A Unified Statistical Framework to Identify Driving Genetic Events in Cancer

Siming Zhao[1], Xin He[1] and Matthew Stephens[1,2]

[1]Department of Human Genetics, University of Chicago

[2]Department of Statistics, University of Chicago

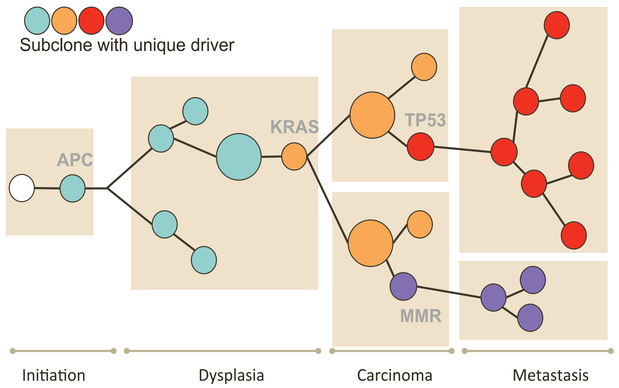
Abstract

Knowledge of driver genes whose mutations lead to tumor-genesis is important for un- derstanding the mechanisms of cancer and for identifying promising drug targets. Despite intensive sequencing and computational efforts, our knowledge of driver genes in most tumor types remain far from complete. Computational methods are essential to leverage cancer sequencing data to identify driver genes. Existing methods typically rely on some pattern of somatic mutations in the genes: driver genes would have more somatic mutations than expected, or the mutations in driver genes are more likely to be deleterious and spatially clustered. Nevertheless, most existing methods would utilize one feature a time. In this work, we develop a novel integrative approach for driver gene discovery that incorporates multiple features of driver genes in a single framework. As part of our approach, we come up with a mixed effect model to capture the baseline mutation rates, that incorporates both global features predictive of mutation rates and local variations of mutation rates. Our model also allows us to put more emphasis on more deleterious mutations (e.g. nonsense mutations) when evalu- ating the mutational burden of a gene. Finally, our model is designed to reward genes with mutations that are closely clustered. We evaluate our approach using TCGA data, and we found that in almost all tumor types, it outperforms MutSigCV, the current leading method for driver gene discovery.

Introduction

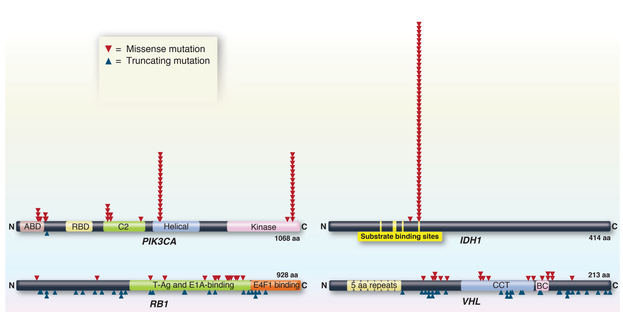
- **Cancer is a disease driven by acquired genetic mutations (somatic mutations).**

Progression of cancer is a multi-stage process. The underlying background mutation process generated mostly passenger mutations and a few mutations with selection advantage will be seen more often by chance. The challenge is to distinguish these driving mutations from the heterogeneous background mutations. [1].



- **Representive TSGs and OCGs with mutations identified in cancer sequencing studies [2].**

Several features has been observed for mutations in tumor suppressor genes and oncogenes. However a unified statistical framework that could use all relevant features to model background and driving mutations has not been proposed.



[1] Michael S et al Lawrence. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–8, jul 2013.

[2] Bert et al Vogelstein. Cancer Genome Landscapes. *Science*, 339(6127):1546 LP – 1558, mar 2013.

Model

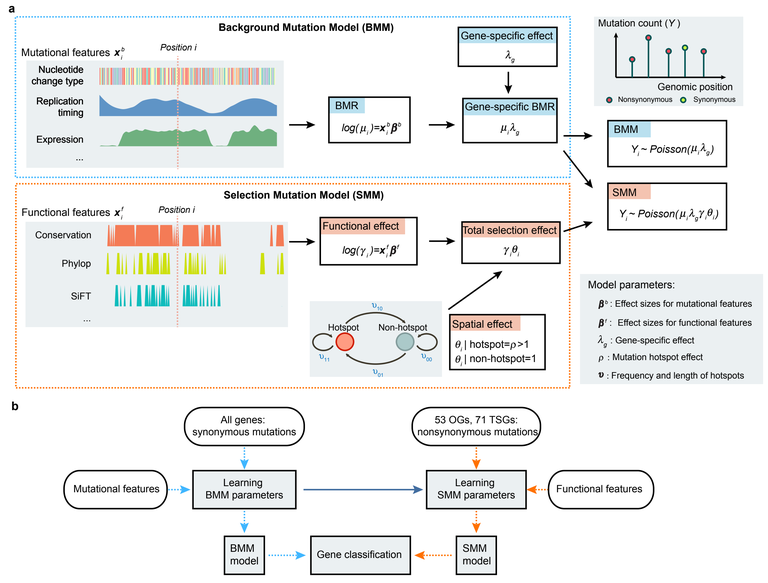


Figure 1 Overview of the model-based framework driverMAPS for cancer driver gene discovery
(a) Base-level Bayesian statistical modeling of mutation count data in driverMAPS. For positions without selection, the observed mutation rate is modeled by Background Mutation Model (BMM). Under BMM, the Background Mutation Rate (BMR)(μ_i) is determined by the log-linear model that takes into account known mutational features and further adjusted by gene-specific effect (λ_g) to get gene-specific BMR ($\mu_i\lambda_g$). For positions under selection, the observed mutation rate is modeled as gene-specific BMR adjusted by selection effect (Selection Mutation Model, SMM). The selection effect has two components: functional effect (γ_i) takes into account functional features of the position by the log-linear model and spatial effect (θ_i) takes into account the spatial pattern of mutations by Hidden Markov Model. For both BMM and SMM, given the mutation rate, the observed mutation count data is modeled by Poisson distribution. Note: we simplify the model to only show mutation rate at position i, ignoring allele specific effect for illustration purposes. See Methods for full parameterization. **(b)** Gene classification workflow. Parameters in BMM are estimated using synonymous mutations from all genes. This set of parameters is fixed when inferring parameters in SMM. To infer parameters in SMM, we use nonsynonymous mutations from known OGs or TSGs. driverMAPS then performs model selection by computing gene-level Bayes Factors to prioritize cancer genes.

Results

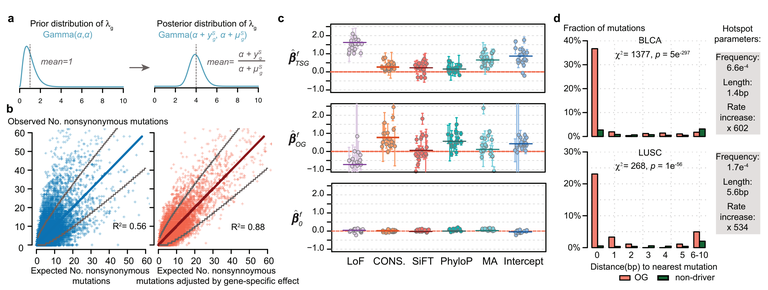


Figure 2 Parameter estimation results for gene-specific, functional and spatial effects
(a) Schematic representation of how fitting synonymous mutation data affects estimation of gene-specific effect (λ_g). Note the difference between the prior and posterior distributions of λ_g. α is a hyperparameter, y_g^S and m_g^S are the observed and expected number of synonymous mutations in gene g, respectively. **(b)** Improved fitting of observed number of nonsynonymous mutations in genes with gene-specific effect adjustment. Data from tumor type SKCM was used. The adjustment here is the posterior mean of λ_g fitting synonymous mutation data ($\frac{\alpha + y_g^S}{\alpha + \mu_g^S}$). Each dot represents one gene. Grey lines indicate upper and lower bounds of 99% confidence interval from Poisson test. The diagonal line has slope =1 and R^2 was calculated using this as the regression line. **(c)** Effect sizes for five functional features and average increased mutation rate for TSGs (top), OGs (middle) and non-driver genes (bottom). Each dot represents an estimate from one tumor type. Horizontal bars represent mean values after shrinkage. All features are binarily coded. LoF, loss-of-function (nonsense or splice site) mutations or not. CONS., amino acid conservation; SiFT, PhyloP and MA, predictions from software SiFT[1], PhyloP[2] and MutationAssessor[3], respectively; intercept, average increased mutation rate. **(d)** Fraction of mutations that has the nearest mutation 0,1,2,.. bp away, where 0bp means recurrent mutations. Data from tumor type BLCA and LUSC was used. The test statistics χ^2 and p values were obtained in the spatial model selection procedure (see method, Table S6). Inferred parameters related to the spatial model are shown on the right.

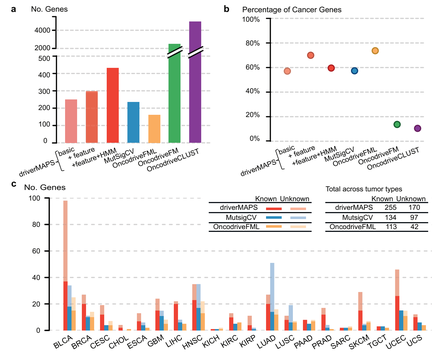


Figure 3 driverMAPS predicts cancer genes with high accuracy and increased power
(a) Total number of predicted driver genes aggregating across all cancer types. driverMAPS (Basic), driverMAPS with no functional features information and no modeling of spatial pattern; driverMAPS (+ feature), driverMAPS with all five functional features in Figure 2, no modeling of spatial pattern; driverMAPS (+feature + HMM), complete version of driverMAPS with all five functional features and spatial pattern. **(b)** Percentage of known cancer genes among predicted driver genes aggregating across all cancer types. **(c)** Number of significant genes at FDR=0.1 stratified by tumor type. For all "Unknown" genes included here, we verified mutations by visual inspection of aligned reads using files from Genomic Data Commons (see Supplementary notes). Total numbers of known and unknown significant genes aggregating across all cancer types are summarized in the table on the topright side.

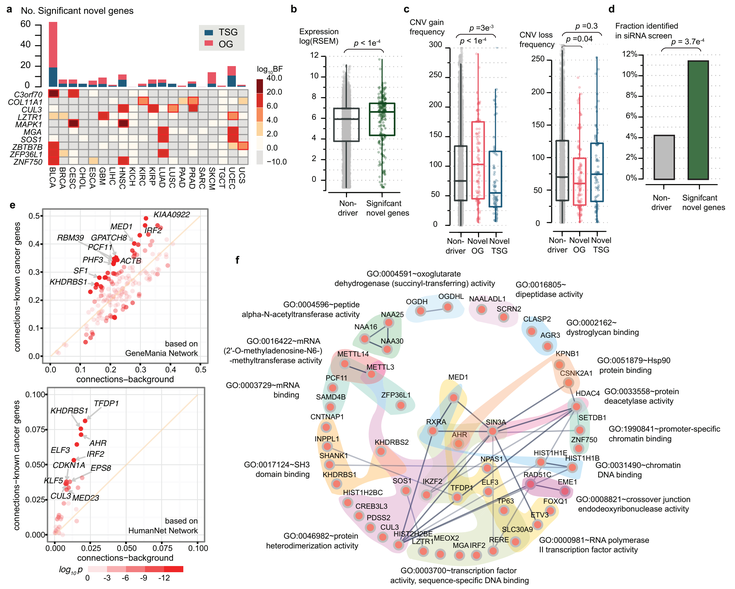


Figure 4 Evaluation of novel cancer genes predicted by driverMAPS
(a) Overview of predicted novel cancer genes. Top, number of novel genes in each cancer type. Bottom, heatmap of Bayes factors (BF) for recurrent novel genes across tumor types. Significant Bayes factors are highlighted by red boxes. **(b-d)** Predicted novel cancer genes show known cancer gene features. For each feature, quantification of the feature level in the novel cancer gene set was compared to the non-driver (neither known or predicted) gene set. The features are gene expression levels[4] stratified by tumor types the novel genes were identified from (b), similarly stratified copy number gain/loss frequencies[4] (c) and fraction of genes identified in a siRNA screen study[5] (d). **(e)** Enriched connectivity of a predicted gene with 713 known cancer genes (Y-axis) compared to with all genes (n=19,512, X-axis). Connectivity of a selected gene with a gene set is defined as the number of connections between the gene and gene set found in a network database divided by the size of the gene set. Each dot represents one of the 159 novel genes with 10 most enriched ones labeled. Color of dots indicates two-sided Fisher exact p value for enrichment. **(f)** Significantly enriched GO-term gene sets (FDR < 0.1, "molecular function" domain) in predicted novel cancer genes. GO-term[6,7] gene sets are indicated by distinct background colors. Links among genes represent interaction based on STRING network database[8] with darker color indicating stronger evidence.

Conclusions

- Here, we present a powerful statistical approach, driverMAPS (Model-based Analysis of Positive Selection) for driver gene identification.
- The key feature of driverMAPS is its modeling of mutation rates at the base-level, reflecting both background mutational processes and positive selection. Its selection model captures elevated mutation rates in functionally important sites using multiple external annotations, as well as spatial clustering of mutations. Its background mutation model accounts for both known covariates and local, gene- specific, variation caused by unknown factors.
- Applying driverMAPS to TCGA data across 20 tumor types identified 159 new potential driver genes. Cross-referencing this list with data from external sources strongly supports these findings.

We want to thank the Research Computing Center(RCC) at University of Chicago for providing wonderful computing resources for this project.