Bioinformatics seminar: Finding cancer driver mutations

Ilya Minkin

Saint Petersburg Academical University

8th December 2012

Motivation

- It is widely accepted that tumorogenesis heavily depends on accumulation of specific mutation
- If we consider genome of a tumor cell, we may see thousands of alterations
- Which mutations cause functional changes that enhance tumor cell proliferation, i.e. "driver"mutations?
- And which mutations are unrelated to this process, i.e. "passenger"mutations?
- This questions is one of the most challenging in cancer genetics

Motivation

- Genes that mutate in wide range of tumors can be confidently classified as "driver"genes
- However, there are many "driver"genes that mutate in $< 1\%$ of tumors
- We can't use traditional methods (studies in model organisms, gene KO, etc) to analyze hundreds of gene candidates
- Thus we need a high-throughput computational method for finding "driver"mutations that is independent on frequency
- The papers are focused on missense mutations, not nonsense/frameshift

CAN-Predict

Distinguishing Cancer-Associated Missense Mutations from Common Polymorphisms, Cancer Research (2007)
Joshua S. Kaminker, Yan Zhang, Allison Waugh, Peter M. Haverty, Brock Peters, Dragan Sebisanovic, Jeremy Stinson, William F. Forrest, J. Fernando Bazan, Somasekar Seshagiri, and Zemin Zhang

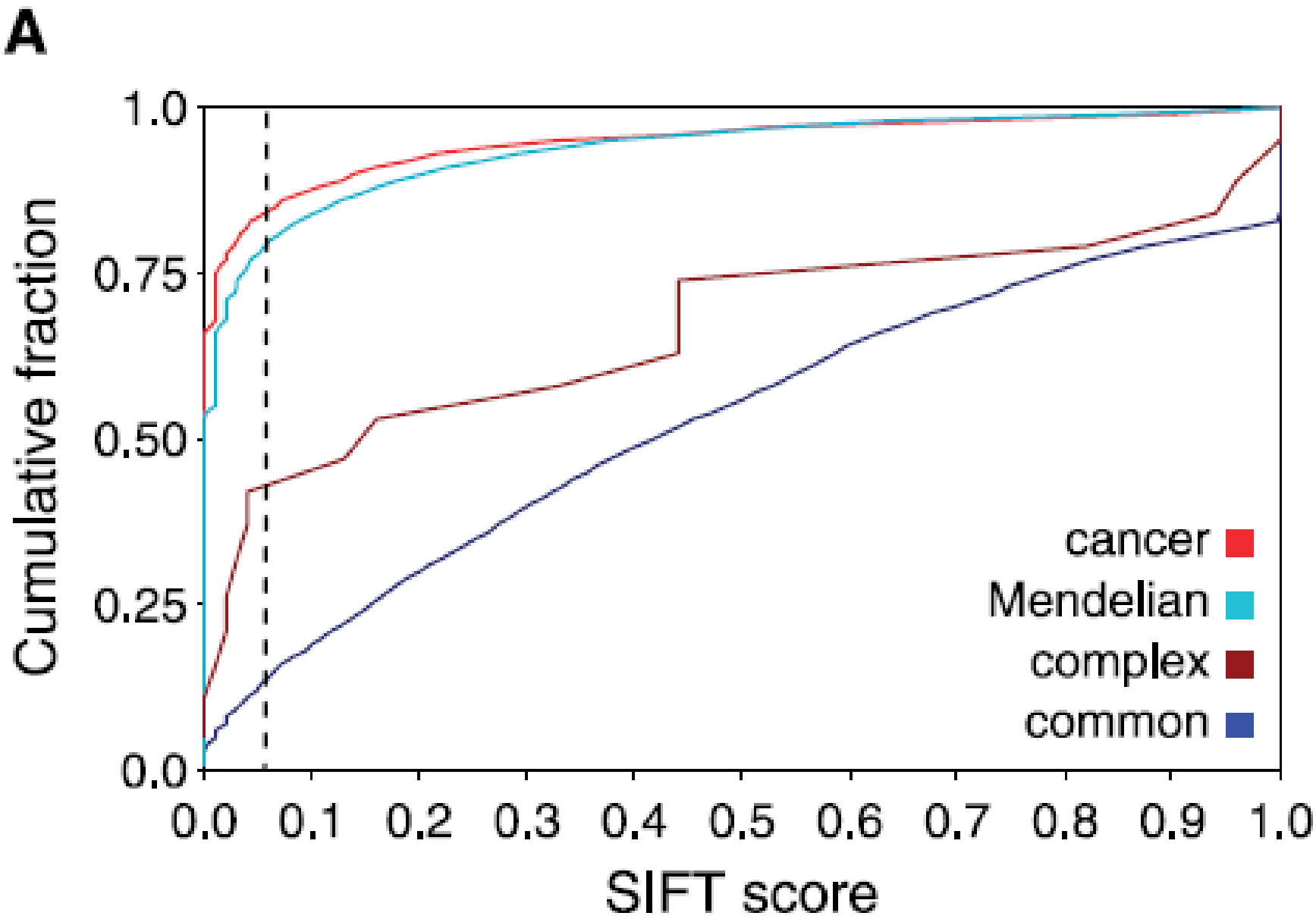
Overview

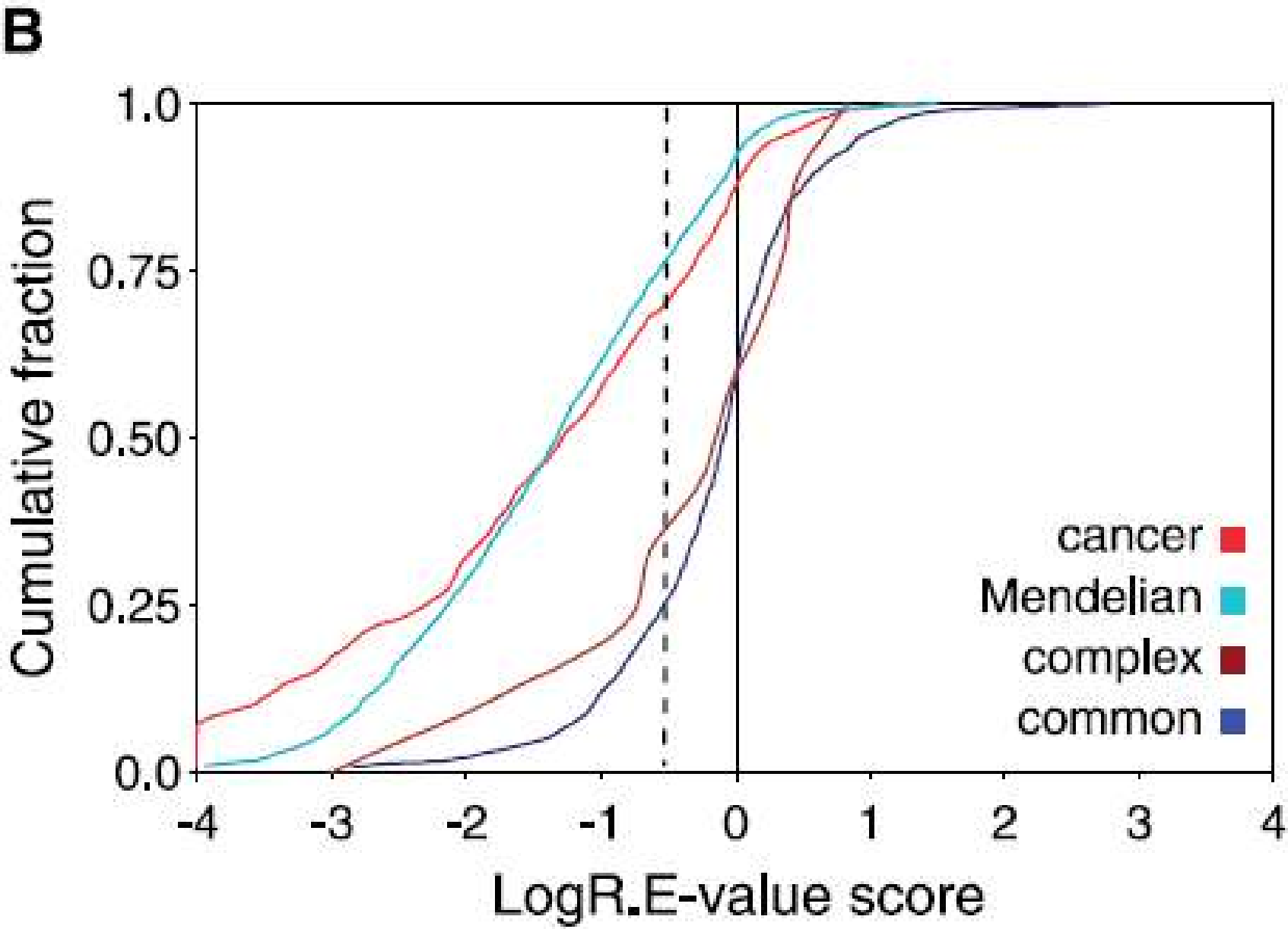
- There are few well-characterized classes of SNPs in the human genome:
- Common missense SNP
- Mendelian disease SNP – variants that cause disease and follow Mendelian pattern of inheritance
- Complex disease SNP – variants that are related to complex disease caused by many genetic and environmental factors
- Cancer driving mutations
- We will try to understand how to distinguish them and consequently build a classifier

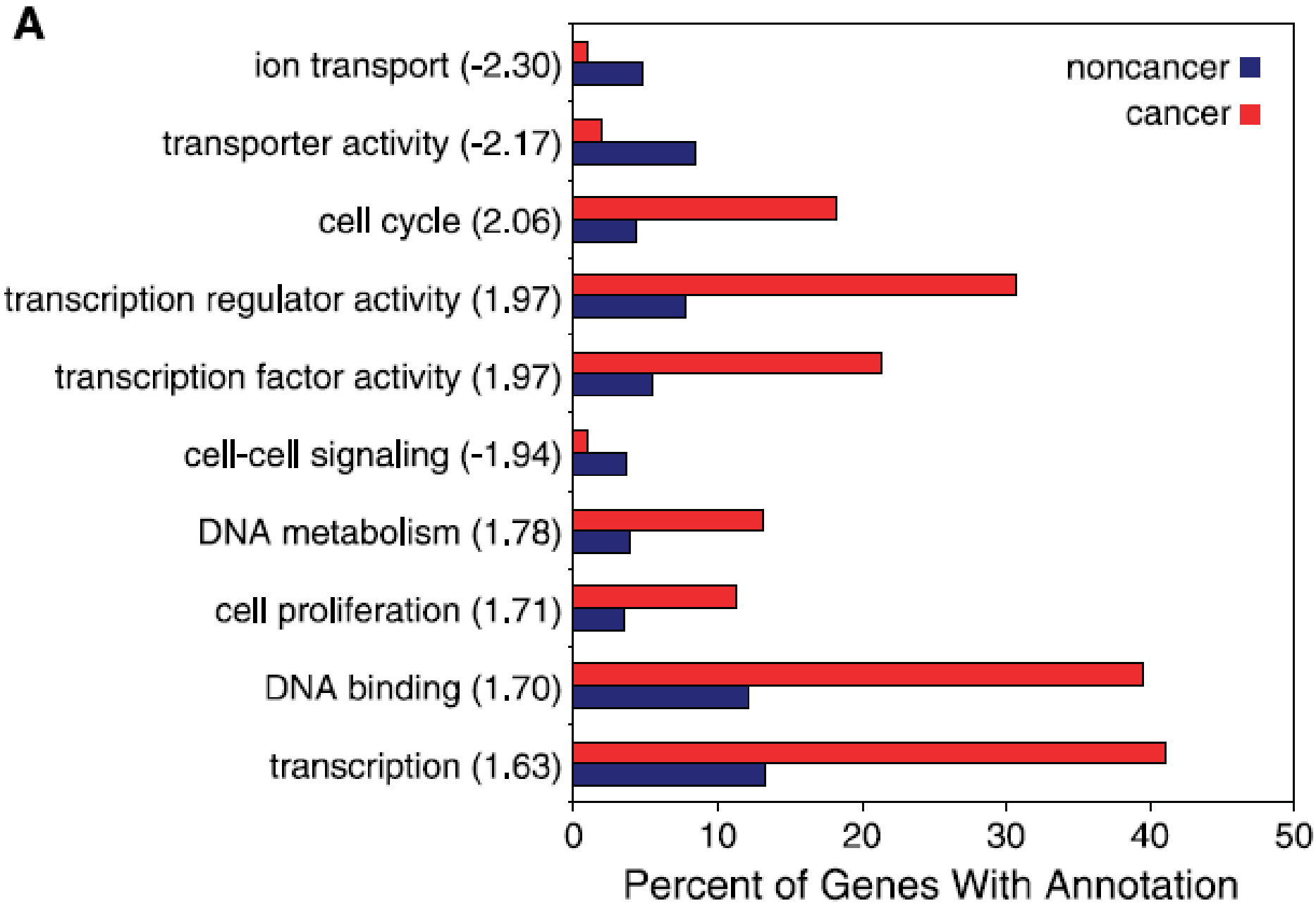
Three characteristics

For each known mutation we obtain:

- 1. SIFT predicts whether an amino acid substitution affects protein function and gives a score
- 2. Align both variant/canonical protein against Pfam database using HMMER. Take the best E-scores and calculate $log_{10}(E_{variant}/E_{canonical})$
- 3. The log-odds scores representing the relative frequency with which a Gene Ontology (GO) term was used to annotate cancer or noncancer gene sets


A
Cumulative fraction
1.0
0.75
0.50
0.25
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
SIFT score
cancer
Mendelian
complex
common


B
1.0
0.75
0.50
0.25
0.0
Cumulative fraction
-4
-3
-2
-1
0
1
2
3
4
LogR.E-value score
cancer
Mendelian
complex
common


A
ion transport (-2.30)
transporter activity (-2.17)
cell cycle (2.06)
transcription regulator activity (1.97)
transcription factor activity (1.97)
cell-cell signaling (-1.94)
DNA metabolism (1.78)
cell proliferation (1.71)
DNA binding (1.70)
transcription (1.63)
noncancer
cancer
0
10
20
30
40
50
Percent of Genes With Annotation

Conclusions

- We see that cancer-driving mutations are similar to Mendelian diseases SNPs
- These three features can be used for classification
- Random forests are used for building the classifier

Random forests

Random forest is an ensemble of decision trees. They are powerful and fast. Each tree is grown as follows:

- If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree
- If there are M input variables, a number m$\ll$M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node
- Each tree is grown to the largest extent possible

Validation

- The Out Of Bag (OOB) error rate is an important measure of accuracy and is calculated by applying each individual classification tree to a subset of training data points not used in the construction of the tree
- OOB = 3.19%
- Cross validation: of 730 variants, only 10 of 581 (1.7%) normal variants were misclassified as cancer and only 13 of 149 (8.7%) cancer variants were misclassified as normal

An application of such classifier is distinguishing relevant cancer-associated mutations from the expected polymorphic variants often identified during sequencing projects

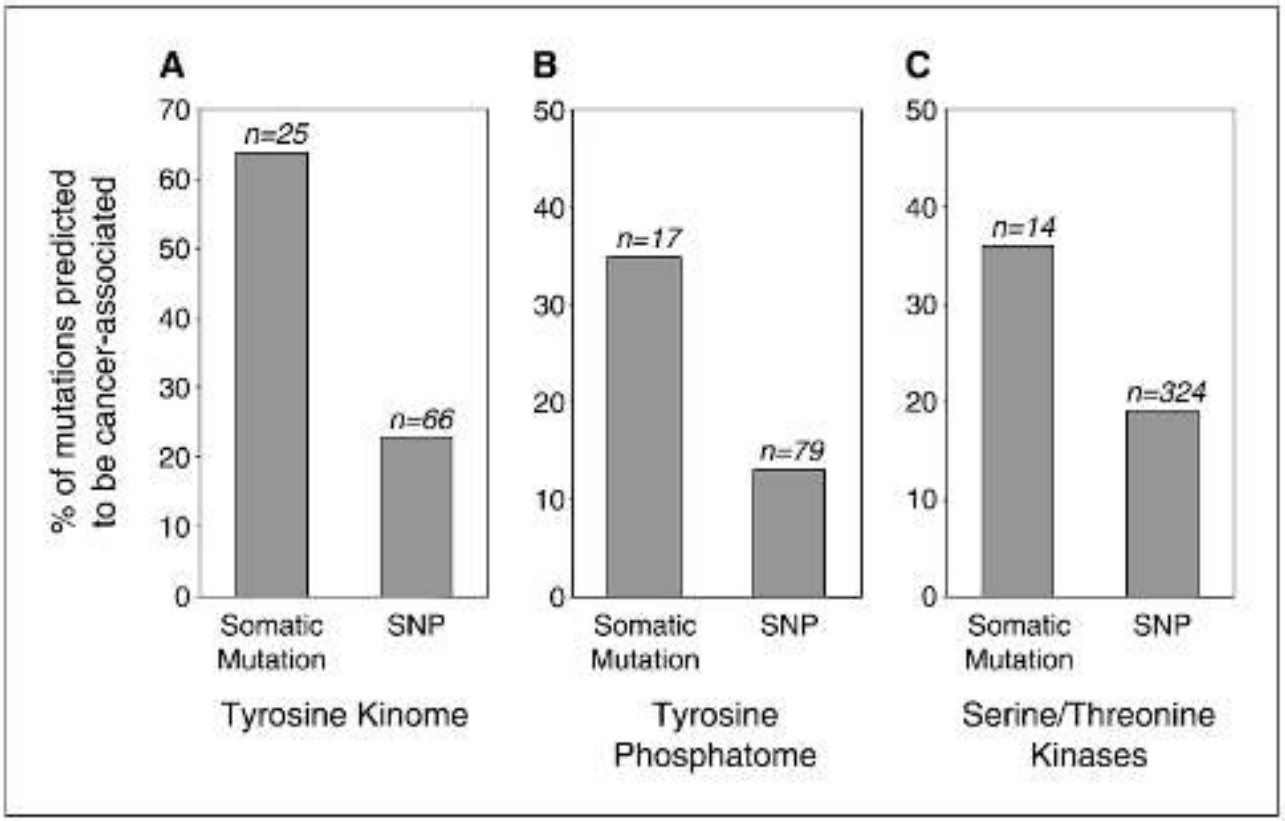


Figure 4. Data sets of somatic mutations are more likely to be predicted to be cancer-associated than variants in dbSNP. Variants in genes of the tyrosine kinome (*A*), tyrosine phosphatome (*B*), or serine/threonine kinases (*C*) were classified using the RF predictor described in the text. Classes of somatic variants and variants isolated from dbSNP are indicated below each panel. Only variants for which there was a GO log-odds score, a LogR.E-value score, and a SIFT score were used in this analysis. The total number of variants is labeled above each column. The percentage of variants predicted to be cancer-associated for somatic and SNP variants is as follows: *A*, 64% and 23%; *B*, 35% and 13%; *C*, 36% and 19%.

Discussion

- There is a classifier that distinguishes between cancer-driving mutations and passenger
- Cross-validation shows that it is pretty accurate
- It was also shown that somatic mutations are more likely to be predicted as cancer-driving compared to common SNPs
- They also tried to predict some novel cancer-driving mutations, but I'm so bored and won't tell about it

CHASM

Cancer-Specific High-Throughput Annotation of Somatic Mutations: Computational Prediction of Driver Missense Mutations, Cancer Research(2009)
Hannah Carter, Sining Chen, Leyla Isik, Svitlana Tyekucheva, Victor E. Velculescu, Kenneth W. Kinzler, Bert Vogelstein, and Rachel Karchin

Previous work

- There were classifiers developed earlier
- Driver mutations are similar to mutations associated with Mendelian disease and may be identifiable by setting constraints on amino acid residues at mutated positions
- Passengers are more similar to nonsynonymous single nucleotide polymorphisms (nsSNP) with high minor allele frequencies (MAF)
- Random forests(CAN-Predict), SVM(Protein kinase-specific classifier)

Previous work

- Previous work used common SNPs as negative training examples
- Existing computational methods could detect differences between somatic missense mutations observed in cancers and high MAF nsSNPs
- But these differences might be less relevant to the discrimination between driver and passenger mutations that occur somatically in tumors

SNPs and passenger mutations

- Although high MAF nsSNPs and passenger mutations have properties in common, they also have differences
- Passenger mutations may or may not have a functional impact on proteins; by definition, they are neutral with respect to cancer cell fitness
- In contrast, high MAF nsSNPs have become fixed in the human genome and must be functionally neutral or have a mild functional impact with respect to normal cell fitness
- Let's generate synthetic set of passenger mutations and train a classifier

Overview

- Random Forest classifier that was trained on 49 predictive features
- Feature selection was done with a protocol based on mutual information
- Driver mutation data set – 2,488 missense mutations previously identified as playing a functional role in oncogenic transformation
- The synthetic passenger mutations were generated by sampling from eight multinomial distributions that depend on dinucleotide context and tumor type
- The final score yielded for each mutation is the fraction of trees that voted for the passenger class.

Candidate features

In total 80 features, such as:

- Change in charge
- BLOSUM 62 substitution score
- 17way exon conservation
- SNP Density (The number of genetic variants or polymorphisms in the exon where the mutation is located)
- ...

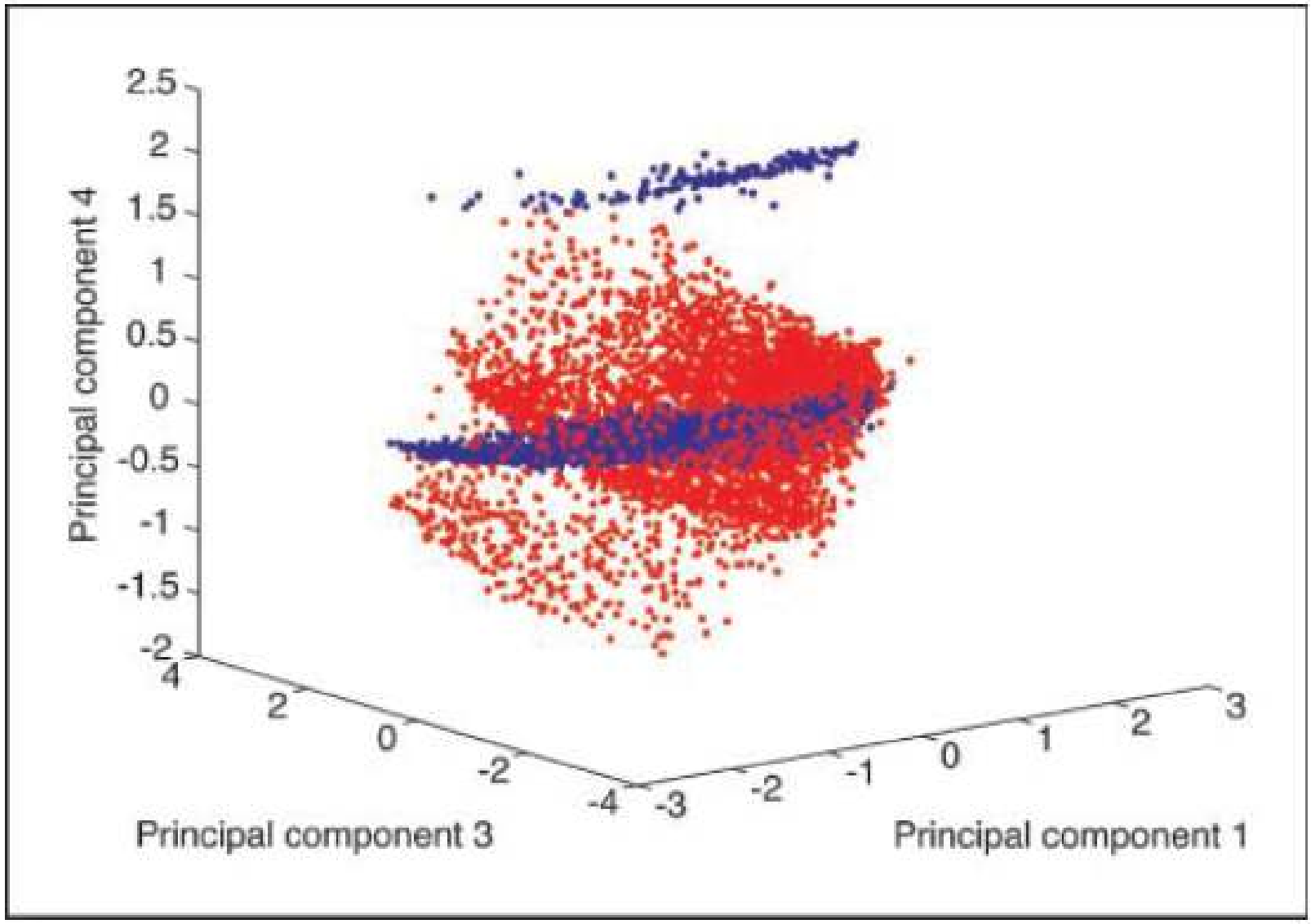


Figure 1. Principal components analysis of nsSNPs versus synthetic passenger mutations. Synthetic passenger mutations (*red*) and high MAF nsSNPs from the HapMap project (*blue*) have substantial overlap in the space defined by principal components one, three, and four, but there are regions in the space occupied only by high MAF nsSNPs and regions occupied only by synthetic passengers.

Probabilistic interpretation of random forest classification scores

- They used the trained Random Forest to compute a classification score for each of 607 glioblastoma multiforme (GBM) missense mutations
- However, these scores are not probabilities and the statistical behavior of the algorithm has not been well-characterized
- It is not evident where to set a trusted score cutoff for purposes of identifying driver mutations

Probabilistic interpretation of random forest classification scores

- For each of the 607 GBM mutants, we test the null hypothesis: the mutant is not functionally related to the growth of the tumor (passenger), versus the alternative hypothesis that it is (driver)
- We obtain a P value for a mutation by comparing its score to the null distribution, which consists of the scores of a filtered set of synthetic passengers that were held out from Random Forest training

Other methods

- PolyPhen classifies missense mutants as "Probably Damaging", "Possibly Damaging" or "Benign" and also provides a continuous measure of a mutation's functional impact, the PSIC score
- SIFT provides a score that ranges between 0 and 1 to report the probability that a missense mutation will be tolerated
- SIFT/PolyPhen consensus
- CanPredict
- KinaseSVM

Supplementary Table 4. *Comparison of CHASM with other methods for missense mutant function prediction.* Performance of each method is shown at its minimum error point. Relative Coverage is the fraction of mutations in the CHASM training set (5749 mutations) that could be classified by each method. AUCs could not be calculated for SIFT/PolyPhen Consensus (Methods).

	CHASM				
	Relative Coverage	Precision	Recall	AUC -ROC	AUC -PR
Training set	NA	0.82	0.58	0.91	0.79
TP53 test set	NA	0.98	0.97	0.996	0.99
EGFR test set	NA	0.92	0.88	0.98	0.96
	PolyPhen PSIC score				
	Relative Coverage	Precision	Recall	AUC -ROC	AUC -PR
Training set	56%	0.33	0.003	0.66	0.29
TP53 test set	59%	0.64	0.65	0.85	0.64
EGFR test set	64%	0.57	0.1	0.68	0.42
	SIFT Score				
	Relative Coverage	Precision	Recall	AUC -ROC	AUC -PR
Training set	42%	0.39	0.49	0.62	0.36
TP53 test set	81%	0.58	0.83	0.81	0.56
EGFR test set	67%	0.09	0.29	0.53	0.09
	SIFT/PolyPhen Consensus				
	Relative Coverage	Precision	Recall	AUC -ROC	AUC -PR
Training set	22%	0.44	0.59	NA	NA
TP53 test set	51%	0.57	0.96	NA	NA
EGFR test set	42%	0.13	0.38	NA	NA

Supplementary Table 5. Method Comparison Statistics. Methods are assessed by their coverage and performance on three datasets, namely the CHASM training drivers and passengers, held out TP53 mutations and passengers and held out EGFR mutations and passengers. The same set of held out passengers, distinct from those used in the training set, are used in both the TP53 and EGFR test sets.

	PolyPhen PSIC score					
	Coverage		Precision	Recall	AUC-ROC	AUC-PR
	drivers	passengers				
Training set	590/1248	2623/4500	0.33	0.003	0.66	0.29
TP53 test set	133/196	332/590	0.64	0.65	0.85	0.64
EGFR test set	132/133	332/590	0.57	0.1	0.68	0.42
	SIFT Score					
	Coverage		Precision	Recall	AUC-ROC	AUC-PR
	drivers	passengers				
Training set	650/1248	1754/4500	0.39	0.49	0.62	0.36
TP53 test set	195/196	444/590	0.58	0.83	0.81	0.56
EGFR test set	41/133	444/590	0.09	0.29	0.53	0.09
	SIFT/PolyPhen Consensus					
	Coverage		Precision	Recall	AUC-ROC	AUC-PR
	drivers	passengers				
Training set	428/1248	858/4500	0.44	0.59	NA	NA
TP53 test set	133/196	266/590	0.57	0.96	NA	NA
EGFR test set	40/133	266/590	0.13	0.38	NA	NA
	CanPredict Score					
	Coverage		Precision	Recall	AUC-ROC	AUC-PR
	drivers	passengers				
Training set	685/1248	1885/4500	0.62	0.63	NA	NA
TP53 test set	190/196	259/590	0.8	0.995	NA	NA
EGFR test set	114/133	259/590	0.58	0.55	NA	NA
	KinaseSVM Score					
	Coverage		Precision	Recall	AUC-ROC	AUC-PR
	drivers	passengers				
Training set	218/1248	123/4500	0.7	0.92	0.71	0.81
TP53 test set	NA	NA	NA	NA	NA	NA
EGFR test set	112/133	29/590	0.92	0.8	0.71	0.95