg_pathways_in_cancer.txt	155	31	50	76	59.70	9.98E-05	Enrichment
kegg_focal_adhesion.txt	131	18	38	58	43.79	1.04E-04	Enrichment
kegg_glycerolipid_metabolism.txt	22	1	3	4	11.97	1.12E-04	Depletion
kegg_pentose_and_glucuronate_interconversions.txt	16	2	3	3	7.77	1.96E-04	Depletion
kegg_steroid_hormone_biosynthesis.txt	25	3	5	4	12.88	2.48E-04	Depletion

nsamp and ngene intersection list

Pathway	nsamp	nannot	nvar	ngene	# samples mutated rand avg	# samples mutated p-value	Enrichment/Depletion	# genes mutated rand avg	# genes mutated p-value	Enrichment/Depletion
kegg_olfactory_transduction.txt	134	37	24	38	192.70	1.36E-19	Depletion	126.11	9.79E-29	Depletion
kegg_neuroactive_ligand_receptor_interaction.txt	129	25	41	44	173.11	1.73E-09	Depletion	85.98	2.38E-19	Depletion
kegg_starch_and_sucrose_metabolism.txt	33	3	4	10	62.33	6.64E-07	Depletion	18.70	3.82E-06	Depletion
kegg_drug_metabolism_other_enzymes.txt	25	3	6	6	42.15	2.03E-04	Depletion	11.77	4.05E-05	Depletion
kegg_glycerolipid_metabolism.txt	22	1	3	4	40.56	1.80E-04	Depletion	11.97	1.12E-04	Depletion

Bonferroni correction = 0.05/186 ≈ 0.0002688172043 = 2.68E-04

LARVA(TCGA Firehose Prostate Exome, KEGG) Genes

- The enriched pathways had a lot of the same players that are seen in many cancers

Gene	nsamp	nvar
TP53	22	4
CTNNB1	9	2
PIK3CA	9	0
PTEN	6	0
BRAF	5	1
CDH1	4	0

- The depletion pathways, however, had nothing in common

kegg_olfactory_transduction.txt

Gene	nsamp	nvar
OR4D5	5	0
OR2M3	4	0
OR4A16	4	0
OR5L2	4	0
CNGA4	3	0
OR10J1	3	0
OR10R2	3	0
OR10Z1	3	0
OR3A1	3	0
OR4A15	3	0

kegg_neuroactive_ligand_receptor_interaction.txt

Gene	nsamp	nvar
GRIN2A	6	0
GRID2	5	0
GABRB1	4	1
GABRG1	4	0
GRIA1	4	0
GABRE	3	0
GRIN2B	3	0
HCRTR2	3	0
GABRR2	2	1
GRM1	2	1

kegg_starch_and_sucrose_metabolism.txt

Gene	nsamp	nvar
UGT1A8	3	0
ENPP1	2	0
GBE1	2	0
GCK	2	0
HK3	2	0
MGAM	2	0
PYGL	2	0
UGT2B11	2	0
UGT2B28	2	0

kegg_drug_metabolism_other_enzymes.txt

Gene	nsamp	nvar
UGT1A8	3	0
CYP2A6	2	0
DPYD	2	0
UGT2B11	2	0
UGT2B28	2	0
UPP1	2	0

LARVA(TCGA Firehose Prostate Exome, HPRD)

- *P*-values weren't very useful here
- Instead, I picked out the interaction pairs that were mutated in at least two samples, and had both partners mutated
- Ranked each node by number of edges
- Investigated any connected components, but all I got was a hairball

Gene	# Partners
TP53	53
GRB2	42
MAPK1	27
CREBBP	21
SHC1	19
HRAS	16
ABL1	14
SP1	11
SMAD4	11
PTK2	10
NOTCH1	10
LRP1	10
CDH1	8
SREBF2	7
RPA1	7
RB1	7
PTPRF	7
PTPN1	7
ITGB2	7
ERBB3	7
EP300	7
ANK1	7
XPA	6
VAV1	6
TUBA4A	6
TTN	6
STAT1	6
PPARA	6
PDGFRB	6
NTRK1	6
F2	6
ERBB2	6
BCAR1	6

LARVA's Scalability

- Question posed a couple of weeks ago about how fast LARVA could analyze a large set of genome variants
- Used the random variant generator to produce 200 simulated whole genomes of 1000 variants each, and timed LARVA's performance
  - Run variants through all annotation sets

LARVA's Scalability

- LARVA-Core did the intersections in ~20 minutes
- LARVA-SAM
  - *nrand*=2000, *ncpu*=60
  - Simulations were complete in roughly 26 hours (excluding ENCODE TF peak data)
- Given the size of the TF peak data, it's estimated that doing LARVA-SAM with the TF peak data would take ~3.5 days
- **Note:** OpenMPI's memory usage for 60 processes is extremely demanding, and for sufficiently large *nrand* it will crash the server. Using a smaller number of processes avoids this problem, but it will increase the running time.

LARVA Validation

- Use LARVA to identify recurrently mutated gene list, and compare to the list in the corresponding literature
- Started with the Berger set of seven prostate genomes
- Paper indicates that SPOP and SPTA1 are mutated in two samples
- LARVA identified these, as well as NBEAL1
  - Literature indicates that this is upregulated in glioma
  - No mention in Berger *et al.* paper
  - Samples: PR-1701, PR-3027
  - Variants: (chr2, 203990092, 203990093), (chr2, 204032021, 204032022)