生物信息学与系统生物学

张世华

中国科学院数学与系统科学研究院

Cancer genomics

An integrated approach to uncover drivers of cancer

De novo discovery of mutated driver pathways in cancer

Background

- Systematic characterization of cancer genomes has revealed a staggering number of diverse aberrations that differ among individuals.
- Therefore, the **functional importance** and **physiological impact** of most tumor genetic alterations remain poorly defined.

Theory

Cell

Cell 143, 1005–1017, 2010

An Integrated Approach to Uncover Drivers of Cancer

Uri David Akavia,[1,2,5] Oren Litvin,[1,2,5] Jessica Kim,[3,4] Felix Sanchez-Garcia,[1] Dylan Kotliar,[1] Helen C. Causton,[1] Panisa Pochanard,[3,4] Eyal Mozes,[1] Levi A. Garraway,[3,4] and Dana Pe'er[1,2,*]

Motivation

- **Each tumor is unique and typically harbors a large number of genetic lesions**
- **Only a few drive proliferation and metastasis.**
- Thus, identifying **driver mutations** (genetic changes that promote cancer progression) and **distinguishing them from passengers** (those with **no** selective advantage) has emerged as a major challenge in the genomic characterization of cancer.

Background(cont.)

- The most widely used approaches are based on the **frequency** that an aberration occurs: if a mutation provides a fitness advantage in a given tumor type, its persistence will be favored, and it is likely to be found in multiple tumors.
- GISTIC

Background(cont.)

- However, there are limitations to analytical approaches based on CNA data alone.
- CNA regions are typically large and contain many genes, most of which are passengers that are indistinguishable in copy number from the drivers.
- CNA data have statistical power to detect only the most frequently recurring drivers above the large number of unrelated chromosomal aberrations that are typical in cancer.

Background(cont.)

- These approaches rarely elucidate the functional importance or physiological impact of the genetic alteration on the tumor.
- **These limitations** highlight the need for new approaches that can integrate additional data to identify drivers of cancer.
- Gene expression is readily available for many tumors, but **how best to combine it with information on CNA is not obvious**.

The work in the **Cell** paper

- **Hypothesis**: driver mutations coincide with a "genomic footprint" in the form of a gene expression signature.
- **Contribution**: to find these signatures and identify likely driver genes located in regions that are amplified or deleted in tumors.
- Each **potential driver gene** is altered in some, but not all, tumors and, when altered, is considered likely to play a contributing role in tumorigenesis.

The work in the Cell paper

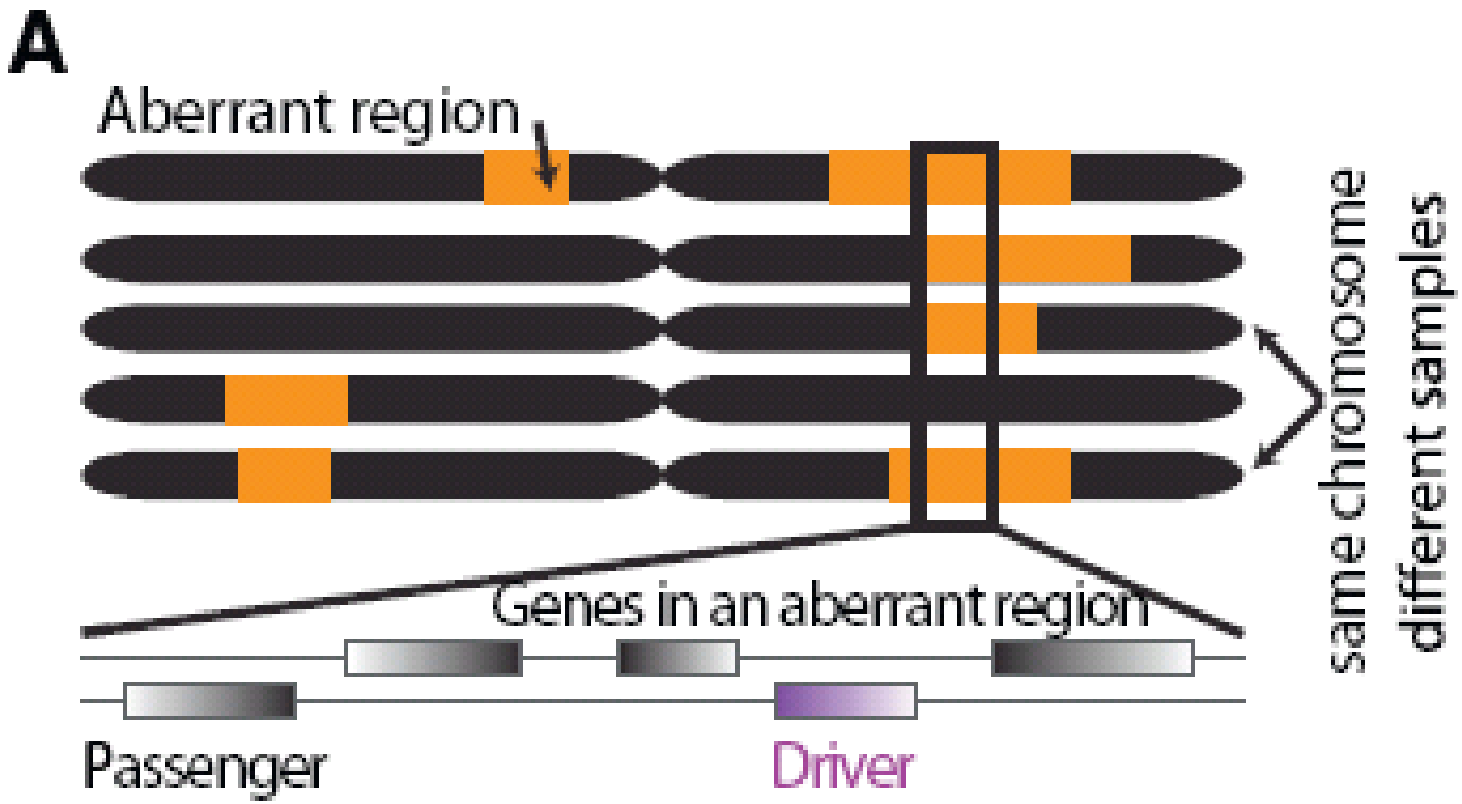
- **Unique**: each driver is associated with a gene module.
- **Transferring** the annotation of the genes in the associated module.

Hypothesis

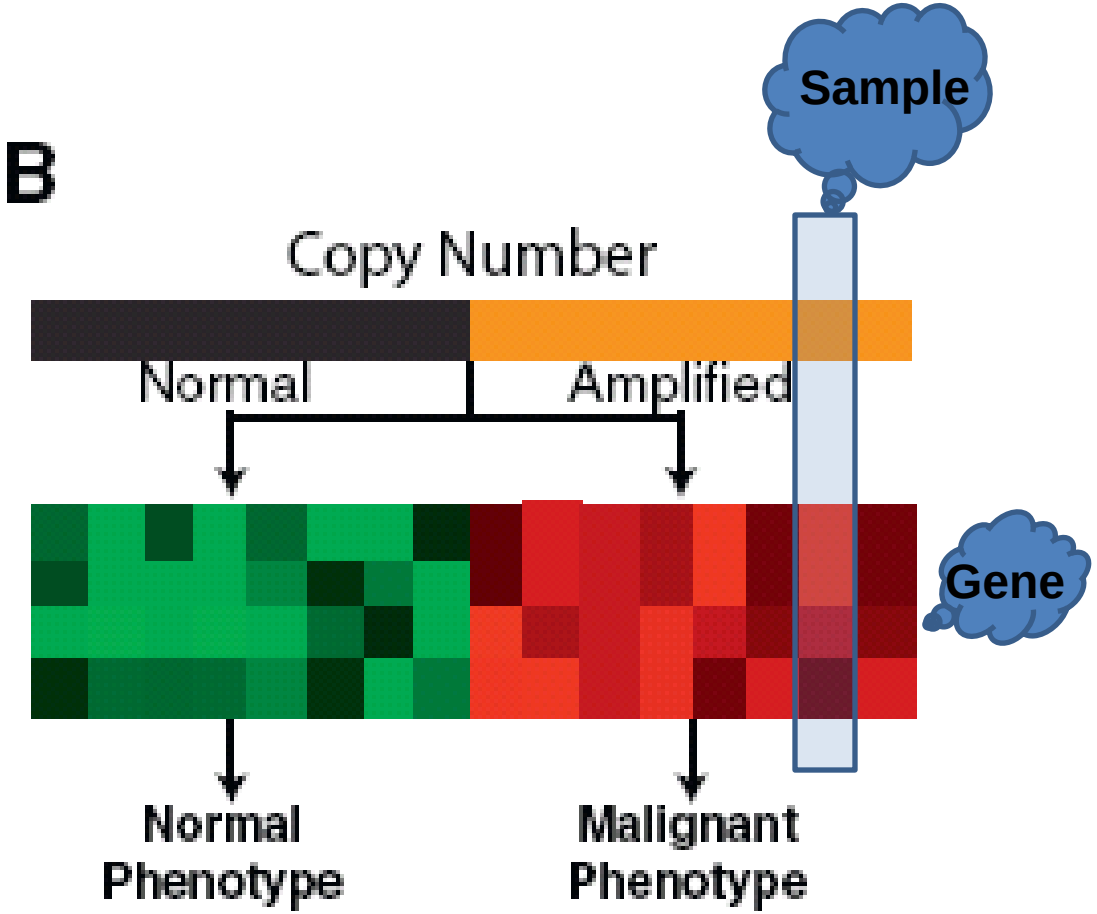
- **Hypothesis**: driver mutations coincide with a "genomic footprint" in the form of a gene expression signature.

- **Assumption 1**
- **Assumption 2**
- **Assumption 3**

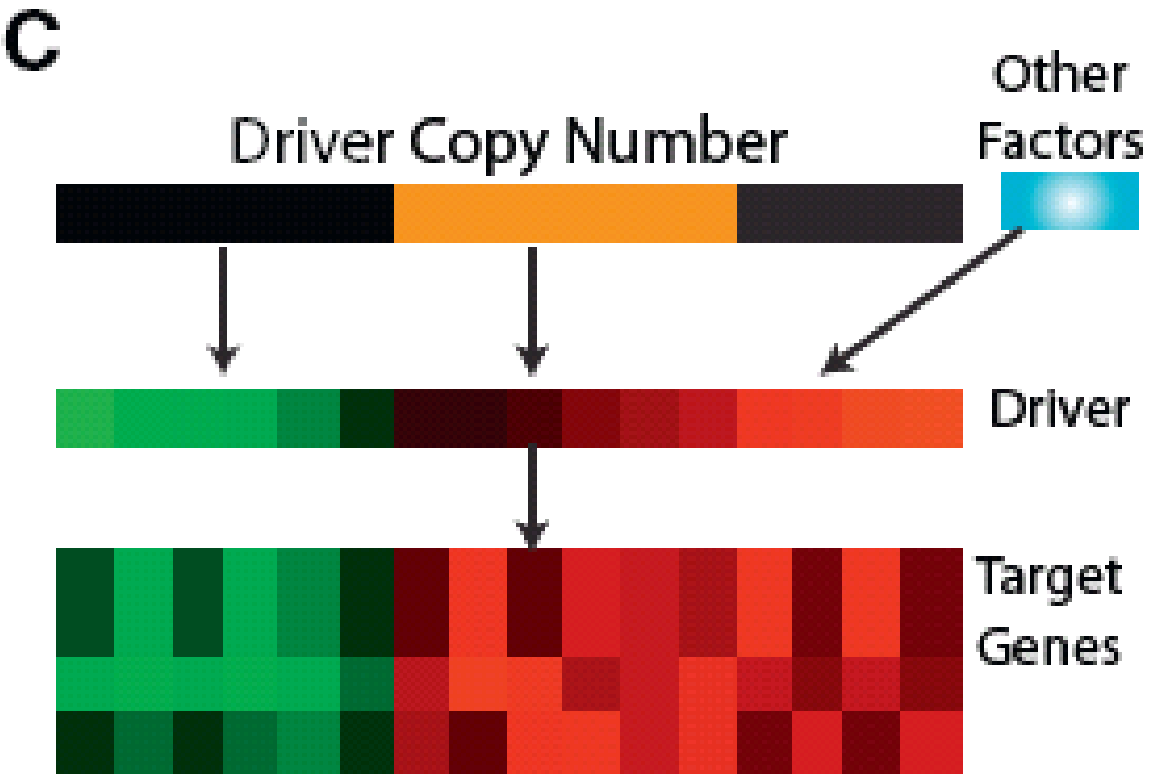
key assumptions 1



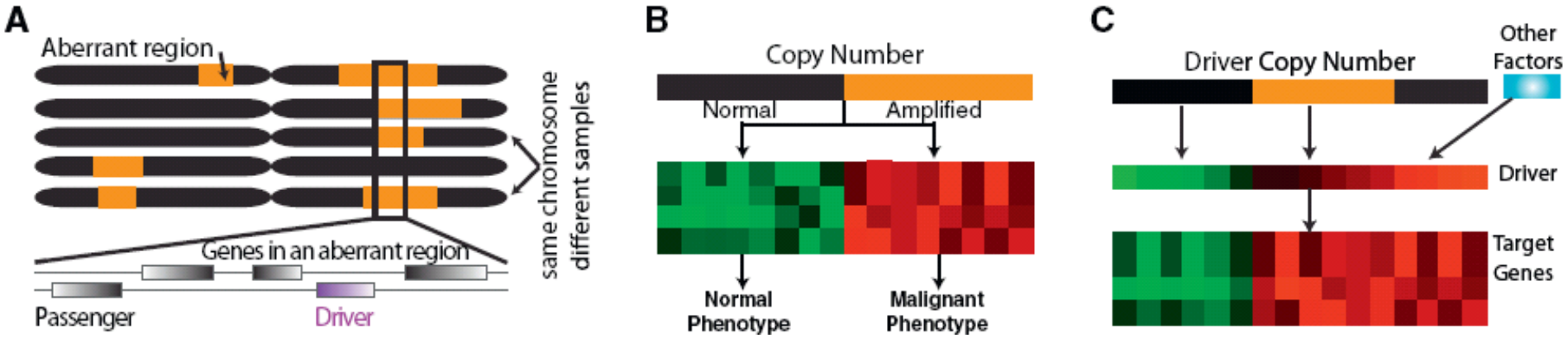
key assumptions 2



key assumptions 3

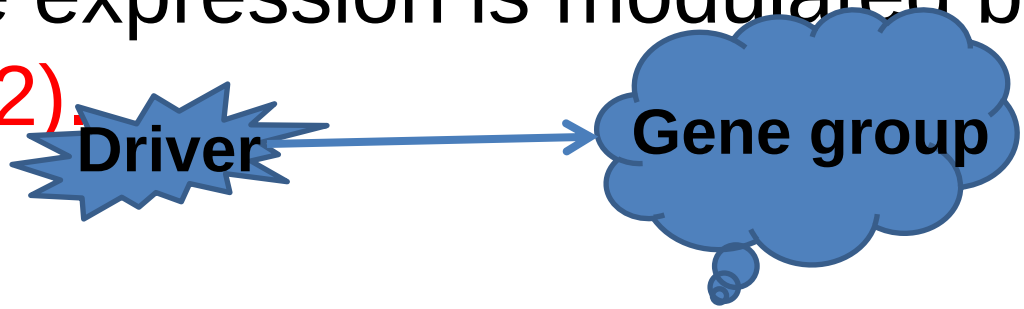


Three key assumptions that can distinguish driver and passenger mutations



Key ideas

- **Driver mutations** are frequently associated with the **abnormal regulation** of processes such as proliferation, differentiation (Key 1)
- A **driver mutation** might be associated with a characteristic gene expression signature or other phenotypic output representing a group of genes whose expression is modulated by the driver (Key 2).



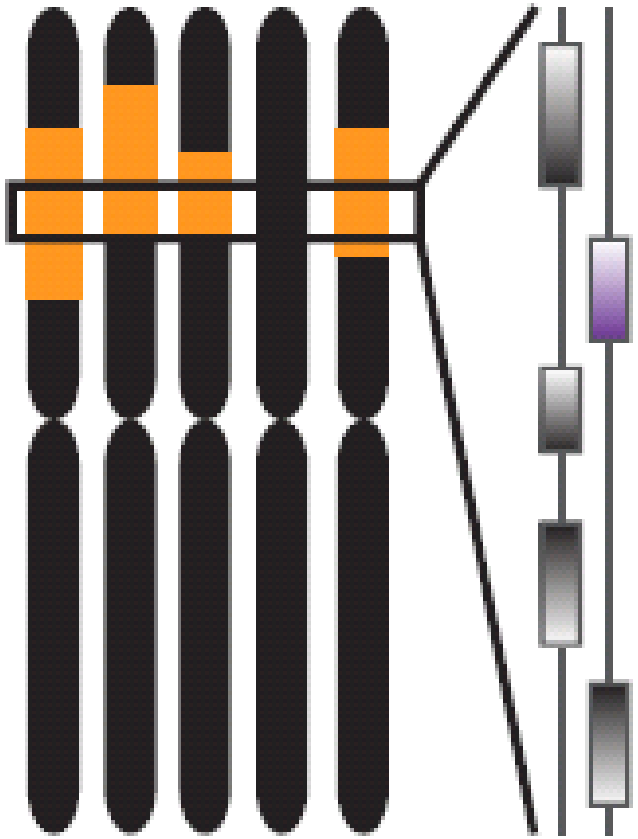
Key ideas

- CNAs do not typically alter the **coding sequence of the driver** and so are expected to influence cellular phenotype via changes in the **driver's expression (Key 3)**.

CONEXIC Learning Algorithm

1. GISTIC

Selection of candidate driver genes (modulators)



Amplified Genes:

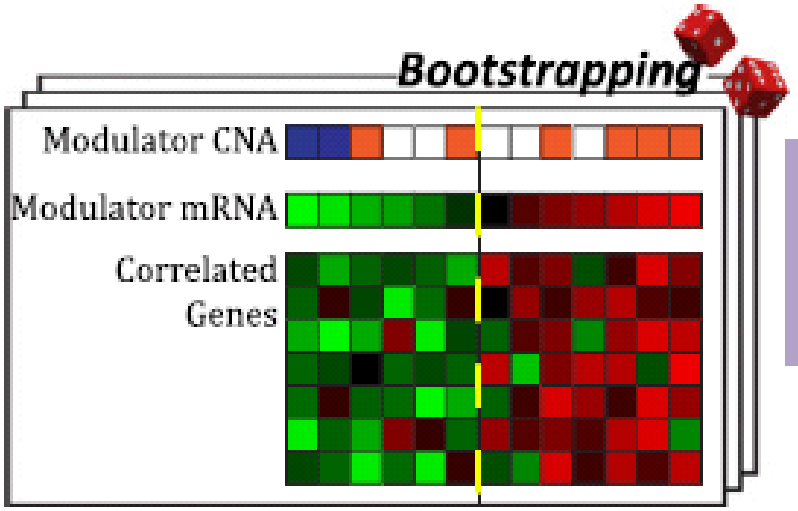
1. CCND1
2. MITF
3.....

Deleted Genes:

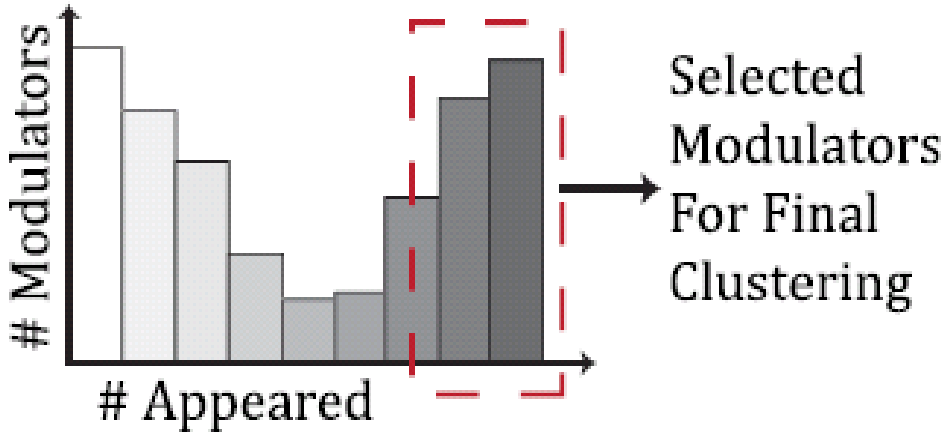
1. CDKN2A
2. KLF6
3.....

CONEXIC Learning Algorithm

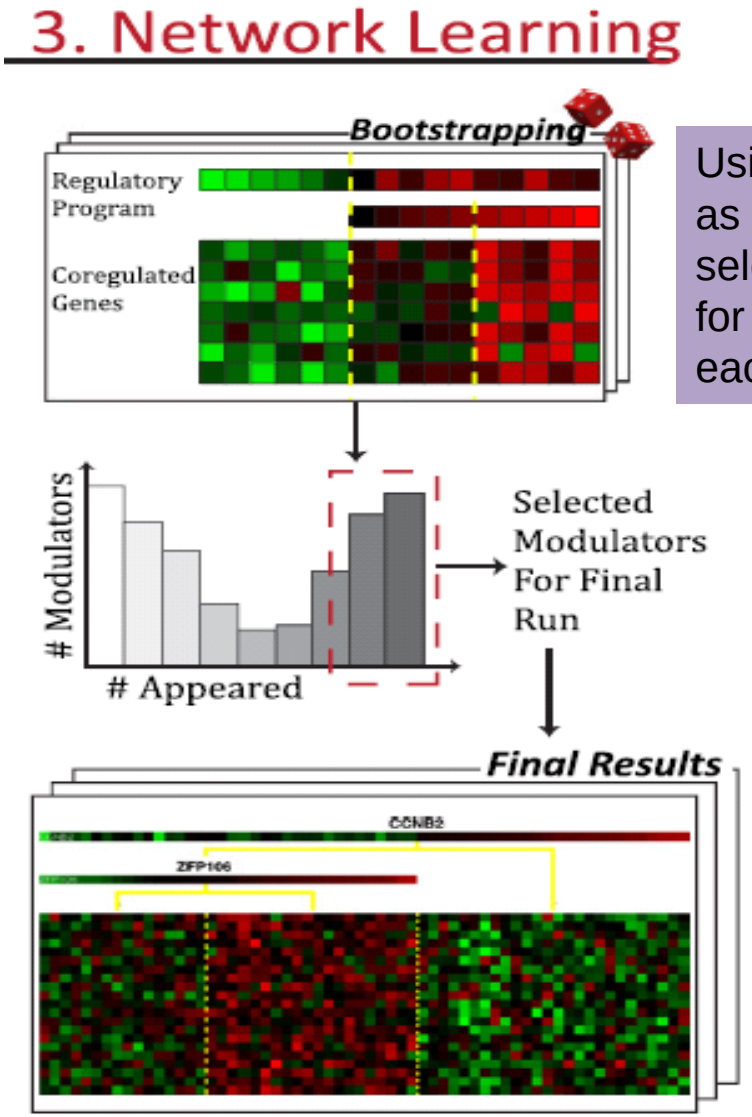
2. Single Modulator



Modules of genes are each associated with the best possible candidate driver, based on gene expression of the gene and the candidate driver.



CONEXIC Learning Algorithm



Using the set of Single Modulator of modules as a starting point, the algorithm refines the selected modulators and modules, now allowing for more than one modulator associated with each module.

Take-home message

Expression of a Driver, Not Its Copy Number, Drives Phenotype

How to analyze cancer mutation data?

De novo discovery of mutated driver pathways in cancer

Fabio Vandin, Eli Upfal and Benjamin J. Raphael

Genome Res. published online June 7, 2011
Access the most recent version at doi:10.1101/gr.120477.111

Cancer is driven by somatic mutations

- Cancer is driven by somatic mutations in the genome that are acquired during the lifetime of an individual.
- These include single nucleotide mutations and larger copy-number aberrations and structural aberrations.

- With the availability of next-generation DNA sequencing technologies, whole-genome or whole-exome measurements of the somatic mutations in large numbers of cancer genomes are now a reality.

- 体细胞突变是发生在正常机体细胞中的突变，比如发生在皮肤或器官中的突变。这样的突变不会传给后代。**体细胞突变**与**种系突变不同**，后者是发生在将成为配子（gametes）（精子和卵子）的细胞中。生殖细胞的突变可传递给后代。

Cancer is driven by somatic mutations

- **Somatic mutations can happen for a variety of reasons**
  - Some appear to be the result of exposure to toxins or radiation which interferes with the cell division process.
  - Others are **spontaneous**, occurring as the result of a random error in the cell division process. Given the length of the genome, occasional mistakes do happen in individual cells, and in fact the body is coded to destroy somatic cells which have mutated, although it is not always successful.

Cancer is driven by somatic mutations

A major challenge is to distinguish the functional "**driver mutations**" responsible for cancer from the random "**passenger mutations**" that have accumulated in somatic cells but that are not important for cancer development.

Cancer is driven by somatic mutations

However, many studies have confirmed that cancer genomes exhibit extensive **mutational heterogeneity**.

- The presence of passenger mutations
- Driver mutations target genes in cellular signaling and regulatory **pathways**.
- there are **numerous combinations** of driver mutations that can perturb a pathway important for cancer.
- This mutational heterogeneity complicates efforts to identify functional mutations by their recurrence across many samples

Driver pathway

- Driver mutations typically target genes in cellular signaling and regulatory **pathways**.
- Several studies begin to examine mutations in the context of cellular signaling and regulatory pathways.

- Two kinds of approach:
  - 1) Based on Known pathways (Boca et al. 2010; Efroni et al. 2011) or genome-scale gene interaction networks (Cerami et al. 2010; Vandin et al. 2011);
  - 2) De novo discovery.

Driver pathway

- Pathway or network analysis of cancer mutations relies on prior identification of the groups of genes in the pathways.
- knowledge of pathways **remains incomplete**!!
- In particular, many pathway databases contain a superposition of all components of a pathway, and information regarding which of these components **are active in particular cell types is largely unavailable**.

Driver pathway

- The above concerns
- Availability of increasing numbers of sequenced cancer genomes

- Is it possible to discover **mutated driver pathways**, **directly** from somatic mutation data collected from large numbers of patients.

Complexity

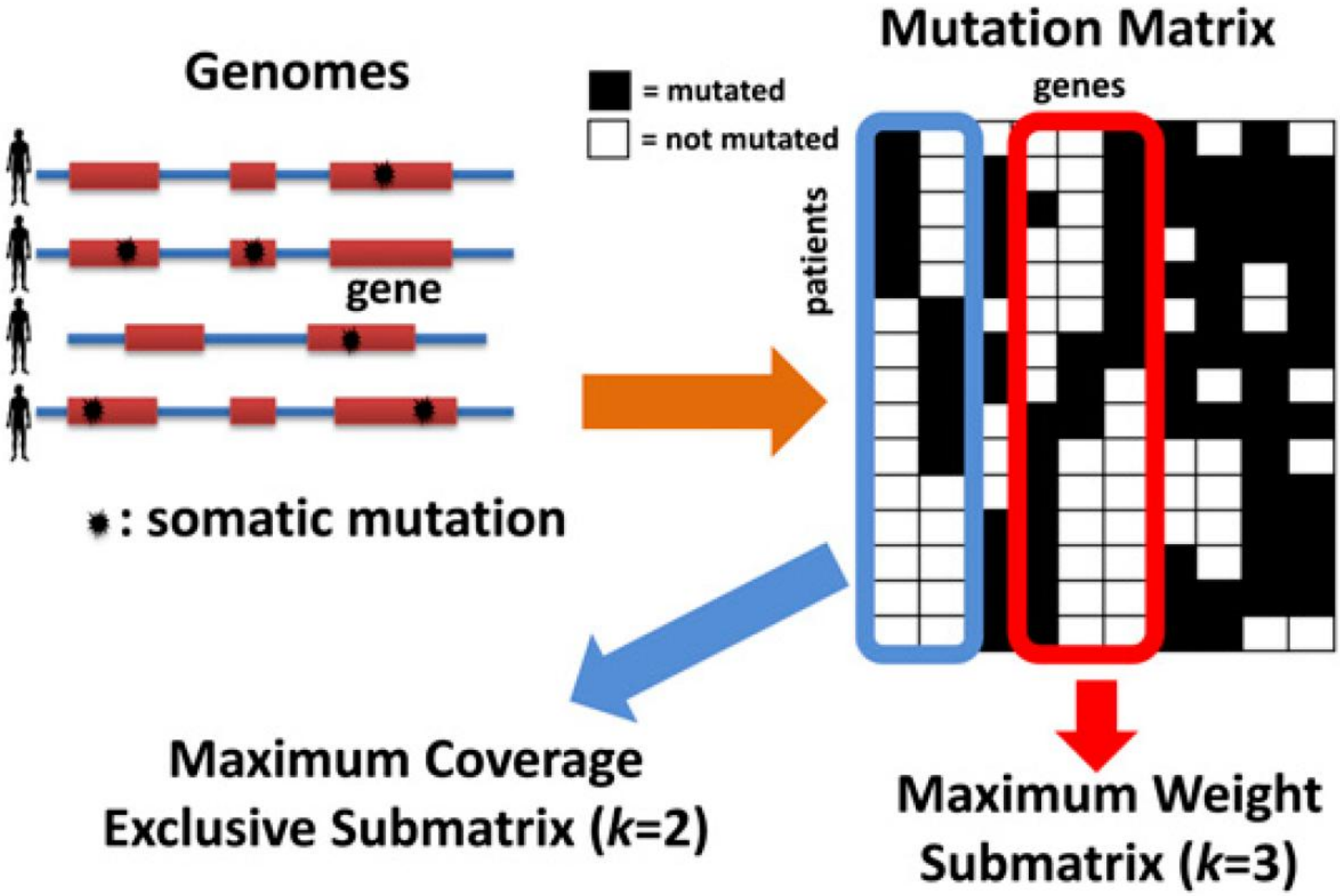
- Large-scale search space!
- Additional constraints are needed!

Hypothesis

The **Drive pathway** usually exhibits such property:

- 1) Coverage: It should be perturbed in a large number of patients;
- 2) Exclusivity: driver mutations are observed in exactly one gene in the pathway in each patient.

Mutation matrix



Maximum Coverage Exclusive Submatrix Problem

Identify sets of genes (columns of the mutation matrix) that are mutated in a large number of patients and whose mutations are mutually exclusive

Maximum Coverage Exclusive Submatrix Problem: *Given an $m \times n$ mutation matrix A and an integer $k > 0$, find a mutually exclusive $m \times k$ submatrix M of k columns (genes) of A with the largest number of non-zero rows (patients).*

This problem is NP-hard!

This problem is too restrictive!

How to slack this problem

- Coverage: Most patients have at least one mutation in M. $\Gamma(M)$
- Approximate exclusivity: Most patients have no more than one mutation in M.

Coverage overlap: $\omega(M) = \sum_{g \in M} |\Gamma(g)| - |\Gamma(M)|$

There is an obvious trade-off.

$$W(M) = |\Gamma(M)| - \omega(M) = 2|\Gamma(M)| - \sum_{g \in M} |\Gamma(g)|$$

Maximum Weight Submatrix Problem

Maximum Weight Submatrix Problem: *Given an $m \times n$ mutation matrix A and an integer $k > 0$, find the $m \times k$ column submatrix $\hat{M}$ of A that maximizes $W(M)$.*

This problem is NP-hard too!

The problem of extracting subsets of genes with particular properties has also been studied in the context of gene expression data.

A greedy algorithm

A greedy algorithm and Gene Independence Model

We propose the following greedy algorithm for the Maximum Weight Submatrix problem.

Greedy (k) :

1. $M = \{g_1, g_2\} \leftarrow$ pair of genes that maximizes $W(\{g_1, g_2\})$.
2. For $i = 3, \ldots, k$ do :
   (a) Let $g^* = \arg\max_g W(M \cup \{g\})$.
   (b) $M \leftarrow M \cup \{g^*\}$.
3. return M.

The time complexity of the algorithm is $O(n^2 + kn) = O(n^2)$. We analyze the performance of the algorithm on mutation matrices generated from the following Gene Independence Model.

Markov chain Monte Carlo (MCMC) algorithm

Initialization: Choose an arbitrary subset M_0 of k genes in $\mathcal{G}$ (the set of all genes).

Iteration: For $t = 1, 2, \ldots$, obtain M_{t+1} from M_t as follows:

1. Choose a gene w uniformly at random from $\mathcal{G}$.
2. Choose v uniformly at random from M_t.
3. Let $P(M_t, w, v) = \min\left[1, e^{cW(M_t - \{v\} + \{w\}) - cW(M_t)}\right]$.[7]
4. With probability $P(M_t, w, v)$ set $M_{t+1} = M - \{v\} + \{w\}$, else $M_{t+1} = M_t$.

The MCMC algorithm was run for 10^7 iterations sampling every 10^4 iterations.

Simulation study

How to simulate the mutation data?

	Optimal gene set (mutation control by p1 and p2)				The rest gene set Mutated using a random model					
If p1 happens	p1	p2	0	0						
	0	p1	p2	0						
	0	0	p1	p2						
	0	p1	p2	0						
If p1 doesn't happen	0	0	0	0						
	p1	p2	0	0						
	0	p1	0	p2						

For SN mutations,

For CNAs,

Simulation study

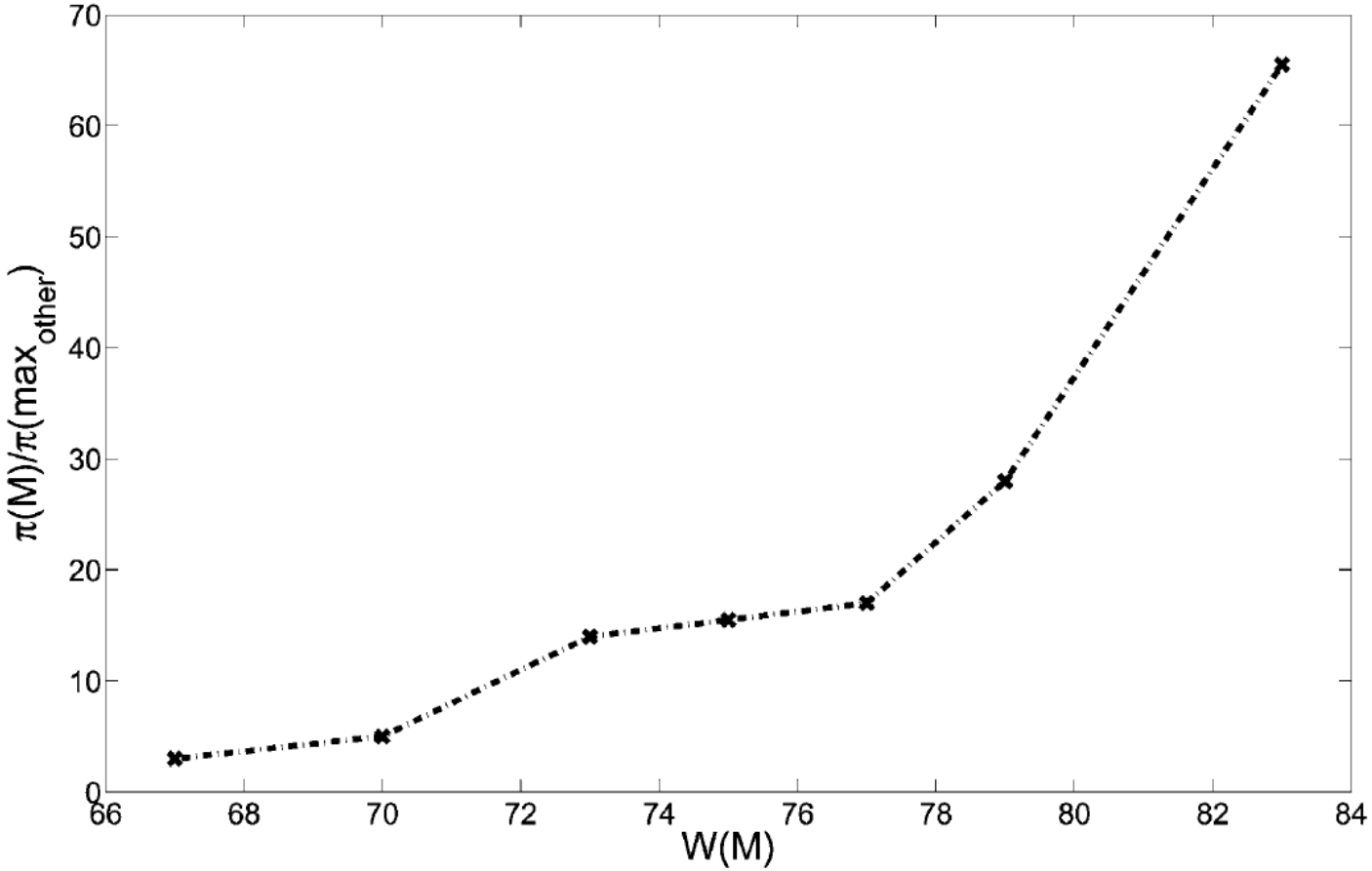


Figure 2. Ratio between the sampled frequency $\pi(M)$ of the maximum weight set, and the maximum frequency $\pi(\max_{other})$ of any other set in the sample for different values of $W(M)$.

Simulation study

Tests for the multiple high weight sets

Table 1. MCMC results on simulated data

	M_1	M_2	$\max_{other}$	avg_{other}
$\tilde{\pi}(\cdot)$	24.5	8.6	0.9	1.6×10^{-4}
$W(\cdot)$	80	78	73	56

$\tilde{\pi}(M_i)$ is the frequency of M_i, $\tilde{\pi}(\max_{other})$ is the maximum frequency with which a set different from M_1 and M_2 is sampled, and $\tilde{\pi}(avg_{other})$ is the average frequency with which a set different from M_1 and M_2 is sampled.

Simulation study: scalability

- 20000 genes vs. 1000 patients

- Based on these results, these algorithms should be useful on whole-exome sequencing studies with a modest number of patients.

Results on cancer mutation data

- The MCMC algorithm was applied to
  - somatic mutations from highthroughput genotyping of 238 oncogenes in 1000 patients of 17 cancer types (Thomas et al. 2007)
  - somatic mutations identified in recent cancer sequencing studies from lung adenocarcinoma (Ding et al. 2008)
  - glioblastoma multiforme (The Cancer Genome Atlas Research Network 2008).
- In the glioblastoma multiforme analysis, we include both **copy-number aberrations** and **single-nucleotide** (or small indel) mutations, while in the lung adenocarcinoma analysis, we consider only single-nucleotide (or small indel) mutations.

- The MCMC algorithm samples sets with frequency proportional to their weights, and thus to restrict attention to sets with high weight, we report sets whose frequency is at least 1%. We also reduce the size of the mutation matrix by combining genes that are mutated in exactly the same patients into larger ‘‘metagenes.’’

Results: multiple cancer types

- Mutation matrix with 298 patients vs. 18 mutation groups.
- They identified a set of eight mutation groups
- There are many sets of size 10 that contain the set of size 8 above. (how to interpret?)

Permutation test to assess the significance of the results: The statistic W(M) vs. the nun distribution by permutated the mutations of each mutation group in M.

Results: multiple cancer types

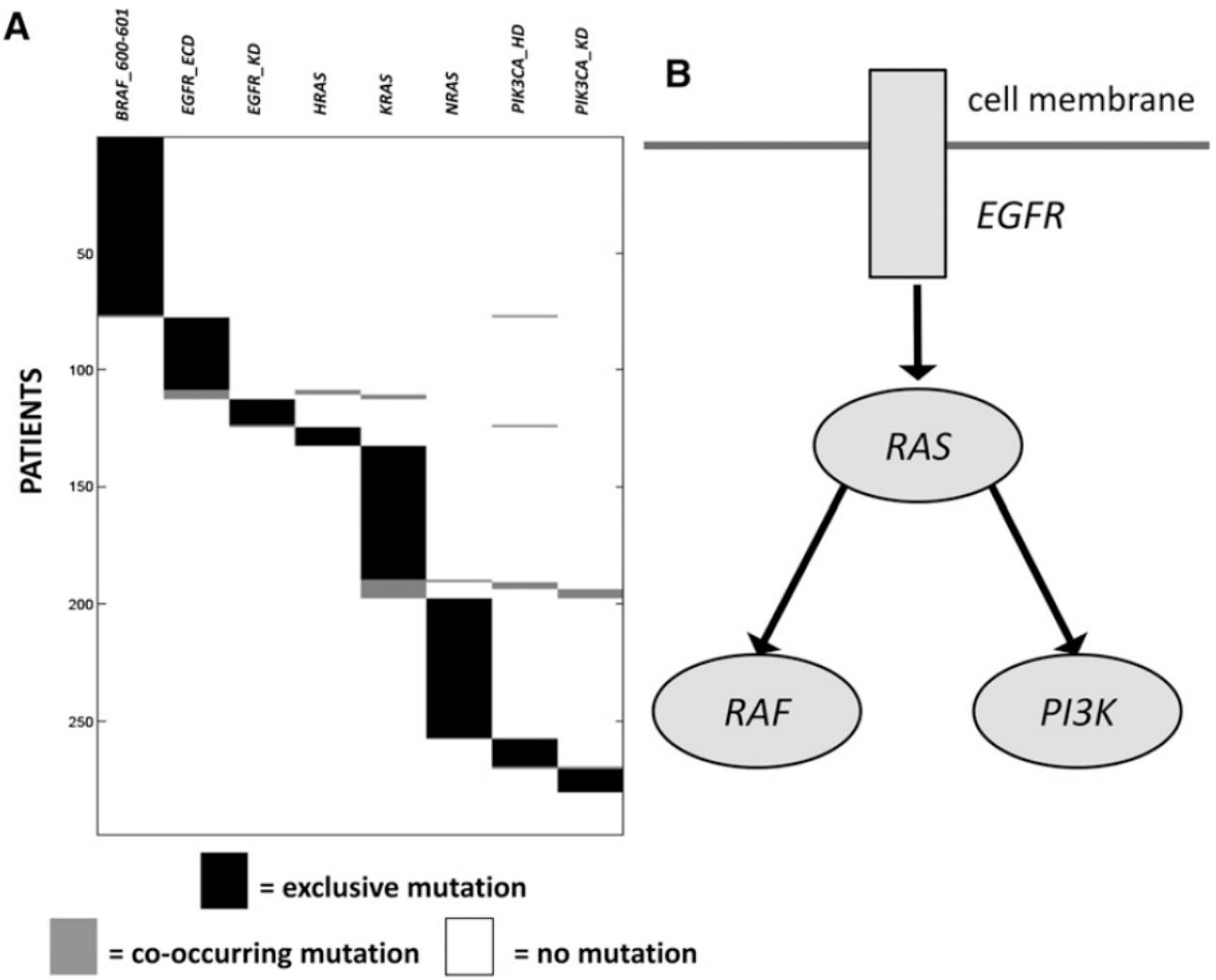


Figure 3. (*A*) High weight submatrix of eight genes in the somatic mutations data from multiple cancer types (Thomas et al. 2007). (Black bars) Exclusive mutations; (gray bars) co-occurring mutations. (*B*) Location of identified genes in known pathway. Interactions in the pathway are as reported in Ding et al. (2008).

Results: lung adenocarcinoma

- Mutation matrix with 188 patients vs. 356 mutation genes.
- (EGFR, KRAS) 99% of the time
  in 90 patients with a coverage overlap w(M) = 0
- (EGFR, KRAS, STK11) is sampled with frequency 8.4%

- The pairs (EGFR, KRAS) and (EGFR, STK11) were reported by Ding et al. (2008). But not (KRAS, STK11)
- (EGFR, KRAS, STK11) is a novel discovery with p-value=0.005

- The three genes EGFR, KRAS, and STK11 are all involved in the regulation of mTOR (Fig. 4)

Results: lung adenocarcinoma

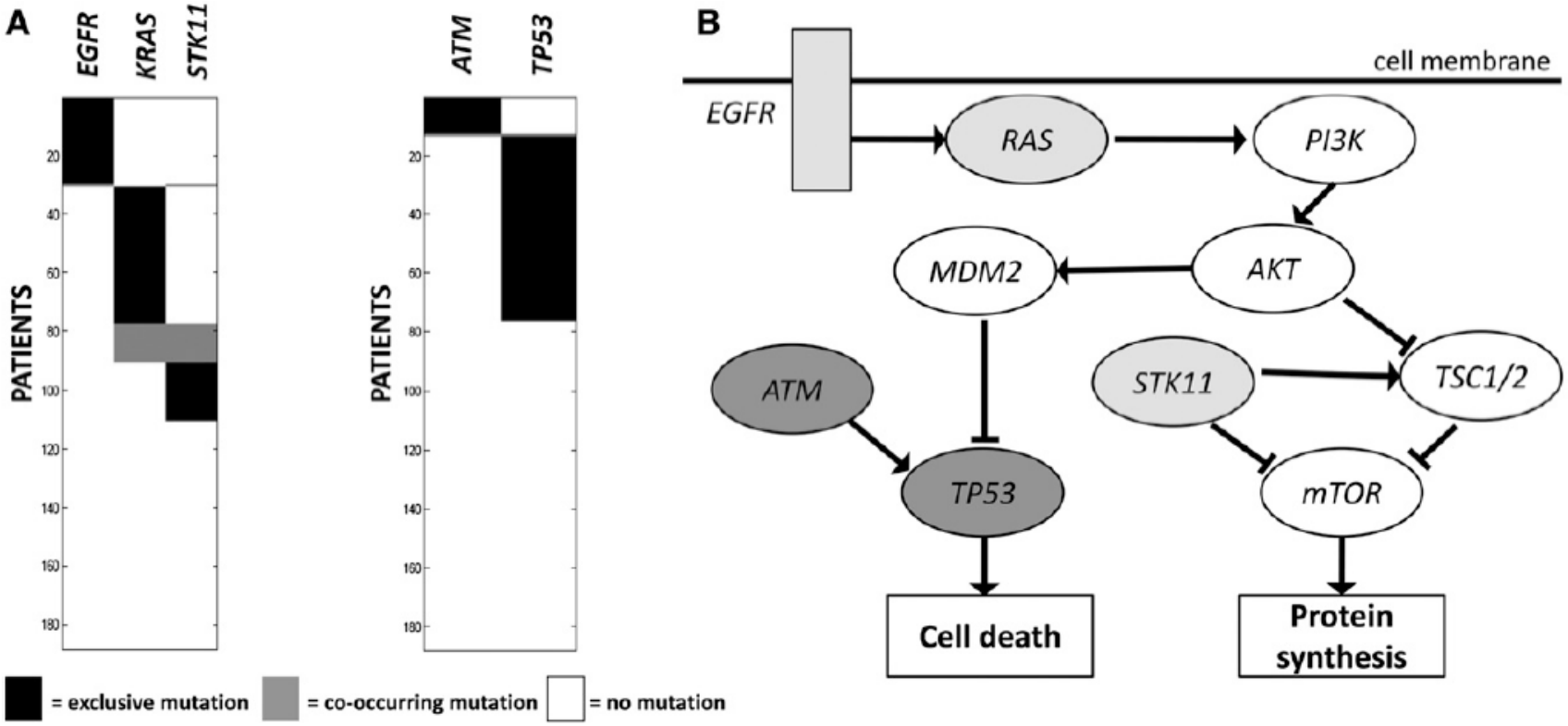


Figure 4. (*A*) High weight submatrices of two and three genes in the lung adenocarcinoma data. (Black bars) Exclusive mutations; (gray bars) co-occurring mutations. Rows (patients) are ordered differently for each submatrix, to illustrate exclusivity and co-occurrence. (*B*) The location of gene sets in known pathways reveals that the triplet of genes codes for proteins in the mTOR signaling pathway (light gray nodes), and the pair (*ATM*, *TP53*) corresponds to interacting proteins in the cell cycle pathway (dark gray nodes). Interactions in the pathway are as reported in Ding et al. (2008).

Results: lung adenocarcinoma

- To identify additional gene sets, we removed the genes EGFR, KRAS, STK11 and ran the MCMC algorithm again on the remaining genes. We sample the pair (ATM, TP53) with frequency 56%, and compute that the weight of the pair is significant ($p < 0.01$).
- both genes are involved in the cell cycle checkpoint control。

- Although the exclusivity of both sets is high, their coverage is low (<60%), suggesting that these gene sets are not complete driver pathways.

Results: Glioblastoma multiforme

- Mutation matrix with 84 patients vs. 601 mutation genes (SN mutation and CNA).
- (CDKN2B, CYP27B1) 18%
- (CDKN2B, a metagene) 10%

- (CDKN2B, RB1, CYP27B1) 10%
- (CDKN2B, RB1, a metagene) 6%

Results: Glioblastoma multiforme

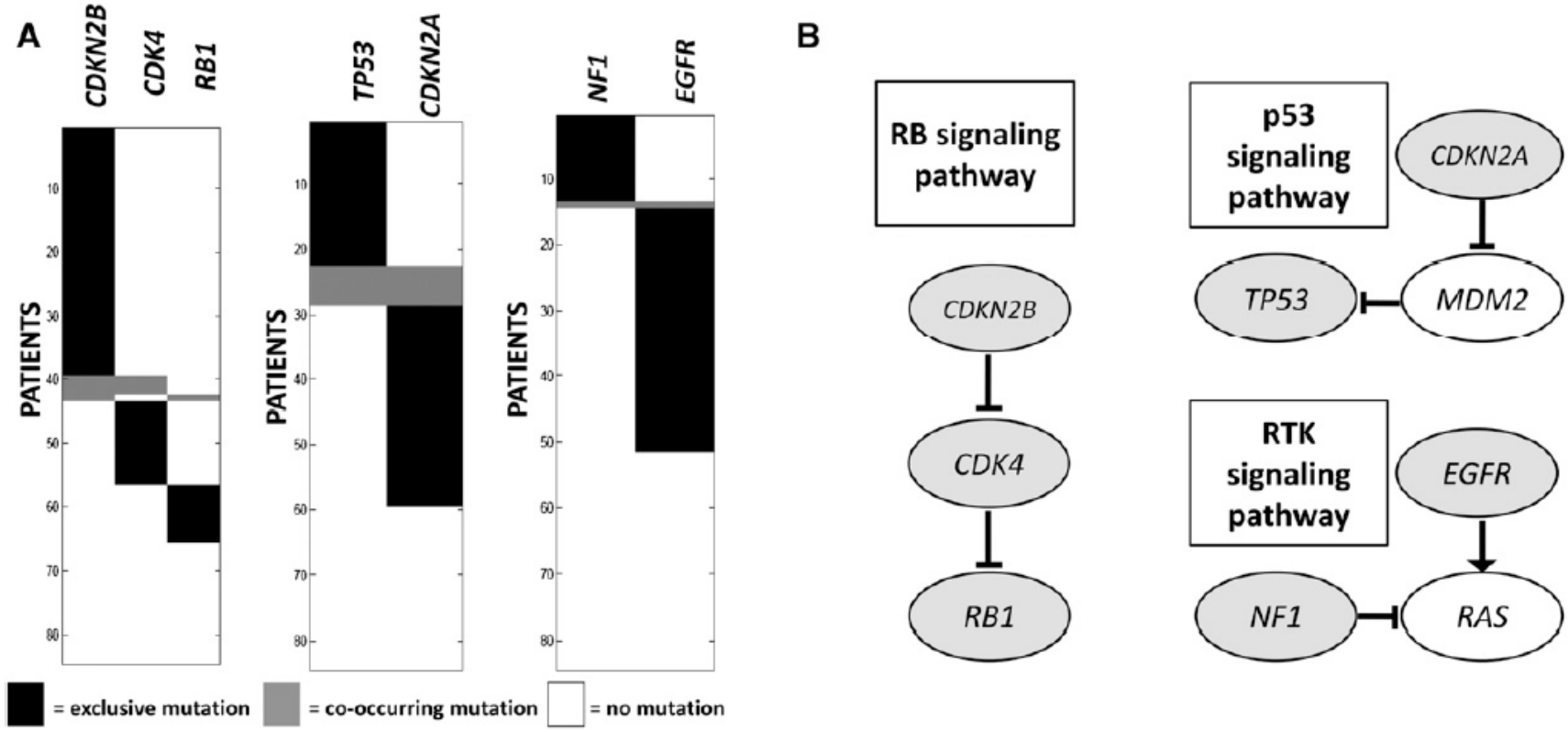


Figure 5. (*A*) High weight submatrices of two and three genes in the glioblastoma data. (Black bars) Exclusive mutations; (gray bars) co-occurring mutations. Rows (patients) are ordered differently for each submatrix, to illustrate exclusivity and co-occurrence. (*B*) Location of identified genes in known pathways. Interactions in pathways are as reported in The Cancer Genome Atlas Research Network (2008).

Results: Glioblastoma multiforme

- The pair (TP53, CDKN2A) is sampled with frequency 30% ($p < 0.01$).
- the pair (NF1, EGFR) sampled with frequency 44% ($p < 0.01$).

Discussion

- The proposed algorithms find sets of genes that are mutated in many samples **(high coverage)** and that are rarely mutated together in the same patient **(high exclusivity).**
- Notably, they discover these pathways de novo from the mutation data without any prior biological knowledge of pathways or interactions between genes.
- **However**, it is also important to note that some of the genes that were measured in these data sets were selected because they were known to have a cancer phenotype, and **thus there is some ascertainment bias in the finding that individual genes** (or groups of genes) are mutated in many samples.

Discussion

- However, in the lung adenocarcinoma and glioblastoma data, **the size of gene sets that we identify is relatively modest.**
- In addition, considering mutation data at the level of individual genes might reduce the power to distinguish driver mutations from passenger mutations.

- **There remain challenges in the identification of somatic mutations** from these data with the incorrect prediction of somatic mutations (false positives) and the failure to identify genuine mutations (false negatives) (Meyerson et al. 2010)

Discussion

- One particular source of false negatives is the **heterogeneity of many tumor samples**, which often include both **normal cell admixture** and **subpopulations of tumor cells** with potentially different sets of mutations.

Discussion and future study

- First, we could include additional information in the scoring of mutations and gene sets.
  - Extending our techniques to use additional information about the functional impact, or expression status, of each mutation is an interesting open problem.
- Second, alternative weight functions W(M) could be considered.
- Finally, the performance of our algorithm in complex situations involving multiple, overlapping high weight sets of genes requires further analysis.
  - It is not yet clear whether such complex situations arise in cancer mutation data.

Discussion and future study

- **What can be done next?**
- **Free discussion**
- **Further reading**

Boca *et al. Genome Biology* 2010, **11**:R112
http://genomebiology.com/2010/11/11/R112

Genome Biology

METHOD Open Access

Patient-oriented gene set analysis for cancer mutation data

Simina M Boca[1], Kenneth W Kinzler[2], Victor E Velculescu[2], Bert Vogelstein[2], Giovanni Parmigiani[3*]