CELL CYCLE
2016, VOL. 15, NO. 1, 22–24
http://dx.doi.org/10.1080/15384101.2015.1120928

Taylor & Francis
Taylor & Francis Group

PERSPECTIVE

Reducing GWAS Complexity

Dennis J. Hazelett[a], David V. Conti[b], Ying Han[b], Ali Amin Al Olama[c], Doug Easton[c], Rosalind A. Eeles[d], Zsofia Kote-Jarai[d], Christopher A. Haiman[b], and Gerhard A. Coetzee[b,d,†]

[a]Bioinformatics and Computational Biology Research Center, Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, CA, USA; [b]Departments of Preventive Medicine and Urology, USC/Norris Cancer Center, Los Angeles, CA, USA; [c]Division of Genetics & Epidemiology, Centre of Cancer Genetic Epidemiology, University of Cambridge, Cambridge, UK; [d]The Institute of Cancer Research & Royal Marsden NHS Foundation Trust, London, UK

ABSTRACT

Genome-wide association studies (GWAS) have revealed numerous genomic 'hits' associated with complex phenotypes. In most cases these hits, along with surrogate genetic variation as measure by numerous single nucleotide polymorphisms (SNPs) that are in linkage disequilibrium, are not in coding genes making assignment of functionality or causality intractable. Here we propose that fine-mapping along with the matching of risk SNPs at chromatin biofeatures lessen this complexity by reducing the number of candidate functional/causal SNPs. For example, we show here that only on average 2 SNPs per prostate cancer risk locus are likely candidates for functionality/causality; we further propose that this manageable number should be taken forward in mechanistic studies. The candidate SNPs can be looked up for each prostate cancer risk region in 2 recent publications in 2015[1,2] from our groups.

ARTICLE HISTORY
Received 19 August 2015
Accepted 12 November 2015

KEYWORDS
GWAS; SNP; Cancer; chromatin; non-coding DNA; enhancer; fine-mapping

Genome-wide association studies (GWAS) of complex phenotypes became more and more powerful as sample sizes of cases and controls increased and meta-analyses were employed. Also, as next generation sequencing techniques became more feasible and increasingly affordable, more and more single nucleotide polymorphisms (SNPs) with lower and lower minor allele frequencies (MAFs) have been identified. Thus, association signals at any given locus have become increasingly complex in large part due to the many candidate risk SNPs, correlated with each other due to linkage disequilibrium (LD). Consequently, it is virtually impossible to assign functionality, let alone causality, to any given SNP at a risk locus. This dispiriting situation is only made more daunting by the unexpected finding that 90% or more of these risk SNPs are located in non-coding DNA.

To address these issues, we and others have used chromatin biofeatures to inform potential functionality on the original discovery SNPs, known to the field as index SNPs, and their many surrogate SNPs, the former revealed by GWAS and the latter defined by r^2 of population-specific LD. Thus, software tools such as FunciSNP,[3] RegulomeDB,[4] Haploreg,[5] Annovar,[6] IGV,[7] and more recently FunSeq[8] and motifbreakR[9] were developed to utilize correlated risk SNPs from the 1000 Genomes Project, co-locating with chromatin annotations (such as obtained from ChIP-seq and nucleosome occupancy data); this significantly reduces the number of candidate functional SNPs. This became necessary since the SNP surrogates were plentiful (for example, there are on average ~500 per prostate cancer risk region at $r^2 \geq 0.5$ to the index SNP - Table 1). Over the last couple of years, we have successfully used this annotation approach for prostate[10] and breast[11] cancer risk regions. Despite the significant reduction in candidate functional SNPs using this approach, the on average ~10 (median = 5) candidate SNPs per prostate cancer risk locus still make a detailed and comprehensive wet-lab analysis of functionality intractable (Table 1). However, more recently a new approach emerged, known as fine-mapping.

With the advent of fine-mapping strategies and corresponding analysis methods, the complexity at any given locus or region can be reduced further. Many new analysis aproaches have been developed to go beyond simple rankings of marginal p-values to statistically identify a putative set of candidate SNPs for further functional analysis. These methods include Bayesian and re-sampling approaches that formally incorporate the uncertainty in estimation and ranking,[1,12,13] model selection approaches to condition on multiple SNPs in the region using either individual-level data[2,14] or marginal test statistics,[15,16] and approaches that formally incorporate prior information, such as SNP annotation, into the final inference.[17,18]

Recently, we and others have fine-mapped prostate cancer risk regions using a multi-ethnic[1] and a single large European population.[2] This reduced the candidate risk SNPs per region (Table 1). Coupled with FunciSNP[3] annotated functionality, the 2 fine mapping studies further reduced the number of common candidates to on average only about 2 SNPs per region (Table 1). This clearly means that without functional annotation or fine-mapping a significant number of false positives at each locus may well lead to non-productive functional analyses that have little to do with risk. Ultimately SNPs must be functional to be causal, but not all functional SNPs are inevitably involved in risk.

CONTACT Gerhard A. Coetzee ✉ coetzee@usc.edu Microbiology and Preventive Medicine, USC/Norris Cancer Center, NTT 6411, Keck School of Medicine, 1441 Eastlake Ave Los Angeles, CA, 90089-9176, USA.

[†]Present address: Van Andel Institute, Grand Rapids, MI, 49503, USA.

Table 1. SNPs per PCa risk region.

Selection Criteria	Average (Median) # of SNPs/region
All SNPs (1K) at 1MB/region	∼50 thousand (MAF $\geq$ 1%)[10]
SNPs $r^2 \geq 0.5$ with index SNP	∼500[10]
SNPs $r^2 \geq 0.5$ in biofeatures	9.8(5)[10]
SNPs significantly fine mapped	12(6)[1] – 22(13)[2]
SNPs significantly fine mapped AND in biofeatures	3.5(2)[1] – 5.4 (3)[2] ∼2 common in the 2 independent studies

Grant support: R01 CA 136924, R01 CA 190182 & UO1 CA 184826

However, even with fine-mapping data at hand, downstream analyses will be complicated when multiple biologically functional SNPs with measurable allelic effects are present at a single locus, of which only one is the major driver of cancer. Also plausible are haplotypes with 2 or more variants partially complementing each other. Finally, an additional complication is that the etiological target tissue of a given risk locus may lie outside the tissue of origin for a particular disease. To give an example, risk for prostate cancers (and others) are likely affected by immune cells in addition to dysregulation in prostate epithelia. This may explain the roughly 1/3 of risk loci where no functionality may be assigned based upon extant epigenomics data in prostate epithelial cell types.[10] To give another example, there may be several plausible tissues of origin, which may be at least partially addressed by assessing candidate loci for enrichment in the epigenomic features of those various tissues. We showed for serous ovarian cancers that finemapped risk SNPs were enriched in regulatory sites of immortalized primary cell lines derived from fallopian tube serous epithelium (FTSE) over those of similar cell lines derived from ovarian serous epithelia, supporting FTSE as the tissue of origin.[19] While this by no means settles the debate on such questions, it may provide an important clue as to the tissue of origin and suggest therapeutic targets for prevention and early intervention measures. Going forward, we anticipate that comparison among multiple candidate tissues (or all known cell types) will become standard in the field. Thus, in the midst of so much uncertainty, it is essential that biological assays give repeatable and reliable measures of these complex interactions. It must also be stressed that functional annotations alone are clearly not comprehensive and that other, as yet unknown, chromatin-related functions likely are to be considered in the future. The ultimate goal would be to identify the true functional/causal SNP (or allele with more than 1 SNP or even multiple causal alleles) at every risk locus. Attainment of this goal may require development of specific assays designed to measure the allelic affect on cancer processes once the variant functionality vis-a-vis the epigenome (or other) has been verified. The eventual utility from this will be the compilation of more informative nomograms for risk assessments and the identification of risk mechanisms.

Disclosure of potential conflicts of interest

No potential conflicts of interest were disclosed.

References

1. Han Y, Hazelett DJ, Wiklund F, Schumacher FR, Stram DO, Berndt SI, Wang Z, Rand KA, Hoover RN, Machiela MJ, et al. Integration of Multiethnic Fine-mapping and Genomic Annotation to Prioritize Candidate Functional SNPs at Prostate Cancer Susceptibility Regions. Hum Mol Genet 2015; 24(19):5603–18.
2. Amin Al Olama A, Dadaev T, Hazelett DJ, Li Q, Leongamornlert D, Saunders EJ, Stephens S, Cieza-Borrella C, Whitmore I, Benlloch Garcia S, et al. Multiple novel prostate cancer susceptibility signals identified by fine-mapping of known risk loci among Europeans. Hum Mol Genet 2015; 24(19):5589–602.
3. Coetzee SG, Rhie SK, Berman BP, Coetzee GA, Noushmehr H. FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. Nucleic Acids Res 2012; 40:e139; PMID:22684628; http://dx.doi.org/10.1093/nar/gks542
4. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 2012; 22:1790–7; PMID:22955989; http://dx.doi.org/10.1101/gr.137323.112
5. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res 2012; 40:D930–4; PMID:22064851; http://dx.doi.org/10.1093/nar/gkr917
6. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010; 38:e164; PMID:20601685; http://dx.doi.org/10.1093/nar/gkq603
7. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat Biotechnol 2011; 29:24–6; PMID:21221095; http://dx.doi.org/10.1038/nbt.1754
8. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. Science 2013; 342:1235587; PMID:24092746; http://dx.doi.org/10.1126/science.1235587
9. Coetzee SG, Coetzee GA, Hazelett DJ. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. Bioinformatics 2015; 31:3847-9; PMID:26272984; http://dx.doi.org/10.1093/bioinformatics/btv470
10. Hazelett DJ, Rhie SK, Gaddis M, Yan C, Lakeland DL, Coetzee SG, Ellipse G-ONc, Practical C, Henderson BE, Noushmehr H, et al. Comprehensive functional annotation of 77 prostate cancer risk loci. PLoS Genet 2014; 10:e1004102; PMID:24497837; http://dx.doi.org/10.1371/journal.pgen.1004102
11. Rhie SK, Coetzee SG, Noushmehr H, Yan C, Kim JM, Haiman CA, Coetzee GA. Comprehensive functional annotation of seventy-one breast cancer risk Loci. PLoS ONE 2013; 8:e63925; PMID:23717510; http://dx.doi.org/10.1371/journal.pone.0063925
12. Wakefield J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. Am J Hum Genet 2007; 81:208–27; PMID:17668372; http://dx.doi.org/10.1086/519024
13. Wakefield J. Bayes factors for genome-wide association studies: comparison with P-values. Genet Epidemiol 2009; 33:79–86; PMID:18642345; http://dx.doi.org/10.1002/gepi.20359
14. Wallace C, Cutler AJ, Pontikos N, Pekalski ML, Burren OS, Cooper JD, Garcia AR, Ferreira RC, Guo H, Walker NM, et al. Dissection of a Complex Disease Susceptibility Region Using a Bayesian Stochastic Search Approach to Fine Mapping. PLoS Genet 2015; 11:e1005272; PMID:26106896; http://dx.doi.org/10.1371/journal.pgen.1005272
15. Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. Identifying causal variants at loci with multiple signals of association. Genetics 2014; 198:497–508; PMID:25104515; http://dx.doi.org/10.1534/genetics.114.167908
16. Chen W, Larrabee BR, Ovsyannikova IG, Kennedy RB, Haralambieva IH, Poland GA, Schaid DJ. Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. Genetics 2015; 200:719–36; PMID:25948564; http://dx.doi.org/10.1534/genetics.115.176107
17. Kichaev G, Yang WY, Lindstrom S, Hormozdiari F, Eskin E, Price AL, Kraft P, Pasaniuc B. Integrating functional data to prioritize causal variants

in statistical fine-mapping studies. PLoS Genet 2014; 10:e1004722; PMID:25357204; http://dx.doi.org/10.1371/journal.pgen.1004722

18. Quintana MA, Conti DV. Integrative variable selection via Bayesian model uncertainty. Statistics in medicine 2013; 32:4938–53; PMID:23824835; http://dx.doi.org/10.1002/sim.5888

19. Coetzee SG, Shen HC, Hazelett DJ, Lawrenson K, Kuchenbaecker K, Tyrer J, Rhie SK, Levanon K, Karst A, Drapkin R, et al. Cell-type-specific enrichment of risk-associated regulatory elements at ovarian cancer susceptibility loci. Hum Mol Genet 2015; 24:3595–607; PMID:25804953