Theoretical and Practical Considerations in Estimation of Heritability on Drug Response from Whole Genome and Molecular Signature SNPs[1]

Running title: Heritability Estimations on SNP Data to Predict Drug Response, W Zhang et al

Wencan Zhang[1], Lin Li[2,] Pingye Zhang[3], Yonghong Zhu[1], Ling Wang[1] and Ray Liu[1]
[1]Takeda Development Center, Deerfield, IL 60016, USA
[2]BioStat Solutions 5280 Corp. Dr., Frederick MD 21703, USA
[3]Merck, 90 E Scott Ave, Rahway, NJ 07065, USA

Abstract

We estimated heritability (h^2) on drug response with SNP data. If we define the true portion of available SNPs on variance explained (VE) as h^2_M and then the SNPs potentially can explain all the genetic variation in the trait ($h^2_M <= h^2$). The VE by genome-wide significant (GWS) SNPs (h^2_{GWS}) may satisfy $h^2_{GWS} <h^2_M <= h^2$. However, in model building, we also consider: 1). use only the SNPs from target genes; 2). use the SNPs with less linkage disequilibrium among top SNPs, and 3). use the SNPs with MAF to be > 0.01. Subsequently, only SNPs with non-zero parameter estimates will be used in a 'molecular signature (MS)'. Then VE by MS (h^2_{MS}) may be even smaller than h^2_{GWS} i.e. $h^2_{MS} < h^2_{GWS} <h^2_M <= h^2$. Within the h^2_{MS} , two VEs can be derived: the placebo (PLA, h^2_{PLA}) and treated (TRT, h^2_{TRT}) , for a drug with significant efficacy, the heritability should satisfy $h^2_{PLA} < h^2_{TRT}$ (within h^2_{MS}). Estimates on real data is reported using GCTA method. Estimates of h^2 on whole genome (0.99± 0.72, p< 0.0001) and on MS to predict drug response from discovery (0.009±0.002, p=0.34 with PLA arm and 0.08±0.05, P=0.002 with TRT arm) and validation (0.013±0.015, p=0.13, with TRT arm) data. To find the missing heritability, further modeling work with higher SNP density, such as WGS, much more SNPs with less stringent p-values and low MAP in the selected model, and other genomic data, such as RNA –seq, and additional clinical information should be explored.

[1] Submitted to The pharmacogenomic Journal.

Introduction

The field of pharmacogenomics is focused on the characterization of genetic factors contributing to the response of patients to pharmacological interventions. In genome-wide association studies (GWAS) of conventional complex traits such as human complex diseases and drug response, a fundamental and yet unsolved question is that of so-called "missing heritability", *i.e.*, the significant and often numerous variants collectively explaining only a small fraction of the total phenotypic variation [1-8]. For example, recent studies show that ~50 variants explain only ~5% of the phenotypic variation for human height, a highly heritable trait with narrow sense heritability of ~80% [5]. While fully resolving the missing heritability remains a challenging task. Shigemizu et al (2014) with real type 2 diabetes data, reported that the best lasso model with cross validated 9 SNPs combined with the clinical factors had 0.073 (0.91%) more AUC than the model with only clinical factors (0.8057 vs 0.7984) [9]. Yang et al. (2011) [10] developed the GCTA software package to estimate the additive genetic variance for a trait using genome-scale single nucleotide polymorphism (SNP) data. This method first estimated relatedness with many thousands of markers and then using estimated relatedness to estimate the additive genetic variance of a trait. If the QCed SNPs adequately capture the relationships among individuals at causative alleles, h^2 with GCTA methodology is equivalent to narrow-sense heritability [10]. Yang el at (2011) shown that the variance explained by each chromosome is proportional to its length, and that SNPs in or near genes explain more variation than SNPs between genes [11]. GCTA also has been used to estimate whole genome heritability in susceptibility to schizophrenia (h^2=0.23)[12], human intelligence (h^2=0.40-0.51) [13], and personality (h^2=0.7)[14]. Zhou et al (2014) estimated heritability of response to metformin (h^2=0.34) [15] . Mirkov, M. U et al (2014) estimated heritability of different outcomes for genetic studies of TNFi response in patients with rheumatoid arthritis (h^2=0.59-0.87) [16]. Li, Q., et al (2017) estimated heritability of Clinical response to the atypical antipsychotic paliperidone in schizophrenic patients (h^2=0.31-0.43)[17].

Pharmacogenomic biomarkers can optimize an individual's therapy; however, the overall role of genetic factors in drug response remains uncertain. The majority of

genetic variants currently used as clinical pharmacogenomic biomarkers affect drug metabolism and transport while fewer biomarkers accurately predict drug response (pharmacodynamics) [18], although a lot of drug response biomarker genomic predictive signature have been developed [19-22]. Siebert et al (2016), reported comparative validation of predictions, the heritability estimates of treatment response to RA disease on the whole genome (h^2=0.36, p=0.05) and on TNF/TNFR pathway with 333 genes (h^2=0.02, p=0.3) from Infliximab+ Adalimumab TRT arm [23]. Heritability estimation on SNP predictive genomic signature on drug response for common disease has not been published in the literature. The objectives of this proposed study are in two folds: 1) Some theoretical considerations on the heritability of molecular signature. 2). Estimations of heritability on whole genome and a molecular signature to predict drug response both in discover and in independent validation cohorts from real genomic and clinical data sets.

Some theoretical considerations

Heritability (h^2) is a genetic parameter used in breeding and genetics works that estimates how much variation in a phenotypic trait in a population is due to genetic variation among individuals in that population. Other causes of measured variation in a trait are characterized as environmental factors, including measurement error. The definition of genomic heritability (h^2_M) is the proportion of variance that can be explained by a linear regression on a massive number of markers [1]. If we define the portion of all available genetic markers explained as h^2_M and then the SNPs potentially can explain all the genetic variation in the trait ($h^2_M \leq h^2$). The difference between the variance explained by genome-wide significant (GWS) SNPs (h^2_{GWS}) and heritability estimate from family studies (h^2) has been called the "missing heritability" and the difference between h^2_{GWS} and h^2_M is called the "hidden" heritability [3]:

$$h^2_{GWS} < h^2_M < h^2 \quad \text{(Wray et al. 2013)} \quad (1)$$

However, in prediction of drug response, we usually need to build up a prediction model (a composite score) from top SNPs identified by GWAS. In model building , we also consider: 1). use only the SNPs from target genes, instead of all SNPs appeared to be statistically significant; 2). use only the SNPs with less LD between top SNPs, and 3). use only the SNPs with MAF to be > 0.01 et al. Since it is not expected that all biomarkers are important to define subgroup membership, the model above may be fit using elastic net [24], which inherently performs feature selection through penalized regression. For biomarkers that are not contributing significantly to the composite score, the corresponding parameter estimates will be zero in the composite score calculation and subsequently only biomarkers with non-zero parameter estimates will be used in the definition of a 'molecular signature (MS)'. For clinical and commercial applications, the molecular signatures usually have few variants.

Then the heritability of the MS should be defined as h^2_{MS} and h^2_{MS} should be even smaller than h^2_{GWS} with a relationship of:

$$h^2_{MS} < h^2_{GWS} < h^2_{M} \leq h^2 \qquad (2)$$

A success molecular signature (MS) should be able to predict the potential drug TRT efficacy and identify the subpopulation with higher response rate. Within the h^2_{MS} , two heritability can be derived: the placebo (PLA, h^2_{PLA}) and treatment (TRT, h^2_{TRT}) , for a drug with significant efficacy, the heritability of TRT cohort should satisfy:

$$h^2_{PLA} < h^2_{TRT} \text{ (within } h^2_{MS}) \qquad (3)$$

Materials and Methods

Data sets

Clinical data: Two phase 3 clinical study data with 1,066 patients were divided into discovery (313 TRT with a marketed drug and 253 with PLA) and validation (an independent data set with 500 patients treated with the drug). Due the original clinical study design, the PLA arm for validation was not available. A continues change from baseline score to measure a non-cancer treatment efficacy variable was used as phenotype in the heritability estimation.

Genomic data: a total of 1,066 samples with PGx informed consent were tested with Illumina OmniExpress chip with ~ 1Million SNP variants.

Data for heritability estimation: An 11 SNP model was developed using a two step approach (see appendix 1) on the change from baseline on the efficacy score which was used in define the responder/non-responder status. Heritabilities were estimated with the whole genome (56,511 SNPs after QC and LD pruning with r^2 LD < 0.1) and the signature (11 SNPs) in discovery data sets (313 TRT and 253 PLA) and an independent validation arm (500 patients in a TRT arm).

Predictive Model (molecular signature) Building

A 11 SNP model was developed using a two step approach: first select top SNPs on univeriate test p-value and second using elastic net [21] regression methodologies with the available clinical and genomic data (a continues index for drug response) to build up the prediction model with 11 SNPs and coefficients. Details of the two-stage predictive model (molecular signature) development for subgroup identification are in appendix 1.

Statistical Method for heritability estimation

We used the program GCTA (Yang et al. 2011) [10] to estimate the proportion of phenotypic variance explained by genotyped SNPs. The GCTA analysis consists of two steps. First, all SNPs are used to calculate the genetic relationship matrix (GRM) among accessions. GCTA uses the accessions included in the analysis as the base population for defining relatedness, such that the

average relatedness between all 'unrelated' pairs of accessions (off-diagonals of GRM) is zero. The GRM is then used as a predictor in a mixed linear model with a trait as the response to estimate h^2. The GCTA method estimates the proportion of additive genetic variance for a trait and thus narrow-sense heritability. Note that the top three principal components derived from the PCA outlined in as well as trial, age, gender, smoking status, and alcohol usage were included in all statistical models as fixed effects.

Results and discussion

The estimations of heritability of data set 1 on whole genome are reported in table 1. Population, sample size, number of SNPs, heritability, standard error and non-zero test p-values are reported. Heritability of the whole genome (with 56,511 SNPs) is estimated in discovery (TRT arm, N=313), discovery (PLA arm, N=253) and validation (TRT, N=500) separately. The estimates are between 76.6 %(PLA) to 99.99 % (TRT) in both discovery and validation arms.

With values between 76.6% and 99.99%, it seems the estimations of the heritability on the drug response in current study from the whole genome on 56,522 SNPs are over estimated. Kumar et al (2016), had quested GCTA's estimates of heritability. GWAS data are necessarily overfit by GCTA and produces high estimates of heritability [25]. In this study, the sample size from the clinical studies are within 500 and the statistical power of the estimation is limited which may explain the over estimated heritability from the whole genome.

Table 1. Heritability estimates from the whole genome on discovery and validation populations

Population	N	Number of SNPs	h^2 %	SE(h^2) %	p-value
Discovery (TRT)	313	56,511	99.99	71.9	0.0005
Discovery (PLA)	253	56,511	76.60	124.2	0.29
Validation (TRT)	500	56,511	99.99	57.8	0.014

The estimations of the heritability on the 11 SNP signature are shown on the table 2. The SNPs were derived through a very complicated statistical procedure (appendix 1). The values are between 0.93% and 8.17% for the discovery cohorts and 1.28% in an independent validation cohort. The discovery cohorts were used to derive the 11 SNP

signature. According to Wray et al (2013), the number one pitfall of predicting complex traits from SNPs is applying the incorrect validation procedure results in over-estimation of the accuracy of the prediction (or overfitting) [3]. Our estimated heritability of 0.93%(PLA) and 8.17%(TRT) from the discovery cohorts for the molecular signature are over estimated for this 11 SNP signature. However, the estimate of 1.28 % from the TRT arm of the validation data set should be an independent one, although it is not statistically significant (p=0.13, table 2). This is likely the true phenotypic variance explained due to the genetic contribution from the 11 SNP signature.

Table 2. Heritability estimates for an 11 SNP molecular signature on discovery and validation populations

Population	N	Number of SNPs	h^2 %	SE(h^2) %	p-value
Discovery(TRT)	313	11	8.17	4.92	0.002
Discovery (PLA)	253	11	0.93	2.30	0.34
Validation (TRT)	500	11	1.28	1.53	0.13

On the other hand, the results show that the number of SNPs should be in the order of hundreds to thousands or even more to allow meaningful representation of the whole genome joint contribution of a particular TRT effect in the predictive model, instead of dozens of SNPs which is a extension of the single gene model from Mendelian genetics.

Results of this study has confirmed findings Siebert et al (2016) [23]. Yang et al [5] found using the estimation of heritabilities on human height are proportional to the percentages of genome covered in the data and concluded the "most of the heritability is not missing but has not previously been detected because the individual effects are too small to pass stringent significance tests and remaining heritability is due to incomplete linkage disequilibrium between causal variants and genotyped SNPs, exacerbated by causal variants having lower minor allele frequency than the SNPs explored to date".

Based on a 50K Illunina BovineSNP50 BeadChip [27] or alike and combined with the information from traditional progeny test, genomic selection has doubled the improved rate in Canadian dairy cattle cows in LPI (lifetime Profit Index) [28] over a period of 2009-2016. The theory was based on an infinitesimal model, all 50K SNPs are used in the estimate of genomic EBV (estimated breeding value) without any hypothesis test to reduce the dimensions and with the methodologies developed Meuwissen et al (2001) [29] and VanRaden (2008) [30]. Although pharmacogenomics may have a different mechanism of action than dairy genetics, we can learn

something and borrow some ideas. If we believe the drug response for common disease is affected by hundreds even thousands of genomic variants with small effects, the current two-stage predictive model approach such as the one used in the development of the 11 SNP model in this paper (appendix 1) may be too stringent to include most causal variants. Significant increase the number of SNPs in the models is recommended. Higher density coverage of the genome, such as WGS with SNPs within disease related genes, drug metabolism genes and pathways should be considered.

Conclusions

In theory, we showed the variance explained by SNP markers from molecular signature, genome wise significance and whole genome should satisfy: $h^2_{MS} < h^2_{GWS} < h^2_{M} \leq h^2$ and within h^2_{MS}, the variance explained by the placebo arm and the TRT arm should satisfy: $h^2_{PLA} < h^2_{TRT}$. With actual data, we found that the heritability for the whole genome were between 76.6% and 99.99% and believed to be over estimated. The heritabilities for an 11 SNP molecular signature were between 0.93% (PLA) and 8.17% (TRT) in discovery cohort and 1.28% in an independent validation cohort. Predictive drug response modeling with SNP data may have limited power with lower heritability found from this 11 SNP signature. Further modeling work with higher SNP density, such as WGS, much more SNPs to cover the whole genome, and other genomic data, such as RNA –seq, and additional clinical information should be explored to address the missing heritability and overall prediction power in predictive modeling on drug response.

Acknowledgements

Useful discussions with Dr. Zheng Zha at Takeda Pharmaceutical Develop Center are highly appreciated.

Conflict of Interest

The disease, drug name, indication and clinical study involved are not mentioned in the manuscript to avoid conflict of interests. Dr. Lin Li is a consultant from Biostat Solutions Inc. Dr. Pingye Zhang was a summer intern at Takeda develop center at Deerfield, IL. USA. All other authors were Takeda employees at the time. There is no conflict of interests.

No writing assistance was utilized in the production of this manuscript.

Reference

1. Gustavo de los Campos, Daniel Sorensen, and Daniel Gianola. Genomic Heritability: What Is It? PLOS Genetics | DOI:10.1371/journal.pgen.1005048 May 5, 2015

2. . S. J. Schrodi, S. Mukerjee, Y. Shan et al. Genetic-based prediction of disease traits: prediction is very difficult, especially about the future. FrontiersinGenetics June (2014) Volume5 Article162 .2

3. Naomi R. Wray, Jian Yang, Ben J. Hayes, Alkes L. Price, Mike E. Goddard, and Peter M. Visscher. Pitfalls of predicting complex traits from SNPs. Nat Rev Genet. 2013 July ; 14(7): 507–515. doi:10.1038/nrg3457.

4. Sang Hong Lee, Naomi R. Wray, Michael E. Goddard, and Peter M. Visscher.
Estimating Missing Heritability for Disease from Genome-wide Association Studies. The American Journal of Human Genetics 88 2011: 294–305, March 11, 2011.

5. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42, 565–569.

6. Visscher, P.M., Yang, J., and Goddard, M.E. (2010). A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). Twin Res. Hum. Genet. 13, 517–524.

7. G. SY Pang, Jingbo Wang, Zihua Wang and C. GL Lee. Predicting potentially functional SNPs in drug-response genes. Phamacogenomics (2009) 10(4), 639-653

8. Y. W. Francis Lam. Scientific Challenges and Implementation Barriers to Translation of Pharmacogenomics in Clinical Practice. ISRN Pharmacology Volume 2013.

9. Daichi Shigemizu, Testuo Abe, Takashi Morizono et al The Construction of Risk Prediction Models Using GWAS Data and Its Application to a Type 2 Diabetes Prospective Cohort. PLoS ONE March 2014 Volume 9 Issue 3 e9254.

10. Yang Jian, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. GCTA: A tool for Genome-wide Complex Trait Analysis. The American Journal of Human Genetics 2011: 88, 76-82.

11. Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M. et al. Genome partitioning of genetic variation for complex traits using common SNPs. Nature Genetics 2011: 43, 519–525.

12. Lee, S.H., DeCandia, T.R., Ripke, S., Yang, J., Sullivan, P.F., Goddard, M.E., Keller, M.C., Visscher, P.M. & Wray, N.R. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nature Genetics 2012: 44, 247–250

13. Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S.E., Liewald, D. et al. Genome-wide association studies establish that human intelligence is highly heritable and polygenic. Molecular Psychiatry 2011: 16, 996–1005.

14. Verweij, K.J.H., Yang, J., Lahti, J., Veijola, J., Hintsanen, M., Pulkki-Raback, L. et al. Maintenance of genetic variation in human personality: testing evolutionary models by estimating heritability due to common causal variants and investigating the effect of distant inbreeding. Evolution 2012: 66, 3238–3251.

15. Zhou, K., Donnelly, L., Yang, J., Li, M., Deshmukh, H., Van Zuydam, N., ... & Colhoun, H. M. (2014). Heritability of variation in glycaemic response to metformin: a genome-wide complex trait analysis. 2014.*The Lancet Diabetes & Endocrinology*, *2*(6), 481-487.

16. Mirkov, M. U., Janss, L., Vermeulen, S. H., van de Laar, M. A., van Riel, P. L., Guchelaar, H. J., ... & Coenen, M. J. (2014). Estimation of heritability of different outcomes for genetic studies of TNFi response in patients with rheumatoid arthritis. 2014. *Annals of the rheumatic diseases*, annrheumdis-2014.
https://pdfs.semanticscholar.org/83c8/29395109810ec5f1060e53071087339a3184.pdf

17. Li, Q., Wineinger, N. E., Fu, D. J., Libiger, O., Alphs, L., Savitz, A., ... & Schork, N. J. (2017). Genome-wide association study of paliperidone efficacy. 2017. *Pharmacogenetics and Genomics*, *27*(1), 7.

18. Wolfgang Sadee. Relevance of 'missing heritability' in pharmacogenomics. Clin Pharmacol Ther. 2012 October ; 92(4): 428–430.

19. C. Lee Ventola, MS. Role of Pharmacogenomic Biomarkers In Predicting and Improving Drug Response. Part 1: The Clinical Significance of Pharmacogenetic Variants. P T. 2013 Sep; 38(9): 545-551, 558-560.

20. Ashraf G. Madian, Heather E. Wheeler, Richard Baker Jones, and M. Eileen Dolan Relating Human Genetic Variation to Variation in Drug Responses. Trends Genet. 2012 October ; 28(10): 487–495.

21. QS Li, C Tian, GR Seabrook, WC Drevets and VA Narayan. Analysis of 23andMe antidepressant efficacy survey data: implication of circadian rhythm and neuroplasticity in bupropion response Citation: Transl Psychiatry (2016) 6, e889;

22. Chhibber, A. et al. Genomic architecture of pharmacological efficacy and adverse events. Pharmacogenomics 2014: 15 (16), 2025–2048.

23. Solveig K. Sieberts , Fan Zhu, Javier Garcı´a-Garcı´a, Eli Stahl, Abhishek Pratap , Gaurav Pandey, et al. Crowdsourced assessment of common genetic contribution to predicting anti-TNF TRT response in rheumatoid arthritis. NATURE COMMUNICATIONS 2016 Aug.7:12460.

24. Zou H. and Trevor, T. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society 2005; Series B,* 67(2), 301-320.

25. Siddharth Krishna Kumara, Marcus W. Feldmana , David H. Rehkopfb , and Shripad Tuljapurkara. Limitations of GCTA as a solution to the missing heritability problem. PNAS February 9, 2016 ; vol. 113 |no. 6 E813

26. Li L, Guennel T, Marshall SL, Cheung LWK. A multi-marker molecular signature approach for TRT-specific subgroup identification with survival outcomes. *The Pharmacogenomics Journal,* 2014; 14(5): 439-45.

27. Illumina. BovineSNP50 Genotyping BeadChip. Feb. 2016.

28. Lynsay Beavers **and** Brian Van Doormaal. Genetic Gain Before and After Genomics. Farms.com, May 26, 2017.

29. T. H. E. Meuwissen, B. J. Hayes and M. E. Goddard. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. April. 2001. Genetics 157: 1819–1829.

30. VanRaden P. Efficient methods to compute genomic predictions. Journal of dairy science. 2008;91:4414–4423.