Journal of Global Biosciences
ISSN 2320-1355
Volume 9, Number 1, 2020, pp. 6737-6747
Website: www.mutagens.co.in
DOI: www.mutagens.co.in/jgb/vol.09/01/090104.pdf
SCAN ME

Research Paper

DYNAMICS OF STATISTICS IN GENOMICS, PROTEOMICS AND TRANSCRIPTOMICS IN EMERGING ERA OF BIOINFORMATICS

Rakesh Ranjan and Saket Vinayak

University Centre of Bioinformatics (Sub-DIC),
T. M. Bhagalpur University, Bhagalpur- 812007,
India.

Abstract

Advances technologies accompanied by a proliferation of genomic data across many scientific disciplines and virtually all disease areas which include latest technologies that can profile genomes, transcriptomes, proteomes and metabolomes raising a plethora of analytic and computational challenges. The general term *Bioinformatics* refers to a multidisciplinary field involving computer scientists, computational biologists, systems biologists, mathematical modelers, and statisticians exploring different facets of the data ranging from organizing, retrieving, storing and subsequent analysis of biological data. Statisticians have a unique perspective and skill set that places them at the center of this process. One of the key attributes that sets statisticians apart from other scientists is their understanding of variability and uncertainty quantification. These are essential considerations in building productive methods for biological discovery and validation, especially for complex, high-dimensional data as encountered in genomics. The experts of statistics are "data scientists" who understand the scientific effect of sampling design decisions on downstream analysis, potential propagation of errors from multi-step processing algorithms, and the potential loss of information from overly reductionistic feature extraction. They are experts in inferential reasoning, which equips them to recognize the importance of multiple testing adjustments to avoid reporting spurious results as discoveries, and to properly design algorithms to search high-dimensional spaces and build predictive models while obtaining accurate measures of their predictive accuracy. Biomedical science have moved to a place where huge data are becoming ubiquitous in research and even clinical practice. This provides great opportunities for the statistical community to play a crucial role in pushing the science forward, as we equip other scientists with the tools they need to extract the valuable information.

Key words: Bioinformatics, Genomics, Proteomics, Transcriptomics, Statistical Modeling.

INTRODUCTION

Statistics have been involved in different aspects of bioinformatics, they have been hesitant to get heavily involved in other discipline. Statisticians are primarily interested in end-stage modeling after all of the data already been collected and preprocessed Baladandayuthapani et.al., 2014). Statistical expertise in the experimental design and low-level processing stages are equally important if not more important than end-stage modeling, since errors and inefficiencies in these steps propagate into subsequent analysis and can preclude the possibility of making new discoveries and scientific conclusions even with the best constructed end-stage modeling strategies(Baggerly et. al., 2010). This has resulted in a missed opportunity for the statistical community to play a larger leading role in bioinformatics that in many cases has been instead assumed by other quantitative scientists and computational biologists and a missed opportunity for biologists as well to more efficiently learn true reproducible biological insights from their data (Bauer et. al., 2010).

Genomics: Primarily DNA-based assays measure genomic events at the DNA level before transcription. Relevant DNA alterations include natural variability in germ line genotype or the DNA sequence across individuals that sometimes affect biological function and disease risk. Germ line or somatic genomic aberrations including various types of mutations including substitutions, insertions, deletions and translocations as well as broader changes in the genome including loss of entire chromosomes or parts or loss of heterozygosity (LOH) involving the loss of one of two distinct alleles originally possessed by the cell. Diploid organisms such as humans have two copies of each autosome (i.e. non-sex chromosomes), but many diseases are associated with aberration in the number of DNA copies in a cell, especially carcinogenic part (Pinkel and Albertson, 2005). Most of the carcinogenic diseases acquire DNA copy number changes manifesting as entire chromosomal changes segment-wise changes in the chromosome or modification of the DNA folded structure. Such cytogenetic modifications during the life of the patient can result in disease initiation and progression by mechanisms wherein disease-suppression genes are lost or silenced, or promoter genes that encourage disease progression are amplified. The detection of these regions of aberration has the potential to impact the fundamental knowledge and treatment of many types of diseases and can play a role in the discovery and development of molecular-based individual therapies. In recent years, cytogeneticists were limited to visually analyzing whole genomes with a microscope by *karyotyping* or *chromosome analysis.* In mid-70's and 80's the development and application of molecular diagnostic methods such as polymerase chain reaction (PCR), Southern blots and flourescence in situ hybirdization (FISH) allowed clinical researchers to make many important advances in genetics, including clinical cytogenetics, moreover these latest techniques have some limitations. Initially, they are time-consuming and labor-intensive and only a limited number and regions of the chromosome can be tested simultaneously. Further, because the probes are targeted to specific chromosome regions, the analysis requires basic knowledge of an abnormality and was of limited use for screening complex karyotypes. Recently scientists have developed techniques that integrate aspects of both traditional and molecular cytogenetic techniques called *chromosomal micorarrays* (Vissers *et al.,* 2010). These high-throughput high-resolution microarrays have allowed researchers to diagnose numerous subtle genome-wise chromosomal abnormalities that were previously undetectable and find many cytogenetic abnormalities in part or all of a single gene. Such information is beneficial for biologists to detect emerging genetic disorders and also provide better understanding of the pathogenic mechanisms of chromosomal aberrations.

There are two types of chromosomal microarrays:

(i) Array-based Comparative genomic hybridization (aCGH arrays) and

(ii) Single nucleotide polymorphism microarrays (SNP arrays).

CGH-based methods were developed to survey DNA copy number variations across a whole genome in a single experiment (Kallioniemi *et al.,* 1992) with CGH, differentially labeled test and reference genomic DNAs are co-hybridized to normal metaphase chromosomes, and fluorescence ratios along the length of chromosomes provide a cytogenetic representation of the relative DNA copy number variation. Chromosomal CGH resolution is limited to 10–20 Mb, thus, any aberration smaller than that will not be detected. Comparative genomic hybridization (aCGH) is a subsequent modification of CGH that provided greater resolution by using microarrays of DNA fragments rather than metaphase chromosomes (Pinkel *et al.,* 1998; Snijders *et al.,* 2001). These arrays can be generated with different types of DNA preparations. One method uses bacterial artificial chromosomes (BACs), each of which consists of a 100-200 kilobase DNA segment. Other arrays are based on complimentary DNA (Pollack et al. 1999) or oligo nucleotide fragments (Lucito *et al.,* 2000). Similar to the CGH analysis, the resultant map of gains and losses is obtained by calculating fluorescence ratios measured via image analysis tools.

SNP arrays are most common types of high-resolution chromosomal microarrays (Mei *et al.,* 2000). SNPs, or single nucleotide polymorphisms are single nucleotides in the genome in which variability across individuals or across paired chromosomes has been observed. Scientistists have identified more than 50 million SNPs in the human genome. SNP arrays take advantage of hybridization of strands of DNA derived from samples with hundreds of probes representing unique nucleotide sequences. As with aCGH, SNP-based microarrays quantitatively determine relative copy number for a region within a single genome. Platform-specific specialized software packages are used to align the SNPs to chromosomal locations, generating genome-wise DNA profiles of copy number alterations and allelic frequencies that can then be interrogated to answer various scientific and clinical questions. Note that unlike aCGH arrays, SNP arrays have the advantage of detecting both copy number alterations as well as LOH events given the allelic fractions, typically referred to as the B-allele frequencies (Beroukhim *et al.,* 2006). They also provide genotypic information for the SNPs, which when considered across multiple SNPs can be used to study different haplotypes. SNP array analysis of germ line samples have been extensively used in genome-wide association studies (GWAS) to find genetic markers associated with various disease of interest (Yau and Holmes, 2009) .

The initial human genome project involved a complete sequencing of a human genome, which took 13 years (1990–2003) and cost roughly $3 billion. Over the last decade significant amendments have been made in the hardware and software undergirding sequencing leading to next generation sequencing (NGS) which can now be used to sequence an entire human genome in less than a day for a cost of about $1000. This data sequencing obtained by applying NGS to DNA, DNAseq, can be used to completely characterize genotypes in GWAS studies and to characterize genetic mutations for cancer tumors and other diseased tissue. Many types of mutations can be characterized including point mutations, deletions, insertions and translocations. It can also be used to estimate copy number variation and LOH throughout the genome. Time and Cost of sequencing is exactly determined by depth of sequencing. When focus is on common mutational variants and copy number determination, low depth sequencing (8x–10x) may be sufficient, but much higher depth is required if rare variants are yet to be detected. Targeted sequencing is performed to focus on specific parts of the genome, e.g. whole exome sequencing for which the gene coding regions only are sequenced.

Proteomics: These emerging technologies allow direct quantification of protein expression as well as post-translational events. Although much more difficult to study than RNA or DNA because their frequency levels span many orders of magnitude. It is important to study proteins as these play a functional role in cellular processes and numerous studies have found that mRNA expression and protein abundance often correlate poorly with each other. Here several important proteomic technologies are available that involve estimating absolute or relative abundance levels, including low to moderate assays that can be used to study small numbers of pre-specified proteins and high-throughput methods that can survey a larger scale of the proteome. Transcriptomics: Earlier work related to the measurement of gene mRNA expression data were based on a "one-gene-at-time" process by using hybridization based methods such as Northern Blots and Reverse transcription polymerase chain reaction (RT-PCR) experiments (Alwine ***et al.*,** 1977). Broadly, the purpose of these low-throughput experiments was to measure the size and abundance of the RNA transcribed for an individual gene using cellular RNA extraction procedures applied to multiple cells from a organism or sample. These experiments were typically time-consuming and involved selection of individual genes to assay expression and were mostly hypothetical. The advent of microarray-based technologies in the mid-1990's then automated these techniques to simultaneously measure expressions of thousands of genes in parallel. This shifted gene expression analyses from mostly hypothesis driven endeavors to hypothesis generating ones that involve an unbiased exploration of the expression patterns of the entire transcriptomics. The major types of analysis can be predominantly classified into three main categories (Seidel, 2008). The first reported works in microarrays involved **spotted microarrays** developed at Stanford University (Schena ***et al.*,** 1998). This process involves printing libraries of PCR products or long oligo nucleotide sequences from a set of genes onto glass slides via robotics and then estimating the gene expression intensities through fluorescent tags (Brown and

Botstein,1999). Other research institutions developed laboratories for printing their own spotted microarrays, which had variable data quality given the challenge of reproducible manufacture of the arrays. **Affymetrix** was the first company to standardize the production of microarrays, becoming the most established and widely-used commercial platform for measure high-throughput gene expression data. Their arrays consist of 25-mer oligo-nucleotides synthesized on a glass chip (Pease ***et al.,*** 1994). As opposed to the single sequence of probes used in spotted microarrays, Affymetrix uses a set of probes to analyze and summarize expression of the genes. Subsequently, other companies, including Illumina, Agilent and Nimblegen, have produced microarrays involving *in situ* synthesis, with each using a different type and length of oligonucleotide as well as photo-chemical process for measurement of gene expression (Blanchard ***et al.,*** 1996). As described in the next section, the development of more cost-effective and efficient sequencing technologies has led to the use of next generation sequencing (NGS) technologies applied to RNA, RNAseq for gene expression. Although each technology has its own characteristics and caveats, the basic read outs contain expression level estimates for thousands on genes on a per-sample basis. This has been used for discovery of the relative fold change in disease versus normal tissues (Alizadeh ***et al.,*** 2000) and among different disease tissues (Ramaswamy ***et al.,*** 2001). However, these technologies have been used to discover molecular signatures that can differentiate subtypes within a given disease that are molecularly distinct (Guinney ***et al.,*** 2015). Clinical applications include but are not limited to development of diagnostic and prognostic indicators and signatures (Cardoso ***et al.,*** 2008; Bueno-de Mesquita ***et al.,*** 2007; Bonato ***et al.,*** 2011).

Integromics: The emerging field of "integromics" is integrative analysis of multi-platform genomics data. The integration of data across diverse platforms has sound biological justifications because of the natural interplay among diverse genomic features. Looking across platforms, attributes at the epigenetic and DNA level such as methylation and copy number variation can affect mRNA expression, which in turn is known to influence clinical outcomes of disease through proteins and subsequent post-translational modifications. Statistically, there are multiple types of data integration methods depending on the scientific question of interest, and the taxonomy can be classified into three broad categories (Kristensen *et al.,* 2014). The first class of methods deal with understanding mechanistic relationships between different molecular platforms for cross-platform interactions such as DNA-mRNA, mRNA-protein etc. The second class of methods involves the identification of latent groups of patients or genes using the multi-platform molecular data and can be cast as either a classification or clustering problems. The third class of methods deals with prediction of an outcome or phenotype for prospective patients.

CONCLUSION:

Statistics has played a significant role in helping develop rigorous design and analysis tools for researchers to use to extract meaningful biological information from the multi-platform genomics data. Their deep understanding of the scientific process has uniquely equipped them to serve a significant role in this venture. One of the key statistical concepts is that unified models that borrow strength across related elements enjoy statistical benefits over piecemeal approaches, leading to more efficient estimation, improved prediction, and greater sensitivity and lower false discovery rates for making discoveries. This borrowing of strength can occur across samples within an object, across data types, and between data and biological knowledge in the literature. This concept at work in peak detection on the mean spectrum, incorporation of copy number and B-allele frequency to determine copy number estimates, borrowing of strength across samples to estimate underlying protein abundances, borrowing strength across samples to identify shared genomic copy number aberrations, incorporating pathway information into models or their natural interrelationships. This principle is also at work in flexible modeling approaches that borrow strength across nearby observations in functional or image data using basis function modeling and regularization priors, a strategy that has been applied to MS, 2DGE, copy number, and methylation data. The concept of regularization is used when smooth functional data in normalization of microarrays, when penalizing regression coefficients in high-dimensional regression models, when denoising spectra before performing peak detection, and when segmenting DNA copy number data. New technologies are continually being developed and introduced at a rapid rate, and there are many new challenges these data will bring. It is believed that statisticians will be involved on the front lines of methods development for these technologies as they are introduced, and that we are involved in all aspects of the science including design, preprocessing and end-stage analysis. Thus, methods and approaches developed on older platforms have some translational importance to the new ones, at least in terms of key issues and the underlying principles behind effective solutions to them. There are a number of areas where more work is clearly needed and future developments are possible. One key area is in integrative analysis. This field is really just getting started and the scientific community is in dire need of new methods for integrating information across multiple platforms to gain more holistic insights into the underlying molecular biology. These methods must balance statistical rigor in building connections, computational efficiency to scale up to big data settings, and interpretability of results so our collaborators can make sense of them. Also, given the extensive efforts in the biological research community to build up knowledge resources that are freely available online, such the recent large-scale federal efforts for unified databases especially in cancer e.g. NCI Genomic Data Commons (Grossman *et al.* 2016). Hence, the statistical community needs to find better ways to incorporate this information into the modeling, which can lead to improved predictions and discoveries as well as enhanced interpretability of the results. Given the interdepencies underlying genetic processes, pathway information is one of the most important types of information that we need to better incorporate.

REFERENCES:

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, (2000). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature. 403(6769):503–511.

Alwine JC, Kemp DJ, Stark GR.(1977) Method for detection of specific rnas in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with dna probes. Proceedings of the National Academy of Sciences. 74(12):5350–5354.

Baggerly KA, Coombes KR (2010) Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. The Annals of Applied Statistics. 3(4):1309–1334.

Baladandayuthapani V, Talluri R, Ji Y, Coombes KR, Lu Y, Hennessy BT, Davies MA, Mallick BK (2014). Bayesian sparse graphical models for classification with application to protein expression data. The Annals of Applied Statistics. 2014;8(3):1443.

Bauer S, Gagneur J, Robinson PN (2010). Going bayesian: model-based gene set analysis of genome-scale data. Nucleic Acids Research. gkq045.

Beroukhim R, Lin M, Park Y, Hao K, Zhao X, Garraway LA, Fox EA, Hochberg EP, Mellinghoff IK, Hofer MD, (2006) Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide snp arrays. PLoS Comput Biol. 2(5):e41.

Blanchard A, Kaiser R, Hood L(1996). High-density oligonucleotide arrays. Biosensors and bioelectronics. 11 (6):687–690. [Google Scholar]

Bonato V, Baladandayuthapani V, Broom BM, Sulman EP, Aldape KD, (2011). Bayesian ensemble methods for survival prediction in gene expression data. Bioinformatics. 27(3):359–367.

Brown PO, Botstein D (1999). Exploring the new world of the genome with dna microarrays. Nature Genetics. 21:33–37.

Bueno-de Mesquita JM, van Harten WH, Retel VP, van't Veer LJ, van Dam FS, Karsenberg K, Douma KF, van Tinteren H, Peterse JL, Wesseling J, (2007). Use of 70-gene signature to predict prognosis of patients with nodenegative breast cancer: a prospective community-based feasibility study The Lancet Oncology. 8(12):1079–1087.

Cardoso F, Van't Veer L, Rutgers E, Loi S, Mook S, Piccart-Gebhart MJ.(2008) Clinical application of the 70-gene profile: the mindact trial. Journal of Clinical Oncology. 26 (5):729–735.

Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM. (2016;) Toward a shared vision for cancer genomic data. New England Journal of Medicine. 375(12):1109–1112.

Guinney J, Dienstmann R, Wang X, de Reynies A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, Bot B, Morris J, Simon I, Gerster S, Fessler E, de Sousa A, Melo F, Missiaglia E, Ramay H, Barras D, Homicsko K, Maru D, Manyam G, Broom B, Boige V, Laderas T, Salazar R, Gray J, Tabernero J, Bernards R, Friend S, Laurent-Puig P, Medema J, Sadanandam A, Wessels L, Delorenzi M, Kopetz S, Vermeulen L, Tejpar S (2015). The consensus molecular subtypes of colorectal cancer. Nature Medicine. 21(11):1350–1356.

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science. 258 (5083):818–821.

Kristensen VN, Lingjærde OC, Russnes HG, Vollan HKM, Frigessi A, Børresen-Dale AL(2014). Principles and methods of integrative genomic analyses in cancer. Nature Reviews Cancer. 14 (5):299–313.

Lucito R, West J, Reiner A, Alexander J, Esposito D, Mishra B, Powers S, Norton L, Wigler M(2000). Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. Genome Research. 10 (11):1726–1736.

Mei R, Galipeau PC, Prass C, Berno A, Ghandour G, Patil N, Wolff RK, Chee MS, Reid BJ, Lockhart DJ (2000). Genome-wide detection of allelic imbalance using human snps and high-density dna arrays. Genome Research. 10 (8):1126–1137.

Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor S(1994). Light-generated oligonucleotide arrays for rapid dna sequence analysis. Proceedings of the National Academy of Sciences. 91(11):5022–5026.

Pinkel D, Albertson DG (2005). Array comparative genomic hybridization and its applications in cancer. Nature Genetics. 37:S11–S17.

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, (1998) High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. Nature Genetics. 20 (2):207–211.

Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO (1999). Genome-wide analysis of dna copy-number changes using cdna microarrays. Nature Genetics. 23 (1):41–46.

Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, (2001). Multiclass cancer diagnosis using tumor gene expression signatures. Proceedings of the Nat. Academy of Sc. 98 (26):15149–15154.

Schena M, Heller RA, Theriault TP, Konrad K, Lachenmeier E, Davis RW (1998). Microarrays: biotechnology's discovery platform for functional genomics. Trends in Biotechnology. 16 (7):301–306.

Seidel C (2008). Introduction to DNA microarrays. Analysis of Microarray Data: A Network-based Approach. 1:1.

Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, (2001) Assembly of microarrays for genome-wide measurement of dna copy number. Nature Genetics. 29 (3):263–264.

Vissers LE, de Vries BB, Veltman JA(2010). Genomic microarrays in mental retardation: from copy number variation to gene, from research to diagnosis. Journal of Medical Genetics. 47 (5):289–297.

Yau C, Holmes C (2009). Cnv discovery using snp genotyping arrays. Cytogenetic and Genome Research. 123(1–4):307–312.