More Cancer Variant Intersections with 1KG Coding Variants

and

What's Going on With the TCGA Datasets?

An eleventh hour double feature from

ll426@gersteinlab

Wednesday, August 1, 2012

Prostate Cancer Exome Variants ∩ 1KG Coding Variants

Prostate cancer

Actual Data

# of cancer exome variants	# of 1KG coding variants	Intersection	Percentage of cancer variants in intersection	Percentage of 1KG coding variants in intersection
216	514,269	34	15.7%	0.00661%

Average of 100 Runs Using Randomized Cancer Variant Positions

# of cancer exome variants	# of 1KG coding variants	Intersection	Percentage of cancer variants in intersection	Percentage of 1KG coding variants in intersection
216	514,269	0.03	0.0139%	0.00000583%

Actual:Random Percentage Ratio
1133

Melanoma Cancer Exome Variants ∩ 1KG Coding Variants

Melanoma SNV (Halaban data)

Actual Data

# of cancer exome variants	# of 1KG coding variants	Intersection	Percentage of cancer variants in intersection	Percentage of 1KG coding variants in intersection
25,489	514,269	771	3.02%	0.150%

Average of 100 Runs Using Randomized Cancer Variant Positions

# of cancer exome variants	# of 1KG coding variants	Intersection	Percentage of cancer variants in intersection	Percentage of 1KG coding variants in intersection
25,489	514,269	5.69	0.0223%	0.00111%

Actual:Random Percentage Ratio
136

What's Up With TCGA Datasets

- Split data into separate studies
  - Compute exome fraction for each study
- Investigate which are backed up in literature
  - Explain where they came from
- Germline variants' dataset sizes were pretty small, and the exome fractions were all over the place
- Focus on somatic variants
  - Things are more solid there

What's Up With TCGA Datasets

- **COAD:** Literature indicates data is exome capture

Cancer	Center	Sequencer	# Mutations	Exome Fraction
COAD	BCM	Illumina	22,147	96.8%
COAD	BCM	SOLiD	9197	99.1%

- **GBM:** No literature support, mutation counts are small, not sure what's going on here

Cancer	Center	# Mutations	Exome Fraction
GBM	BCM	450	85.1%
GBM	MIT	436	81.1%
GBM	WUSTL	436	91.5%

What's Up With TCGA Datasets

- **LAML:** No literature support, mutation counts are small, not sure what's going on here

Cancer	Center	Sequencer	# Mutations	Exome Fraction
LAML	WUSTL	Illumina GA	724	83.1%
LAML	WUSTL	Illumina HiSeq	9	77.7%

- **OV:** Literature indicates data is exome capture, but the datasets highlighted with red stars look suspect (green stars are OK)

Cancer	Center	Sequencer	# Mutations	Exome Fraction
*OV	BCM	-	2,456	97.2%
*OV	MIT	Illumina	12,615	48.4%
*OV	MIT	Unknown	20	90.0%
*OV	WUSTL	ABI	1	100.0%
*OV	WUSTL	Illumina	6192	90.1%

What’s Up With TCGA Datasets

- **READD:** Literature indicates data is exome capture

Cancer	Center	Sequencer	# Mutations	Exome Fraction
READ	BCM	Illumina	1,716	97.6%
READ	BCM	SOLiD	8,768	99.2%

- **Summary:** Datasets with literature support and aren’t doing something funky include 2 COAD, 2 OV, and 2 READ studies

Multiple Myeloma

- According to paper, there's both whole genome and exome data
- Sample IDs in paper supplement don't match with sample IDs in data file
- If data separated by sample, exome fraction averages ~30%
  - Inconclusive