Published in final edited form as:

Chem Biodivers. 2012 May ; 9(5): 935–944. doi:10.1002/cbdv.201100384.

Integration of Diverse Statistical Evidence of Gene-Trait Association in Systems Biology Studies

Cheng Cheng^{a)}

Cheng Cheng: cheng.cheng@stjude.org

^{a)}Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, 38105, USA (phone: +1-901-595-2935; fax: 901-595-8843)

Abstract

The rapid advancement of high-throughput genomic assay technologies has generated large amounts of diverse genomic data in disparate human populations and diseases. These data provide a unique opportunity for biomedical investigators to systematically study multifaceted aspects of genes' involvement in the biological processes underlying important traits from the systems biology perspective. An important component in such a study is the inference that integrates diverse lines of statistical evidence for gene-trait association from the observed trait values and the massive numbers of measured genomic features. A novel integrated statistical analysis procedure is developed in this paper and is illustrated by an application in studying childhood leukemia.

1. Introduction

The rapid advance of genomic biotechnologies has enabled investigators to measure or derive, on the whole-genome scale, diverse genomic characteristics (referred hereafter as *genomic features*) of large sets of individuals, in their germline genomes as well as tumor genomes of cancer patients. In a relatively short time, genome-wide data on SNPs, copy number variations, mRNA expressions, microRNA expressions, and methylation levels can be obtained for a large cohort of individuals in a typical biomedical research institution. The mountains of data on diverse genomic features in large patient cohorts have provided researchers a unique opportunity to investigate the genomic basis of disease development, treatment response, and clinical outcome from systems biology standpoint, by which a comprehensive understanding of the underlying molecular biological processes can be obtained.

The general biological model (assumption) for this type of study is that a gene's (locus') effect on a complex trait is not a result of the action of the gene alone, but rather by combined actions of various genomic features related to the gene and their interactions. Although each genomic feature may have very different biological meaning and function from the others, their joint actions form part of the underlying biological process affecting the trait. For example, mRNA expression can be regulated by microRNA binding and affected by SNPs in the transcription binding sites; transcription signaling can be affected by methylation at or near a transcription binding site; a SNP in the microRNA can change its transcript binding ability; etc. The combined effects result in protein level changes ultimately affecting the trait. There has been ample evidence in the cancer genomics literature supporting this model. Elucidation of how the genes may affect the trait under this model requires a systems biology study. One approach is the integrated genome-wide association study (IGWAS) in which sets of diverse genomic features are tested for their joint association with traits of interest.

A statistical inference procedure for IGWAS is developed in this paper (cf. Experimental Part). The concepts of composite genomic factor (CGF), gene-centric composite genomic factor (gCGF) and pathway-centric CGF (pCGF) herein effectively facilitates the derivation of integrated statistical tests for gene-trait and pathway-trait associations by integrating evidence from genomic features related to the gene or pathway. The methodology is demonstrated by an IGWAS of a clinical trait measuring *in vivo* resistance of *de novo* childhood leukemia to remission induction chemotherapy.

2. Results and Discussion

2.1. An application to studying childhood leukemia

Contemporary chemotherapy of childhood acute lymphoblastic leukemia (ALL) typically consists of remission induction, consolidation, and maintenance phases. Disease (ALL) burden is typically measured in the middle and at the end of remission induction. Flow cytometry and PCR are the technologies to measure the percentage of leukemic cells in bone marrow or peripheral blood, called minimal residual disease (MRD). Not only MRD is one of the best clinical prognostic markers, biologically it is an *in vivo* measure of the disease resistance to chemotherapy. Genes taking part in cell cycle progression and regulation, and DNA replication and repair have been identified in gene expression profiling studies of MRD [1, 2]. As a proof of concept, a genome-wide scan of gCGFs consisting of gene expressions and SNP genotypes (see Experimental Part) for association with the remission induction day-19 MRD (referred as “D-19 MRD” hereafter) is performed and presented below.

Association between D-19 MRD and the gCGFs were tested by the procedure described in the Experimental Part. Expression-trait and SNP-trait associations were tested using Spearman's correlation. The statistical significance threshold for multiple tests of the 11,227 gCGFs was determined by the profile information criterion I_p [3] which tries to strike a balance between the levels of false positive and false negative errors. The P value cutoff value was 0.0057; at this threshold 237 gCGFs were statistically significant.

The genes captured included those in cell cycle progression and regulation (Figures 1–3, plus BUB1, BUB1B, CCNA2, CCNB2, CDC2, CDC20, etc.), cell division regulation and oncogenic genes (Figure 4), and other functions such as drug metabolism (ABCA1), DNA replication/repair (POLD1), and JNK interacting protein in mitochondria (SH3BP5; Figure 5).

D-19 MRD is a complex clinical trait reflecting *in vivo* resistance to initial remission induction chemotherapy. The results seem to support the notion of multifaceted nature of the genes' involvement underlying this complex trait. Among the genes involved with cell cycle progression, expression level and a *cis* SNP of CDK2, NUSAP1, and ITGA6 were associated with the trait (Figure 1). Only expression level of ARHGAP19, KIF14, and KIF11 (Figure 2), and only *cis* SNP of ARHGAP6 and KIF3B were associated with D-19 MRD; whereas expression and a *cis* SNP of KIF2C were moderately associated with the trait (Figure 3). In all cases (except perhaps CDK2) there was no evidence that the *cis* SNP actually regulates the expression level, suggesting that blast mRNA expression levels and germline SNPs affect D-19 MRD through separate paths. Among the oncogenic or cell division genes, there was only evidence that *cis* SNPs were associated with the trait (Figure 4); whereas different patterns could be seen in genes of other functions (Figure 5).

2.2. Discussion

Notably none of the SNPs shown here reached the genome-wide significance level and would likely be dismissed in a usual GWAS. The genome-wide trait-gCGF association scan

has helped bring these SNPs to attention. This is particularly noteworthy in light of the frequent “winner’s curse” (top hits are often false positives) phenomenon in large scale association scans.

CGF is a useful concept that helps derive statistical inference in effective and accurate integration of diverse lines of statistical evidence on genes’ (or pathways’, functional groups’) involvement in the biological process underlying a complex trait. In this paper, a genome-wide inference procedure for trait-gCGF association was developed and demonstrated by an application to childhood ALL.

A CGF is conceptually different from a gene set in gene set enrichment analysis (GSEA; [6]). GSEA integrates evidence of trait association with a single type of genomic features across a set of genes; whereas inference based on CGF integrates trait association evidence across disparate types of genomic features for single genes/loci or sets of genes (pathways or functional groups). Notably, CGF is a very general concept, and a GSEA gene set can be regarded as a CGF; therefore the inference method developed in this paper can be applied to GSEA as well. These two methods are inferentially different. The inference in a GSEA is comparative; that is, significance of a gene set is determined by the rank order of the P values of the features in the set, thus is relative to all the other P values. CGF inference however, is self-contained: as described in Experimental Part, significance of a CGF solely depends on the unit genomic factors’ (UGF) P values inside the CGF, no comparison (rank ordering) to other UGFs or CGFs is required. This “self-containedness” property may be more desirable in genome-wide scan studies for reducing levels of false positive errors because rank order among the genomic features does not directly reflect the absolute strength (effect size) of each trait-feature association.

It should be noted that in constructing the trait-gCGF association test statistic (cf. Experimental Part), the UGF P values are not regarded as measures of statistical significance, but rather as normalized measures of strength of trait association. This provides a means to combine evidence of gene-trait association much more flexible than those using the actual statistics of testing trait-UGF associations. This allows one to choose the most appropriate test for each type of UGF. Different types of UGF may require different types of statistics to test the trait association; thus disparate UGFs lead to disparate test statistics, and it is difficult to combine, for example, t statistics with chi-square statistics having different degrees of freedoms. The UGF P values of the test statistics are versions of the test statistics meaningfully normalized to the same scale.

Still caution must be taken when combining the P values. UGF disparity can result in different P value behaviors substantially affecting the performance of the test statistic. For example, in a large CGF, the significance can be driven by one or few very strong trait-associated UGFs inside it; on the other hand the association signal from one or few UGFs may be buried in the noise from the UGFs not or very weakly associated with the trait. The former situation is not of major concern because a more detailed look into the CGF after the genome-wide scan can reveal other possible association signals, such as in the cases of CDK2, NUSAP1, ITGA6 (Figure 1), POLD1 and SH3BP5 (Figure 5). The latter situation is currently dealt with using the extreme (max) statistic. This seems to have helped detect gCGFs containing relatively modest trait-UGF associations (e.g., KIF2C; Figure 3).

The tests (P values) of trait-UGF association within a CGF are often stochastically dependent; thus the “effective” number of UGF tests is some $K' < K$ (the size of the CGF). Statistical power of testing the trait-CGF association can be improved by replacing K with K' . However K' is typically (almost always) unknown, but can be estimated by a re-sampling analysis. Additionally, one could borrow from false discovery rate methods [3,4,5]

to estimate the proportion of $U(0,1)$ UGF P values inside a CGF, and down weight the contributions of the P values in the $U(0,1)$ portion in the mixture distribution. Future research will include incorporating these changes into the construction of the trait-CGF association test statistics.

3. Conclusion

The concept of CGF can effectively facilitate the derivation of integrated statistical analysis for testing gene-trait associations encompassing diverse types of genomic features. As demonstrated by the application to childhood leukemia study, the proposed statistical procedure for genome-wide tests of CGF-trait association can effectively discover genomic features biologically meaningful to the subject matter, and point out interesting genomic features not reaching the genome-wide significance level and likely be dismissed in a usual GWAS.

Acknowledgement

The research is supported in part by the grants U01GM 92666, U19HL065962 and cancer center support grant CA21765 from NIGMS, NHLBI, and NCI of the National Institutes of Health, and by the American Lebanese Syrian Associated Charities (ALSAC). Thanks are due Drs. Dario Campana, James Downing, Ching-Hon Pui, William Evans and Mary Relling at St. Jude Children's Research Hospital for their gracious collaborations that inspired this work.

Experimental Part

Basic concepts: A *unit genomic factor* (UGF) is in general a genomic feature that can be represented by a single variable. SNPs, mRNA, microRNA, and methylation expression probes (or "probeset" in case of Affymetrix GeneChips), are examples of UGFs. In a very general term, a *composite genomic factor* (CGF) is a set of UGFs. Biologically meaningful CGFs can be constructed on the bases of known or putative genes and pathways. A gene (or locus) CGF (gCGF) of a given gene (or locus) is a set of UGFs with known or putative association with the gene; for example, a set consisting of mRNA expressions of the gene, *cis* SNPs, and microRNAs binding to the transcripts of the gene form a gCGF of the gene. A gCGF of a gene is a multifaceted representation of the gene. A pathway CGF (pCGF) can be defined as the set union of gCGFs of (all the) genes in the pathway; like a gCGF, a pCGF gives a multifaceted representation of a pathway.

In a genome-wide scan, gCGFs are constructed for all the genomic loci of interest (often those interrogated on the genome-wide arrays), giving a genome-wide coverage. A genome-wide, integrated gene-trait association analysis is carried out by testing the association between each gCGF and the trait. The association between the trait and each UGF in a gCGF provides a line of evidence for the gene's involvement in the processes that result in the biological variations in the trait. Association between the gene and the trait can be assessed using the gCGF by aggregating the lines of evidence provided by the element UGFs with proper test statistics. A specific method is described below.

Denote a gCGF in general as a set of variables, $\{V_1, V_2, \dots, V_K\}$; each variable V measures a UGF. For example, $K=12$, V_1, V_2 are mRNA expressions, V_3 through V_{11} are genotypes of *cis* SNPs, and V_{12} is the expression of a microRNA with a known binding site on a transcript. Note the gCGFs of two different loci do not necessarily contain the same number of UGFs, but may share UGFs. K is used to denote generically the number of UGFs in a CGF.

Statistical test: The genome-wide trait-gCGF association scan takes the following steps: (1) Test each UGF for its association with the trait, represent the result by weight of statistical evidence, i.e., P value. This step can be done by established methods (t test, ANOVA, linear/logistic/Cox regression, Cochran-Armitage test, nonparametric tests, etc.). (2) For a gCGF $\{V_1, V_2, \dots, V_K\}$, let $P_1, P_2, \dots, P_K$ be the P values of the element UGFs by the association tests in Step (1). Compute a statistic aggregating the evidence of association from the UGFs, which is a function of the P values: $S=S(P_1, P_2, \dots, P_K)$. Repeat this for each gCGF in the study; thus each gCGF has an association test statistic computed. (3) Determine the statistical significance (P value) of each gCGF statistic under the *complete null hypothesis* that no UGF in the gCGF is associated with the trait, by a proper “null distribution” of the test statistic. Now each gCGF has a statistical significance for its association with the trait. (4) Assess the statistical significance of all the gCGFs (genome-wide) using established massive multiple hypothesis test procedures [3, 4, 5]; rank order the gCGFs by this assessment and report results. Methods for Steps (1) and (4) are already available.

In Step (2), several considerations must go into the construction of the test statistic $S(P_1, P_2, \dots, P_K)$. First, it should be sensitive to evidence of association reflected by small UGF P values, and avoid undue extremity when no evidence of association exists; second, it should account for the heterogeneity among the UGFs that may result in disparate strengths (effect sizes) of trait-UGF associations and P value distributions; third, it should possess a relatively simple probability distribution under the complete null hypothesis so that statistical significance can be determined with efficient computation. A specific choice is

$$S=S(P_1, \dots, P_K)=\max\{\Phi^{-1}(1-P_1), \dots, \Phi^{-1}(1-P_K)\},$$

where Φ^{-1} is the standard Normal quantile (inverse distribution) function. Here large values of S indicate statistical significance of trait-gCGF association. For Step (3), the probability distribution of S can be derived under the complete null hypothesis (the null distribution) as follows. Assume that each test of the trait-UGF association is unbiased so that each P_j ($j=1, 2, \dots, K$) follows the $U(0,1)$ distribution under the no-association null hypothesis; and assume that the P values are stochastically independent. Then the null distribution of S has the cumulative distribution function (cdf) $F(x)=[\Phi(x)]_K$, with Φ the cdf of the standard normal $N(0,1)$. The P value of an observed value of $S=S_{obs}$ is calculated as $P=1-F(S_{obs})$ if $S_{obs}>0$; $P=1$ otherwise. Intuitively one would worry that simply taking the maximum as the test statistic S may lead to an elevation of the false positive error probability if the size K of the CGF is large. Fortunately, the null distribution is adaptive to K : it shifts toward the right as K increases (Figure 6); thus at say $K=500$, to obtain the same level of significance (P value) as at $K=50$, the observed statistic has to be much more extreme.

Gene expression and SNP Data: In the application (cf. Section 2) gene expression (GE) data on 189 subjects were generated using Affymetrix U133A GeneChips in *de novo* leukemic blasts, as described previously [1, 2]. Of these, 187 had trait data. Germline SNP genotypes were available in a subset of 160 subjects, and were determined by Affymetrix 100K and 500K, and SNP6.0 arrays as described previously [7]; after a quality control filtering [7], 481,820 SNPs remained for analysis. Of the 160 subjects with SNP data, 159 had trait data. In step (1), GE-trait associations were tested using all GE arrays of the 187 subjects, and SNP-trait association were tested using all SNP arrays of the 159 subjects.

Each gCGF consisted of two types of genomic features, mRNA (transcript) expression in *de novo* leukemic blasts, and germline SNP. This study included all genes/loci that have at least one *cis* SNP (defined as either inside or within 5kb up or down stream of the locus) on the

SNP arrays and at least one probeset on the GeneChip. Totally 11,227 such loci were identified and the gCGF of each locus consisted of all *cis* SNPs genotyped by the SNP arrays and mRNA probesets of the locus. Furthermore, the *cis* SNPs with no variation in the study cohort were then excluded; after this, the mean (median, range) number of probesets and SNPs in a gCGF are 1.66 (1, 1 to 15) and 12.5 (5, 0 to 778) respectively.

References

1. Flotho C, Coustan-Smith E, Pei D, Cheng C, Song G, Pui CH, Downing JR, Campana D. Blood. 2007; 110:1271. [PubMed: 17456722]
2. Flotho C, Coustan-Smith E, Pei D, Iwamoto S, Song G, Cheng C, Pui CH, Downing JR, Campana D. Blood. 2006; 108:1050. [PubMed: 16627760]
3. Cheng C, Pounds S, Boyett JM, Pei D, Kuo M-L, Roussel MF. Statistical Applications in Genetics and Molecular Biology. 2004; 3:Article 36. <http://www.bepress.com/sagmb/vol3/iss1/art36>.
4. Benjamini Y, Hochberg Y. J R Stat Soc Ser B. 1995; 57:289.
5. Storey JD, Taylor J, Siegmund D. J R Stat Soc Ser B. 2003; 66:187.
6. Subramanian A, Tamayoa P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirova JP. Proc Natl Acad Sci. 2005; 102:15545. [PubMed: 16199517]
7. Yang J, Cheng C, Yang W, Pei D, Cao X, Fan Y, Pounds S, Trevino LR, French D, Campana D, Evans WE, Pui CH, Devida M, Bowman WP, Camitta BM, Willman C, Davies SM, Borowitz MJ, Carroll WL, Hunger SP, Relling MV. Journal of American Medical Association. 2009; 301:393.

Figure 1.

gCGF of three genes involved in cell cycle progression and regulation. Left panel: relationship between D-19 MRD and the most significant probeset is plotted on rank scale; the ties in D-19 MRD resulted in the smallest average rank of 37 shown by the circles on the bottom. Middle panel: Box plot of D-19 MRD distribution (on rank scale) across genotypes of the most significant *cis* SNP. The SNP genotype is represented as 0=AA, 1=AB, 2=BB. Right panel: Box plot of the distribution of the most significant probeset (on rank scale) across genotypes of the most significant *cis* SNP.

Figure 2.
 gCGF of three genes involved in cell cycle progression and regulation. The layout is the same as Figure 1. The statistical significance of these genes was driven by gene expression but not by SNP.

Figure 3. gCGF of three genes involved in cell cycle progression and regulation. The layout is the same as Figure 1. SNPs drove the statistical significance of ARHGAP6 and KIF3B; whereas for KIF2C both expression and a SNP are moderately associated with D-19 MRD.

Figure 4. gCGF of three genes involved in cell division and tumor genesis. The layout is the same as Figure 1. Statistical significance of these genes was driven by SNPs.

Figure 5. gCGF of three genes involved in a few other biological functions. The layout is the same as Figure 1.

Figure 6. Probability density function (pdf) of the test statistic under the null hypothesis for different CGF sizes: $K=500$ (solid line), 100 (dashed), and 50 (dotted).