Leveraging Omics Biomarker Data in Drug Development: With a GWAS Case Study

Weidong Zhang

Abstract Biomarkers have proven powerful for target identification, understanding disease progression, drug safety and treatment responses in drug development. Recent development of omics technology has offered great opportunities for identifications of omics biomarkers at low cost. Although biomarkers have brought many promises to drug development, steep challenges arise due to high dimensionality of data, complexity of technology and lack of full understanding of biology. In this article, the application of omics data in drug development will be reviewed. A genome wide association study (GWAS) will be presented.

Keywords Biomarker · Omics · Simulation · GWAS

1 Introduction

1.1 Overview of Biomarker in Drug Development

Precision medicine has gained great popularity in the last decade. In 2015, a total of $215 million investment was budgeted to develop national databases after President Barack Obama announced a 'Precision Medicine Initiative'. The goals of this initiative are two-folds: (a) to focus on precise cancer drug development and (b) to build a database with knowledge of biomarkers that can be used for a broader range of diseases [4].

Biomarkers are indispensable assets to precision medicine and overall drug development. A biomarker can be defined as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" [3]. Biomarkers have been identified as important factors to improve probability of success in drug development. From a recent analysis performed by Thomas et al. 9,985 phase transition trials from

W. Zhang (✉)
Pfizer Inc., Cambridge, MA, USA
e-mail: weidong.zhang2@pfizer.com

R. Liu and Y. Tsong (eds.), *Pharmaceutical Statistics*,
Springer Proceedings in Mathematics & Statistics 218,
https://doi.org/10.1007/978-3-319-67386-8_22

2006 to 2015 were analyzed. Phase transitions are defined as either a drug candidate advances into the next phase of development or is suspended by the sponsor. It was shown that the success rate from Phase I to approval was increased to ~25% when selection biomarkers were used as compared to ~8% for those programs without [2].

In this article, an overview of the biomarker discovery and omics biomarker technologies will be presented. The statistical considerations in omics biomarker analysis will be discussed. A GWAS case study will be presented for illustration of application of omics technology.

1.2 Classification of Biomarkers

Depending on their functions, biomarkers can be classified into predictive biomarkers [8], prognostic biomarkers, pharmacodynamic (PD) biomarkers and surrogate biomarkers. A predictive biomarker predicts a patient's clinical response to the treatment he/she received. Predictive biomarkers are of particular interest in precision medicine due to the fact that a predictive biomarker can be used to identify a patient population that potentially respond or respond better to the new treatment or avoid side effects of a treatment. A recent successful story was reported by Tesaro, Inc, in which there was a study that patients who carried the germline BRCA mutation had progression-free survival (PFS) of 21 months after receiving niraparib as compared to 5.5 months in the control group (Tesaro 2017). A prognostic biomarker, however, can predict a patient's clinical outcome in a way that is independent of any treatment. An example of a prognostic biomarker can be found in a report by Paik et al. in which case a 21-gene recurrence score was used to predict breast cancer recurrence and overall survival in node-negative, tamoxifen-treated breast cancer [15]. A prognostic biomarker may not be used to predict treatment response. However, it may be helpful to a physician to decide whether chemotherapy should be prescribed for high risk patients or avoided by low risk patients. Many biomarkers, however, may be both prognostic and predictive biomarkers in nature, for example, in breast cancer estrogen receptor (ER) can be used as a prognostic biomarker because ER negative patients have a higher risk of relapse than ER-positive patients. On the other hand, the anti-estrogen tamoxifen is more effective in preventing breast cancer recurrences in ER-positive patients than in ER-negative patients, which constitutes ER as a predictive biomarker. Predictive biomarkers will be focused in most of the discussions of this article due to their unique value in patient stratification in clinical trial design.

A PD biomarker can be used to quantify drug modulation and demonstrate principle of mechanism. Frequently, PD biomarkers are useful tools in early clinical trials such as phase 1 to provide guidance for dose selection. PD biomarkers are critical to demonstrate three pillars (target exposure, target binding and target modulation) in drug discovery. It was shown that trials with successful demonstration of these three pillars had much high overall successful rate in the subsequent proof of concept (POC) studies [14].

A surrogate biomarker may be used as a substitute for a clinical endpoint of interest. According to the Biomarker Working Group [3], a surrogate endpoint is defined as "a biomarker intended to substitute for a clinical endpoint. A clinical investigator uses epidemiological, therapeutic, pathophysiological, or other scientific evidence to select a surrogate endpoint that is expected to predict clinical benefit, harm, or lack of benefit or harm". For example, many imaging markers such as total brain volume, hippocampal volume, etc. have been used as surrogate markers in Alzheimer's disease since those imaging markers seem to correlate well with disease progression [11]. However, Fleming and DeMets [7] pointed out that correlation does not automatically guarantee a surrogate status. In some circumstances, a drug may be efficacious on the marker that correlates well with the clinical endpoint but may not have any effect on the clinical endpoint of interest.

1.3 Overview of Omics Biomarker and Cutting-Edge Technologies

Omics technologies refer to the new advanced technologies that are primarily used for the global detection of genes (genomics), mRNA (transcriptomics), proteins (proteomics) and metabolites (metabolomics) in a specific biological sample. Omics biomarkers are typically high-dimensional as illustrated in Fig. 1.

For example, gene expression profile technologies can measure abundance of all the genes (~25 k) in the transcriptome for each sample, which gives scientists an unbiased view of the global biological landscape. The omics technologies started to emerge from the late 20th century when Microarray was first available for gene profiling of whole transcriptome and whole genome genotyping. The early DNA microarray consists of a solid glass surface and a collection of DNA fragments, known as probes or oligos attached to the surface. A probe is a fragment of a section of a gene that can be used to uniquely hybridize a cDNA or cRNA from a fluorescent molecule labeled target sample. The fluorescent intensity of a probe-target hybridization is quantified to determine the abundance of DNA molecules in the target sample. The microarray technology has evolved greatly over the last decade; however, it suffers from major drawback such as dependence on known genes, relatively low sensitivity and low dynamic range. Early in the 21st century, the next generation sequencing (NGS) technologies started to show new promises by offering variety of novel methods for genomics study. Over the last decade, turnaround time and cost of sequencing have been substantially reduced as a result of the advancement of this new technology. It was estimated that the cost of sequencing a genome dropped from $100 million in 2001 to $1,245 in 2015 Wetterstrand [22], and the turnaround time was shortened from years in the late 90 s to days including analysis in 2016 [13]. As of today, NGS technology has been widely applied to a variety of biomedical research areas including transcriptome profiling, identification of new RNA splice variant, genome-wide genetic variants identification, genome-wise epigenetic modification

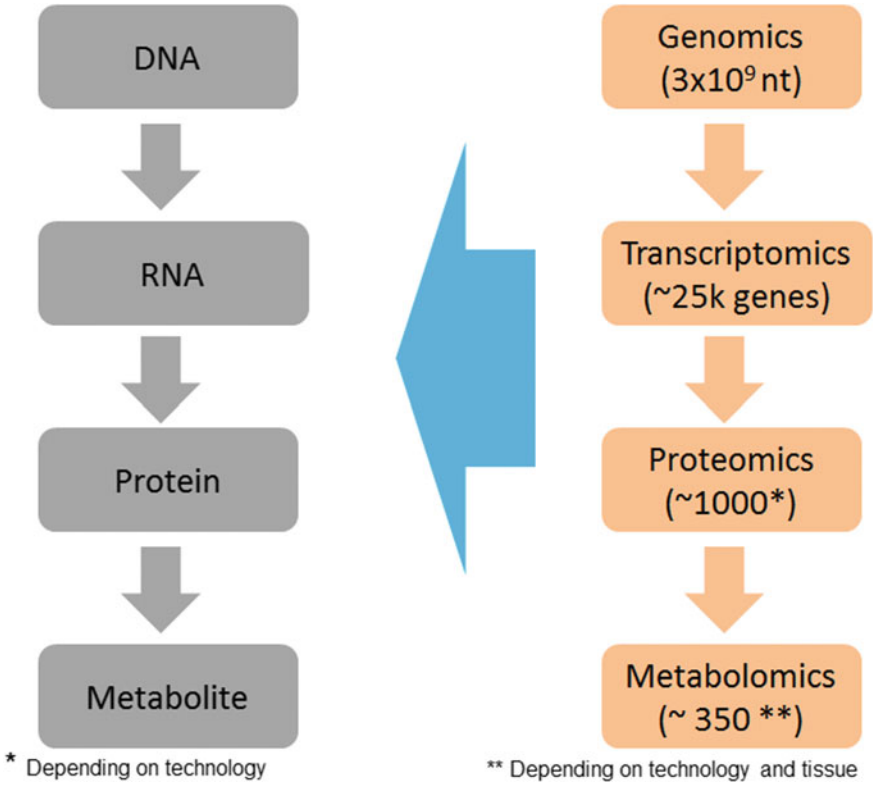

Fig. 1 Omics provide paramount view of biological cascade

and DNA methylation profiling etc. In particular, NGS technology is a promising tool for cancer research, given the "disorder of genome" nature of cancer disease. In cancer research, NGS has significantly enhanced our ability to conduct comprehensive characterization of cancer genome to identify novel genetic alterations, and has significantly helped to dissect tumor complexity. Coupling with sophisticated computational tools and algorithms, significant achievements have been accomplished for breast cancer, ovarian cancer, colorectal cancer, lung cancer, liver cancer, kidney cancer, head and neck cancer, melanoma, acute myeloid leukemia (AML) etc. [18].

Choice of technologies should be made based on the goal of the study. Unbiased high dimensional technology gives maximum information but may not be an efficient choice if the pathway under study is relatively well understood. For example, in oncology, many times scientists want to focus on a select set of genes, gene regions, or amplicons that have known associations with cancer, in which case targeted sequencing panel may be used instead of whole exome or whole genome. In pharmacology, genes from a specific pathway, e.g. JAK-STAT pathway may be of

interest to study drug modulation for JAK inhibitors, and a Taqman low density array (TLDA) panel may be sufficient instead of whole transcriptome.

2 Considerations of Statistical Analysis

Analysis of high dimensional omics biomarker needs special statistical considerations. Conventional statistics focus on problems with large number of experimental units (n) as compared to small number of features or variables (p) measured from each unit. High dimensional biomarker data are often large in p and small in n. For example, in GWAS in a clinical trial, about one million single nucleotide polymorphisms (SNPs) can be collected using microarray from each subject with the number of subjects ranging from dozens to hundreds. Many statistical methods have been developed in analysis of high-dimensional omics data. Typical methods include clustering analysis for pattern discovery, and univariate or multivariate regression and supervised and unsupervised classification analysis to predict disease status [9]. For expression based omics data such as gene expression, proteomics, metabolomics etc., dimension reduction is considered as the first step before subsequent analysis. Dimension reduction techniques include descriptive statistical approach such as coefficient variation (CV) filtering, by which biomarkers with low CV are removed from subsequent regression/ANOVA analysis. This approach is particularly useful when computing power is limited. However, the CV filtering step is typically skipped with today's high computational capacity, and instead, a univariate regression analysis is used for both dimension reduction and inference.

Univariate single biomarker analysis is popular due to the simplicity and interpretation benefit, but is often criticized for being oversimplifying biology by including only one biomarker in the regression model. Multivariate and multiple regressions consider multiple biomarkers in a model become more popular for being able to take into account (1) Complexity of disease mechanism requires an integrated information from multiple biomarkers to explain more biological variations. (2) Interactions between biomarkers cannot be modeled with single biomarker analysis. (3) Correlation and dependency among biomarkers cannot be handled with single biomarker analysis.

Another challenging area in statistical analysis of high-dimensional omics data is how to control false discovery rate (FDR), especially with presence of correlation structure among biomarkers. Family-wise error rate (FWER) adjustment techniques such as Bonferroni correction calculate the probability of making at least one type I error, often considered too conservative. FDR based approaches control the probability of false discoveries from the "positive" findings (rejected null hypotheses). Therefore, FDR procedures are more powerful than FWER but at the cost of high type I errors. Common FDR based methods include Benjamini and Hochberg (BH) method [1] and q value method [19]. The BH method first finds the largest k such that $\mathrm{P}(m) \leq k/m * \alpha$, where m stands for m tests and α is a predefined FDR level. Second, the null hypothesis for each H(i)with $\mathrm{i} = 1 \ldots \mathrm{k}$ are rejected. The q value

method calculates q values that are considered as quantification of false discovery rate. Both the qvalue and BH methods allow dependence of testing. However, the qvalue method may provide more power than BH method, and has been widely used in many omics studies [19].

For GWAS, determination of genome-wide significance threshold is difficult due to as many as millions of statistical testing and complex genetic linkage disequilibrium (LD) structures. Many procedures have been proposed including Bonferroni, FDR, Sidak, and permutation etc. however, it was suggested that a $p = 5 \times 10^{-8}$ can be used for genome wide significance and $p = 1 \times 10^{-7}$ can be used as a suggestive threshold at practical level [16, 17]. Fadista et al. recently studied different scenarios and suggested that P-value thresholds should take into account impact of LD thresholds, MAF and ancestry characteristics. Further, they confirmed a p value threshold of 5×10^{-8} was appropriate for European population with MAF > 5%. However, they suggested that the P-value threshold needs to be more stringent with European ancestry with low MAF (3×10^{-8} for MAF >= 1%) due to the increasing number of variants and the lower LD between less frequent variants [6].

3 A Case Study—A Novel Bootstrap Based Model Average Approach for GWAS Using Outbred Mice

In a study conducted by Zhang et al. [23], a total of 288 outbred mice were used to identify genetic polymorphisms that may be associated with phenotypes such as High-density lipoprotein (HDL), Systolic blood pressure (SBP), Triglyceride (TG), Glucose (GLU) and Albumin Creatinine Ratio (ACR). Outbred mice are similar to human population with regard to genetic diversity but offer great accessibility. The genotype were measured using Affymetrix® Mouse Diversity Array covering ~620 k SNPs. Population structure was first evaluated by calculating correlations between SNP pairs within 50 Mb sliding window across the whole genome. A kinship matrix between the individual animals was calculated based on identity by state among the 44,428 SNPs using Efficient Mixed-Model Association (EMMA) [10]. Single-locus association genome scans were performed by ANOVA and EMMA taking into account population structures. To assess genome-wide significance of the association statistics, a novel simulation technique was used as illustrated in the following steps:

(1) Each phenotype was transformed using van derWaerden's scores [5].
(2) Genetic and residual variances of the transformed data for each phenotype were estimated using EMMA. For each phenotype, 288 trait values were generated by sampling from a multivariate normal distribution using the mvrnorm function in R with covariance matrix defined by the estimated kinship.
(3) The observed trait values were reordered based on the rank orders of the simulated values. By doing so, permutation was performed on the original data that retains the correlation structure implied by the kinship matrix.

(4) A genome scan using the permuted trait values and recorded the largest –log(p) scores. This was repeated 100 times. A generalized extreme value distribution was fitted to these scores and significance thresholds were derived from the quantiles of this distribution [11].

It is well known that the biological process is a complex system that involves multiple components. To obtain realistic estimates of effect sizes, multilocus analysis was performed using forward stepwise regression with bootstrap resampling [21]. First, 100 data sets were generated by sampling with replacement from the 288 animals. Forward stepwise regression on each resampled data set was performed to obtain a multilocus regression model with 20 SNPs. The choice of 20 is arbitrary just to ensure that the number of SNPs in the regression model is more than the number that could significantly influence the phenotype. A resample model inclusion probabilities (RMIP) for each SNP m was calculated as

$$RMIP_m = \frac{1}{R}\sum_{r=1}^{R} i_{rm}$$

where R = 100 is the number of resampled data sets $i_{rm} = 1$ if at least one SNP within ±w Mb of SNP m was included in the model of sample r, otherwise $i_{rm} = 0$. We varied the window size w from ±0.5 Mb to ±4 Mb.

Precision of the locations of the GWAS hits was not well understood. A simulation approach was used in this study to assess the genome-wide average precision of mapping in this population. The steps are illustrated as follows:

(1) A SNP was randomly selected from the genome and trait values were simulated assuming that SNP selected was the causal locus.
(2) Simulate an effect size corresponding to the same percentage of total variance explained as the HDL QTL on Chromosome 1. Phenotype values were sampled from a multivariate normal distribution using mvrnorm in R with correlation structure defined by the kinship matrix and the genetic and residual variances were the same as those estimated for HDL.
(3) The selected SNP was removed from the data and a genome scan was performed using EMMA. The distance between the SNP with highest –log(p) and the target SNP was recorded.
(4) The process from (2) to (3) was repeated 1000 times, and the distribution of distances from the peak to the target SNP was computed.

The significance thresholds were evaluated by simulation and unrestricted permutation, and was applied to each of the following three methods for measuring association: the trend test, ANOVA test and EMMA. The estimated genome-wide significance thresholds for glucose, HDL cholesterol, systolic blood pressure, and triglycerides were similar across all of these combinations. Values ranged from 5.12 to 5.90, but no single method or trait was consistently higher or lower than another.

Two highly significant loci associated with HDL were identified from chromosome 1 and 5. There seemed to be an association with SBP on proximal Chromosome

10 at 7 Mb that exceeded the genome-wide 0.05 thresholds for the simple trend and ANOVA tests, however, it was not significant for the EMMA test. The logACR trait was the most variable of the five traits two loci seemed to be significant on Chromosome 5 at 147 Mb and Chromosome 11 at 88 Mb using the 0.05 thresholds from either simple trend test or the ANOVA test. The results from multilocus genome-wide scans using forward stepwise variable selection on bootstrapped samples showed that RMIP for the two loci Chromosome 1 at 173 Mb and Chromosome 5 at 126 Mb for HLD were 100% but the hit on Chromosome 1 at 181 Mb was never included as an independent QTL in the multilocus analysis, which indicate this method may be useful for prioritization of GWAS hits.

The simulated precision analysis showed that a GWAS hit in this population with a large effect, e.g. as large as the effect of the HDL hit on chromosome 1, can be localized within 1.34 Mb of the greatest association peak. This approach could be expanded to a range of effect sizes in any genotyped population sample including human GWA studies.

This study demonstrates that the GWA analysis employed here can be successfully applied to outbred mice populations to identify genetic variants underlying complex traits.

4 Summary

Omics technology and genomics data have proven to be powerful tools in drug development. Complexity of the biology, technology and high dimensionality of omics data require extensive attention on novel analytical methodology development. Using an example in GWAS, it can be shown that simulation-based method offers many advantages in regards to prioritizing multiple GWAS hits, determination of genome wide threshold considering population structure, and estimation of precision of GWAS hits. With whole-genome sequencing becoming a new norm for genotyping, transcriptome profiling and many other genomic quantification applications, additional challenges associated with handling data quality control, interaction modeling and integration of multiple types of biomarkers will manifold more complex. With collective efforts from the statistical and other analytical communities, significant progresses have been made and will greatly facilitate using these omics information to elucidate disease mechanisms.

References

1. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. **57**(1), 289–300 (1995)
2. Biomarkers Definition Working Group Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. Clin Pharmacol Therapeutics. **69**, 89–95 (2001)

3. Collins, F.S., Varmus, H.: A new initiative on precision medicine. N. Engl. J. Med. **372**(9), 793–795 (2015)
4. Conover, W.J.: Practical Nonparametric Statistics. John Wiley Chichester, New York (1999)
5. Fadista, J., Manning, A., Florez, J., Groop, L.: The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. Eur. J. Hum. Genet. **24**, 1202–1205 (2016)
6. Fleming, T.R., DeMets, D.L.: Surrogate end points in clinical trials: are we being misled? Ann. Intern. Med. **125**(7), 605–613 (1996)
7. Gosho, M., Nagashima, K., Sato, Y.: Study designs and statistical analyses for biomarker research. Sensors **12**, 8966–8986 (2012)
8. Johnstone, I., Titterington, D.: Statistical challenges of high-dimensional data. Phil. Trans. R. Soc. A **367**, 4237–4253 (2009)
9. Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., et al.: Efficient control of population structure in model organism association mapping. Genetics **178**, 1709–1723 (2008)
10. Katz, R.: Biomarkers and Surrogate Markers: an FDA Perspective. NeuroRx **1**(2), 189–195 (2004)
11. Knijnenburg, T.A., Wessels, L.F., Reinders, M.J., Shmulevich, I.: Fewer permutations, more accurate P-values. Bioinformatics **25**, i161–i168 (2009)
12. Meienberg, J., Bruggmann, R., Oexle, K., Matyas, G.: Clinical sequencing: is WGS the better WES? Hum. Genet. **135**, 359–362 (2016)
13. Morgan, P., Van Der Graaf, P.H., Arrowsmith, J., Feltner, D.E., Drummond, K.S., Wegner, C.D., Street, S.D.: Can the flow of medicines be improved? Fundamental pharmacokinetic and pharmacological principles toward improving Phase II survival. Drug. Discov. Today **17**, 419–424 (2012)
14. Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., et al.: A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N. Engl. J. Med. **351**(27), 2817–2826 (2004)
15. Panagiotou, O.A., Ioannidis, J.P.: Genome-wide significance project. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. Int. J. Epidemiol. **41**(1), 273–86 (2012)
16. Pe'er, I., Yelensky, R., Altshuler, D., Daly, M.: Estimation of the multiple testing burden for Genomewide association studies of nearly all common variants. Genet. Epidemiol. **32**, 381–385 (2008)
17. Shyr, D., Liu, Q.: Next generation sequencing in cancer research and clinical application. Biol. Proced. Online **15**, 4 (2013)
18. Storey, J.D.: A direct approach to false discovery rates. J. Roy. Stat. Soc. **64**, 479–498 (2002)
19. TESARO's Niraparib Significantly Improved Progression-Free Survival for Patients With Ovarian Cancer in Both Cohorts of the Phase 3 NOVA Trial (2016). http://ir.tesarobio.com/releasedetail.cfm?releaseid=977524
20. Thomas, D.W., Burns, J., Audette, J., Carroll, A., Dow-Hygelund, C., Hay, M.: Clinical Development Success Rates 2006–2015. June 2016. https://www.bio.org/sites/default/files/Clinical%20Development%20Success%20Rates%202006-2015%20-%20BIO,%20Biomedtracker,%20Amplion%202016.pdf
21. Valdar, W., Holmes, C.C., Mott, R., Flint, J.: Mapping in structured populations by resample model averaging. Genetics **182**, 1263–1277 (2009)
22. Wetterstrand, K.A.: DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) (2016). www.genome.gov/sequencingcostsdata. Accessed 23 Dec 2016
23. Zhang, W., Korstanje, R., Thaisz, J., Staedtler, F., Harttman, N., Xu, L., Feng, M., Yanas, L., Yang, H., Valdar, W., Churchill, G.A., DiPetrillo, K.: Genome-wide association mapping of quantitative traits in outbred mice. G3: Genes Genomes Genet. **2**(2), 167–174 (2012)