Leveraging putative enhancer-promoter interactions to investigate two-way epistasis in Type 2 Diabetes GWAS

Elisabetta Manduchi

Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, 3700 Hamilton Walk, Philadelphia, PA, 19104, USA and Division of Human Genetics and Endocrinology, The Children's Hospital of Philadelphia, 3615 Civic Center Boulevard, Philadelphia, PA 19104, USA
Email: manduchi@pennmedicine.upenn.edu

Alessandra Chesi

Division of Human Genetics and Endocrinology, The Children's Hospital of Philadelphia, 3615 Civic Center Boulevard, Philadelphia, PA 19104, USA
Email: chesia@email.chop.edu

Molly A. Hall

Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, 3700 Hamilton Walk, Philadelphia, PA, 19104, USA
Email: hallma@mail.med.upenn.edu

Struan F.A. Grant

Division of Human Genetics and Endocrinology, The Children's Hospital of Philadelphia, 3615 Civic Center Boulevard, Philadelphia, PA 19104, USA
Email: grants@email.chop.edu

Jason H. Moore*

Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, 3700 Hamilton Walk, Philadelphia, PA, 19104, USA
Email: jhmoore@exchange.upenn.edu

We utilized evidence for enhancer-promoter interactions from functional genomics data in order to build biological filters to narrow down the search space for two-way Single Nucleotide Polymorphism (SNP) interactions in Type 2 Diabetes (T2D) Genome Wide Association Studies (GWAS). This has led us to the identification of a reproducible statistically significant SNP pair associated with T2D. As more functional genomics data are being generated that can help identify

* Correspondence

potentially interacting enhancer-promoter pairs in larger collection of tissues/cells, this approach has implications for investigation of epistasis from GWAS in general.

Keywords: epistasis; GWAS; type 2 diabetes; enhancers; genetic encoding.

1. Introduction

In the context of GWAS, epistasis refers to SNP interactions associated with a particular complex trait. There are different schools of thought regarding the role of epistasis in human genetics. The majority of human GWAS to date have focused on detecting main effects, i.e. single SNP associations. This is naturally the first aspect to explore, and many would argue that this is the most relevant since the vast majority of SNP contributions to a given trait are largely additive.[1] However, a different opinion is that epistasis is a non-negligible component of human genetic architecture, possibly accounting for the predicted 'missing heritability',[2-3] based on the observations that biological systems are regulated by complex biomolecular networks and epistasis plays an important role in model organisms.[4]

Exploring epistasis in a typical GWAS is complicated by several factors. One major challenge arises from the large search space and subsequent limited computational and statistical power. Genes and regulatory elements typically form complex networks, thus epistatic interactions could well involve several SNPs. But, even if one wanted to focus on two-way interactions on genotyped SNPs, a typical GWAS involves at least half a million such markers, so the number of possible pairs is greater than 10^{11}. Another difficulty is that different SNPs relevant to a complex trait may have different mechanisms of action so the choice of genetic encoding affects interaction detection (Hall *et al.*, in preparation). Given all of this, it is therefore uncommon to find statistically significant epistatic interactions in a GWAS data set. Moreover, it is even less common for such observations to be reproduced in a replication data set.

In this work we illustrate an example where the combination of suitable biological filters and a data-driven weighted encoding approach has led us to a statistically significant pairwise interaction in a T2D discovery data set which we were able to go on to replicate in an independent data set. Statistical epistasis is different from biological epistasis[5-7] so establishing whether or not this pair corresponds to an actual biological mechanism associated with T2D will require additional experimentation. However we are using this example to highlight a possible avenue of epistasis investigation which exploits the increasing availability of functional genomics data sets aimed at exploring regulatory and physical interactions among genomic features, such as ChIP-Seq or high-throughput chromatin capture data sets. Our approach is illustrated in the next section.

2. Filters and Encodings

Figure 1 outlines our workflow, which involves two main components:

2.1. *Defining biological filters based on functional genomics data*

Due to the large search space, the first step in an epistasis analysis is to reduce the number of models (i.e. candidate interacting SNP sets) to analyze. To this end, both computational and biological filter approaches have been previously proposed in the literature.

Computational approaches include methods such as ReliefF and its derivatives,[8-10] MDR,[11] and "greedy" approaches which first identify SNPs with main effects (significant or marginally significant) in a GWAS and then use models involving only these SNPs.[12-13] The latter is certainly a reasonable approach; however it will miss potential interactions which involve SNPs with no main effects (i.e. it will miss what is referred to as 'pure and strict epistasis'[14]).

Biological filters may exploit biological annotations (derived from curation of low or high-throughput experiments) or analyses of functional genomics data sets to reduce the search space. For example, knowledge about the biological relevance of the Ras/MAPK pathway to Autism Spectrum Disorders has been used to limit the search space of SNP pairs analyzed for interactions to those where one of the SNPs is in a Ras/MAPK pathway gene.[15] In this work we too use biological filters, but of a different type as described below.

We sought to exploit the increasing availability of functional genomics data sets elucidating genomic features with likely regulatory functions in different tissues and cell lines. This was motivated by the recognized importance of regulatory networks in genomic studies.[16-17] The regulation of gene expression is complex, but a fundamental component lies in enhancers, i.e. non-coding regions in the genome which may affect the expression of distal genes through chromatin looping. We reasoned that natural candidates for two-way interactions are SNP pairs from interacting enhancer-promoter regions in tissues or cell lines relevant to the trait being studied. Based on this, we have selected appropriate interacting regions from the EnhancerAtlas.[18] This resource provides collections of enhancer-gene interactions for several tissues and cell lines, derived from the integration of almost 4000 high throughput experimental data sets from resources including the UCSC genome browser,[19] NCBI GEO,[20] Cistrome database,[21] ENCODE project data portal,[22] Epigenome Roadmap data portal[23] and eRNA.[24]

2.2. *Using weighted encoding*

When we consider a single SNP, we want to encode the biological action in the way that it likely functions, so if a genotype has no alternate alleles, the risk would be 0 and if it has two such alleles the risk would be 1. For a heterozygous genotype, coding it as recessive assumes it has no risk (equal to homozygous referent), coding it as dominant assumes it yields full risk (the same as two alternate alleles), and coding it as additive is right in between. Yet in biology, a heterozygous genotype may act anywhere in this range from recessive to dominant. This has been heavily discussed in the literature for single SNP associations and the consensus has been that additive encoding will capture the largest amount of genetic effects.[25]

PLATO software[26] (https://ritchielab.psu.edu/files/RL_software/plato-manual-2.1.pdf) allows for different choices of encoding. One of them is a data-driven approach to compute an appropriate SNP-specific encoding weight for the heterozygous genotype. In order to describe the latter, we first need to define what is meant by 'codominant' encoding. As described in the

PLATO software manual, in this encoding each marker uses two variables as a dummy encoding; the "Het" variable is 1 only when the marker is heterozygous, and the "Hom" variable is 1 only when the marker is homozygous alternate. In weighted encoding, for each marker, the result from a univariate model (with appropriate covariates) is used to determine an encoding from marker state to the set {0, x, 1}, where x is chosen such that the model with the encoded allele is identical to the codominant model. Data-driven weighted encoding was tested on simulated data sets spanning a comprehensive array of underlying interactions of genetic models, concluding that it had a better performance than the other encodings based on a combination of power and type I error (Hall *et al.*, in preparation).

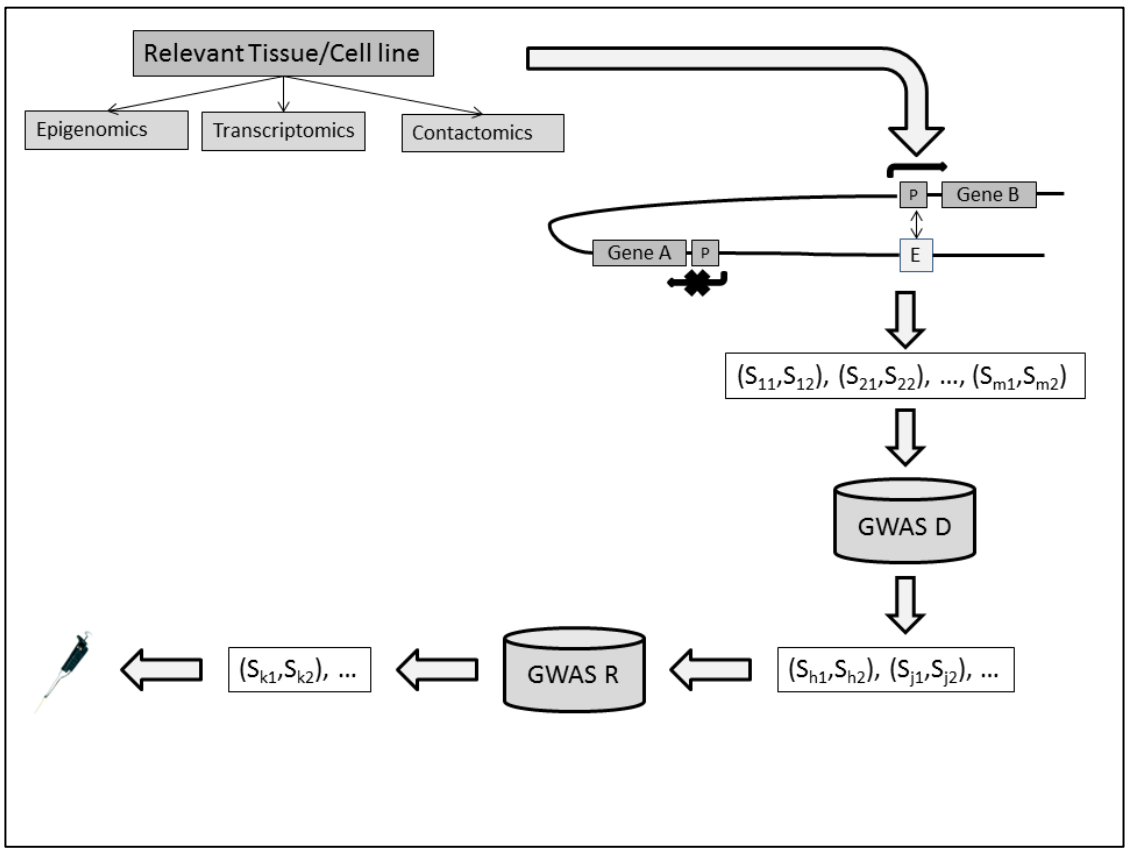

Figure 1. Functional genomics data sets of different kinds, for tissues or cells relevant to the trait of interest, are used to identify putative active and interacting enhancer-promoter pairs. Pairs of SNPs harbored in these interacting regions are extracted and analyzed for epistasis in a discovery GWAS. Significant pairs from this analysis are then examined in one or more replication GWAS to identify candidates for subsequent follow-up work.

3. Methods

3.1. *GWAS data sets*

In this work we utilized three T2D GWAS data sets. As discovery data set, we used GWAS data generated by the Welcome Trust Case Control Consortium (WTCCC),[27] precisely derived from the data sets EGAD00000000009 and EGAD00000000021.

As replication data sets we used two GWAS studies from the database of Genotypes and Phenotypes (dbGaP).[28]

1. The GENEVA Genes and Environment Initiatives in Type 2 Diabetes (Nurses' Health Study (NHS)/Health Professionals Follow-up Study (HPFS)). Data were downloaded from the dbGaP web site, under phs000091.v2.p1 **(**https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000091.v2.p1).
2. The Finland-United States Investigation of NIDDM Genetics (FUSION) Study. Data were downloaded from the dbGaP web site, under phs000867.v1.p1 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000867.v1.p1)**.**

3.2. *Data pre-processing*

3.2.1. *WTCCC data set*

Quality Control (QC) followed standard recommendations.[29] PLINK 1.9 (https://www.cog-genomics.org/plink2/) was used to filter subjects based on ambiguous gender assignment and missing call rate (threshold=95%) and duplicated individuals where removed using a PI_HAT threshold of 0.8. Then SNPs were filtered based on missing call rate (threshold=95%), Hardy-Weinberg Equilibrium tests (HWE, p=0.00001) and Minor Allele Frequency (MAF, threshold=0.01). Ambiguous SNPs (A/T or G/C) were removed. After QC we had data for 4916 subjects (1960 cases and 2956 controls) over 341,531 SNPs. These data were used to compute the first 10 Principal Components (PC) using PLINK v1.9, after Linkage Disequilibrium (LD) pruning.

Phasing was performed with SHAPEIT[30] and imputation with IMPUTE2[31] using 1000 Genomes phase 1 version 3 (http://www.internationalgenome.org) as reference panel. For imputed SNPs with information score >0.70, genotype was assigned according to the best probability if this was >0.90.

3.2.2. *GENEVA data set*

QC was first separately applied to each of the two panels (NHS and HPFS) and followed standard recommendations.[29] In each case PLINK 1.9 was used to filter subjects based on ambiguous gender assignment and missing call rate (threshold=95%). Then SNPs were filtered based on missing call rate (threshold=95%), differential missing call rates between cases and controls (p=0.00001) and MAF (threshold=0.05). The resulting data from the two panels were then merged and SNP filtering was applied again as above, finally subjects were filtered again based on missing call rate. This resulted in a QC-ed data set for 5485 subjects (2524 cases and 2961

controls) over 656,226 SNPs. The first 10 PCs were computed as for WTCCC, after LD pruning, with PLINK v1.9.

We did not impute this data set as it was only used to verify two SNP pairs resulting from the analyses in the discovery data set. If a SNP in a pair was not genotyped we used a proxy obtained from HaploReg,[32] selecting the proxy as a genotyped SNP having the highest r^2 with the SNP in our pair. For each pair to verify, we extracted the data corresponding to that pair and filtered individuals with missing genotypes on those two SNPs.

3.2.3. *FUSION data set*

QC was performed similarly to the GENEVA data set, yielding a data set of 1706 subjects (919 cases and 787 controls), with over 301,195 SNPs. PCs were also computed as above.

3.3. *Candidate pairs selection*

Among the cell lines and tissues for which enhancer-gene interactions were available in EnhancerAtlas at the time of these analyses, HCT116 and pancreas were the most relevant to T2D. For each of these two biological sources we proceeded as follows to identify candidate SNP pairs for interaction analyses.

The files from EnhancerAtlas link enhancers to ENSEMBL (ensemblgenomes.org) transcripts. We identified the regions spanning from 1000bp upstream to 500bp downstream of the Transcription Start Site (TSS) of each transcript as its promoter region. We extracted all SNPs in the enhancer and resulting promoter regions using Biofilter[33-34] (https://ritchielab.psu.edu/software/biofilter-download-1). Of these SNPs, we retained those which were either genotyped or imputed with an information score >0.70 and a probability >0.90 in the discovery data set.

We then performed LD pruning separately within each enhancer and within each promoter using a 0.8 threshold for r^2. For each enhancer-gene interaction from EnhancerAtlas we then paired up each resulting SNP in the enhancer with each resulting SNP in the corresponding promoter and removed all pairs where the two SNPs had an $r^2>0.6$. After this processing we had 11,395 pairs for HCT116 and 1,220 for pancreas.

3.4. *Two-way interaction analyses*

We analyzed separately the HCT116 and the pancreas pair collections for interactions using our discovery data set. To this end we utilized the PLATO software mentioned above, with data-driven weighted encoding and logistic regression, adjusting for gender and the first 10 PCs. This adjustment was applied after assessing the association between these covariates and the phenotype in the discovery data set. PLATO computes p-values for interactions based on the Likelihood Ratio Test (LRT) between the full model ($\text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \times X_2 +$ covariates) and the reduced model ($\text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 +$ covariates), where π is the probability of Y=1 for

a binary outcome Y and X_i is the encoded genotype at SNP_i i= 1,2. PLATO outputs unadjusted p-values, as well as Bonferroni and FDR multiple-testing adjusted p-values.

For the GENEVA replication analyses we only tested the pairs detected as significant in the discovery data set. Again we used PLATO to run the interaction analyses and we adjusted for the 10 PCs, but needed not to adjust for gender. This was determined after assessing the association between gender and the first 10 PCs with the phenotype in this data set.

For the FUSION replication analyses we tested the pair that was detected as significant in the discovery data set and the GENEVA data set, proceeding as above but adjusting for gender and the first 3 PCs. This was again determined after assessing the association between gender and the first 10 PCs with the phenotype in this data set.

4. Results

The PLATO analysis on the SNP pairs derived from the pancreas regulation data did not yield any significant pair. The PLATO analysis on the pairs derived from HCT116 yielded two pairs with FDR <10%: the pair (rs1474445, rs78495961) on chr7 (p-value=1.00394e-05; FDR adjusted p-value=0 .082295) and the pair (rs8008440, rs12882535) on chr14 (p-value= 1.44441e-05; FDR adjusted p-value=0 .082295).

We followed up only these two pairs in the GENEVA replication data set. We used rs7777433 as a proxy for rs78495961 (r^2=0.92) and rs8007341 as a proxy for rs12882535 (r^2=0.99), since we did not have genotypes for these two SNPs in GENEVA. The first pair (rs1474445, rs7777433) did not have a significant p-value in GENEVA (p=0.682361), whereas the second pair (rs8008440, rs8007341) was significant (p-value= 0.0181457; Bonferroni adjusted p-value <0.04).

Finally we followed up in the FUSION data set the pair (rs8008440, rs12882535) replicated in the GENEVA data set. The FUSION p-value was higher in this case, but remained borderline significant (p=0.155562). The combined p-value across the three data sets using the 'sumz' method in the 'metap' R package (https://cran.r-project.org/web/packages/metap/), with weights proportional to the square root of the sample sizes was 4.199818e-06, which is significant at the 0.05 level after Bonferroni correction by the number of HCT116 SNP pairs (11,395). The regression coefficients of the full model for each data set are reported in Table 1.

The two SNPs forming the pair are not in LD (r^2<0.01). One SNP (rs12882535) is in the promoter of the *OR6S1* gene while the other (rs8008440) is in a predicted EnhancerAtlas HCT116 enhancer (for *OR6S1*) located within an intron of *ANG*.

Table 1. Regression coefficients in the full model for each of the three data sets. The un-adjusted LRT p value (full model versus reduced model) and sumz combined p value are also reported. For the GENEVA data set rs8007341 was used as a proxy for SNP1.

	SNP1	SNP2	β_SNP1	β_SNP2	β_SNP1×SNP2	LRT p	comb p
WTCCC	rs12882535	rs8008440	0.053	-0.923	-7.672	1.44E-05	4.20E-06
GENEVA			0.001	0.465	-1.152	0.018	
FUSION			0.131	100.158	-199.465	0.156	

5. Discussion

In this work we have utilized functional genomics-based filters to identify candidate SNP pairs to be analyzed for epistatic interactions associated with T2D, based on GWAS data. This led to the two significant pairs (FDR<0.1) in our discovery data set. We followed up these two pairs in a replication data set and one of them (rs8008440, rs12882535) remained significant. It is typically difficult to replicate epistasis results at the SNP level, so this was encouraging. We also looked at this pair in a third data set. Here the pair did not quite reach significance, but its p-value remained relatively low. This third data set had half the number of subjects as the other two and the MAFs of the two SNPs were somewhat different (0.47 and 0.08 in this data set versus 0.45 and 0.12 in the previous two). This may explain the increase in p-value, especially considering that detection of statistical epistasis is very sensitive to MAF.[35] Overall, the meta-analysis adjusted combined p-value across the three data sets is significant. In both the GENEVA and FUSION data sets the regression coefficients of the full model correspond to an antagonistic type of interaction. However, considering that we used data-driven weighted encodings, comparisons of these coefficients across models are hard to interpret, especially when interaction terms are involved.

The pair we identified consists of a SNP in the promoter of *OR6S1* and a SNP in a putative enhancer for this gene based on HCT116 data. *OR6S1* belongs to the family of G-Protein-Coupled Receptors (GPCRs). Besides the EnhancerAtlas evidence supporting its expression in HCT116, expression of *OR6S1* was detected both in colon and pancreas in various microarray and RNAseq surveys reported by GeneCards (www.genecards.org). The presence and role of taste and olfactory receptors in the gut has been discussed in recent papers,[36-38] which indicate the importance of these chemosensors in detecting luminal contents and inducing the modulation of systemic metabolism, including glucose homeostasis. Indeed some GPCRs have recently received attention as new therapeutic targets for type 2 diabetes.[39]

HCT116 is a human cell line derived from colonic carcinoma, which has been used in other studies on T2D[40] and, together with pancreas, was the most relevant for T2D among the sources for which EnhancerAtlas provided enhancer-gene interaction data at the time of our download. We did not detect significant interactions among SNP pairs derived from the pancreas data.

The T2D statistical epistasis association that we found is consistent with the recent literature support for the relevance of GPCRs to T2D. Nevertheless, ascertaining whether or not this pair represents actual biological epistasis requires follow-up experiments. This work should be taken as a proof of concept. One of the limitations of our study was that albeit whole pancreas and HCT116 are relevant tissues/cells for T2D, they are not ideal. Having enhancer-promoter data from biosources such as the pancreatic beta-cell and the enteroendocrine L cell would improve the chance to detect epistatic pairs associated with T2D. As more data become available for these particular biosources, both in the public domain and in our labs, we can generate richer and more specific sets of candidate pairs to explore. The types of relevant data include epigenetic information (e.g. open chromatin, enhancer and promoter histone marks) and gene expression data. We are working on a generalization of our workflow which incorporates relevant

epigenomics data as well as high-throughput chromatin conformation capture data (contactomics) to identify additional tissue-specific active and interacting genomic regions from which to select SNP pairs (Manduchi E., Grant S., Moore J., in preparation) .

Our intent is to exploit the availability of different types of omics data sets and to incorporate state of the art analysis methods into a reasonable workflow to explore a typically difficult-to-detect phenomenon such epistasis. Based on our results we believe that this type of workflow is promising and besides significantly reducing the search space it could yield results that are much easier to interpret. Since it is applicable to any GWAS, as we build on data sets which are more extensive in terms of type of functional genomics data and trait-specific tissues, it has the potential to unveil interactions associated to many traits, with a higher likelihood of being reproducible and biologically meaningful.

In this work we were interested in assessing reproducibility and hence utilized a Discovery/Replication data set paradigm. An alternative way to take advantage of independent GWAS for a trait is to combine them in a meta-analysis framework, which can help increase power. We are investigating this approach also in the context of exploring results at different resolutions, besides SNP-pair level (Manduchi E., Grant S., Moore J., in preparation).

6. Availability of Data and Materials

The GWAS data sets analyzed in study are available, upon application, at: https://www.wtccc.org.uk/ (WTCCC, use the data set IDs provided in the main text), https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000091.v2.p1 (GENEVA), and https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000867.v1.p1 (FUSION).

7. Funding

Funding for this work was provided by NIH grants LM010098, DK112217, ES013508, the CHOP's Spatial and Functional Genomics Initiative, and the Daniel B. Burke Endowed Chair for Diabetes Research.

8. Authors' Contributions

EM, JHM, and SFAG conceived the study, identified appropriate data sets, and obtained the necessary access. EM designed the workflow, performed the analyses, and prepared the initial manuscript draft. AC performed QC and imputation on the WTCCC GWAS data set. MAH helped with PLATO settings and weighted encoding. All authors contributed to and approved the final version of the manuscript. SFAG and JHM supervised this work.

9. Acknowledgments

This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award

076113, 085475 and 090355. The Consortium and/or Individual Investigators bear no responsibility for the further analysis or interpretation of these data, over and above that published by the Consortium.

The authors thank P.Schmitt, S. Dudek and C. Calafut for their assistance with the Biofilter and PLATO software installation and administration.

References

1. W.G. Hill, M.E. Goddard, P.M. Visscher, *PLoS Genet*. **4(2)**, e1000008 (2008).
2. E.E. Eichler, J. Flint, G. Gibson, A. Kong, S.M. Leal, J.H. Moore, J.H. Nadeau, *Nat. Rev. Genet.* **11**, 446 (2010).
3. O. Zuk, E. Hechter, S.R. Sunyaev, E.S. Lander, *PNAS USA* **109**, 1193 (2012).
4. T.F.C. Mackay, *Nat. Rev. Genet*. **15(1)**, 22 (2014).
5. J.H. Moore, S.M. Williams, *Bioessays* **27(6**), 637 (2005).
6. J.H. Moore, S.M. Williams, *The American Journal of Human Genetics* **85**, 309 (2009).
7. P.C. Phillips PC, *Nat Rev Genet.* **9(11)**, 855 (2008).
8. M. Robnik-Šikonja, I. Kononenko, *Mach. Learn*. **53**, 23 (2003).
9. J.H. Moore, B.C. White, in E. Marchiori, J.H. Moore, J.C. Rajapakse (eds), *Evolutionary computation, machine learning and data mining in bioinformatics*, Springer, Berlin, pp.166 (2007).
10. J.H. Moore, *Methods Mol Biol.* **1253**, 315 (2015).
11. M.D. Ritchie, L.W. Hahn, N. Roodi, L.R. Bailey, W.D. Dupont, F.F. Parl, J.H. Moore, *Am. J. Hum. Genet.* **69**, 138 (2001).
12. L. Qi, R.M. van Dam, F.W. Asselbergs, F.B. Hu, *Diabetic Medicine* **24**, 1187 (2007).
13. S.S. Verma, J.N. Cooke Bailey, A. Lucas, Y. Bradford, J.G. Linnemann, M.A. Hauser, L.R. Pasquale, P.L. Peissig, M.H. Brilliant, C.A. McCarty, J.L. Haines, J.L. Wiggs, T.R. Vrabec, G. Tromp, M.D. Ritchie, eMERGE Network, NEIGHBOR Consortium, *PLOS Genetics* **12(9)**, e1006186 (2016).
14. R.J. Urbanowicz, A.L.S. Granizo-Mackenzie, J. Kiralis, J.H. Moore, *BioData Min.* **7**, 8 (2014).
15. I. Mitra, A. Lavillareuix, E. Yeh, M. Traglia, K. Tsang, C.E. Bearden, K.A. Rauen, L.A. Weiss, *PLOS Genetics* **13(1)**, e1006516 (2017).
16. R. Cowper-Sal Iari, M.D. Cole, M.R. Karagas, M. Lupien, J.H. Moore, *Wiley Interdiscip Rev Syst Biol Med.* **3(5)**, 513 (2011).
17. E.A. Boyle, Y.I. Li, J.K. Pritchard, *Cell* **169(7)**, 1177 (2017).
18. T. Gao, B. He, S. Liu, H. Zhu, K. Tan, J. Qian, *Bioinformatics* **32(23)**, 3543 (2016).
19. W. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, D. Haussler, *Genome Res.* **12(6)**, 996 (2002).
20. T. Barrett, S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C.L.

Robertson, N. Serova, S. Davis, A. Soboleva, *Nucleic Acids Res.* **41(Database issue)**, D991 (2013).
21. B. Qin, M. Zhou, Y. Ge, L. Taing, T. Liu, Q. Wang, S. Wang, J. Chen, L. Shen, X. Duan, S. Hu, W. Li, H. Long, Y. Zhang, X.S. Liu, *Bioinformatics* **28(10)**, 1411 (2012).
22. ENCODE Project Consortium, *Nature* **489(7414)**, 57 (2012).
23. Roadmap Epigenomics Consortium, et al., *Nature* **518(7539)**, 317 (2015).
24. R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li, L. Schwarzfischer, D. Glatz, J. Raithel, B. Lilje, N. Rapin, F.O. Bagger, M. Jorgensen, P.R. Andersen, N. Bertin, O. Rackham, A.M. Burroughs, J.K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhata, S. Maeda, Y. Negishi, C.J. Mungall, T.F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C.O. Daub, P. Heutink, D.A. Hume, T.H. Jensen, H. Suzuki, Y. Hayashizaki, F. Müller; FANTOM Consortium, A.R. Forrest, P. Carninci, M. Rehli, A. Sandelin, *Nature* **507(7493)**, 455 (2014).
25. G.M. Clarke, C.A. Anderson, F.H. Pettersson, L.R. Cardon, A.P. Morris, K.T. Zondervan, *Nat Protoc.* **6(2)**, 121 (2011)
26. M.A. Hall, J. Wallace, A. Lucas, D. Kim, A.O. Basile, S.S. Verma, C.A. McCarty, M.H. Brilliant, P.L. Peissig, T.E. Kitchner, A. Verma, S. Pendergrass, S. Dudek, J.H. Moore, M.D. Ritchie, *Nature Communications* **in press**, (2017).
27. Welcome Trust Case Control Consortium, *Nature* **447(7145)**, 661 (2007)
28. K.A. Tryka, L. Hao, A. Sturcke, Y. Jin, Z.Y. Wang, L. Ziyabari, M. Lee, N. Popova, N. Sharopova, M. Kimura, M. Feolo, *Nucleic Acids Res.* **42(Database issue)**, D975 (2014).
29. C.A. Anderson, F.H. Pettersson, G.M. Clarke, L.R. Cardon, A.P. Morris, K.T. Zondervan, *Nat. Protoc.* **5(9)**, 1564 (2010).
30. O. Delaneau, J. Marchini, J.F. Zagury, *Nat. Methods* **9(2)**, 179 (2012).
31. J. Marchini, B. Howie, S. Myers, G. McVean, P. Donnelly, *Nature Genetics* **39**, 906 (2007).
32. L.D. Ward, M. Kellis, *Nucleic Acids Res.* **40(D1)**, D930 (2011).
33. W.S. Bush, S.M. Dudek, M.D. Ritchie, *Pac. Symp. Biocomput*, 368 (2009).
34. S.A. Pendergrass, A. Frase, J. Wallace, D. Wolfe, N. Katiyar, C. Moore, M.D. Ritchie, *BioData Min.* **6(1)**, 25 (2013).
35. C.S. Greene, N.M. Penrod, S.M. Williams, J.H. Moore JH, *PLoS One*. 2009 **4(6)**, e5639 (2009).
36. F. Reiman, G. Tolhurst, F.M. Gribble, *Cell Metabolism* **15(4)**, 421 (2012).
37. C. Sternini, L. Anselmi, E. Rozengurt, *Curr Opin Endocrinol Diabetes Obes.* **15(1)**, 73 (2008).
38. I. Kaji, S. Karaki, A. Kuwahara, *Curr. Pharm. Des.* **20(16)**, 2766 (2014).
39. F. Reimann, F. Gribble, *Diabetologia* **59**, 229 (2016).
40. Q. Xia, A. Chesi, E. Manduchi, B.T. Johnston, S. Lu, M.E. Leonard, U.W. Parlin, E.F. Rappaport, P. Huang, A.D. Wells, G.A. Blobel, M.E. Johnson, S.F. Grant, *Diabetologia* **59(11)**, 2360 (2016).