Assessing gene length biases in gene set analysis of Genome-Wide Association Studies

Peilin Jia

Departments of Biomedical Informatics and Psychiatry,
Vanderbilt University Medical Centre,
Nashville, Tennessee 37232, USA
E-mail: peilin.jia@vanderbilt.edu

Jian Tian

Department of Computer Science,
Vanderbilt University,
Nashville, Tennessee 37232, USA
E-mail: jian.tian@vanderbilt.edu

Zhongming Zhao*

Departments of Biomedical Informatics,
Psychiatry, and Cancer Biology,
Vanderbilt University Medical Centre,
Nashville, Tennessee 37232, USA
E-mail: zhongming.zhao@vanderbilt.edu
*Corresponding author

Abstract: Genome-Wide Association Studies (GWAS) have rapidly become a major genetics approach to studying complex diseases. Although many susceptibility variants and genes have been uncovered by single marker analysis, gene set based analysis is emerging as a very promising approach aiming to detect joint association of a set of genes with disease. In the available gene set based methods, it is often the smallest *P* value of the Single Nucleotide Polymorphisms (SNPs) in a gene region is used to represent the gene-level association signal. This approach may introduce strong bias of association signal towards long genes. In this study, we propose a resampling strategy by randomly generating genomic intervals across the accessible genomic region to estimate the background distribution of *P* values at the gene level. Comparing with the gene-wise *P* value in real data, the proportion of random intervals could be used to assess the bias that might be introduced by gene length and in turn to help the investigators choose the appropriate gene set analysis algorithms in their GWAS datasets. Our method uses only summarised GWAS data with no need of permutation, thus, it is computationally efficient. A computer program is freely available for the users.

Keywords: GWAS; genome-wide association studies; pathway enrichment analysis; gene set; gene length; bias.

Reference to this paper should be made as follows: Jia, P., Tian, J. and Zhao, Z. (2010) 'Assessing gene length biases in gene set analysis of Genome-Wide Association Studies', *Int. J. Computational Biology and Drug Design*, Vol. 3, No. 4, pp.297–310.

Biographical notes: Peilin Jia received her PhD in Bioinformatics from the Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, China. She is currently a post-doctoral researcher in the Department of Biomedical Informatics and Psychiatry, Vanderbilt University School of Medicine. Her research interests include development and application of bioinformatics methods and algorithms in Genome-Wide Association Studies of complex diseases, integration of multi-dimensional resources from large scale analysis of genomic and genetic data, common and rare variant identification from next-generation sequencing data in complex diseases, and knowledge discovery using data mining methods.

Jian Tian is currently an Undergraduate Student in the Department of Computer Science at the Vanderbilt University. He is interested in computational biology and genomic medicine.

Zhongming Zhao received his PhD in Human and Molecular Genetics from the University of Texas Health Science Center at Houston and MD Anderson Cancer Center, Houston, Texas in 2000. Currently, he is an Associate Professor in the Departments of Biomedical Informatics, Psychiatry, and Cancer Biology at the Vanderbilt University Medical Center. His research interests include bioinformatics and systems biology approaches to studying complex diseases, genome-wide or large-scale analysis of genetic variation and methylation patterns, next-generation sequencing data analysis, comparative genomics and biomedical informatics.

1 Introduction

Genome-Wide Association Studies (GWAS) have rapidly become a major genetics approach to studying complex diseases, under the hypothesis of "Common Disease-Common Variant (CD-CV)" (Lohmueller et al., 2003). In a typical GWA study, half to a few million SNPs across the human genome are systematically tested in hundreds to thousands of samples for their association with complex diseases or traits. Recent success of GWA studies has uncovered many disease susceptibility genes or variants. However, more effort is much needed to mine abundant genetic association signal from various available GWAS data due to two reasons. First, only a few variants or genes, sometimes even none, could reach genome-wide significance ($P < 5 \times 10^{-8}$) in a typical GWA study (Jia et al., 2011b). Second, complex diseases might be caused by many genetic variants with weak or moderate risk, but they interact with each other to have major risk. Therefore, Gene Set Enrichment Analysis (GSEA), especially pathway-based enrichment analysis, of GWAS data has recently become one of the most promising approaches. It is proposed to provide alternative insights than the traditional single marker analysis (Wang et al., 2010) and has been proved to be promising in many studies (Jia et al., 2010;

Wang et al., 2007, 2009). For this approach, several methods have been applied in a variety of diseases, such as GSEA (Wang et al., 2007), traditional hypergeometric test (Jia et al., 2010), ALIGATOR (Holmans et al., 2009), GRASS (Chen et al., 2010), and dmGWAS (Jia et al., 2011a). The pipeline that implements such approaches typically includes the following procedures.

- to perform an association test of single markers such as the basic allelic test and the Cochran-Armitage trend test
- to map SNPs in the GWAS dataset to the corresponding genes
- to annotate or group genes into gene sets such as biological pathways (Kanehisa et al., 2008) or Gene Ontology (GO) categories (Ashburner et al. 2000)
- to perform statistic test of association significance at the gene set level (Dinu et al., 2007; O'Dushlaine et al., 2009, 2010; Perry et al., 2009; Subramanian et al., 2005; Tian et al., 2005; Wu et al., 2010).

A key procedure in such a pipeline is how to map SNPs to genes and estimate the summarised *P* values at the gene level (i.e., gene-wise *P* values). Although many methods have been proposed in recent literatures, including Fisher's combined method (Luo et al., 2010; Peng et al., 2010), the Simes' method (Peng et al., 2010; Simes, 1986; Wang et al., 2007), and gene-wise FDR correction (Luo et al., 2010), the most popular one is to simply choose the SNP in the gene region whose *P* value is the smallest and use the *P* value to represent the significance level of the gene (Wang et al., 2007, 2009). This way of obtaining gene-wise *P* values has been shown to be sensitive in many studies (Jia et al., 2010; Wang et al., 2007).

In the gene set analysis of GWAS datasets, investigators often use gene-wise *P* values of the genes mapped in the same pathway to detect the enriched (joint) genetic signal at the pathway level. Although the smallest-*P*-value method is effective to represent gene-wise *P* value and it is easy to implement, there are several issues in this method that may introduce biases for estimating gene-wise *P* values; such biases have been frequently ignored in previous studies. For example, assuming that SNPs are evenly distributed across the genome, a long gene is expected to have more SNPs included in a GWAS dataset and, thus, has a higher chance to have significant markers – here significant markers denote small nominal *P* values (e.g., $P < 0.05$ or $P < 0.01$) from the GWAS dataset. In the real experiments, genes having more SNPs included in the commercial or custom microarray genotyping chips (e.g., Affymetrix and Illumina chips) are expected to have better chance to identify significant SNPs; and such genes are generally longer than the others. Another issue that further complicates the bias is the local Linkage Disequilibrium (LD) structure, which implicitly determines the 'effective' SNPs or independent SNPs for each gene, rather than the absolute number of SNPs per gene. Thus, the chance of a gene to be related to a significant SNP is not only related to its length and the number of SNPs included in the chips, but also the local LD environment (Holmans et al., 2009; Jia et al., 2011b).

In most GSEA, gene length biases are typically adjusted by permutation data, which is generated by swapping the case and control labels in the original genotyping data. However, this process is computationally intensive and time consuming. Since each

of the available GSEA methods has its advantages and disadvantages, an efficient way would be to assess the gene length bias in advance to determine its effect, and then select the appropriate approach accordingly to avoid repeating try-and-error processes.

In this study, we propose a method to estimate the gene length bias from any GWAS dataset, aiming to provide pre-analysis assessment of potential biases and, subsequently, to help the investigators determine which method would be appropriate in their gene set (e.g., pathway) enrichment analysis of the GWAS dataset. We propose to use the summary GWAS data to perform genome-wide resampling and estimate the background *P* value distribution for a query gene (i.e., any gene from the human genome and included in the GWAS dataset). For complex diseases that long genes are expected to be commonly involved, this evaluation is valuable to help the investigators select appropriate algorithms. For example, neurodevelopmental genes, which tend to be longer than other human genes, have been commonly implicated in psychiatric disorders, under a hypothesis of neutro-physiopathology hypothesis (Sun et al., 2010). Since our method does not rely on permutation, it is computationally efficient. We developed a computer program and made it freely available for the investigators for the assessment of gene length biases.

2 Materials and methods

2.1 GWAS datasets

We used the GWAS dataset for schizophrenia from the Genetic Association Information Network (GAIN) (Manolio et al., 2007). GAIN is a public–private partnership of the Foundation for the National Institutes of Health and has funded several GWA studies such as schizophrenia (Shi et al., 2009) and major depression disorder (Sullivan et al., 2009). The schizophrenia GWAS data was available in dbGaP and was approved for our use by the GAIN DAC through the National Human Genome Research Institute. The data was genotyped using the Affymetrix Genome-Wide Human SNP 6.0 array. We used only unrelated European ancestry samples. The following criteria for inclusion of individuals and markers were performed:

- individual samples were removed if the missing genotype rate was >5%
- SNPs were excluded if the missing genotype rate were >5%, or Minor Allele Frequency (MAF) <0.05.

After the quality control, there were 1158 schizophrenia cases and 1378 controls and ~651,000 SNPs that were used in this analysis.

2.2 Statistic tests

We used Cochran-Armitage Trend test to compute the significance of association of each SNP with schizophrenia. According to previous studies, there was no significant stratification found in the GAIN samples of European ancestry (Shi et al., 2009). The genomic control inflation factor of this dataset was 1.07, further confirming the data quality.

2.3 Gene coordinates and gene length data

Gene coordinate information was extracted using the 'seq_gene.md' file downloaded from the NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/mapview/, build 36.3). We used only protein-coding genes, as annotated in the file "Homo_sapiens.gene_info" that was also available at the same NCBI ftp site. A total of 19,739 protein-coding genes were identified and their corresponding coordinates in the human assembly were used for mapping SNPs in the GAIN GWAS dataset.

2.4 Evaluating gene length biases in individual genes using random genomic intervals

We propose a strategy to evaluate the gene length bias for any protein-coding gene in the human genome. Figure 1 illustrates our strategy. A SNP is mapped to a gene if it is located within the gene region, or within 20 kb immediate upstream or downstream of the gene. For analysis methods that involve gene-level data, a summarised *P* value for each gene is necessary to represent the gene. We use the SNP having the smallest *P* value in the gene for this purpose. Although there are several other ways to represent the summarised *P* values on the gene level (Jia et al., 2010; Luo et al., 2010; Wang et al., 2007), the smallest-*P*-value strategy is the most commonly used one in GSEA and has been shown to be sensitive compared to other methods such as the Simes' method (Wang et al., 2007).

To evaluate whether the summarised *P* value of a gene is biased towards its gene length, we propose to use a resampling based method across the whole human genome. The strategy has the following three steps.

Step 1. Identify accessible genomic regions. As most GWA studies are conducted on commercial genotyping chips, the SNPs that can be genotyped are limited by the applied platform and are generally distributed on each chromosome excluding the telomere and the centromere regions. We thus define the regions that can be genotyped as "accessible genomic regions", which will serve as the "genomic region pool" for resampling. For each specific GWAS dataset, the accessible genomic regions might differ slightly after quality control. This may be implemented by different exclusion criteria in different case studies. The final accessible genomic regions are expected to have up to 46 regions for the human genome, with each chromosome being separated into two possible disconnected regions by their chromosomal arms (p and q arms). However, the p arms of some chromosomes might be too short to be a practical accessible genomic region. Specifically in this study, we demonstrated our strategy by using the GAIN GWAS dataset for schizophrenia, which was generated by the Affymetrix Genome-Wide Human SNP 6.0 array. In this array, no SNP markers are found in four chromosomes' p-arms (chromosomes 13, 14, 15, and 22). As a result, there are in total 42 accessible regions that could be served as the resampling pool.

Step 2. Generation of random genomic intervals for a specific gene. Given a query gene with length l and a summarised *P* value (P_{real}), to assess whether the P_{real} value is biased towards its gene length, we randomly generate genomic intervals with the same length of the gene from the accessible genomic regions, and then compute a summarised *P* value for each interval in the same way as used for the real case. Specifically, a genomic interval is generated by first randomly selecting a region out of the accessible genomic

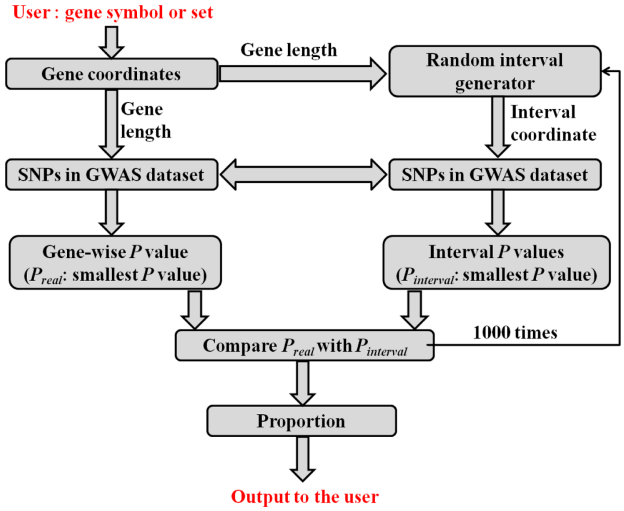
regions, followed by randomly selecting an interval in the region with the length l plus an extended 20 kb in both ends – a total of $l + 40$ kb, as it is done for the real case of each gene. These resampling intervals thus form a background distribution of P values for the query gene.

Step 3. Empirical P value – proportion of more significant intervals than the real case. For the resampling intervals, we count the number of intervals that have more significant summarised P values than the real case and divide it by the number of resampling intervals that successfully cover SNPs in the GWAS dataset, i.e.,

$$\text{Prop} = 100\% \times \frac{\#\{P_{\text{interval}} < P_{\text{real}}\}}{\#\{\text{effective resampling}\}},$$

where P_{interval} is the summarised P value for each interval, P_{real} is the summarised P value for the gene, and the effective resampling denotes the random resampling intervals after excluding those that fall in desert regions without covering any GWAS SNPs.

Figure 1 Work flow to assess the gene length biases in a GWAS dataset. The details are provided in the text (see online version for colours)



2.5 *Evaluating gene length biases in gene sets*

We further evaluate gene length biases in a set of genes, rather than individual genes. We demonstrate this strategy by using the pathways downloaded from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al., 2008), a popularly used pathway resources in GWAS pathway analysis. There are 214 pathways used in our analysis (as of September 7, 2010). For each pathway, the gene length distribution is explored and the median value of all gene lengths within a pathway is computed for comparison of these KEGG pathways. We select one representative pathway with many long genes and examine its gene length bias using the random genomic interval method.

3 Results and discussion

3.1 Implementation of gene length bias evaluation strategy

The details of the gene length bias evaluation method were provided in Section 2.4. Below we presented a pseudo-code to further illustrate the algorithm and then we implemented our random genomic interval method into a computer program that is publicly available for the user.

```
// initialization
query.genes = new Vector()
foreach query.gene in query.genes do
{
        query.gene = getGeneCoordinate()
}
endfor

// get coordinates for accessible genomic regions
accessible.regions = new Vector()
qcSNPsOnChip = new Vector()
foreach chromosome do
{
        chromosome.qcSNPsOnChip = getQcSNPs()
        coordinateMin1 = chromosome.qcSNPsOnChip.getMin1()
        coordinateMax2 = chromosome.qcSNPsOnChip.getMax2()
        chromosome.qcSNPsOnChip.searchForBiggestJunction()
        coordinateMax1 = chromosome.qcSNPsOnChip.getMax1()
        coordinateMin2 = chromosome.qcSNPsOnChip.getMin2()
        accessible.regions.addNewRegions(chromosome, coordinateMin1,
            coordinateMax1)
        accessible.regions.addNewRegions(chromosome, coordinateMin2,
            coordinateMax2)
}
endfor

// resampling
foreach query.gene in query.genes do
{
        gene.pvalue = query.gene.getSmallestPvalue()
        round = 0
        real.round = 0
        significant.intervals = 0
        while (round < 1000) do
        {
                random.regions = accessible.regions.getRandomRegion()
                random.interval = random.regions.getRandomInterval()
                if(random.interval.containsSNPs)
                {
                        real.round++
                        random.pvalue = random.interval.getSmallestPvalue()
                        if(random.pvalue < gene.pvalue) significant.intervals++
                }
        }
        query.gene.proportion = significant.intervals/real.round
}
endfor
```

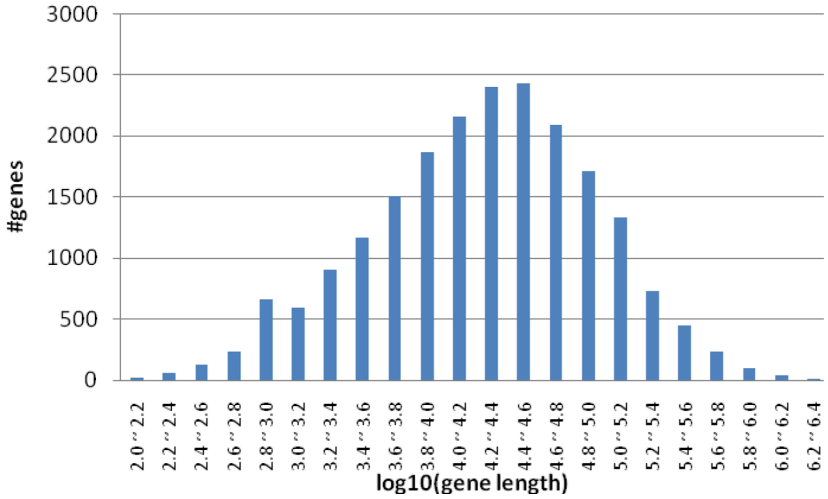
Briefly, the GWAS dataset is initialised by coordinating the accessible genomic regions and the genes inputted by the user. Then, a random interval generator is formulated to generate genomic intervals by a given length. For each query gene, an iteration of 1000 times of resampling is executed, each of which creates a random interval by the generator and then assigns the smallest *P* value of the SNPs mapped in the interval (Figure 1). Finally, the proportion is computed for each query gene by summarising the effective rounds that successfully cover SNPs. Note that the 'effective' rounds may not be exactly 1000 times because in some cases, an interval may fall in desert regions without covering any SNPs in the GWAS dataset. This is reasonable and reflects the real design of the GWAS chips, in which there are many genes that are not covered.

We executed the above pseudo-code by JAVA, an object-oriented programming language. The computer program is freely available on our website: http://bioinfo.mc.vanderbilt.edu/software.html

3.2 Gene length distribution

Of the ~20,000 human protein-coding genes, we first explored the length distribution. As expected, the human gene lengths varied greatly – ranging from 0.1 kb to 2300 kb – and they were heavily screwed towards the right tail (i.e., long genes) (data not shown). This observation indicates that long genes are un-equivalently distributed among all the human genes. In summary, the longest 10% of genes accounted for 54.4% of total length of the human genes. We thus plotted the distribution of logarithm transformed gene length in the human genome. The distribution is shown in Figure 2.

Figure 2 Length distribution of protein-coding genes in the human genome. X-axis is $\log_{10}$(gene length) and Y-axis is the number of genes (see online version for colours)



3.3 Evaluating biases in individual genes: short genes, intermediate genes and long genes

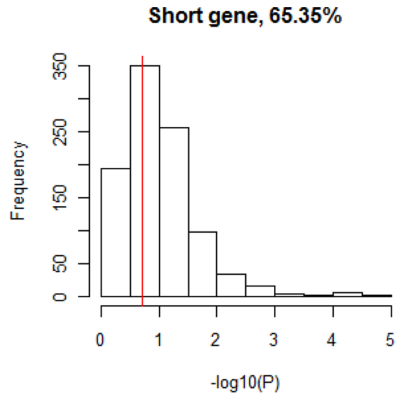
We selected four representative genes to examine the influence of gene length bias in those genes' summarised *P* values. These four genes are:

- one gene randomly selected from the lower 25% quantile in gene length distribution (Figure 2), representing short genes
- two genes from the middle of the distribution (gene length within 25–75% quantile of the distribution), representing intermediate genes
- one gene from the upper 25% quantile, representing long genes.

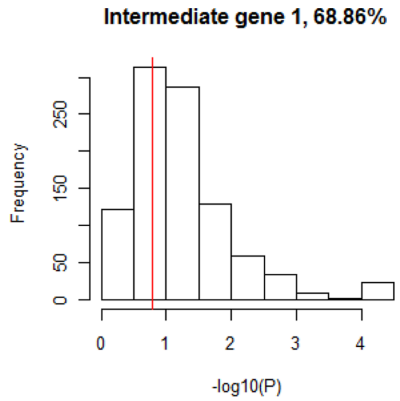
We performed resampling analysis as in Section 2.4, according to their gene lengths.

As expected, the long gene had the highest proportion of resampling intervals, indicating that they had more significant summarised *P* values than the real cases (Prop = 97.03%), while for the short and intermediate genes, the proportion of more significant intervals were <70% (Figure 3). The results revealed that for a long gene region, it tends to cover SNPs with more significant *P* values.

Figure 3 Assessing gene length biases in four representative human genes. Y-axis represents frequency of random genomic intervals having the same length of the gene selected. X-axis represents –log(*P*), where *P* is the smallest *P* value within each of the random interval: (A) A short gene; (B) and (C) Two intermediate genes and (D) A long gene (see online version for colours)

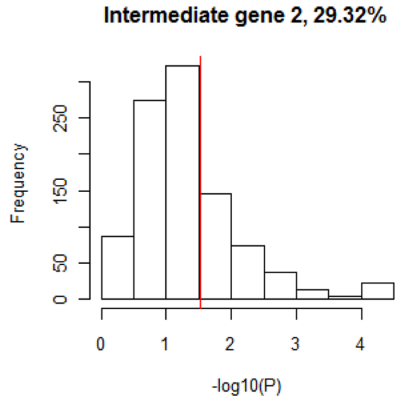


(A)

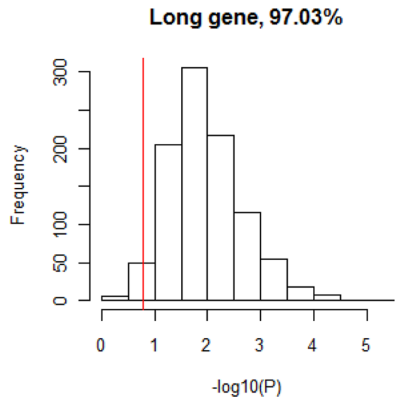


(B)

Figure 3 Assessing gene length biases in four representative human genes. Y-axis represents frequency of random genomic intervals having the same length of the gene selected. X-axis represents –log(*P*), where *P* is the smallest *P* value within each of the random interval: (A) A short gene; (B) and (C) Two intermediate genes and (D) A long gene (see online version for colours) (continued)



(C)



(D)

3.4 Evaluating biases in gene sets: KEGG pathways

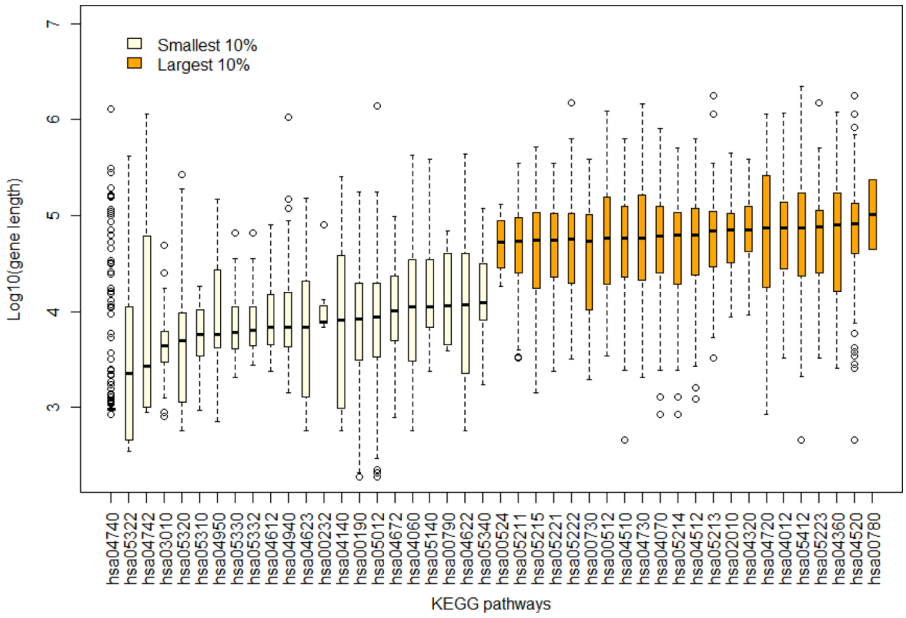
We next examined the gene length distribution of all the KEGG pathways. We found great variation among KEGG pathways in term of gene length distribution and the median values of all gene lengths in the pathways. Among the 214 KEGG pathways, their median values of gene lengths varied from 947 bp to 141,700 bp, with the median being 28,140 bp.

To better present the trend, we plotted the bottom 10% pathways that had the smallest median values of gene length, as well as, the top 10% pathways that had the largest median values (Figure 4). The pathway "biotin metabolism (hsa00780)", which had the largest median value of gene length, contained only two genes (*HLCS* and *BTD*) and was very likely driven by the long gene *HLCS*. The other long-gene pathways included several neuron related pathways such as axon guidance (hsa04360), long-term potentiation (hsa04720) and dorso-ventral axis formation (hsa04320), consistent with the

previous knowledge that neuron-related genes tend to be long (Sun et al., 2009). This also indicates that for complex diseases that neuron-related genes are expected to be involved, caution needs to be taken to estimate the gene length bias before any gene set based analysis of GWAS dataset (Jia et al., 2011b; Sun et al., 2009).

We took the axon guidance pathway (hsa04360) as an example to demonstrate the gene length bias. There were 129 genes annotated to this pathway according to the KEGG database, 49 of which had their gene length located within the top 10% gene length distribution of all human genes. Thus, a substantial portion of the genes in this pathway is long. We performed the random interval estimation for each of these genes and found that 18 out of the 49 long genes had high proportion (Prop > 70%) of randomly selected intervals having P values (P_{interval}) smaller than the real case (P_{real}). This result clearly indicated that these genes are biased towards their gene length.

Figure 4 Gene length distribution of the top 10% shortest and top 10% longest KEGG pathways in term of their median gene length (see online version for colours)



4 Conclusion

As several hundreds of GWAS datasets have been made available, and many more are being generated, gene set based analysis, especially pathway based analysis, of GWAS datasets is emerging rapidly as a powerful approach to uncovering genetic signal for many complex diseases or traits. So far, investigators often select the smallest P value among all the SNPs in a gene region to represent gene-wise association significance, and then test enriched association signal in a gene set by combing gene-wise P values of the gene set. This study addresses a potential strong bias in such an approach. We examined the gene length biases in all human protein-coding genes and in the KEGG pathways. We proposed a strategy – random genomic interval analysis – to assess the

gene length bias in GWAS dataset through resampling in the accessible genomic regions and generating random genomic intervals to estimate the background distribution of gene-wise *P* values. We demonstrated the strategy in a schizophrenia GWAS dataset and showed that pathways with high proportion of long genes tend to be biased towards gene length. This will provide insights for researchers to choose appropriate methods in follow up analysis. We developed a computer program and made it publicly available for the investigators to assess gene length biases for both individual genes and gene sets.

Acknowledgements

We would like to thank Satishkumar Ganakammal for his assistance in figure preparation. This work was partially supported by the National Institutes of Health Grant Nos. R21AA017437 and T15LM007450, Vanderbilt's Specialised Program of Research Excellence in GI Cancer grant P50CA95103, the VICC Cancer Center Core grant P30CA68485, 2009 NARSAD Maltz Investigator Award to Z.Z., and the Department of Psychiatry at the Vanderbilt University. The funding agencies had no further role in the design, implementation, or generation of this research report.

References

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) 'Gene ontology: tool for the unification of biology. The gene ontology consortium', *Nat. Genet.*, Vol. 25, pp.25–29.

Chen, L.S., Hutter, C.M., Potter, J.D., Liu, Y., Prentice, R.L., Peters, U. and Hsu, L. (2010) 'Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data', *Am. J. Hum. Genet.*, Vol. 86, pp.860–871.

Dinu, I., Potter, J.D., Mueller, T., Liu, Q., Adewale, A.J., Jhangri, G.S., Einecke, G., Famulski, K.S., Halloran, P. and Yasui, Y. (2007) 'Improving gene set analysis of microarray data by SAM-GS', *BMC Bioinformatics*, Vol. 8, p.242.

Holmans, P., Green, E.K., Pahwa, J.S., Ferreira, M.A., Purcell, S.M., Sklar, P., Owen, M.J., O'Donovan, M.C. and Craddock, N. (2009) 'Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder', *Am. J. Hum. Genet.*, Vol. 85, pp.13–24.

Jia, P., Wang, L., Meltzer, H.Y. and Zhao, Z. (2010) 'Common variants conferring risk of schizophrenia: a pathway analysis of GWAS data', *Schizophr. Res.*, Vol. 122, pp.38–42.

Jia, P., Zheng, S., Long, J., Zheng, W. and Zhao, Z. (2011a) 'dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks', *Bioinformatics*, Vol. 27, No. 1, pp.95–102.

Jia, P., Wang, L., Meltzer, H.Y. and Zhao, Z. (2011b) 'Pathway-based analysis of GWAS datasets: effective but caution required', *Int. J. Neuropsychopharmacology*, Advanced Online Access December 16, 2010, DOI: 2010.1017/S1461145710001446.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. and Yamanishi, Y. (2008) 'KEGG for linking genomes to life and the environment', *Nucleic Acids Res.*, Vol. 36, pp.D480–D484.

Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S. and Hirschhorn, J.N. (2003) 'Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease', *Nat. Genet.*, Vol. 33, pp.177–182.

Luo, L., Peng, G., Zhu, Y., Dong, H., Amos, C.I. and Xiong, M. (2010) 'Genome-wide gene and pathway analysis', *Eur. J. Hum. Genet.*, Vol. 18, pp.1045–1053.

Manolio, T.A., Rodriguez, L.L., Brooks, L., Abecasis, G., Ballinger, D., Daly, M., Donnelly, P., Faraone, S.V., Frazer, K., Gabriel, S., Gejman, P., Guttmacher, A., Harris, E.L., Insel, T., Kelsoe, J.R., Lander, E., McCowin, N., Mailman, M.D., Nabel, E., Ostell, J., Pugh, E., Sherry, S., Sullivan, P.F., Thompson, J.F., Warram, J., Wholley, D., Milos, P.M. and Collins, F.S. (2007) 'New models of collaboration in genome-wide association studies: the Genetic Association Information Network', *Nat. Genet.*, Vol. 39, pp.1045–1051.

O'Dushlaine, C., Kenny, E., Heron, E., Donohoe, G., Gill, M., Morris, D. and Corvin, A. (2010) 'Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility', *Mol. Psychiatry*, Advanced Online Access February 16, 2010.

O'Dushlaine, C., Kenny, E., Heron, E.A., Segurado, R., Gill, M., Morris D.W. and Corvin, A. (2009) 'The SNP ratio test: pathway analysis of genome-wide association datasets', *Bioinformatics*, Vol. 25, pp.2762, 2763.

Peng, G., Luo, L., Siu, H., Zhu, Y., Hu, P., Hong, S., Zhao, J., Zhou, X., Reveille, J.D., Jin, L., Amos, C.I. and Xiong, M. (2010) 'Gene and pathway-based second-wave analysis of genome-wide association studies', *Eur. J. Hum. Genet.*, Vol. 18, pp.111–117.

Perry, J.R., McCarthy, M.I., Hattersley, A.T., Zeggini, E., Weedon, M.N. and Frayling, T.M. (2009) 'Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach', *Diabetes*, Vol. 58, pp.1463–1467.

Shi, J., Levinson, D.F., Duan, J., Sanders, A.R., Zheng, Y., Pe'er, I., Dudbridge, F., Holmans, P.A., Whittemore, A.S., Mowry, B.J., Olincy, A., Amin, F., Cloninger, C.R., Silverman, J.M., Buccola, N.G., Byerley, W.F., Black, D.W., Crowe, R.R., Oksenberg, J.R., Mirel, D.B., Kendler, K.S., Freedman, R. and Gejman, P.V. (2009) 'Common variants on chromosome 6p22.1 are associated with schizophrenia', *Nature*, Vol. 460, pp.753–757.

Simes, R.J. (1986) 'An improved Bonferroni procedure for multiple tests of significance', *Biometrika*, Vol. 73, pp.751–754.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P. (2005) 'Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles', *Proc. Nat. Acad. Sci. USA*, Vol. 102, pp.15545–15550.

Sullivan, P.F., de Geus, E.J., Willemsen, G., James, M.R., Smit, J.H., Zandbelt, T., Arolt, V., Baune, B.T., Blackwood, D., Cichon, S., Coventry, W.L., Domschke, K., Farmer, A., Fava, M., Gordon, S.D., He, Q., Heath, A.C., Heutink, P., Holsboer, F., Hoogendijk, W.J., Hottenga, J.J., Hu, Y., Kohli, M., Lin, D., Lucae, S., Macintyre, D.J., Maier, W., McGhee, K.A., McGuffin, P., Montgomery, G.W., Muir, W.J., Nolen, W.A., Nothen, M.M., Perlis, R.H., Pirlo, K., Posthuma, D., Rietschel, M., Rizzu, P., Schosser, A., Smit, A.B., Smoller, J.W., Tzeng, J.Y., van Dyck, R., Verhage, M., Zitman, F.G., Martin, N.G., Wray, N.R., Boomsma, D.I. and Penninx, B.W. (2009) 'Genome-wide association for major depressive disorder: a possible role for the presynaptic protein piccolo', *Mol. Psychiatry*, Vol. 14, pp.359–375.

Sun, J., Jia, P., Fanous, A.H., Oord, E.v.d., Chen, X., Riley, B.P., Amdur, R.L., Kendler, K.S. and Zhao, Z. (2010) 'Schizophrenia gene networks and pathways and their applications for novel candidate gene selection', *PLoS ONE*, Vol. 5, p.e11351.

Sun, J., Jia, P., Fanous, A.H., Webb, B.T., van den Oord, E.J., Chen, X., Bukszar, J., Kendler, K.S. and Zhao, Z. (2009) 'A multi-dimensional evidence-based candidate gene prioritization approach for complex diseases-schizophrenia as a case', *Bioinformatics*, Vol. 25, pp.2595–2602.

Tian, L., Greenberg, S.A., Kong, S.W., Altschuler, J., Kohane, I.S. and Park, P.J. (2005) 'Discovering statistically significant pathways in expression profiling studies', *Proc. Nat. Acad. Sci. USA*, Vol. 102, pp.13544–13549.

Wang, K., Li, M. and Bucan, M. (2007) 'Pathway-based approaches for analysis of genomewide association studies', *Am. J. Hum. Genet.*, Vol. 81, pp.1278–1283.

Wang, K., Li, M. and Hakonarson, H. (2010) 'Analysing biological pathways in genome-wide association studies', *Nat. Rev. Genet.*, Vol. 11, pp.843–854.

Wang, K., Zhang, H., Kugathasan, S., Annese, V., Bradfield, J.P., Russell, R.K., Sleiman, P.M., Imielinski, M., Glessner, J., Hou, C., Wilson, D.C., Walters, T., Kim, C., Frackelton, E.C., Lionetti, P., Barabino, A., Van Limbergen, J., Guthery, S., Denson, L., Piccoli, D., Li, M., Dubinsky, M., Silverberg, M., Griffiths, A., Grant, S.F., Satsangi, J., Baldassano, R. and Hakonarson, H. (2009) 'Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn disease', *Am. J. Hum. Genet.*, Vol. 84, pp.399–405.

Wu, M.C., Kraft, P., Epstein, M.P., Taylor, D.M., Chanock, S.J., Hunter, D.J. and Lin, X. (2010) 'Powerful SNP-set analysis for case-control genome-wide association studies', *Am. J. Hum. Genet.*, Vol. 86, pp.929–942.