1-10-2008

DESIGN AND ANALYSIS ISSUES IN GENOME-WIDE SOMATIC MUTATION STUDIES OF CANCER

Giovanni Parmigiani
The Sydney Kimmel Comprehensive Cancer Center, Johns Hopkins University & Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, gp@jimmy.harvard.edu

Simina Boca
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Jimmy Lin
The Johns Hopkins University, School of Medicine

Kenneth W. Kinzler
The Johns Hopkins University, School of Medicine, Oncology Center

Victor E. Velculescu
The Johns Hopkins University, School of Medicine, Oncology Center

See next page for additional authors

Authors
Giovanni Parmigiani, Simina Boca, Jimmy Lin, Kenneth W. Kinzler, Victor E. Velculescu, and Bert Vogelstein

Design and analysis issues in genome-wide somatic mutation studies of cancer.

Giovanni Parmigiani, Simina Boca, Jimmy Lin,
Kenneth W Kinzler, Victor Velculescu, Bert Vogelstein

Ludwig Center for Cancer Genetics and Therapeutics,
and The Howard Hughes Medical Institute
at The Johns Hopkins Kimmel Cancer Center, Baltimore, MD 21231, USA

Abstract

The availability of the human genome sequence and progress in sequencing and bioinformatic technologies have enabled genome-wide investigation of somatic mutations in human cancers. This article briefly reviews challenges arising in the statistical analysis of mutational data of this kind. A first challenge is that of designing studies that efficiently allocate sequencing resources. We show that this can be addressed by two-stage designs, and demonstrate via simulations that even relatively small studies can produce lists of candidate cancer genes that are highly informative for future research efforts. A second challenge is to distinguish mutated genes that are selected for by cancer (drivers) from mutated genes that have no role in the development of cancer and simply happened to mutate (passengers). We suggest that this question is best approached as a classification problem and discuss some of the difficulties of more traditional testing-based approaches. A third challenge is to identify biologic processes affected by the driver genes. This can be achieved by gene set analyses. These can reliably identify functional groups and pathways that are enriched for mutated genes even when the individual genes involved in those pathways or sets are not mutated at sufficient frequencies to provide conclusive evidence as drivers.

Key words: Mutation analysis, Cancer genomics

1 Introduction

Genome-wide somatic mutation studies. The discovery of genes mutated in human cancers has provided key insights into the mechanisms underlying tumorigenesis and has proven useful for the design of targeted therapeutic approaches [1]. Recently, the availability of the human genome sequence and progress in sequencing and bioinformatic technologies have enabled genome-wide investigation of somatic mutations in human cancers [2,3]. These studies exemplify an emerging trend that includes other large- scale sequencing efforts of cancer genomes (ref. to Stratton). Analysis focuses on the comparison between the sequences found in tumor samples and those of the originating normal tissues. The goal of this comparison is to identify regions of the genome that differ frequently enough to warrant further investigation of potential causal mechanisms. So far this comparison has focused primarily on coding sequences of well-annotated genes.

Passengers and drivers. Cancer arises as the result of successive clonal expansions driven by cells that acquire a selective advantage through mutations. Generally, alterations are the result of errors that arise during the process of DNA replication during cellular expansion. These errors are associated with mistakes during DNA polymerization or with external agents, such as carcinogens, and may or may not provide a selective advantage to the affected cell. As a result, before it undergoes a new mutation that provides a selective advantage, a cell will typically accumulate other alterations that are neutral with respect to selection. Mutations that are disadvantageous may also occur, but these will selected against during tumorigenesis and will not be present in clonal expansions of tumor cells. Genome-wide somatic mutation studies will therefore identify two types of mutations: the "drivers" — those providing a selective advantage, and the "passengers" — those neutral to the selection process [4]. Genes capable of harboring driver mutations are referred to as "driver genes". Similarly, genes which are not driver genes are referred to as "passenger genes". One of the major goals of the analysis of data from genome-wide somatic mutation studies is the ranking of genes based on the likelihood that they may be drivers.

Mountains and hills. If one represents likely driver genes as relief features on a map, the resulting landscapes will contain a small number of major mountains, representing genes that are mutated in the majority of cancers, and a much larger number of hills representing the genes that are mutated at relatively low frequency. This general genomic landscape is a common feature of both breast and colorectal tumors [3]. So far, cancer research has focused on the gene mountains. The ability to analyze the sequence of virtually all protein-encoding genes in cancers has shown that the vast majority of mutations in cancers, including many that are highly likely to be drivers, do not occur

in such mountains, emphasizing the heterogeneity and complexity of human neoplasia. This new view of cancer is consistent with the idea that a large number of mutations, each associated with a small fitness advantage, drive tumor progression [5,6]. But while the number of potential driver genes is large, changes appear to occur in a more limited number of "driver" pathways [7,1,8–11].

This landscape has important implications for the statistical design and analysis of genome-wide studies of somatic mutations in cancer.

2 Design

Choice of tumor samples. The successive bottlenecks that characterize the evolution of a tumor are driven by mutations that tend to occur at different stages — mutations of certain pathways are typically important earlier on, while others are more likely to occur at a later stage. For example, in colorectal cancer, mutations that are associated with adenoma formation are typically different from those that contribute to the progression of those adenomas to carcinomas. Therefore, initial studies whose primary goal is to efficiently identify the largest number of drivers with the fewest samples have concentrated on advanced disease samples. Another important consideration has been the exclusion of samples with such widespread genetic alterations that the information provided is minimal. For example, in their genome-wide mutation analysis of colorectal cancer Sjöblom *et al.* [2] excluded samples with mismatch repair mutations. An important goal of future studies should be that of characterizing from an epidemiological standpoint the frequency of tumors wherein a given gene or pathway plays the role of a driver. This will require different designs, larger sample sizes, and consideration of the patient population to which the estimated frequencies should be applied.

Multi-stage Sampling. Despite recent progress in sequencing technologies, genome-wide analysis of somatic mutations in cancer remains a major undertaking. Time and cost considerations should be a factor in determining the scope of a study and the sample sizes. Significant gains in efficiency can be achieved by multi-stage approaches, in which an initial "discovery" phase is performed first on a genome-wide scale, followed by a "validation" phase in which genes that emerge as candidates from the discovery phase are evaluated in additional samples. For example Sjöblom *et al.* [2] performed a genome-wide analysis on all the genes in the CCDS database on 11 samples, followed by analysis of all the genes that were mutated at least once in 24 additional samples. Wood *et al.* [3] adopted a similar design, integrated by a third phase designed to provide a more accurate estimate of mutation frequencies on a yet smaller set of genes. In the Wood study about 4% of the genes in colon can-

cer, and about 5% in breast cancer were found to be mutated in the discovery stage and thus sequenced in the validation stage.

Simulation of Mutation Analysis Data. To facilitate design and analysis of mutation studies, we developed software to perform *in silico* experiments which exactly replicate the experimental procedure. They represent mutations found in a hypothetical genome with a known composition of driver and passenger genes. It is reasonable to assume that the likelihood of a random mutation will apply to individual nucleotides, and that the precise base that is mutated as well as its neighbors are important when evaluating the probability of a mutational event. We will refer to this as the mutation's "context". Therefore, each gene's probability of accumulating a random mutation will depend on its size and nucleotide composition. Also, in real experiments, if quality control criteria are applied to sequencing results, the number of nucleotides successfully sequenced is generally less than 100%. The actual fraction, or "coverage" should be a consideration.

In the simulation presented here we considered each gene in turn and simulated, for each nucleotide context, the number of mutations from a binomial distribution with success probability equal to either a) the context-specific passenger rate or b) a randomly selected rate, higher than the passenger rate. These rates were drawn from a distribution of mountains and hills that mimicked what was observed in real experiments. To generate mutations in driver genes, we used the empirically observed rates of the 160 genes found to be mutated in colorectal cancers in both the discovery and the validation screen of Wood *et al.* [3]. The number of available nucleotides in each context was based on the RefSeq database. For the binomial calculation, the gene sizes were adjusted using the proportion of nucleotides successfully sequenced in Wood *et al.* [3] for that particular gene and by the number of samples (K_d) available in the discovery screen. We then considered all genes for which at least one mutation was generated, and repeated the process with K_v samples to simulate the validation screen. The software used for the simulations presented here is available as part of a package called `CancerMutationAnalysis` [3]. Users can specify passenger and driver rates, sample sizes, gene sizes and composition, gene-specific counts of successfully sequenced nucleotides and other variables.

Sensitivity and Positive Predictive Values in one and two-stage designs. Using this tool, we assessed the tradeoffs associated with choosing the sample sizes in one- and two-stage studies. To concisely capture the effectiveness of a specific choice, we focus on the properties of lists composed of the top T most promising genes, where "most promising" is defined in terms of the likelihood ratio test [12] for the null hypothesis that the gene is mutated at the same rate as the passenger mutation rate. We report the sensitivity — the proportion of genes included in the top T among all drivers, and the Positive Predictive Value (PPV) — the proportion of drivers among the top T genes.

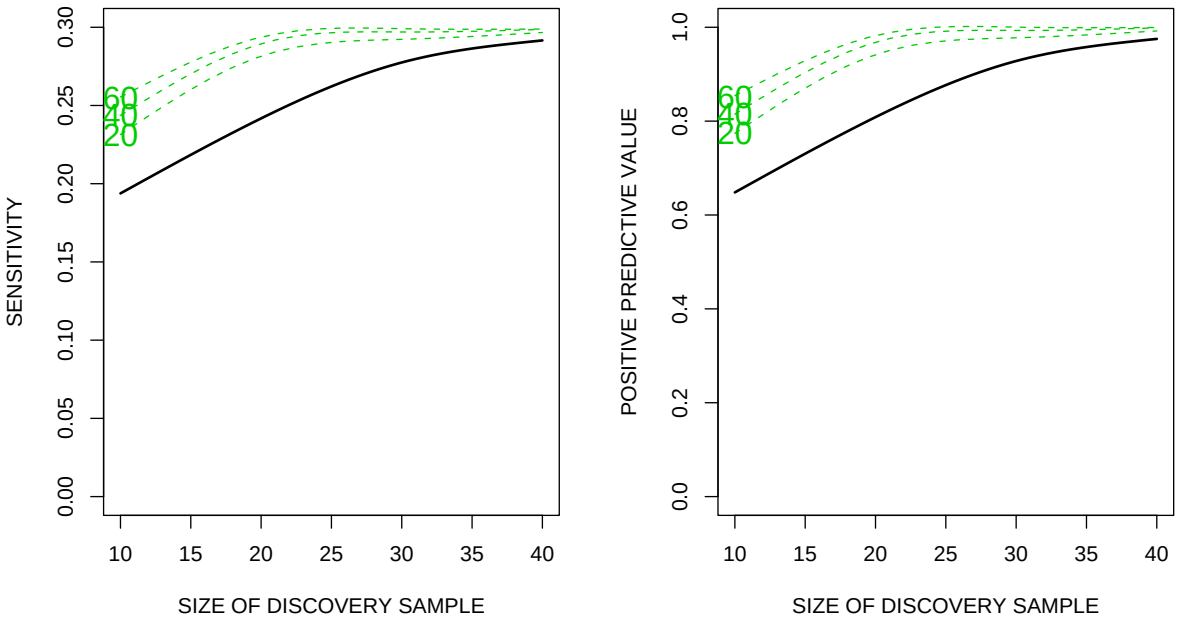

Fig. 1. Sensitivity (left) and Positive Predictive Value (right) of alternative sample sizes. The black continuous line represents a single-stage design, while the green dotted lines represent two-stage designs with different validation sample sizes (20, 40, or 60). Because we considered lists of the top 300 genes, and simulated data assuming that there are 1000 drivers in the genome, the sensitivity can be at most 0.3.

The sensitivity is related to statistical power, but it is not the same for two reasons: it is a probability across a set of genes, rather than a probability for a single gene over multiple experiments, and the driver mutations rates are allowed to vary.

Figure 1 illustrates results for lists composed of the top 300 genes. Data are simulated assuming that there are 1000 drivers in the genome. Both the sensitivity and the Positive Predictive Value reach their maximum values of 0.3 and 1.0 at relatively small sample sizes. Even a relatively small experiment with 10 discovery and 20 validation samples has a PPV in excess of 75% [Would it be worth change the scale on Fig 1 right so that Y axis is from 0 to 1.0? I initially did not realize the value on the Y axis started at 0.5 and I was led to think that the values were much lower]. When the list size is 150, similar to Wood *et al.*, the PPV of a study with 10 discovery and 20 validation samples is 98%.

The passenger rates used in Figure 1 correspond to the intermediate scenario of Wood *et al.* [3]. At higher rates the PPV and sensitivity are lower, though

even small studies remain informative. For example, the PPV of a study with 10 discovery and 20 validation samples remains around 70%. An important assumption in this analysis is that the passenger rates are the same across genes and samples. If these rates were actually to vary across genes, larger samples would be required to achieve similar performance.

Studies of optimal two-stage genotyping in population-based association studies using SNPs have suggested that two-stage designs halve the cost for a given power in that context [13]. The analyses presented here suggest that in mutation analysis studies the gains are likely to be significantly greater. For example, from Figure 1, the sensitivity of a study with 20 discovery samples and no validation (that is, a one-stage study) is comparable to that of a study with 10 discovery and 20 validation samples, while the sequencing effort involved is approximately 55% of the original effort, assuming that about 5% of the genes included in the discovery screen will be sequenced in the validation screen.

Finally, the distribution of driver rates used in the simulations presented in Figure 1 cover a broad range. We also examined the ability of a study to identify the larger hills or "major drivers". We assume there are 150 major drivers and their rates are drawn from the distribution of the top 20 candidate colorectal cancer genes in Wood *et al.* [3]. The sensitivity of the list of top 150 genes in a study with 10 discovery and 20 validation samples is 59% and the PPV is 58%. A sensitivity of 80% is achieved by studies with 30 discovery and 60 validation samples, and a sensitivity of 83% with 40 discovery and 60 validation samples.

3 Analysis

Goals. The overarching goal of data analysis in somatic mutation studies of cancer is to prioritize the research that follows. Two tasks are especially important: to provide quantitative measures useful for ranking the genes that are most worthy of further investigation, and to point to pathways or other gene classes whose analysis may reveal important mechanistic evidence or suggest therapeutic approaches. In this section we review statistical challenges related to these two tasks.

Passenger Mutation Rates. An important role in statistical analyses is played by the rate at which passenger mutations appear in cancer samples. This is a difficult quantity to estimate empirically, because the rate refers to hypothetical cell populations that underwent the same mutagenic exposures and clonal bottlenecks as a real cancer, but where those bottlenecks occurred for reasons other than selection. Wood *et al.* [3] approximated this situation by

studying portions of the genomes of cancers that are a priori highly unlikely to harbor regions whose mutation would provide an advantage. In this way they obtained a lower bound for the passenger rates. Independently, estimates of the passenger mutation rates were also obtained through the quantification of synonymous missense mutations. As the majority of synonymous changes are expected not to be selected for or against during tumorigenesis, such changes can be used as a tool to estimate passenger mutation rates. The analysis of synonymous mutations provided two estimates of the non-synonymous (NS) mutation rate. One estimate was based on the ratio of non-synonymous to synonymous mutations in the human germline determined from the HapMap project, and was considered to be a minimum because the ratio of non-synonymous to synonymous coding region mutations may be higher in the germline than in tumors due to greater negative selection for NS mutations in the germline. An additional estimate was derived by calculating the expected ratio of non-synonymous to synonymous changes after accounting for codon usage of RefSeq genes and the different mutation spectra observed in colorectal and breast cancers. This estimate was considered a maximum because it does not take into account the fact that nonsynonymous mutations that retard cell growth will be selected against during tumorigenesis. The fraction of such nonsynonymous alterations that retard cell growth may be quite large as studies in yeast suggest that alterations of up to 40% of protein coding genes can lead to quantitative growth defects [**?**].

Passenger rates vary considerably from tumor to tumor, undoubtedly determined by their intrinsic mutability and the number of generations and bottlenecks through which they have evolved.

Sorting Drivers from Passengers. To prioritize future studies it is useful to assign, to each of the genes in which mutations are found, a score that captures whether it is more plausibly a driver or a passenger. Statistically this question can be formulated, as a first approximation, as that of classifying genes by whether they have mutation rates higher than the passenger rate. A useful framework for this analysis is that of classification: in our case, classification of genes into passengers and drivers. Probabilistic classification is especially useful as it provides for each gene a probability of being a driver. Wood *et al.* [3], for example, use an Empirical Bayesian approach adapted from Efron *et al.* [14] to derive these probabilities. The key feature underlying this approach is the *in silico* generation of a study identical to the one performed, except that all mutations occur at the passenger rates, i.e., there are no driver genes. For each gene a score is then computed for both the observed and *in silico* data. The distribution of these scores in the real experiment is then thought of as a mixture of passengers' scores, drawn from the distribution generated *in silico*, and drivers' scores, drawn from a different and unknown distribution. For each gene, the probability of belonging to each of these two mixture components provides the classification probability.

Alternative approaches proceed by testing, for each gene, the null hypothesis that the mutation rate is the same as the passengers rate. One challenge in this context is to devise an appropriate multiple testing adjustment. Traditional frequentist approaches have serious limitations. Firstly, when data are collected in a two-stage approach only genes that harbored, say, at least one mutation in the discovery screen are analyzed in the validation screen. As a result of this, p-values are very computationally-intensive to evaluate. Second, p-values will be 1 for all genes in which no mutations are found. This makes it impossible to provide adjustments that account for the size and coverage of those genes, which constitute the vast majority. This can lead to an excessive multiple testing correction of the p-value calculations and an underestimate of the number of genes mutated at higher than passenger rates.

Gene sets. A third challenge is to identify biologic processes affected by the driver genes. To address this question one can examine their putative roles based on sequence similarity, membership in known functional groups and pathways, and potential interactions with other proteins. These analyses can reliably identify functional groups and pathways that are enriched for mutated genes, even when the individual genes involved in those pathways are not mutated sufficiently often to provide conclusive evidence. Statistically the goal is the evaluation of a set of genes as a single candidate "driver". A simple approach along these lines is to consider an entire Functional Gene Set as a pool of nucleotides at risk of somatic mutations and apply the same techniques used for individual genes directly to the whole pool [11]. This is a sensitive approach and is easy to implement. Possible drawbacks include an excessive emphasis on sets that include a single gene mutated at very high frequency, and lack of consideration of the sizes of the genes in which mutations were and were not found. Alternatively, genes can be sorted by a score that reflects the likelihood of being mutated at rates higher than the passenger rates, and a test used to compare the scores inside and outside of Functional Gene Sets, similarly to what was previously used for microarray expression analysis [15,16].

4 Conclusion

Above, we briefly reviewed the main analysis challenges arising in genome-wide studies of somatic mutations in cancer. We showed via simulations that even with relatively small sample sizes, two-stage designs can be highly informative for future studies, and briefly reviewed the lessons we learned about such analyses from the efforts of Sjöblom *et al.* [2] and Wood *et al.* [3].

In statistical analyses of mutation frequency alone, the "drivers" are equated to the genes that are mutated at higher frequencies than the passengers. This

is not the same as being is a true cancer gene. The former can be precisely defined and investigated using cancer genome sequencing studies. The latter, while interpretable in many ways, implies some additional independent validation of causality. We believe that sequencing data can, at best, only point to candidates worthy of further study.

The above lessons are equally applicable to studies employing the new generations of high throughput massively parallel sequencing technologies as they are to the classic Sanger sequencing methods that formed the bases for the current analysis. However, along with the promise that these new technologies offer, they also present unique challenges. For example, the above cited studies were only possible because of the efficient and proven strategies for eliminating technical false positives that have been developed over the 30 year history of Sanger sequencing. Similar strategies will have to be developed for these new approaches. Additionally, all of the leading new technologies rely on digital sequencing (i.e., single molecule sequencing) which will both simplify and complicate mutational analyses. Such digital approaches require a significant oversampling to ensure that both alleles of a diploid sample are assessed in order to avoid technical false negatives. At the same time, the digital nature and required oversampling of these new approaches may allow application of an unprecedented statistical rigor to the evaluation of sequencing data.

References

[1] B. Vogelstein, K. W. Kinzler, Cancer genes and the pathways they control., Nat Med 10 (8) (2004) 789–799.
URL http://dx.doi.org/10.1038/nm1087

[2] T. Sjöblom, S. Jones, L. D. Wood, D. W. Parsons, J. Lin, T. D. Barber, D. Mandelker, R. J. Leary, J. Ptak, N. Silliman, S. Szabo, P. Buckhaults, C. Farrell, P. Meeh, S. D. Markowitz, J. Willis, D. Dawson, J. K. V. Willson, A. F. Gazdar, J. Hartigan, L. Wu, C. Liu, G. Parmigiani, B. H. Park, K. E. Bachman, N. Papadopoulos, B. Vogelstein, K. W. Kinzler, V. E. Velculescu, The consensus coding sequences of human breast and colorectal cancers., Science 314 (5797) (2006) 268–274.
URL http://dx.doi.org/10.1126/science.1133427

[3] L. D. Wood, D. W. Parsons, S. Jones, J. Lin, T. Sjöblom, R. J. Leary, D. Shen, S. M. Boca, T. Barber, J. Ptak, N. Silliman, S. Szabo, Z. Dezso, V. Ustyanksky, T. Nikolskaya, Y. Nikolsky, R. Karchin, P. A. Wilson, J. S. Kaminker, Z. Zhang, R. Croshaw, J. Willis, D. Dawson, M. Shipitsin, J. K. V. Willson, S. Sukumar, K. Polyak, B. H. Park, C. L. Pethiyagoda, P. V. K. Pant, D. G. Ballinger, A. B. Sparks, J. Hartigan, D. R. Smith, E. Suh, N. Papadopoulos, P. Buckhaults, S. D. Markowitz, G. Parmigiani, K. W. Kinzler, V. E. Velculescu, B. Vogelstein, The genomic landscapes of human breast and colorectal cancers., Science 318 (5853) (2007) 1108–1113.
URL http://dx.doi.org/10.1126/science.1145720

[4] H. Davies, C. Hunter, R. Smith, P. Stephens, C. Greenman, G. Bignell, J. Teague, A. Butler, S. Edkins, C. Stevens, A. Parker, S. O'Meara, T. Avis, S. Barthorpe, L. Brackenbury, G. Buck, J. Clements, J. Cole, E. Dicks, K. Edwards, S. Forbes, M. Gorton, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, D. Jones, V. Kosmidou, R. Laman, R. Lugg, A. Menzies, J. Perry, R. Petty, K. Raine, R. Shepherd, A. Small, H. Solomon, Y. Stephens, C. Tofts, J. Varian, A. Webb, S. West, S. Widaa, A. Yates, F. Brasseur, C. S. Cooper, A. M. Flanagan, A. Green, M. Knowles, S. Y. Leung, L. H. J. Looijenga, B. Malkowicz, M. A. Pierotti, B. T. Teh, S. T. Yuen, S. R. Lakhani, D. F. Easton, B. L. Weber, P. Goldstraw, A. G. Nicholson, R. Wooster, M. R. Stratton, P. A. Futreal, Somatic mutations of the protein kinase gene family in human lung cancer., Cancer Res 65 (17) (2005) 7591–7595.
URL http://dx.doi.org/10.1158/0008-5472.CAN-05-1855

[5] R. K. Thomas, A. C. Baker, R. M. Debiasi, W. Winckler, T. Laframboise, W. M. Lin, M. Wang, W. Feng, T. Zander, L. MacConaill, L. E. Macconnaill, J. C. Lee, R. Nicoletti, C. Hatton, M. Goyette, L. Girard, K. Majmudar, L. Ziaugra, K.-K. Wong, S. Gabriel, R. Beroukhim, M. Peyton, J. Barretina, A. Dutt, C. Emery, H. Greulich, K. Shah, H. Sasaki, A. Gazdar, J. Minna, S. A. Armstrong, I. K. Mellinghoff, F. S. Hodi, G. Dranoff, P. S. Mischel, T. F. Cloughesy, S. F. Nelson, L. M. Liau, K. Mertz, M. A. Rubin, H. Moch, M. Loda, W. Catalona, J. Fletcher, S. Signoretti, F. Kaye, K. C. Anderson, G. D. Demetri, R. Dummer, S. Wagner,

M. Herlyn, W. R. Sellers, M. Meyerson, L. A. Garraway, High-throughput oncogene mutation profiling in human cancer., Nat Genet 39 (3) (2007) 347–351.
URL http://dx.doi.org/10.1038/ng1975

[6] N. Beerenwinkel, T. Antal, D. Dingli, A. Traulsen, K. W. Kinzler, V. E. Velculescu, B. Vogelstein, M. A. Nowak, Genetic progression and the waiting time to cancer., PLoS Comput Biol 3 (11) (2007) e225.
URL http://dx.doi.org/10.1371/journal.pcbi.0030225

[7] P. Duesberg, R. Li, Multistep carcinogenesis: a chain reaction of aneuploidizations., Cell Cycle 2 (3) (2003) 202–210.

[8] P. Stephens, S. Edkins, H. Davies, C. Greenman, C. Cox, C. Hunter, G. Bignell, J. Teague, R. Smith, C. Stevens, S. O'Meara, A. Parker, P. Tarpey, T. Avis, A. Barthorpe, L. Brackenbury, G. Buck, A. Butler, J. Clements, J. Cole, E. Dicks, K. Edwards, S. Forbes, M. Gorton, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, D. Jones, V. Kosmidou, R. Laman, R. Lugg, A. Menzies, J. Perry, R. Petty, K. Raine, R. Shepherd, A. Small, H. Solomon, Y. Stephens, C. Tofts, J. Varian, A. Webb, S. West, S. Widaa, A. Yates, F. Brasseur, C. S. Cooper, A. M. Flanagan, A. Green, M. Knowles, S. Y. Leung, L. H. J. Looijenga, B. Malkowicz, M. A. Pierotti, B. Teh, S. T. Yuen, A. G. Nicholson, S. Lakhani, D. F. Easton, B. L. Weber, M. R. Stratton, P. A. Futreal, R. Wooster, A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer., Nat Genet 37 (6) (2005) 590–592.
URL http://dx.doi.org/10.1038/ng1571

[9] C. Greenman, P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, S. Edkins, S. O'Meara, I. Vastrik, E. E. Schmidt, T. Avis, S. Barthorpe, G. Bhamra, G. Buck, B. Choudhury, J. Clements, J. Cole, E. Dicks, S. Forbes, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, A. Jenkinson, D. Jones, A. Menzies, T. Mironenko, J. Perry, K. Raine, D. Richardson, R. Shepherd, A. Small, C. Tofts, J. Varian, T. Webb, S. West, S. Widaa, A. Yates, D. P. Cahill, D. N. Louis, P. Goldstraw, A. G. Nicholson, F. Brasseur, L. Looijenga, B. L. Weber, Y.-E. Chiew, A. DeFazio, M. F. Greaves, A. R. Green, P. Campbell, E. Birney, D. F. Easton, G. Chenevix-Trench, M.-H. Tan, S. K. Khoo, B. T. Teh, S. T. Yuen, S. Y. Leung, R. Wooster, P. A. Futreal, M. R. Stratton, Patterns of somatic mutation in human cancer genomes., Nature 446 (7132) (2007) 153–158.
URL http://dx.doi.org/10.1038/nature05610

[10] P. A. Jones, S. B. Baylin, The epigenomics of cancer., Cell 128 (4) (2007) 683–692.
URL http://dx.doi.org/10.1016/j.cell.2007.01.029

[11] J. Lin, C. M. Gan, X. Zhang, S. Jones, T. Sjöblom, L. D. Wood, D. W. Parsons, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, G. Parmigiani, V. E. Velculescu, A multidimensional analysis of genes mutated in breast and colorectal cancers., Genome Res 17 (9) (2007) 1304–1318.
URL http://dx.doi.org/10.1101/gr.6431107

[12] C. Greenman, R. Wooster, P. A. Futreal, M. R. Stratton, D. F. Easton, Statistical analysis of pathogenicity of somatic mutations in cancer., Genetics 173 (4) (2006) 2187–2198.
URL http://dx.doi.org/10.1534/genetics.105.044677

[13] J. M. Satagopan, R. C. Elston, Optimal two-stage genotyping in population-based association studies., Genet Epidemiol 25 (2) (2003) 149–157.
URL http://dx.doi.org/10.1002/gepi.10260

[14] B. Efron, R. Tibshirani, J. D. Storey, V. Tusher, Empirical Bayes analysis of a microarray experiment, Journal of the American Statistical Association 96 (2001) 1151–1160.

[15] I. Chowers, D. Liu, R. H. Farkas, T. L. Gunatilaka, A. S. Hackam, S. L. Bernstein, P. A. Campochiaro, G. Parmigiani, D. J. Zack, Gene expression variation in the adult human retina., Hum Mol Genet 12 (22) (2003) 2881–2893.
URL http://dx.doi.org/10.1093/hmg/ddg326

[16] V. K. Mootha, C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhom, D. Altshuler, L. C. Groop, PGC-1α-responsive genes involved in oxidative phosphorylatioin are coordinately downregulated in human diabetes, Nature Genetics 34 (2003) 267–273.