BMC Proceedings

BioMed Central

Proceedings

Open Access

Integration of gene ontology pathways with North American Rheumatoid Arthritis Consortium genome-wide association data via linear modeling

Jérémie JP Lebrec*[1,2], Tom WJ Huizinga[2], René EM Toes[2], Jeanine J Houwing-Duistermaat[1] and Hans C van Houwelingen[1]

Addresses: [1]Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, P.O. Box 9604, 2300 RC Leiden, The Netherlands and [2]Department of Rheumatology, Leiden University Medical Center, Leiden, P.O. Box 9604, 2300 RC Leiden, The Netherlands

E-mail: Jérémie JP Lebrec* - j.j.p.lebrec@lumc.nl; Tom WJ Huizinga - t.w.j.Huizinga@lumc.nl; René EM Toes - r.e.m.toes@lumc.nl; Jeanine J Houwing-Duistermaat - j.j.houwing@lumc.nl; Hans C van Houwelingen - j.c.vanhouwelingen@lumc.nl
*Corresponding author

from Genetic Analysis Workshop 16
St Louis, MO, USA 17-20 September 2009

Published: 15 December 2009

BMC Proceedings 2009, **3**(Suppl 7):S94 doi: 10.1186/1753-6561-3-S7-S94

This article is available from: http://www.biomedcentral.com/1753-6561/3/S7/S94

Abstract

We describe an empirical Bayesian linear model for integration of functional gene annotation data with genome-wide association data. Using case-control study data from the North American Rheumatoid Arthritis Consortium and gene annotation data from the Gene Ontology, we illustrate how the method can be used to prioritize candidate genes for further investigation.

Background

By definition, heritable complex traits like rheumatoid arthritis (RA) are caused by several mutations with inherently small effect size. Some of those mutations have low frequency and therefore can not be reliably detected, even in large studies. Nevertheless, it is plausible that causal mutations for a given disease belong to a set of common biological pathways. Such external biological information about gene function is now widely available (e.g., gene ontogeny (GO) [1] and KEGG [2]). Methods for integration of this biological knowledge were first developed by microarray analysts in order to improve power and reproducibility of transcriptome studies. Recently, some attempts [3,4] have been made to transfer this approach into the genomics field. The aim of such data integration is twofold. Firstly, this approach could be used to identify whether some predefined gene sets are over-represented in genome-wide association studies (GWAS) and thus are likely to play a role in disease etiology. Secondly, the available biological information could be used to supplement the candidate disease gene hierarchy provided by the GWAS data. We focus on the second of those two objectives using case-control GWAS data from the North American Rheumatoid Arthritis Consortium (NARAC) provided by Genetic Analysis Workshop 16 (GAW16). Currently existing methods are still in their infancy [4,5] (see Sohns et al. [6] in this issue for an application of the method). We have previously proposed a modeling framework for integration of linkage data [7] with pathway information and have applied it to a linkage scan of the NARAC sample set [8] using 47 GO biological processes known to be involved in rheumatoid arthritis [9]. In this paper, we have now derived a similar model for integration of GWAS signals and gene set or pathway information that will enable researchers to prioritize candidate genes for further investigation.

Methods

Data definition

The pathway gene annotation provides information at the gene level. Formally, all *n* human genes may or may not be involved in the *n* pathways considered relevant to the disease of interest. This information is gathered in the $n \times p$ design matrix *n* where $X_{ij} = 1$, if gene *n* belongs to pathway *n* and $X_{ij} = 0$, otherwise. In turn, the GWAS signal is measured at the single-nucleotide polymorphism (SNP) level. As in Wang et al. [3], we chose to summarize the GWAS signal at the gene level so as to drastically reduce the dimension of the problem. Each SNP was assigned to the nearest gene (to more than one gene when genes overlap) and was analyzed using the Cochrane-Armitage trend test after application of standard filtering criteria (genotype call rate > 95%, Hardy-Weinberg equilibrium *p*-value > 0.001, minor allele frequency > 0.01, and exclusion of seven individuals with sex classification inconsistent with genetic data). For each gene, the signal at the most significantly associated SNP was selected. Note that the chosen SNP-to-gene annotation strategy may lead to incorrect annotation if a SNP is in linkage disequilibrium with another SNP within a more distant gene or plays a role in the regulation of that gene. The effect size and the related asymptotic variance of the maximal SNP for each gene were then estimated by logistic regression using the number of SNP alleles as a continuous variable in the model. Note that the sign of the effect size was arbitrarily chosen as positive to ensure the incorporation of simple parameters of pathway effects in Model (1). For each gene indexed by *n*, we therefore obtained a GWAS signal denoted γ_i equal to the positive allelic effect (on the log odds scale) of the maximal SNP mapping to the gene and a related variance σ_i^2 given by the asymptotic estimate.

A hierarchical model for integration of GWAS signal with pathway information

As defined above, γ_i is an estimate of the underlying gene effect μ_i. We further assume that $\gamma_i \sim N(\mu_i, \sigma_i^2)$ and although the corresponding observations at nearby genes may be correlated, we use a diagonal working covariance. The association signal at the SNP level has mean equal to 0 on the log odds scale if the SNP is unrelated to disease, the γ_i values have positive expectation because they are taken as the positive log odds effect of the most significant SNP mapped to each gene, and their distribution is skewed and departs from normality. In addition, while it is not entirely clear how σ_i^2 (defined as the variance of a SNP effect) relates to the variance of the γ_i signal (taken as the maximal signal), it remains closely related to it and is used in our approximate model. In order to introduce pathway effects in the model, we denote by β_j the effect of pathway *j* and further assume that given pathway effects, genes act independently and are distributed as

$$\mu_i \sim \mathrm{N}\,(\mu_0 + \sum_j X_{ij}\beta_j, \tau^2), \tag{1}$$

where μ_0 is an overall average gene effect (this parameter includes the positive bias that results from the definition of γ_i as a positive quantity) and τ^2 denotes the between-gene variance. Marginally, the summarized GWAS signal γ_i is thus distributed as

$$\gamma_i \sim N(\mu_0 + \sum_j X_{ij}\beta_j, \sigma_i^2 + \tau^2). \tag{2}$$

Given τ^2, estimates $\hat{\mu}_0$, $\hat{\beta}$ of μ_0, β are given by the normal equations and thus have an explicit form. Maximum-likelihood estimates of all parameters can thus easily be obtained by numerical maximization of the profile likelihood.

Posterior gene effects

The previous model can be given an (empirical) Bayesian interpretation. The prior gene effects μ_i reflecting pathway effects are updated in the light of new GWAS data γ_i. The posterior gene effects are distributed as

$$\mu_i |\, \mathrm{y}_i \sim N\left(\left(\sigma_i^{-2} + \tau^{-2} \right)^{-1} \left(\sigma_i^{-2}\gamma_i + \tau^{-2} \sum_j X_{ij}\beta_j \right), \left(\sigma_i^{-2} + \tau^{-2} \right)^{-1} \right) \tag{3}$$

and the standardized posterior gene effects provide an overall ranking for genes. Note that each posterior gene effect is a simple weighted average of the prior gene effect (based on pathway information) and the GWAS signal in the vicinity of the gene. The relative weight of these two sources of information is governed by the ratio σ_i^2 / τ^2.

Results

Integration of GWAS and pathway information

While we originally intended to use the same 47 GO biological processes derived in Aidinis et al. [9] and used in Lebrec et al. [7], we have restricted our analysis to 27 pathways provided in the c5 MSigDB gene set definition [10] to ensure comparability with other members of our GAW16 working group. We carried out a forward-stepwise regression procedure in Model (2) in order to select the most relevant pathways among the 27 considered. The fitting process is shown in Table 1. Model selection was based on the Akaike information criterion (AIC). Biological processes associated with a positive effect include GO:0009605 (response to external stimulus), GO:0030036 (actin cytoskeleton organization and biogenesis), GO:0006508 (proteolysis), and GO:0009100 (glycoprotein metabolic process). Other negatively associated pathways have genes in common with those four positive pathways and therefore help in refining their effects. For instance, *TP53* is a component of both GO:0009605 ($\hat{\beta}$ = + 0.018) and GO:0006996 ($\hat{\beta}$ = - 0.015), while *ABCF1* is only found in GO:0009605. The corresponding prior gene effects are therefore +0.018 for *ABCF1* and +0.018 - 0.015 = 0.003 for *TP53*. Table 2 shows the number of genes in each GO set and the number of genes in common between them. We then ranked all genes according to the standardized posterior gene effect computed based upon the distribution in Eq. (3) using the empirical Bayes estimates $\hat{\beta}$ and $\hat{\tau}^2$. Genes with the highest standardized posterior gene effects are located in the HLA region on chromosome 6 where *p*-values of association are the most significant. Due to the strong effect of the GWAS signal in that region, any prior effect based on pathway information is overruled by the GWAS signal. It is however possible to tune the relative influence of GWAS signal and pathway effects by increasing the σ_i^2/τ^2 ratio in the posterior gene effects calculation. We focused on the 1% most highly ranking genes when the tuning is extreme (σ_i^2/τ^2 was set to 100), i.e., when only pathway effects determine posterior gene effects. Seventeen consensus genes arose (they were defined as being part of both extreme priority lists obtained using the genes ranked in the top 1% when either σ_i^2/τ^2 is as initially estimated or when σ_i^2/τ^2 is set to 100): *ABCF1*, *AGER*, *AIF1*, *BACE2*, *C2*, *CAPG*, *CCL1*, *CCL18*, *CD40*, *CDKN1A*, *DOCK2*, *L3MBTL4*, *NEBL*, *NFX1*, *RND3*, *RPS19*, and *TNXB*. The list generated is unlikely to be random because it contains at least four genes reported to be associated with RA. Under the assumption that 100 such genes have already been identified among ≈20,000 human genes, the probability of detecting four or more of those 100 genes in 17 genes drawn at random is $\approx 10^{-6}$. It therefore seems that it contributes additional information as compared to the results from GWAS alone. *AIF1* and *CD40* were recently confirmed as susceptibility genes for RA [11,12] but *CD40* (rs1569723, nominal trend test *p*-value $\approx 10^{-4}$) would not have been prioritized as an

Table 1: Forward-stepwise selection in NARAC data with 27 GO biological processes

					Pathway effect estimates for consecutive models obtained when adding terms as shown in column 1										
Model[a]	**Deviance**	**AIC**	$\hat{\tau}^2$ [b]	$\hat{\beta}$								**Selected Model**			
Intercept		-65279.5	0.001916	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	**0.16**	0.16	0.16	0.16
+GO:0009605[c]	4.27	-65285.99	0.001915		0.017	0.017	0.017	0.018	0.018	0.017	0.017	**0.018**	0.021	0.022	0.018
+GO:0006996	2.29	-65288.57	0.001913			-0.010	-0.010	-0.014	-0.014	-0.015	-0.015	**-0.015**	-0.014	-0.014	-0.014
+GO:0006968	1.75	-65290.08	0.001911				-0.025	-0.025	-0.024	-0.024	-0.024	**-0.024**	-0.024	-0.024	-0.034
+GO:0030036	1.62	-65291.31	0.001909					0.020	0.035	0.035	0.035	**0.035**	0.035	0.035	0.035
+GO:0007015	3.07	-65295.45	0.001907						-0.055	-0.055	-0.055	**-0.055**	-0.055	-0.055	-0.055
+GO:0006508	1.53	-65296.50	0.001904							0.013	0.013	**0.013**	0.013	0.013	0.013
+GO:0009100	0.99	-65296.47	0.001904								0.016	**0.017**	0.017	0.017	0.017
+GO:0006629	1.07	**-65296.61**	**0.001903**									**-0.008**	-0.008	-0.008	-0.008
+GO:0006950	0.46	-65295.54	0.001902										-0.005	-0.005	-0.007
+GO:0009617	0.49	-65294.52	0.001902											-0.019	-0.007
+GO:0006952	0.85	-65294.23	0.001902												0.012

[a] GO definitions: GO:0009605, response to external stimulus; GO:0006996, organelle organization and biogenesis; GO:0006968, cellular defense response; GO:0030036, actin cytoskeleton organization and biogenesis GO:0007015, actin filament organization; GO:0006508, proteolysis; GO:0009100, glycoprotein metabolic process; GO:0006629, lipid metabolic process.
[b] $\hat{\tau}^2$, the between-gene variance estimate
[c] Bold font indicates elements included in the selected model.

Table 2: Number of genes in common between 8 GO biological processes retained in model for NARAC SNP data[a]

Model[b]	GO:0009605	GO:0006996	GO:0006968	GO:0030036	GO:0007015	GO:0006508	GO:0009100	GO:0006629
$\hat{\beta}$	0.018	-0.015	-0.024	0.035	-0.055	0.013	0.017	-0.008
GO:0009605	283	15	14	9	1	8	1	13
GO:0006996		424	0	97	23	11	1	10
GO:0006968			49	0	0	0	0	1
GO:0030036				97	23	1	0	1
GO:0007015					23	0	0	0
GO:0006508						166	6	6
GO:0009100							76	11
GO:0006629								279

[a] Row 2 displays the corresponding pathway effect estimate in selected model.
[b] GO definitions: GO:0009605, response to external stimulus; GO:0006996, organelle organization and biogenesis; GO:0006968, cellular defense response; GO:0030036, actin cytoskeleton organization and biogenesis GO:0007015, actin filament organization; GO:0006508, proteolysis; GO:0009100, glycoprotein metabolic process; GO:0006629, lipid metabolic process.

RA-associated locus according to the NARAC GWAS data alone.

Discussion

Our data and methodology point towards an added value of combining pathway information with GWAS data to possibly assist in the prioritization of candidate genes for further study. However, we resorted to quite a few approximations to cast the problem into the gaussian framework of Model (2). In analogy to the *t*-test, we can realistically hope that the forward stepwise fitting process that we used is robust to the normality assumption. It is beyond the scope of this manuscript to fully assess the robustness of the methodology but the equal variance assumption in Model (1) should especially be challenged. Taking the SNP with maximum association as a response in Model (2) might favor pathways with large genes, as suggested by one reviewer, this bias might be reduced by inclusion of gene size as a predictor in Model (2). This modification of the regression hardly changed the selected model (data not shown) and left top ranking genes unaltered. It is of note that the proportion of between-gene variance explained by the retained model (≈1%) is much smaller than in the case of linkage data (≈50%) [6]. In spite of the fact that SNP data is at a higher resolution than gene-level data, we have chosen a gene-level summary signal due to computational limitations. This may represent a simplistic model that may result in large amount of information being discarded. It will therefore be highly relevant to run this algorithm with SNP-level information for association signals together with pathway information at SNP level to exploit all the available data being generated. Given the wealth of biological information available, it is enticing to use Model (2) with hundreds of candidate pathway predictors. However, the risk of over-fitting would be high. In our opinion, restricting the set of initial candidate pathways to a limited number represents a better approach.

Conclusion

The empirical Bayesian linear model for GWAS and pathway data derived in this study offers a simple and flexible framework for data integration. Because new tools for candidate gene prioritization are rapidly arising [4,5,13], comparison between methods and their relative usefulness for researchers to prioritize genes is warranted.

List of abbreviations used

AIC: Akaike information criterion; GAW16: Genetic Analysis Workshop 16; GO: Gene ontogeny; GWAS: Genome-wide association study; NARAC: North American Rheumatoid Arthritis Consortium; RA: Rheumatoid arthritis; SNP: Single-nucleotide polymorphism.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

HCVH conceived of the study and participated in its design. JJD-H helped with the statistical analysis. TWJH and REMT participated in the design of the study. JJPL finalized the statistical analysis and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences. The work of JJPL was funded by ZonMW grant 40-00812-98-03016 and the EULAR Gidael project.

This article has been published as part of *BMC Proceedings* Volume 3 Supplement 7, 2009: Genetic Analysis Workshop 16. The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/3?issue=S7.

References

1. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R and Gene Ontology Consortium: **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Res* 2004, **32:**D258–D261.
2. Kanehisa M, Goto S, Kawashima S, Okuno Y and Hattori M: **The KEGG resource for deciphering the genome.** *Nucleic Acids Res* 2004, **32:**D277–D280.
3. Wang K, Li M and Bucan M: **Pathway-based approaches for analysis of genome-wide association studies.** *Am J Hum Genet* 2007, **81:**1278–1283.
4. Lewinger JP, Conti DV, Baurley JW, Triche TJ and Thomas DC: **Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation.** *Genet Epidemiol* 2007, **31:**871–882.
5. Chen GK and Witte JS: **Enriching the analysis of genome-wide association studies with hierarchical modeling.** *Am J Hum Genet* 2007, **81:**317–404.
6. Sohns M, Rosenberger A and Bickeböller H: **Integration of a priori gene set information into genome-wide association studies.** *BMC Proc* 2009, **3(Suppl 7):**S95.
7. Lebrec JJ, Nishchenko I, Wijk van der HJ, Huizinga TW and van Houwelingen HC: **A polygenic model for integration of linkage and pathway information.** *Genet Epidemiol* 2009, **33:**198–206.
8. Amos CI, Chen WV, Lee A, Li W, Kern M, Lundsten R, Batliwalla F, Wener M, Remmers E, Kastner DA, Criswell LA, Seldin MF and Gregersen PK: **High-density SNP analysis of 642 Caucasian families with rheumatoid arthritis identifies two new linkage regions on 11p12 and 2q33.** *Genes Immun* 2006, **7:**277–286.
9. Aidinis V, Carninci P, Armaka M, Witke W, Harokopos V, Pavelka N, Koczan D, Argyropoulos C, Thwin MM, Möller S, Waki K, Gopalakrishnakone P, Ricciardi-Castagnoli P, Thiesen HJ, Hayashizaki Y and Kollias G: **Cytoskeletal rearrangements in synovial fibroblasts as a novel pathophysiological determinant of modeled rheumatoid arthritis.** *PLoS Genet* 2005, **1:**e48.
10. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES and Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci USA* 2005, **102:**15545–15550.
11. Harney SM, Vilariño-Güell C, Adamopoulos IE, Sims AM, Lawrence RW, Cardon LR, Newton JL, Meisel C, Pointon JJ, Darke C, Athanasou N, Wordsworth BP and Brown MA: **Fine mapping of the MHC class III region demonstrates association of AIF1 and rheumatoid arthritis.** *Rheumatology (Oxford)* 2008, **47:**1761–1767.
12. Raychaudhuri S, Remmers EF, Lee AT, Hackett R, Guiducci C, Burtt NP, Gianniny L, Korman BD, Padyukov L, Kurreeman FA, Chang M, Catanese JJ, Ding B, Wong S, Helm-van Mil van der AH, Neale BM, Coblyn J, Cui J, Tak PP, Wolbink GJ, Crusius JB, Horst-Bruinsma van der IE, Criswell LA, Amos CI, Seldin MF, Kastner DL, Ardlie KG, Alfredsson L, Costenbader KH, Altshuler D, Huizinga TW, Shadick NA, Weinblatt ME, de Vries N, Worthington J, Seielstad M, Toes RE, Karlson EW, Begovich AB, Klareskog L, Gregersen PK, Daly MJ and Plenge RM: **Common variants at CD40 and other loci confer risk of rheumatoid arthritis.** *Nat Genet* 2008, **40:**1216–1223.
13. Hutz JE, Kraja AT, McLeod HL and Province MA: **CANDID: A flexible method for prioritizing candidate genes for complex human traits.** *Genet Epidemiol* 2008, **32:**779–790.