Gene Mutations and Different Computational Techniques to Identify Driver Genes

Amandeep Kumar* and Damanpreet Singh**
*Research Scholar,Department of Computer Science & Engineering, Punjab Technical University, Jalandhar
aman.sp86@gmail.com
**Associate Professor, Department of Computer Science &Engineering, SLIET Longowal, Sangrur
damanpreetsingh@sliet.ac.in

Abstract: Cancer genomics research aim at identifying cancer-related genes and their contribution to cancer initiation and development. With the rapid development of high-throughput sequencing computational techniques, huge volume of cancer genomics data have been generated. Understanding this data poses great challenges to computational biologists. One of such key challenges is to distinguish driver mutations from passenger mutations. Distinguishing driver mutations which contribute to cancer development, from passenger mutationsthat have accumulated in somatic cells but without functional consequences is main aim of computational cancer genomics.In this article, we aim to review the recent development of computational models and algorithms for discovering driver genes or modules in cancer.

Keywords: Cancer, Mutations, Driver genes, Bioinformatics, Tools

Introduction

Cancers develop as a result of alterations that have occurred in the Deoxyribonucleic acid (DNA) sequence of the genomes of cancer cells. Over the past quarter of a century much has been learnt about these alterations and the abnormal genes that operate in human cancers[1].Understanding the mechanism of carcinogenesis has been a great challenge for human. With the rapid advance in deep sequencing technologies, many large-scale cancer projects have generated huge amount of cancer genomics data (e.g., The Cancer Genome Atlas(TCGA) , the International Cancer Genome Consortium (ICGC) , the Cancer Cell Line Encyclopedia(CCLE) , and the Therapeutically Applicable Research to Generate Effective Treatments (TARGET))[2].

All cancers arise as a result of somatically acquired alterations in the DNA sequence of genome of cancer cells. It does not mean, however, that all the somatic alterations present in a cancer genome have been involved in development of the cancer [3]. All cancers begin when one or more genes in a cell are mutated, or changed. This creates an abnormal protein or no protein at all. An altered protein provides different information than a normal protein, which can cause cells to multiply uncontrollably and become cancerous[1]. Genes and chromosomes can mutate in either somatic or germinal tissue, and these changes are called somatic and germinal mutations, respectively. Many genetic changes in the genomes of somatic cells initiate and promote tumor growth and cancer genomics researchers are now aiming at detecting all of these cancer driver mutations[4]. Driver genes have been defined as those for which the non-silent mutation rate is significantly greater than a background mutation rate(BMR) estimated from silent mutations[5]. The number of driver mutations, and hence the number of abnormal cancer genes, in an individual cancer is a main conceptual factor of cancer development, but is not well established. It is highly likely that most cancers carry more than one driver and that the number varies between cancer types. However, it is difficult to distinguish driver mutations from a great number of passenger mutations because of mutational heterogeneity, which is the key factor to deal with the problem of cancer treatment. Finding these important somatic mutation or driver mutation is of great benefit to the gene therapy of cancer patients [6].

The detection of driver mutations from passenger mutations and germinal polymorphisms usually starts with sequencing of matched tumor and normal DNA samples from cohorts of cancer patients. Identified after comparing sequences against the human reference genome are somatic mutations, those present in only the tumor samples, and germiline mutations, those present in the tumor and the matched normal samples. A crucial next step is to prioritize the somatic mutations and find driver mutations that are responsible for cancer development. Over the last three decades, many analytical tools have been developed to help predicting the relationships between somatic mutations and cancer phenotypes[1].Mutation rates can vary frequently within a cancer type, often owing to the degree of exposure to an environmental mutagen, or dependent on the particular genes that are altered (for example, tumours with mutations in mismatch repair genes will have higher mutation rates). Second, the mutation spectra also vary within cancer types [5].The rates of different mutational processes vary among tumors and cancer types. Though numbers vary widely, most cancers carry 1000 to 20,000 somatic point mutations and a few to hundreds of insertions, deletions, and rearrangements. Pediatric brain tumors and leukemia typically have the lowest

numbers of mutations, whereas tumors induced by exposure to mutagens, such as lung cancers (tobacco) or skin cancers (UV rays), present the highest rates. Although these are common figures, some cancers acquire dramatically increased mutation rates due to the loss of repair pathways or chromosome integrity checkpoints[7].

Types of Genomic Alterations in Somatic Cells

The biology of cancer is driven by the following type of mutations:

- single nucleotide variants (SNVs),
- small insertions and deletions (indels),
- copy number alterations (CNAs),
- fusion genes,
- chromosomal/structural rearrangements,
- epigenetic reprogramming[8]

SNVs are sequence alterations that involve a single nucleotide and they are the most abundant variants observed in sequencing data. Synonymous SNVs (sSNVs) do not change protein sequences, whereas nonsynonymous single nucleotide variants (nsSNVs) change protein sequences[1].

Indels usually refer to insertions or deletions of short (1bp to 50 bp) nucleotide sequences in a genome.

Base Insertion - new bases have been inserted into the sequence, e.g.:

CTGGAG --> CTGGTGGAG

Here, an extra nucleotide is added to the sequence. This could happen when a strand “wrinkles”, allowing room for an extra nucleotide.

Base deletion - a section (one ore more bases) of the sequence is lost:

CTGGAG -> CTAG

As with an insertion, a “wrinkle” in the DNA strand can cause one or more nucleotides to be skipped during DNA replication [16].

DNA **copy number alterations** (CNAs) are an important component of genetic variation, affecting a greater fraction of the genome than single nucleotide polymorphisms (SNPs).CNAs are structurally variant regions in which copy number differences have been observed between two or more genomes [9].CNAs are gains or losses of DNA segments, which can result in non-diploid copies of DNA segments in a genome. There are many ways to classify copy number alterations, depending on their sizes and types of alterations. For instance, aneuploidies usually refer to losses or gains affecting whole chromosomes; whereas, CNAs usually refer to alteration of segments between 1kbp and 1Mbp in length[10].

Chromosomal rearrangements refer to gross changes in the structure of a chromosome due to duplication (increase of the number of copies of a chromosomal region), inversion (partial rotation of a chromosomal segment), deletion, translocation (one part of a chromosome attaches to another chromosome) and transpositions (short DNA segments moves from one position to the next position).

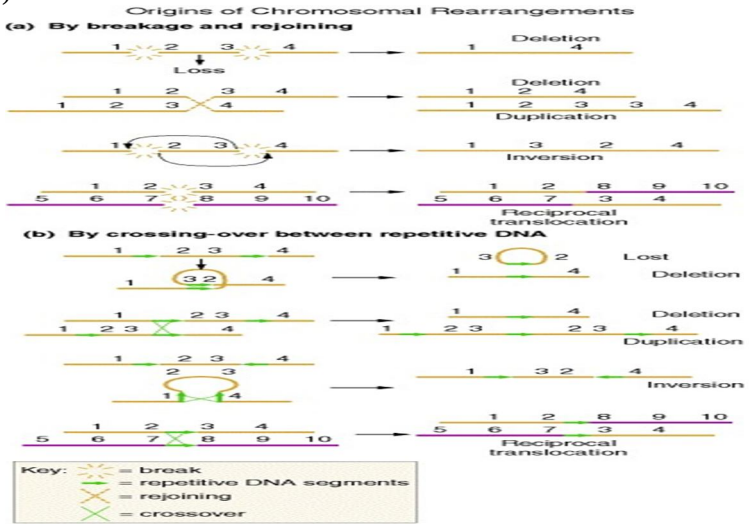

Fig 1. Chromosomal rearrangements[16]

A fusion genecan be introduced either by inter-chromosomal rearrangements, which combine two or multiple genes from different chromosomes into one fused gene, or by intra-chromosomal rearrangements such as deletion, inversion or duplication of large DNA segments on the same chromosomes. To date, many algorithms have been developed to enable the discovery of not only nsSNVs, but also impaired genes and pathways associated with cancer initiation, progression and development.

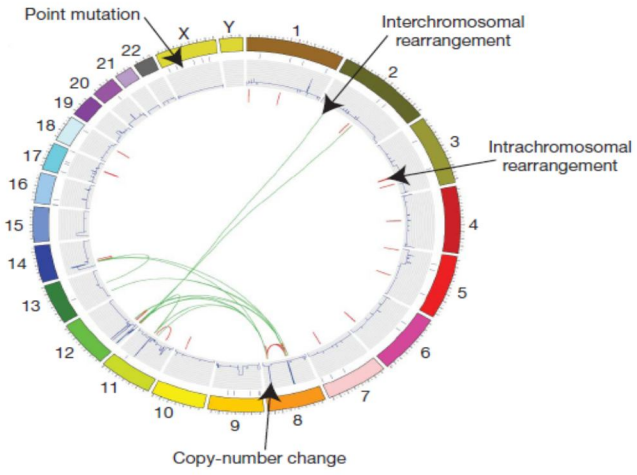

Figure 2. Somatic mutations in a single cancer genome.[3]

Review of Computational Techniques to Identifying Driver Genes

A common type of methods for identifying driver genes is based on gene mutational frequency (frequency based methods) [2].The basic principle of these methods is to test individual genes whether they are mutated in a significant number of cancer patients than expected by chance.A key step is to estimate the background mutation rate (BMR) to quantify the accumulation of random passenger mutations. Proper estimation of BMR is a key factor affecting the power of this type of methods. An overestimation of BMR fails to identify true recurrent mutations (false negatives), whereas an underestimation would lead to too many false positives. Early frequency-based methods assume a single constant background rate across the genome for all samples [2]. However, recent studies demonstrate that BMR is not constant across the genome [11].Moreover, a number of features (other than mutation frequency) could affect the mutation rate including mutation types, sequence context, gene-specific features mutation specific scores that assess functional impact and so on.

Several approaches towards the identification of driver genes have been implemented through computer software. With mutation data given as input these tools are able to return with a list ofgenes identified as drivers[12].

Therefore, recent studies have developed a number of frequency-based methods which adopt one or more of these features to get a more accurate BMR estimation. For example, both MuSiC[13] and MutSigCV[12] employ the mutation types and sample-specific mutation rates. MutSigCV also allows for the inclusion of gene-specific features such as the expression level and replication timing.. Several methods adopt these new features with improved sensitivity and specificity [13]. OncodriveCLUST uses the evidence of positional clustering to identify oncogenes [11]. MADGiC (Modelbased Approach for identifying Driver Genes in Cancer) is an unified empirical Bayesian model-based approach which uses all the above features to identify driver genes [14].

Genes do not work isolated but they interact through complex cellular reactions whose normal dynamics are altered in cancer. Based on these interactions, they are organized in groups, often called pathways[4].

MuSiC (Mutational Significance in Cancer)

Massively parallel sequencing technique and the associated rapidly decreasing sequencing costs have enabled systemic analyses of mutations in huge cohorts of cancer cases.

The main aim of MuSiC is to separate the significant events which are likely drivers mutations for disease from the passenger mutations present in mutational discovery sets using a variety of statistical methods[13].

MuSiC currently consists of seven analysis modules and an eighth execution module, ‘‘MuSiC Play,’’ which runs each analysis module sequentially. MuSiC Play parses the input and output of each modules and then produces a composite summary of all executed modules. Table lists the type of analysis performed and the types of variants included by each individual MuSiC module. Descriptions of the specific analysis algorithms performed by each module are given below[13].

Table 1: Analyses performed and the variants included for each MuSiC module[13]

MuSiC module	Analysis type	Variants included
SMG test	Statistical test	Optional
PathScan	Statistical test	Optional
Mutation relation test	Statistical test	Optional
Clinical correlation test	Statistical test	Optional
Proximity Analysis	Mathematical query	Optional
COSMIC/OMIM Analysis	Database query	Optional
Pfam Annotation	Database query	All

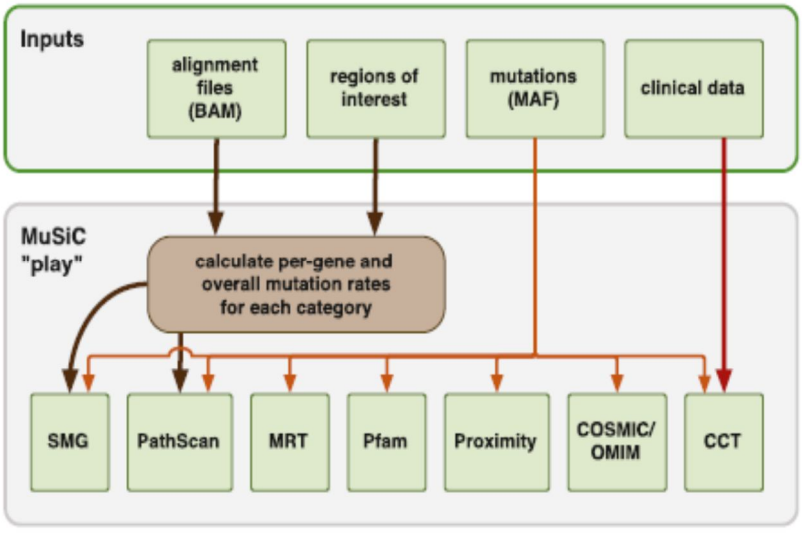

Figure 3. MuSiC flow diagram [13]

MutSigCV (Mutation Significance Covariant)

MutSiganalyzes mutations discovered in DNA sequences, to detect genes that were mutated more often than expected by chance given background mutation processes.

MutSigCV starts from the observation that the data is very sparse, and there are usually few silent mutations in a gene for its BMR to be estimated with any confidence. MutSigCV improves the BMR estimatation by pooling data from 'neighbor' genes in covariate space. These neighbor genes are selected on the basis of having similar genomic features to the central gene: features such as DNA replication time, chromatin state, and general level of transcription activity. These genomic factors have been observed to strongly correlate (co-vary) with background mutation rate[11].

MADGiC (Model based Approach for Identifying Driver Genes in Cancer)

MADGiC is an integrative model that inculcate posterior probabilities for improved inference for driver gene uniquely detect. The empirical Bayesian framework provides a natural way to incorporate different features together that were previously considered in isolation. In order to evaluate the utility of incorporating functional impact scores in the model, as well as assess what could be gained with a score that was better able to distinguish between passenger and driver mutations, MADGiC was evaluated under three different functional impact profiles: (i) ignoring functional impact, (ii) realistic impact—SIFT score profiles and (iii) high impact—passenger scores drawn from Beta(1,1.5) and driver scores set equal to one[14].

OncodriveCLUST

OncodriveCLUST Uses the binomial cumulated distribution from nsSNVs, stop-gain SNVs, and splice site mutations and coding silent mutations to calculate gene clustering score.

OncodriveCLUST, a computational technique to identify genes with a significant bias towards mutation clustering within the protein sequence[1]. This technique generates the background model by assessing coding-silent alterations, which are assumed not to be under positive selection and thus may reflect the baseline tendency of mutations to be clustered. OncodriveCLUST analysis of the Catalogue of Somatic Mutations in Cancer retrieve different genes enriched by the Cancer Gene Census, prioritizing those with dominant phenotypes but also highlighting some recessive cancer genes, which shows wider but delimited mutation clusters. Assessment of datasets from The Cancer Genome Atlas demonstrate that

OncodriveCLUST selected cancer genes that were nevertheless missed by methods based on frequency and functional impact criteria.
The OncodriveCLUST method consists five steps, First, single-nucleotide protein-affecting mutations (i.e. non-synonymous, stop and splice site mutations) are retrieved of each gene across a cohort of tumors are evaluated looking for those protein residues having a number of mutations barely expected by chance . Second, these positions are thereafter grouped to form mutation clusters and are identified as potentially meaningful cluster seeds . Third, these positions are grouped to form clusters, each cluster is scored with a figure proportional to the percentage of the gene mutations that are enclosed within that cluster and inversely related to its length joining positions. Fourth, once these clusters are obtained, they are completed by including the positions within or adjacent to each cluster that contains mutations in addition to those considered in the second step. Finally, a score is computed for each cluster with the background model to obtain a significance value. Background model is obtained performing the same steps than above but assessing only coding silent mutations. [15].

Table 2: Summary of features of methods to identify driver genes[14]

Methods	Mutation Type	Frequency	Gene-Specific Background	Functional Impact	Spatial Patterning
MADGiC	✓	✓	✓	✓	✓
MuSiC	✓	✓			
MutSigCV	✓	✓	✓		
OncodriveCLUST	✓				✓

Discussion

In summary, the explanation of genes involved in cancer is a challenging task that requires the combined use of approaches based on different criteria. In this regard, we show that OncodriveCLUST complements well other existing methods and should be taken into account for the identification of cancer drivers. Moreover, Genes do not work isolated but they interact through complex cellular reactions whose normal dynamics are altered in cancer. Based on these interactions, they are organized in groups, often called pathways

Conclusion

Cancer is a complex disease and different computational methods may discover different drivers at different levels (e.g. mutation-level, gene-level and module-level). Due to lack of golden standard approach, their performance varies by dataset, which make them incomparable because of nonreproductive results. In this review we have survey the different computational techniques to identify the driver genes on the basis of different features. These demonstrat that BMR is not constant across the genome. Moreover, a number of features (other than mutation frequency) could affect the mutation rate including mutation types, sequence context , gene-specific features mutation specific scores that assess functional impact and so on. Many mutations occur in different genes among different patients. Such mutational heterogeneity in cancer genomes is another important factor affecting the performance of frequency-based methods. This heterogeneity may be a consequence of the presence of passenger mutations in each cancer genome. To ease this process, the development of better computational algorithms is in urgent need in order to overcome some of weakness of the current methods.

References

[1] B. Djotsa Nono, K. Chen, and X. Liu, “Computational Prediction of Genetic Drivers in Cancer,” no. February, 2016.
[2] J. Zhang and S. Zhang, “The Discovery of Mutated Driver Pathways in Cancer : Models and Algorithms,” pp. 1–11, 2016.
[3] M. R. Stratton, P. J. Campbell, and P. Andrew F, “The cancer genome,” Nature, vol. 458, no. 7239, pp. 719–724, 2009.
[4] M. Dimitrakopoulos and N. Beerenwinkel, “Computational approaches for the identification of cancer genes and pathways.,” Wiley Interdiscip. Rev. Syst. Biol. Med., vol. 9, no. February, pp. 1–18, 2016.
[5] Hu and S. Wang, “A Robust Method for Identifying Mutated Driver Pathways in Cancer.”
[6] S. Wang and C. Hu, “Multi-population Genetic Algorithm for Identifying Mutated Driver Pathways in Cancer,” vol. 9, no. 4, pp. 291–304, 2016.
[7] Martincorena and P. J. Campbell, “Somatic mutation in cancer and normal cells,” Science (80-.)., vol. 349, no. 6255, pp. 1483–1489, 2015.
[8] P. Futreal et al., “A census of human cancer genes,” Nat Rev Cancer, vol. 4, no. 3, pp. 177–183, 2004.
[9] Shlien and D. Malkin, “Copy number variations and cancer.,” Genome Med., vol. 1, no. 6, p. 62, 2009.
[10] L. A. Loeb, K. R. Loeb, and J. P. Anderson, “Multiple mutations and cancer.,” Proc. Natl. Acad. Sci. U. S. A., vol. 100, no. 3, pp. 776–81, 2003.
[11] M. S. Lawrence et al., “Mutational heterogeneity in cancer and the search for new cancer-associated genes.,” Nature, vol. 499, no. 7457, pp. 214–8, 2013.

[12] Tokheim, N. Papadopoulis, K. W. Kinzler, B. Vogelstein, and R. Karchin, “Evaluating the Evaluation of Cancer Driver Genes,” bioRxiv, vol. 113, no. 50, p. 60426, 2016.
[13] N. D. Dees et al., “MuSiC: Identifying mutational significance in cancer genomes,” Genome Res., vol. 22, no. 8, pp. 1589–1598, 2012.
[14] K. D. Korthauer and C. Kendziorski, “MADGiC: A model-based approach for identifying driver genes in cancer,” Bioinformatics, vol. 31, no. 10, pp. 1526–1535, 2014.
[15] Tamborero, A. Gonzalez-Perez, and N. Lopez-Bigas, “OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes,” Bioinformatics, vol. 29, no. 18, pp. 2238–2244, 2013.
[16] Anthony JF Griffiths, William M Gelbart, Jeffrey H Miller, and Richard C Lewontin,"Chromosome Mutations" in Modern Genetic Analysis, New York, W.H. Freeman,1999, Ch. 8,pp 97-110