Patient-centric Mutation Burden + Pathway Analysis (aka NIMBus v2)

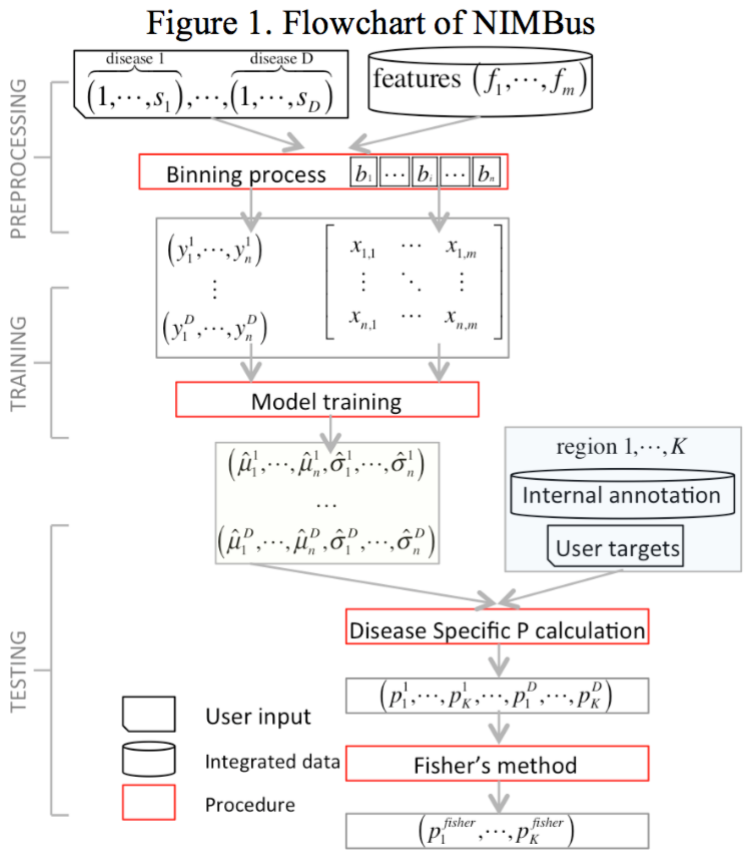
Jason Liu

Jing Zhang

March 29, 2016

Previously: NIMBus

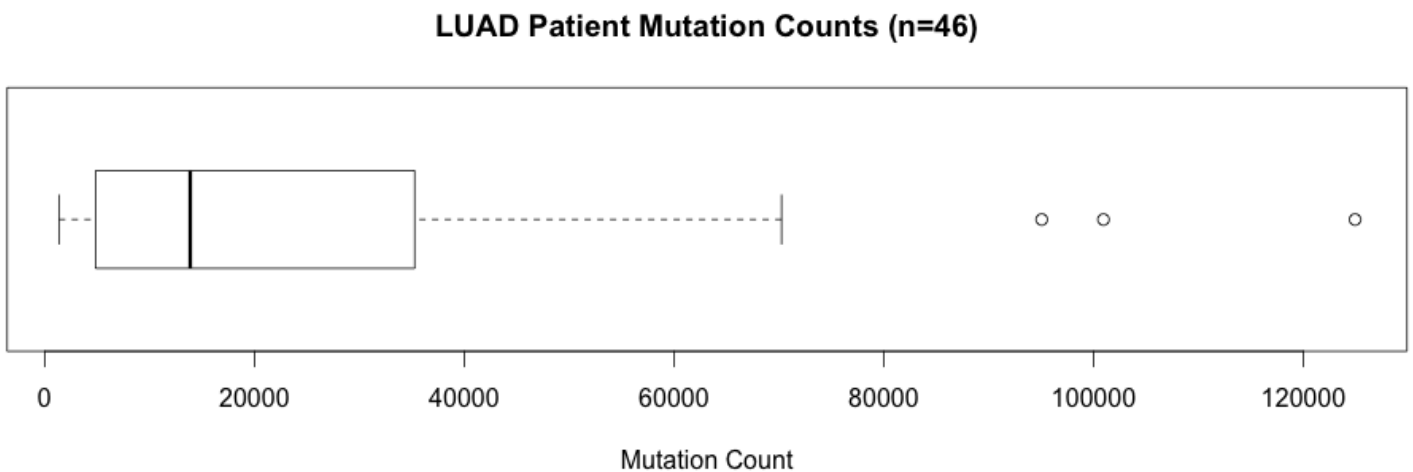
- One negative-binomial model for each cancer type used
- NegBin because patients have different background mutation rate



Data

4 cancer types

- BRCA (n=119)
- GACA (n=100)
- LICA (n=88)
- LUAD (n=46)



Patient-centric Binomial Model

For each patient:

$Y_i \sim \text{Binomial}(n, p_i)$

- $n = 1$ Mbp bin
- $Y_i =$ mutation count in bin i
- $p_i =$ background mutation rate in bin i

$\Pr\{Y_i = y_i\} = \binom{n}{y_i} p_i^{y_i} (1 - p_i)^{n - y_i}$

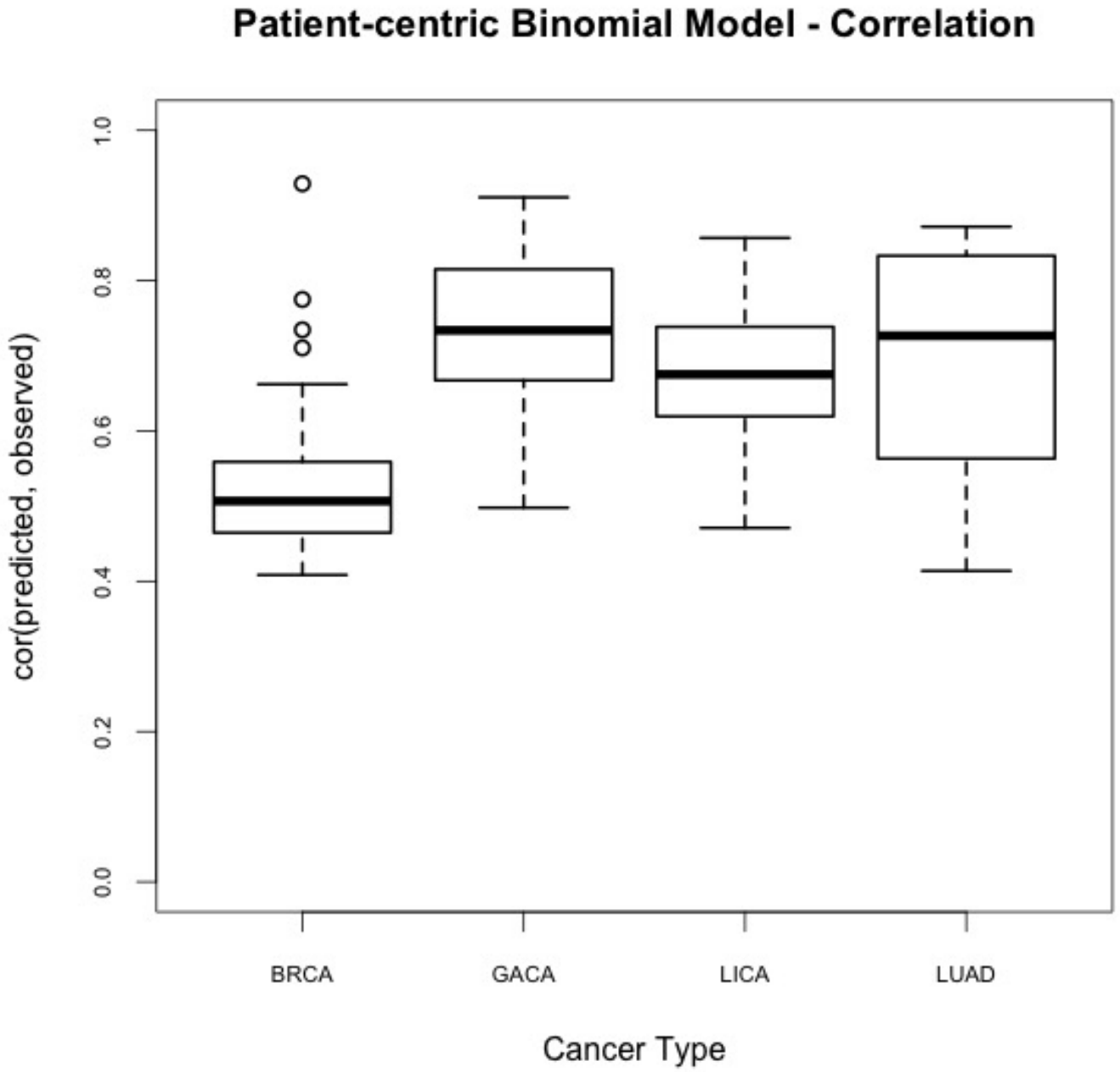
Binomial Regression

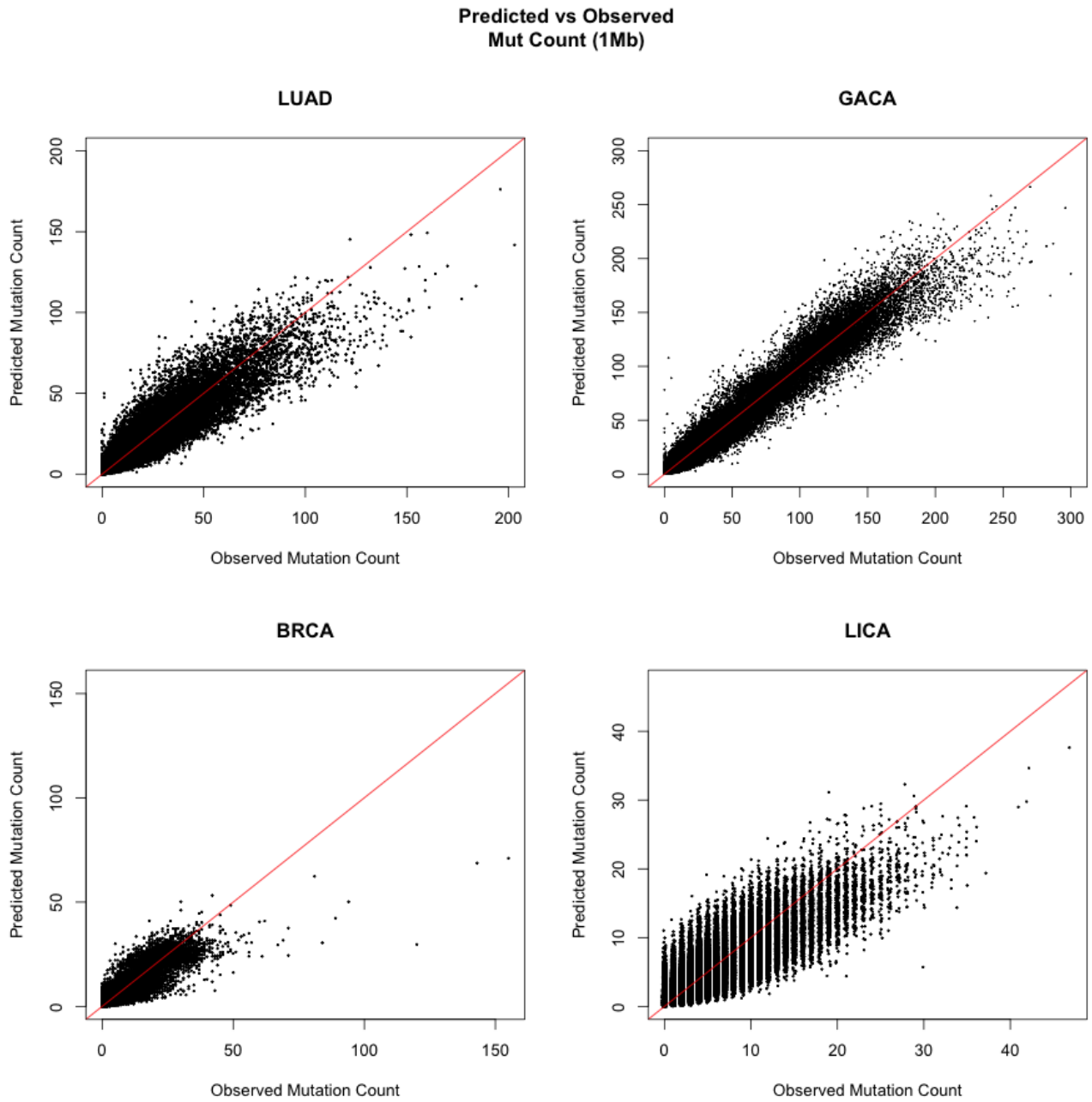
Perform regression for each patient:

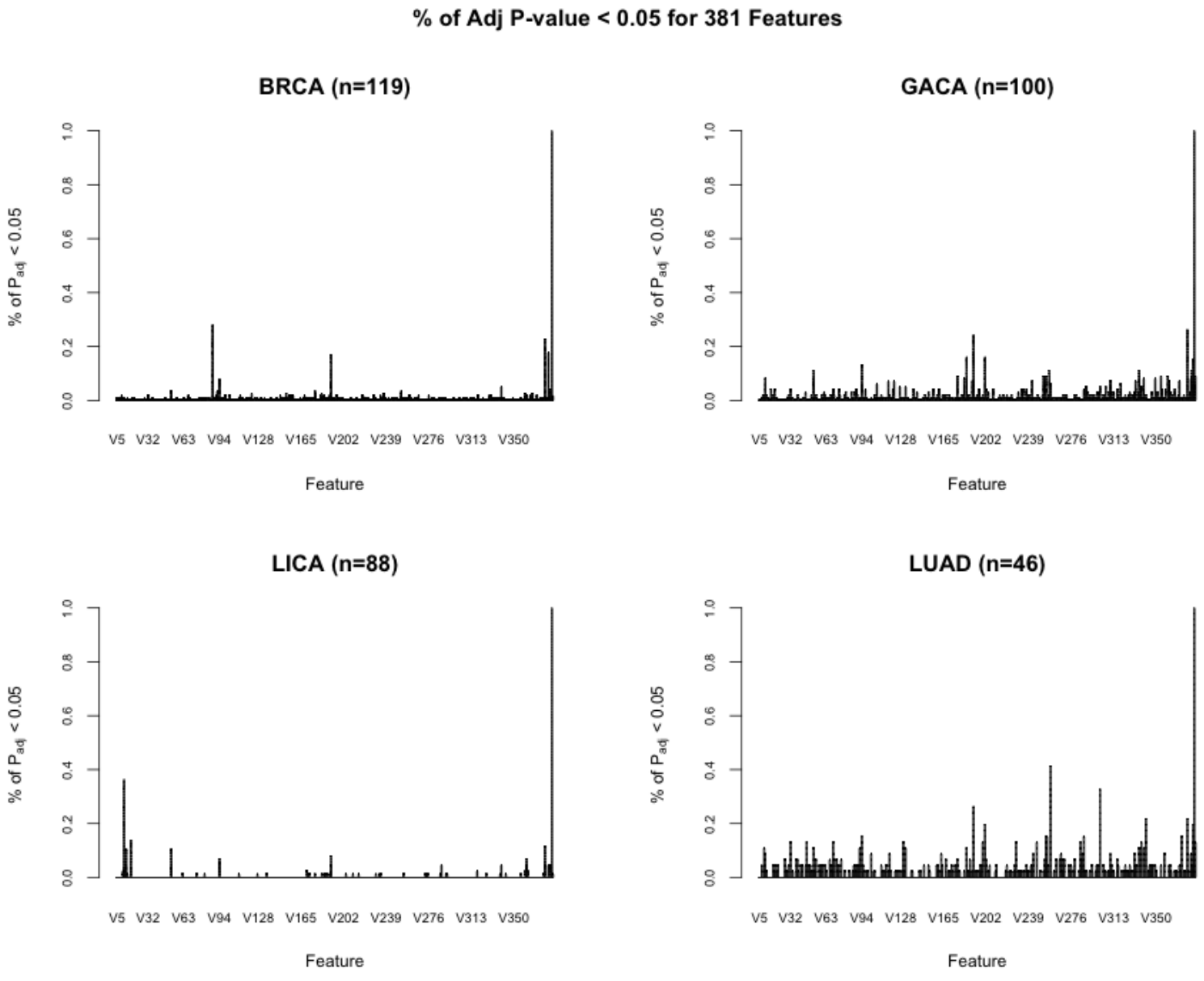
$$\text{logit}(p_i) = \vec{\boldsymbol{x_i}}' \vec{\boldsymbol{\beta}}$$

- $\text{logit}(p_i) = \log \frac{p_i}{1-p_i}$
- $\vec{\boldsymbol{x_i}} =$ Covariate matrix, 381 features
- $\vec{\boldsymbol{\beta}} =$ vector of regression coefficients

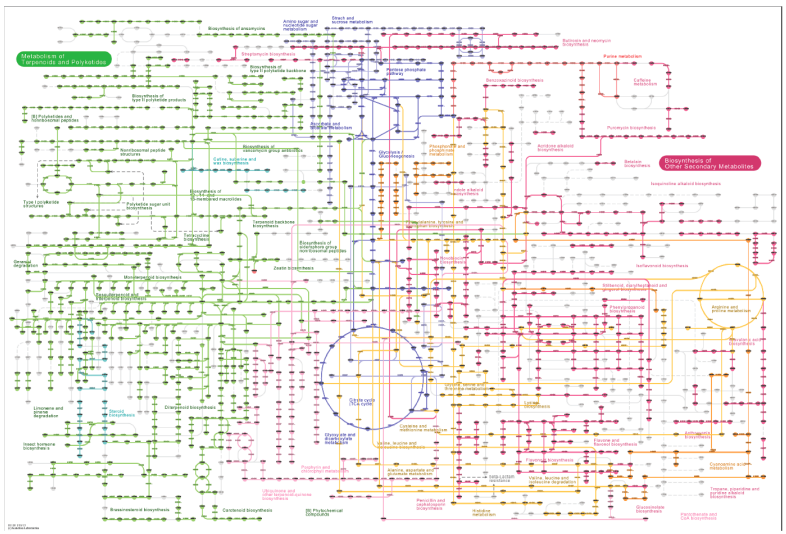
After regression we can calculate $\hat{Y}_i = n \cdot \hat{p}_i$


Patient-centric Binomial Model - Correlation
cor(predicted, observed)
0.0
0.2
0.4
0.6
0.8
1.0
BRCA
GACA
LICA
LUAD
Cancer Type


Predicted vs Observed
Mut Count (1Mb)
LUAD
Predicted Mutation Count
Observed Mutation Count
0
50
100
150
200
GACA
Predicted Mutation Count
Observed Mutation Count
0
50
100
150
200
250
300
BRCA
Predicted Mutation Count
Observed Mutation Count
0
50
100
150
LICA
Predicted Mutation Count
Observed Mutation Count
0
10
20
30
40


% of Adj P-value < 0.05 for 381 Features
BRCA (n=119)
GACA (n=100)
LICA (n=88)
LUAD (n=46)
% of P_{adj} < 0.05
0.0
0.2
0.4
0.6
0.8
1.0
V5 V32 V63 V94 V128 V165 V202 V239 V276 V313 V350
Feature

Application to KEGG Pathway Analysis using Poisson Binomial



Poisson Binomial

N independent Bernoulli

Parameters: $\vec{\boldsymbol{p}} \in [0,1]^N$

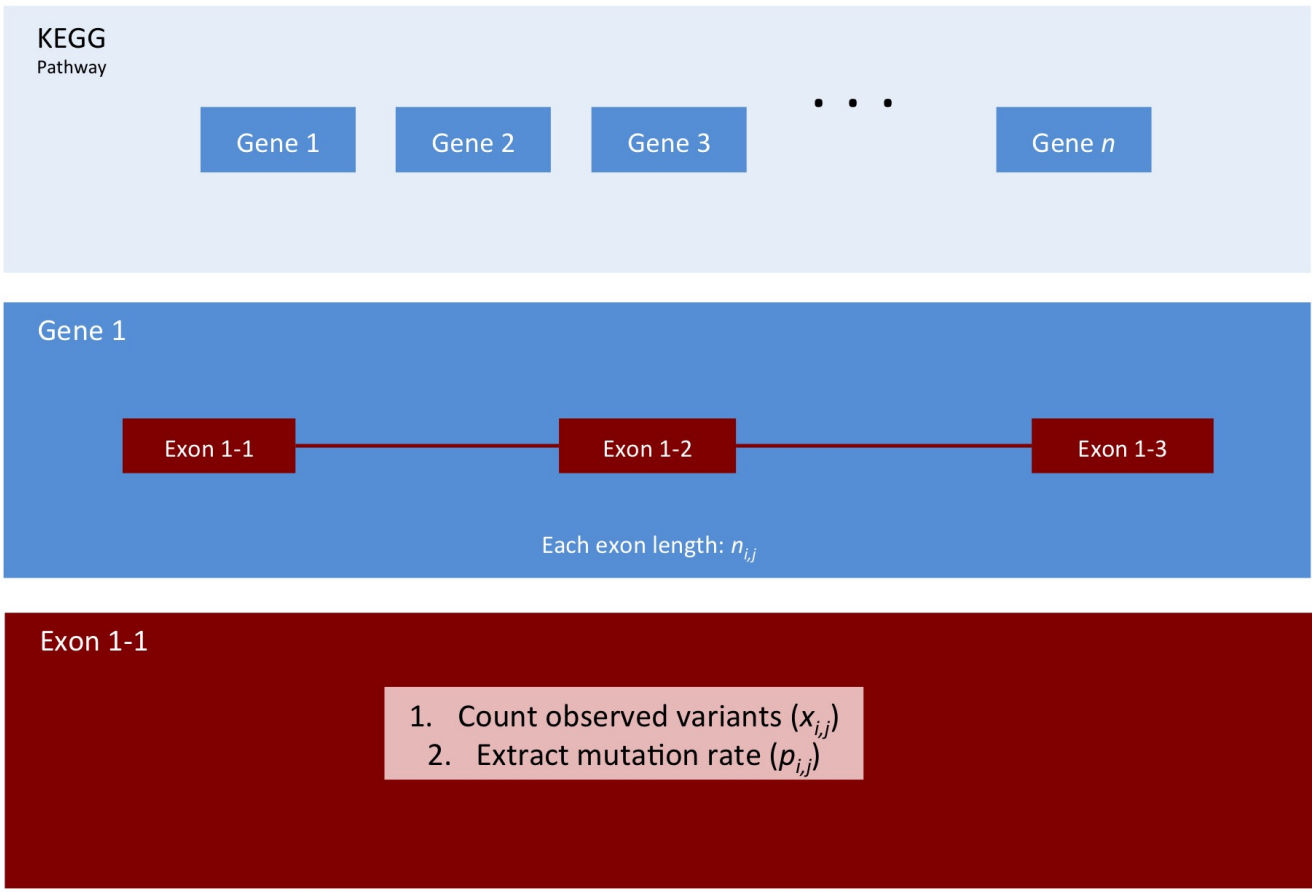
pmf:

$$\Pr(K = k) = \sum_{A \in F_k} \prod_{i \in A} p_i \prod_{j \in A^c} (1 - p_j)$$

F_k is set of all subsets of k integers that can be selected from $1 : n$

Poisson Binomial

Claim: Pathway Variants $\sim$ Poisson Binomial



Poisson Binomial

Result: Each base pair in pathway is an independent Bernoulli

- $N = \sum n_{i,j}$
- $X = \sum x_{i,j}$
- $\vec{\boldsymbol{p}} = \langle p_{1,1}, \ldots, p_{1,J}, \ldots, p_{I,1}, \ldots, p_{I,J} \rangle$

Where each $p_{i,j}$ appears $n_{i,j}$ times.

p-value $= 1 - \Pr(\hat{X} < X)$

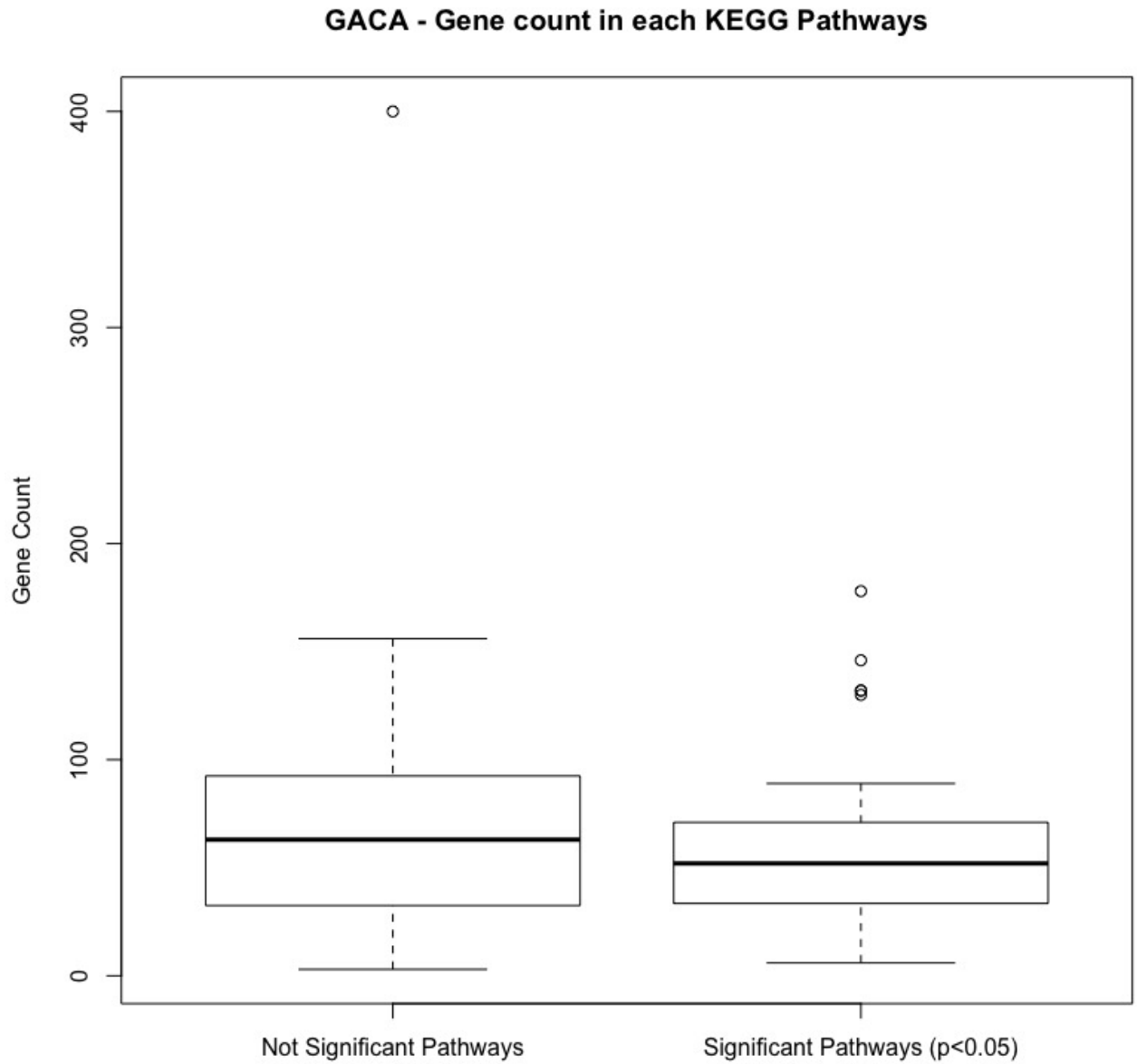
`R: poibin`

Gastric Cancer

Top Recurrent + Significant ($p_{\text{adj}} < 0.05$) Pathways

KEGG Pathway	# Patients	Description
KEGG:05219	6	Bladder cancer
KEGG:05216	5	Thyroid cancer
KEGG:05218	4	Melanoma
KEGG:03410	4	Base excision repair

Gastric Cancer



KEGG:05219

	1	2	3	4	5	6		
ARAF							Gene ID: 369	Proto-oncogene; cell growth
BRAF							Gene ID: 673	Proto-oncogene; kinase regulation
CDH1	1	3	1	1	1	1	Gene ID: 999	Cell-cell adhesion protein*
CDK4							Gene ID: 1019	Kinase; G1 cell cycle progression
E2F1							Gene ID: 1869	TF; cell cycle and TS control
EGF							Gene ID: 1950	Epidermal growth factor
EGFR						1	Gene ID: 1956	Epidermal growth factor receptor
ERBB2						1	Gene ID: 2064	EGF receptor of tyrosine kinase
MDM2							Gene ID: 4193	Proto-oncogene; promote tumor form
MYC							Gene ID: 4609	Oncogene; progression, apoptosis
NRAS							Gene ID: 4893	Oncogene; membrane protein
RAF1							Gene ID: 5894	Proto-oncogene; homolog of raf
RB1							Gene ID: 5925	TSG; negative regulator of cell cycle
THBS1				1			Gene ID: 7057	Adhesive protein; cell-cell interaction
TP53	1	1	1		1		Gene ID: 7157	Tumor suppressor protein
VEGFA							Gene ID: 7422	Vascular endothelial growth factor

Further Analysis

- Other pathways: Gene Ontology, Reactome, etc.
- Apply to TF binding sites, PPI networks
- More annotations, extend to noncoding regions

Acknowledgments

Jing Zhang

Lucas Lochovsky

Donghoon Lee

Mark Gerstein