



UNIL | Université de Lausanne

Unicentre

CH-1015 Lausanne

<http://serval.unil.ch>

---

Year : 2017

## Going beyond GWAS: New methods to interpret association signals

Lamparter Félix David

Lamparter Félix David, 2017, Going beyond GWAS: New methods to interpret association signals

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive <http://serval.unil.ch>

Document URN : urn:nbn:ch:serval-BIB\_OA6E811551354

### **Droits d'auteur**

L'Université de Lausanne attire expressément l'attention des utilisateurs sur le fait que tous les documents publiés dans l'Archive SERVAL sont protégés par le droit d'auteur, conformément à la loi fédérale sur le droit d'auteur et les droits voisins (LDA). A ce titre, il est indispensable d'obtenir le consentement préalable de l'auteur et/ou de l'éditeur avant toute utilisation d'une oeuvre ou d'une partie d'une oeuvre ne relevant pas d'une utilisation à des fins personnelles au sens de la LDA (art. 19, al. 1 lettre a). A défaut, tout contrevenant s'expose aux sanctions prévues par cette loi. Nous déclinons toute responsabilité en la matière.

### **Copyright**

The University of Lausanne expressly draws the attention of users to the fact that all documents published in the SERVAL Archive are protected by copyright in accordance with federal law on copyright and similar rights (LDA). Accordingly it is indispensable to obtain prior consent from the author and/or publisher before any use of a work or part of a work for purposes other than personal use within the meaning of LDA (art. 19, para. 1 letter a). Failure to do so will expose offenders to the sanctions laid down by this law. We accept no liability in this respect.



**UNIL** | Université de Lausanne

Faculté de biologie  
et de médecine

**Département de Biologie computationnelle**

**Going beyond GWAS: New methods to interpret association signals**

**Thèse de doctorat ès sciences de la vie (PhD)**

présentée à la

Faculté de biologie et de médecine  
de l'Université de Lausanne  
par

**David Felix LAMPARTER**

Biologiste diplômé ou Master de l'EPFZ

**Jury**

Prof. Matthias Stuber, Président  
Prof. Sven Bergmann, Directeur de thèse  
Prof. Zoltán Kutalik Codirecteur  
Prof. Alexandre Reymond, expert  
Prof. Marc Robinson-Rechavi, expert  
Prof. Lude Franke, expert

Lausanne, 2017



UNIL | Université de Lausanne

Faculté de biologie  
et de médecine

**Ecole Doctorale**

Doctorat ès sciences de la vie

# Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

<b>Président · e</b>	Monsieur Prof. Matthias <b>Stuber</b>
<b>Directeur · rice de thèse</b>	Monsieur Prof. Sven <b>Bergmann</b>
<b>Co-directeur · rice</b>	Monsieur Prof. Zoltan <b>Kutalik</b>
<b>Experts · es</b>	Monsieur Prof. Alexandre <b>Reymond</b>
	Monsieur Prof. Marc <b>Robinson-Rechavi</b>
	Monsieur Prof. Lude <b>Franke</b>

le Conseil de Faculté autorise l'impression de la thèse de

**Monsieur David Lamparter**

Master in Statistics and Biology de l'Ecole polytechnique de Zürich

intitulée

**Au delà du GWAS: nouvelles méthodes  
pour l'interprétation des associations de signaux**

Lausanne, le 9 février 2017

pour le Doyen  
de la Faculté de biologie et de médecine

Prof. Matthias Stuber

Prof. Matthias STUBER  
Directeur du CIBM-CHUV  
RAD-CHUV-BH07  
Rue du Bugnon 46  
CH-1011 Lausanne  
Tél. +41 21 314 75 34  
matthias.stuber@chuv.ch

## **Acknowledgements**

First and foremost I would like to thank Professor Sven Bermann and Professor Kutalik for their supervision and support during my PhD work. Both have given me freedom to explore, yet supported me wholeheartedly throughout the process, be it with initial setting of the course, project design, as well as the actual work of method and experiment design, all the way through to the writing process. Further, I want to thank all the members of their teams for providing a great work environment. I would like to specially mention Daniel Marbach, who taught me a lot, while also becoming a good friend. Further, Rico Rueedi whose eye for clear scientific writing I often relied on. I further would like to thank Béryll Boyer-Bertrand and Aurélien Macé for their friendship and making sure, that I would jump all the administrative hurdles in time. I also would like to thank Sina Rüeger for her continued support in all sorts of matters, such as providing me with a roof over my head, helping me get past setbacks and celebrating successes with me to name just a few.

I would like to thank my family, especially my mother, who supported me through all the ups and downs that led me here. Also I want to thank Zoé Blanc who supported me with love and patience throughout even as I asked sacrifices from her to complete this work.

## Abstract (english)

The aim of genetics is to understand the genetic basis of traits by linking genetic variability to phenotypic variability. In recent years, progress in the field of complex human trait genetics led to the discovery of thousands of common genetic variants robustly associated with complex human traits through genome-wide association studies (GWAS). However, it is currently unclear how to best to tackle the challenge of interpreting variants in the context of the biology involved. My work explored various avenues to help in this challenge.

One strategy for interpretation is pathway analysis, where prior biological knowledge is formalized into sets of genes with annotated functions and results from genetics studies are searched for enrichments. Using this approach, one can connect the biological processes to the investigated trait. For this purpose, I developed a methodology to calculate pathway enrichments from GWAS results in an efficient way and in agreement with statistical principles. As a first step, the methodology combines results for SNPs in a gene region into a single gene wise p-value, with methods that are both fast and have a high level of numerical precision. The speed allows controlling the pathway enrichment step for potential correlation between genes leading to statistically correct p-values. This methodology was implemented in a software tool called Pascal. Its performance was tested on a large set of GWAS results and compared favorably to other methods. Efforts were made to ensure that the software would be easy to use by a wider community.

Another challenge in the interpretation of GWAS results is to understand the reasons a genetic variant leads to changes in phenotype. Most uncovered variants seem to impact gene regulation. Therefore, understanding chromatin architecture will be crucial to understand the regulatory consequences of genetic variants. One feature of eukaryotic chromatin is that it can take the form of a compacted state making it inaccessible to most regulatory factors. To help elucidate which factors play a role in moving between compacted and an open state, I developed a new method of integrative data analysis for transcription factor motif, DNase1 hypersensitivity and gene expression data. Transcription factor motif and DNase1 hypersensitivity were combined to calculate chromatin accessibility scores. These in turn were associated to gene expression using a linear mixed modeling approach. Applying this method on large public datasets predicted a set of candidate chromatin accessibility regulators. This set was heavily enriched in 'pioneer factors': factors that can bind and open compacted chromatin, suggesting that the approach did indeed uncover regulators of chromatin accessibility.

A major hindrance to the interpretation of human variants uncovered by GWAS is that it is not possible to perform genetic manipulations to validate and build on the findings. Therefore, investigations using model organisms remain relevant. To further the understanding of the genetics of fly growth control, I helped in the statistical analysis of a GWAS data set in an outbred fly population. The study is noteworthy for its extensive environmental control and follow-up experiments on candidate genes.

## Abstract (français)

La génétique cherche à comprendre la base génétique de caractères observables, dits phénotypes, en liant la variabilité génétique à la variabilité phénotypique. Ces dernières années, les progrès apportés à la génétique des phénotypes complexes ont amené à la découverte de milliers de variations génétiques associées significativement à des phénotypes humains complexes, au moyen de l'étude d'association pangénomique, communément appelée GWAS (de l'anglais Genome-Wide Association Study). Cependant, interpréter ces associations dans leur contexte biologique reste un défi. Mon travail a consisté à explorer différentes possibilités pour y répondre.

Une des stratégies pour relier un phénotype étudié aux processus biologiques est l'analyse par voies moléculaires, où l'on recherche un enrichissement des associations GWAS parmi l'ensemble de groupes de gènes de fonctions cohérentes. Cette analyse permet ainsi de relier des processus biologiques au phénotype étudié. Dans ce but, j'ai développé une méthode calculant avec efficacité l'enrichissement des voies moléculaires des associations GWAS. La méthode combine les résultats des polymorphismes dans la région d'un gène en une probabilité pour un gène, au moyen de méthodes rapides et précises. La vitesse permet de contrôler l'étape d'enrichissement des voies pour une corrélation potentielle entre les gènes, menant à des probabilités statistiquement correctes. J'ai implémenté cette méthode dans le logiciel Pascal. Sa performance a été testée sur un large jeu de résultats GWAS et il surpasse les autres méthodes. Des efforts ont été faits afin d'assurer que le logiciel soit facile d'utilisation pour la communauté scientifique.

Un autre défi lié à l'interprétation des résultats GWAS est de comprendre les raisons pour lesquelles une variation génétique résulte en un changement phénotypique. La plupart des variations découvertes semblent affecter la régulation des gènes. Ainsi, comprendre l'architecture de la chromatine est crucial pour appréhender les conséquences régulatrices de ces variations. Une des caractéristiques de la chromatine des eucaryotes est qu'elle peut être compactée, la rendant inaccessible à la plupart des facteurs de régulations. Pour trouver quels sont les facteurs jouant un rôle dans le passage entre états compactés et ouverts, j'ai développé une nouvelle méthode intégrant l'analyse des motifs de facteurs de transcription, l'hypersensibilité de la Dnase1 et les données d'expression des gènes. Les deux premiers critères ont été combinés pour calculer des scores d'accessibilité de la chromatine. Ils ont ensuite été associés à l'expression des gènes en utilisant un modèle linéaire mixte. L'application de cette méthode sur des larges données publiques a prédit des régulateurs candidats d'accessibilité de la chromatine. Ce jeu était enrichi en "facteurs pionniers", qui s'accrochent et ouvrent la chromatine compactée, suggérant que cette approche a en effet permis de découvrir des régulateurs d'accessibilité à la chromatine.

Un obstacle majeur dans l'interprétation des variations humaines découvertes par GWAS est qu'il n'est pas possible de réaliser des manipulations génétiques permettant de valider ces découvertes, d'où l'utilisation d'organismes modèles. Afin de comprendre davantage la génétique du contrôle de croissance des mouches, j'ai aidé à l'analyse statistique de données de GWAS d'une population de mouches consanguines. L'étude est remarquable pour son contrôle environnemental étendu et ses expériences de suivi sur les gènes candidats.

# Contents

<b>1. Introduction</b>	<b>1</b>
1.1. Recent development in genetics of complex traits . . . . .	1
1.2. Gene set analysis . . . . .	7
1.3. Confounding control in large-scale biology . . . . .	10
1.4. Modern approaches to building genome wide protein-DNA binding maps . . . . .	12
<b>2. Pathway Analysis for GWAS data</b>	<b>14</b>
<b>3. Identifying chromatin accessibility regulators</b>	<b>35</b>
<b>4. Investigation of the genetic control of Drosophila body size</b>	<b>56</b>
<b>5. Discussion</b>	<b>87</b>
5.1. How it all fits together: Interpretation of GWAS findings . . . . .	87
5.2. Methodology overview of each project . . . . .	89
5.3. Confounding: A problem revisited in each project . . . . .	91
5.4. Causality . . . . .	93
5.5. Future work . . . . .	94
<b>A. Estimating the contribution of pairwise interactions to the genetic variance</b>	<b>97</b>
A.1. A computationally efficient estimation strategy . . . . .	97
A.2. Power considerations . . . . .	99
A.3. Power of the pairwise interaction covariance matrix . . . . .	100
<b>B. Connection between random effects score test and the <i>Pascal</i> sum statistic</b>	<b>105</b>
<b>C. Estimating heritability within a genic region via maximum likelihood</b>	<b>109</b>
<b>D. Supplementary Information for Chapter 2</b>	<b>111</b>

<b>E. Supplementary Information for Chapter 3</b>	<b>127</b>
E.1. Supplementary Figures . . . . .	127
E.2. Supplementary Methods . . . . .	139
<b>F. Supplementary Information for Chapter 4</b>	<b>141</b>



# 1. Introduction

The following pages document work that I have done with the help of others in the domain of human genetics and integrative analysis of functional genomics under supervision of Prof. Sven Bergmann and Prof. Zoltán Kutalik. After a general introduction, aimed to prepare the reader, three projects are presented, two of which I spearheaded and one where I made substantial contributions. All three of these efforts yielded publications. The first, titled '*Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics*', presented *Pascal*, a tool for pathway analysis as a downstream step GWAS analysis [1]. The second, titled '*Genome-wide association between transcription factor expression and chromatin accessibility reveals chromatin accessibility regulators*', presented a method to integrate expression data and chromatin accessibility [2]. The third, titled '*A Genome-Wide Analysis Reveals Novel Regulators of Growth in Drosophila melanogaster*', presented a GWAS analysis of drosophila body size metrics [3]. Some final remarks will reflect on the commonalities and differences among the projects, set the projects in their historical context and give some additional insights on the methodological aspects of the work presented.

## 1.1. Recent development in genetics of complex traits

Recently, the drop in the cost of production of large amounts of biological and medical data have opened up ways to conduct biomedical research in a hypothesis-free way. During the years 2000-2010 this development was best epitomized by the microarray [4]. This technology made it possible to monitor gene expression comprehensively. It also allowed to cheaply genotype individuals for common variants, leading to the discovery of numerous common genetic variants affecting traits as diverse as height, facial appearance or expression of a particular gene through GWAS (genome wide association studies) [5, 6, 7]. In GWAS, the phenotype in question is regressed onto the genotypic state for a common genetic variant in the sample of individuals to

decide whether this variant is associated with the phenotype or not. This approach is applied to a large fraction of SNPs (single nucleotide polymorphisms) present in the population to get an unbiased collection of SNPs showing a significant association.

The early successes of this approach were helped by the fact that humans are a relatively inbred species having low genetic diversity compared to global population size, which is likely due to population bottlenecks in the relatively recent past. Indeed, of all  $3 * 10^9$  possible single nucleotide variants in the human genome, less than  $10^7$  are observed as a common polymorphism (meaning that the frequency of the allele less common in the general population (the minor allele frequency; MAF) is above 5%)<sup>1</sup> [9]. Another consequence of the relatively low number of cross-over events since the last evolutionary bottleneck, is that neighbouring common variants tend to be in linkage disequilibrium (LD), a measure of correlation between nearby SNPs [10]. The presence of LD implied that much fewer SNPs actually needed to be assayed and could function as tagging SNPs for other un-assayed (and potentially unknown) variants. The power to detect an association between a SNP and a trait depends on the allele frequency and the effect size in the population under study. For traits under low selection pressure (and weak pleiotropy with traits under selection), allele frequency and effect size should be close to independent because evolutionary pressure will not drive large effect variants to different frequency in the population. Therefore, the variants most strongly associating in a GWAS experiment should be common variants. Although the assumption of independence turned out to be violated for most traits, the dependence is often weak enough that variants most easily detected have relatively high frequency [11]. Microarrays measuring 500'000-2 Mio common SNPs in parallel were well positioned to detect these variants. Large investments went into genotyping the common variants of cohorts. Simplicity of the statistical modeling (the workhorse being simple linear and logistic regression) allowed for efficient meta analysis strategies of results from various cohorts, leading to studies involving tens of thousands to hundreds of thousands of individuals.

---

<sup>1</sup>As an illustration, we counted the number of autosomal SNPs , with frequency above 0.05 in the european panel from the phase 1 1000 genomes project[8]. We found  $6.6 * 10^6$  SNPs with frequency above this threshold.

Although many SNPs were found to have a reproducible impact on human traits, the impact of a single variant was usually minuscule and summing the variance explained by all SNPs found in GWAS only yielded a fraction of the genetic variation that was expected based on the results of twin studies. The huge difference in genetic variance explained by GWAS results and expected from twin studies was termed the missing heritability [12]. There were three popular hypotheses put forward to explain this fact.

First, there was the hypothesis that low frequency variants (with MAF below 5%) were responsible for most of the missing heritability. In hindsight this hypothesis has the difficulty that although there are many more low frequency variants than common variants known, low frequency variants have much lower variance in the population than common variants. Assuming complete independence between allele frequency and effect size, the fraction of heritability associated with each SNP group (i.e. high MAF SNPs and low MAF SNPs) would be approximately proportional to the sum of the variances of all SNPs in the group. From sequencing data one can show that the high MAF group dominates in this respect<sup>1</sup>. For traits under low selection pressure, one therefore needs to assume very strong pleiotropic effects for the low frequency variants to dominate. While there are of course traits under very strong selection pressure (for instance rare Mendelian diseases that severely hamper fitness), for many traits such as diseases of old age, strong selection pressure does not seem likely.

The second hypothesis centered around interaction effects. This hypothesis is that most of the variation would be explained by interactions of SNPs that could not be found by regressing single genotype values onto phenotypes. One appeal of this hypothesis was that it corresponded nicely with a view of biology where gene products worked in concert to yield an effect. A forceful

---

<sup>1</sup>While it is difficult to estimate the variance of a single low frequency SNP, the sum of all variances of all low frequency SNPs should be stably estimable. Note that as a SNPs MAF goes to 0, the variance of the SNP goes to its mean (because  $p(p-1) \rightarrow p$  as  $p \rightarrow 0$ ). We can therefore estimate the sum of variances of low MAF SNPs by the average number of low frequency SNPs an individual in the population carries. When comparing the variance contributed by low MAF SNPs (below 2% frequency) to the variance contributed by high MAF SNPs (above 2%) using the 1KG phase 3 release data, we see that the contribution by common SNPs is 11.7 times greater.

critique of this hypothesis is that the additive fraction of heritability estimated from family studies was also much larger than what was estimated from GWAS data. Since two non-linked alleles carried by a parent only have a 25% chance to be inherited together, whereas each allele by itself has a 50% chance of being inherited, one typically only inherits 25% of pairwise interaction effects active in a parent but 50% of marginal effects. Higher order interaction effects dissipate even faster across generations. Family data, therefore, allows to identify the contribution of only the marginal effects by looking at phenotype correlations across multiple generations or comparing monozygotic to dizygotic twins. While the estimated additive contributions vary from trait to trait, GWAS results fell well short of explaining a substantial portion of the additive heritability.

The third hypothesis was that effect sizes were in general very small and that current cohort sizes were just too small to detect them. This hypothesis was bolstered by a seminal paper which pointed out that, for height, genetic similarity of individuals measured by microarray was correlated with phenotypic similarity [13]. The extent of this correlation was commensurate with microarrays at the time tracking around 50% of the additive fraction, and that it was likely, that adding common SNPs not well tagged yet would further increase this fraction.

Since then, the notion that increasing sample size would be more cost effective to uncover new variants involved in common traits than measuring low frequency variants or interactions has broadly held. For example, an upcoming study compares results for body heights in humans obtained with regular genome wide microarrays to results obtained with exome chips. These chips are designed to have deep coverage of low frequency and common variants in exome regions. Again, the conclusions supported the importance of sample size over coverage although some additional variants were found (Marouli et al. Nature, accepted). A further example is a recent investigation into the genetic architecture of type 2 diabetes where whole genome and exome sequencing was used to investigate the idea that low frequency variants contribute substantially to the missing heritability. Again it was found that the data supported a model, where common

variants contributed the bulk to heritability [14].

The cost-effectiveness of increasing sample size to look for additional signal in the additive component has also to do with the economics and logistics involved in performing low frequency variant and interaction studies for large sample sizes. However, these factors might change in the near future. With regards to investigation of interactions, the current model of performing association studies in a distributed fashion by meta-analysis is ill-suited. The reason is that every analysis group would need access to powerful computational infrastructure and exchanging the results between cohorts would need the exchange of very large data sets. Studies exploring interactions are often performed on single cohorts and therefore tend to have limited power [15]. Systematic investigation of interactions therefore have been relatively few and results are often controversial. For instance, a recent paper investigated interactions using gene expression as the trait of interest. Expression seems well placed for investigating interactions because the size of single cohorts are large compared to effect sizes [16]. The study reported 30 SNP-SNP interactions with effects on 19 genes. This finding was challenged by Wood et al. suggesting that many of the apparently interacting SNPs were both weakly tagging a third variant that was not assayed in the original study [17]. However, there are efforts underway to assemble very large cohorts in centralized databases such as the UK Biobank or the Estonian genome project [18, 19]. These resources make it easier to investigate interactions, be it interactions between genotypes or between environment and genotypes. This is particularly interesting for traits with a relatively large interaction component such as human intelligence and BMI [20, 21, 22]. Computational and statistical power constraints however suggest that for interactions between genotypes, only pairwise interaction will be amenable to comprehensive standard regression analysis in the near future. (For an approach to determine what fraction of the heritability pairwise interactions contribute see Appendix A). Apart from uncovering novel associations, interaction studies allow to investigate interactions which might yield further biological insights as to how genes act in concert.

Also the costs for sequencing, the method of choice for performing low frequency variant studies, have been historically much higher than genotyping costs, but might be cost effective in the near future allowing for easier investigation of traits under strong selection pressure. Alternatively, custom genotyping chips focused mainly on exome regions can be used to genotype variants that are hard to impute using data from regular genotyping chips. These chips are cost effective compared to sequencing. They have recently been used to investigate low frequency variants influencing height using data from a large cohort and compared to results obtained from regular genotyping chips (Marouli et al. Nature, accepted). Rare variant analysis has the potential to improve detection of causal genes. Common variants affecting the phenotype often fall into regulatory regions, which can make it hard to associate them to a specific gene. Rare variants with strong impact are often found in the gene body, making it easier to determine the causal gene. Also, low frequency variants typically have much lower LD than common variants, making it easier to pinpoint causal variants. Therefore, studies of low frequency variants might be interesting to further investigate genomic regions that are prioritized by GWAS, by providing evidence which genes in the region are involved in the phenotype.

In short, the field of human genetics has changed enormously in last decade. Gene candidate studies have been replaced by truly comprehensive approaches. These have helped to reproducibly uncover variants with impact on human traits and disease. However, the revolution is far from over. The assembly of large centralized repositories of genetic data seems like a preview of how genetics might be done in the near future when genetic analysis informing medical decisions could be commonplace. The questions answered might range from informing on drug dosage requirements and drug sensitivities to isolate particular genetically at risk patient populations [23, 24]. Large genetic datasets gathered in the process will allow to create ever more precise and sophisticated models of how genetic variants impact individual humans.

## 1.2. Gene set analysis

Since the beginning of high throughput biology, gene set analysis (often also referred to as pathway analysis) has been used as a means to extract biological information from experiments, in particular from differential expression experiments. Standard differential expression analysis looks at each gene in isolation asking for each gene the question whether its expression levels differ in cases versus controls. Results can be hard to interpret if the study is underpowered to detect reproducible differential expression on the gene level. However, the results are also difficult to interpret if a large numbers of genes are differentially expressed, because this can make it hard to understand the biological context.

Gene set analysis is a strategy trying to answer both of these challenges. In gene set analysis, predefined sets of genes are tested as a group. While each member of the gene set might not yield a significant  $p$ -value, pooling the effects of all members can still lead to significant results. This has the potential to extract significant results even from underpowered experiments, although poor definition of biologically relevant gene sets can hamper this. Furthermore, knowledge about the predefined gene sets can help in the interpretation of the results. Thus, gene set analysis became a popular tool to analyse differential expression experiments. One strategy to define a  $p$ -value for a gene set is to check whether more genes in the gene set are nominally significant (i.e. above a given threshold) than is expected when drawing a random sample of genes. This strategy is called hypergeometric enrichment [25]. This approach can be generalized. To check enrichment, one can take some transformation of the gene-wise  $t$ -statistics (say the square to upweight outliers) and sum the resulting scores for member genes to get a gene set score. Then, one calculates multiple gene set scores for randomly sampled gene sets of the same size. The  $p$ -value can be approximated by taking the fraction of random samples leading to a higher or at least as high a gene set score as the original gene set.

While this strategy was popular, it had the fundamental drawback that it does not control for

correlation between genes. Because genes within the same gene set are functionally related, they also tend to have expression values that are more correlated than average gene pairs. Correlation between expression values of two genes directly leads to correlation of the differential expression statistics. However, when genes are randomly sampled one implicitly assumes that genes are interchangeable in terms of correlation they have to other genes. One way to remedy this problem is to look at expression arrays as the sampling unit instead of the genes. In this approach one calculates the gene set scores analogously as above. One then generates random permutations of the annotation labels in the experiment between cases and controls and calculates the gene set score for each permutation. The  $p$ -value can be approximated by taking the fraction of permuted samples leading to at least as high a gene set score as the original gene set. This approach takes the correlation between genes into account: Correlations between genes remain the same in permuted samples, since the only thing that gets permuted is the annotation of the microarray experiments. However, this approach tests a different null hypothesis (hereafter often referred to simply as ‘null’) than that tested in the gene randomization approach. Within the context of a differential expression experiment, gene randomization tests the null that the genes in the genes set are on average not more differentially expressed than a random sample of genes. The annotation permutation approach tests the null that there is no differentially expressed genes at all in the gene set. This means that the permutation approach can yield significance even if only one gene in the gene set shows strong signal and is independent of the results of genes outside of the gene set. It also means that the results are very sensitive to overdispersion of the  $t$ -statistics due to confounding whereas the gene randomization approach is less sensitive to this problem. Because of that, the two approaches are also called competitive (for the gene permutation approach) and non-competitive (for the annotation permutation approach). To combine the strengths of both approaches, sampling strategies were developed to take correlation into account while still performing a competitive test [26, 27, 28]. Perhaps the most straight-forward approach is due to Efron et al.: Permuted gene set scores are rescaled to account for the amount of overdispersion seen for the gene wise scores [27].



When new gene set analysis approaches for new kinds of data are devised, an important question to ask is whether one wants to have a competitive or non-competitive test. If one opts for a competitive test, it is vital to make sure that one accounts for potential correlation. In the chapter on pathway analysis (chapter 2), we will see an example of a competitive test where correlation is accounted for by a merging strategy between genes that are physical neighbours on the genome. This is possible in this case because correlation in GWAS is local in the genome.

A further lesson that can be drawn from experience of differential expression analysis pertains to the use of comparable statistics. Often, differential expression analysis report fold changes between the two conditions while ignoring the variation within conditions for this gene. The reason for this was that variation within conditions was often regarded as too noisy leading to less stable results. This practice was criticized by Efron et al. since it leads to biased results for null genes[29]. Genes with high variance would have large spreads in fold change when randomly partitioned into conditions. This means that two genes both with zero effect might have not equal probability of showing a false positive depending on the variation. Scaling with the standard deviation ameliorates this problem. However genes with the same non-zero effect but non-equal variability will not be detected with equal power due to scaling. When testing gene sets with a competitive approach any bias in the composition of genes within a gene set with regards to a factor influencing power can also bias enrichment results. In general it is regarded as crucial to control such biases with regards to null genes because they will be very frequent. Therefore, one typically uses test statistics that all show the same distribution in the case of absence of signal. In the chapter on *Pascal*, we will use such statistics as a starting point for our competitive test. However, it has to be pointed out that any such statistic can lead to a difference in power for non-null genes. This bias can persist at the gene set enrichment stage if gene sets are biased with regards to such a confounding factor.

### 1.3. Confounding control in large-scale biology

Another subject that is of vital importance in high-throughput biology is confounding control. It is a natural assumption for a high-throughput experiment, where thousands to millions of hypotheses are tested, that for most of them the null hypothesis holds. However, for many high throughput experiments, one could clearly see that the bulk of test statistics did not follow the null but rather showed over- or under-dispersion [30]. One potential culprit for this effect can be that that asymptotic arguments for convergence to normality of the test statistics do not hold yet as the actual sample size might be low particularly for low frequency variants. The first attempts to deal with this problem, was to scale the test statistics in such a way that the bulk of the data again followed the prescribed null distribution. In GWAS this method was introduced under the name of genomic control, whereas in differential expression analysis a similar method was introduced under the name of empirical null estimation [30, 31].

It was realized that the main reason for overdispersion was unaccounted confounding: hidden variables that showed an association with response variable as well as an association with many explanatory variables. This lead to methods that tried to estimate these underlying variables from the data [32]. It is clear that the confounding variables, that lead to overdispersion of  $p$ -values, need to associate with the bulk of the independent variables that are tested and therefore explain on average a lot of variation of the independent variables. The principal components of the sample covariance matrix therefore are likely to associate very strongly with the confounding variables leading to inflation and controlling for the largest principal components in a regression approach could therefore control confounding. This indeed turned out to be the case in GWAS applications. Furthermore, it turned out that the largest principal components in the case of GWAS with mixed populations would associate with geographic origin, which is the main source of confounding in GWAS [33].

A further advance in the control of population structure came with the use of linear mixed models.

These models have their origin in animal studies which are typically much smaller in size and often have a known confounding structure (through known kinship), making computationally demanding mixed models both feasible and pertinent. This strategy allows to apply linear regression to test the null for a particular SNP while at the same time allowing for the fact all other SNPs have a nonzero effect. These models therefore model overdispersion directly. To avoid overfitting, the assumption is made that these nonzero effects are normally distributed around zero and the variance of this distribution is estimated. This is a general strategy of mixed models where instead of fitting all effects separately only the distribution of effects are estimated leading to a model with much lower number of parameters. These models are computationally more demanding to fit but have advantages. First, they naturally incorporate lower order principal components. Additionally, even in the case of no confounding whatsoever, these models can be statistically more powerful. The reason is that the random effect can explain part of the variability of the phenotype making it easier to detect associations. To improve power further strategies that fit mixed effects models with more realistic distributional assumptions for the random effect may be used [34]. Further improvements in power can be gained by excluding the locus under investigation from contributing to the distribution of random effects.

In the chapter on fly growth control (chapter 4) we see an example use of mixed models as confounding control strategy. The chapter on chromatin state regulators (chapter 3) contains a use of such models in a new context. There, a complex correlation structure makes mixed modeling advantageous. With regards to chapter 2, there is a connection between the sum statistic used in *Pascal* and testing a random effects model for the gene region in question. We show this connection in Appendix B. We follow this thread to propose a way to estimate the pathway heritability via maximum likelihood from  $z$ -scores alone in Appendix C.

#### 1.4. Modern approaches to building genome wide protein-DNA binding maps

Just as the drop in the costs of DNA sequencing and genotyping have led to new strategies for investigating the genetics of phenotypic traits a comprehensive untargeted fashion, it has also allowed to scale up classical molecular biology assays to genome-wide levels. Examples of this are the Chromatin-immunoprecipitation (ChIP) assay and the DNase1 hypersensitivity (DHS) assay [35, 36]. Classical ChIP can be used to answer the question whether a protein of interest, be it a transcription factor or a specifically modified histone, is binding to a DNA region of interest. After reversibly cross-linking DNA and proteins, a specific protein antibody is used to bind and extract the protein under investigation. If the DNA region of interest is extracted along with the protein, one concludes that the region is bound by the protein. Cheap and reliable sequencing has allowed to move from targeting specific chromatin regions to investigating the whole genome. This approach is called ChIP-seq and since its inception almost 10 years ago it has become a very heavily used technique in genomics studies [37, 38, 39]. Another classical assay that has upscaled to genome wide levels is the DNase1 hypersensitivity assay. DNase1 hypersensitivity (DHS) relies on high sensitivity of certain DNA regions to digestion with the endonuclease DNase1. These regions of hypersensitivity to the enzyme were first discovered in drosophila heat shock genes and SV40 [40, 41, 42]. In parallel, Dnase1 footprinting and other *in vitro* protection assays were developed, to show sequence-specific DNA binding of a particular protein at base-pair resolution [43]. The method relied on the fact that protein binding would protect DNA from digestion by DNase1 and allowed to delineate where transcription factors and other DNA binding proteins such as histones would bind. While this was at first an *in vitro* such as histones would bind. While this was at first an *in vitro* method, it was subsequently shown that DHS sites were indeed depleted in histones, the most abundant DNA binding protein complex [44]. *In vivo* footprinting methods allowing to define specific protein binding at bp resolution followed suit [45]. Again, cheap and accurate sequencing allowed for these methods to be scaled up to the genome wide level [46, 47]: Chromatin is digested by DNase1 and restriction site fragments are sequenced in

a genome wide manner. Low sequencing depth uncovers DHS sites, regions of higher DNase1 sensitivity at the resolution of about 100 bp. Higher sequencing depths allow to find footprints of transcription factor binding as dips and stereotypical patterns in the DNase cleavage pattern at single bp resolution [48, 49]. Combining transcription factor binding motifs together with general DNase hypersensitivity data allows to predict with high accuracy where transcription factors are binding for a large fraction of transcription factors. However, whereas general DHS sites seem to be well correlated with transcription factor binding [49, 50], the increased precision yielded by detailed footprinting depends on the strength of binding interaction and will not work for factors with low DNA residence time with rapid on-off cycles [51, 52]. Nevertheless, one genome-wide DNase1 assay complemented with TF-motif information allows to approximate a comprehensive genome-wide map of transcription factor binding to DNA that would take hundreds of individual ChIP-Seq experiments to build.

The modern capabilities to generate comprehensive maps of protein-DNA binding histone and DNA modifications as well as expression has led to multiple concerted efforts to build such maps in a large array of cell types, cell lines, and tissues and use them to define functional elements in the genome. The first large-scale endeavour of this kind was the ENCODE project [53]. ENCODE mainly focused on cell lines relevant in research. The subsequent ROADMAP project in turn focussed on tissue samples [54]. These public resources are used extensively as reference data for new biological experiments, and continuous method development allows to refine the built maps and add new dimensions to them [51]. The functional elements called are also helpful in fine-mapping loci identified through GWAS [55, 56].

## 2. Pathway Analysis for GWAS data





Genome-wide association studies (GWAS) typically generate lists of trait- or disease-associated SNPs. Just as with differential expression studies, GWAS studies can be difficult to interpret either because the study was underpowered or the number of variants found to associate with the trait is overwhelming. For some highly complex traits, the number of relevant variants uncovered ranges in the hundreds. This situation calls again for strategies to derive biological insight.

The following chapter presents a paper that was published in PLOS Computational Biology. It focused on *Pascal* (Pathway scoring algorithm), a tool designed and implemented by me with the help of my co-authors. The tool allows for gene and pathway-level analyses of GWAS association results without the need to access the original genotypic data. *Pascal* was designed to be fast, accurate and to have high power to detect relevant pathways. Importantly, the lessons learned from differential expression analysis were not forgotten: Correlation between gene scores and their impact on pathway scores was addressed. Also, just as using t-statistics instead of fold changes for pathway analysis in differential gene expression can diminish subtle biases on the pathway level, we aimed to minimize these biases by using genewise  $p$ -values as a starting point for pathway scores. Simulation of realistic scenarios was used to show that pathway  $p$ -values were indeed well calibrated.


The paper also shows results of extensive testing of the approach on a large collection of real GWAS association results and saw better discovery of confirmed pathways than with other popular methods.

RESEARCH ARTICLE

# Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics

David Lamparter<sup>1,2</sup>, Daniel Marbach<sup>1,2</sup>, Rico Rueedi<sup>1,2</sup>, Zoltán Kutalik<sup>1,2,3</sup>,  
Sven Bergmann<sup>1,2</sup>

**1** Department of Medical Genetics, University of Lausanne, Lausanne, Switzerland, **2** Swiss Institute of Bioinformatics, Lausanne, Switzerland, **3** Institute of Social and Preventive Medicine (IUMSP), Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland

 These authors contributed equally to this work.

‡ ZK and SB also contributed equally to this work.

\* [Sven.Bergmann@unil.ch](mailto:Sven.Bergmann@unil.ch); [Zoltan.Kutalik@unil.ch](mailto:Zoltan.Kutalik@unil.ch).



 OPEN ACCESS

**Citation:** Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S (2016) Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLoS Comput Biol* 12(1): e1004714. doi:10.1371/journal.pcbi.1004714

**Editor:** Jennifer Listgarten, Microsoft Research, UNITED STATES

**Received:** June 16, 2015

**Accepted:** December 17, 2015

**Published:** January 25, 2016

**Copyright:** © 2016 Lamparter et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** Download location of meta analysis results are given in [S2 Table](#). Meta analysis pertaining to the colaus data set is given at (<http://www2.unil.ch/cbg/index.php?title=PascalTestData>). The software is available (<http://www2.unil.ch/cbg/index.php?title=Pascal>).

**Funding:** The CoLaus study was and is supported by research grants from GlaxoSmithKline (<https://www.gsk.com/>), the Faculty of Biology and Medicine of Lausanne, and the Swiss National Science Foundation (<http://www.snf.ch>) (grants 33CSCO-122661, 33CS30-139468 and 33CS30-148401). ZK received financial support from the Leenaards

## Abstract

Integrating single nucleotide polymorphism (SNP) p-values from genome-wide association studies (GWAS) across genes and pathways is a strategy to improve statistical power and gain biological insight. Here, we present *Pascal* (Pathway scoring algorithm), a powerful tool for computing gene and pathway scores from SNP-phenotype association summary statistics. For gene score computation, we implemented analytic and efficient numerical solutions to calculate test statistics. We examined in particular the sum and the maximum of chi-squared statistics, which measure the strongest and the average association signals per gene, respectively. For pathway scoring, we use a modified Fisher method, which offers not only significant power improvement over more traditional enrichment strategies, but also eliminates the problem of arbitrary threshold selection inherent in any binary membership based pathway enrichment approach. We demonstrate the marked increase in power by analyzing summary statistics from dozens of large meta-studies for various traits. Our extensive testing indicates that our method not only excels in rigorous type I error control, but also results in more biologically meaningful discoveries.

## Author Summary

Genome-wide association studies (GWAS) typically generate lists of trait- or disease-associated SNPs. Yet, such output sheds little light on the underlying molecular mechanisms and tools are needed to extract biological insight from the results at the SNP level. Pathway analysis tools integrate signals from multiple SNPs at various positions in the genome in order to map associated genomic regions to well-established pathways, i.e., sets of genes known to act in concert. The nature of GWAS association results requires specifically tailored methods for this task. Here, we present *Pascal* (Pathway scoring algorithm), a tool that allows gene and pathway-level analysis of GWAS association results without the need

Foundation (<http://www.leenaards.ch>), the Swiss Institute of Bioinformatics (<https://www.isb-sib.ch/>) and the Swiss National Science Foundation (31003A-143914, 51RTP0\_151019). SB received funding from the Swiss Institute of Bioinformatics, the Swiss National Science Foundation (grant FN 310030\_152724 / 1) and SystemsX.ch through the SysGenetiX project. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

to access the original genotypic data. *Pascal* was designed to be fast, accurate and to have high power to detect relevant pathways. We extensively tested our approach on a large collection of real GWAS association results and saw better discovery of confirmed pathways than with other popular methods. We believe that these results together with the ease-of-use of our publicly available software will allow *Pascal* to become a useful addition to the toolbox of the GWAS community.

## Introduction

Genome-wide association studies (GWAS) have linked a large number of common genetic variants to various phenotypes. For most common phenotypes, high-powered meta-analyses have revealed tens to hundreds of single nucleotide polymorphisms (SNPs) with robust associations. However, deriving biological knowledge from these associations is often challenging [1,2]. Many genes function in multiple biological processes and it is typically not clear which of these processes is related to the phenotype in question.

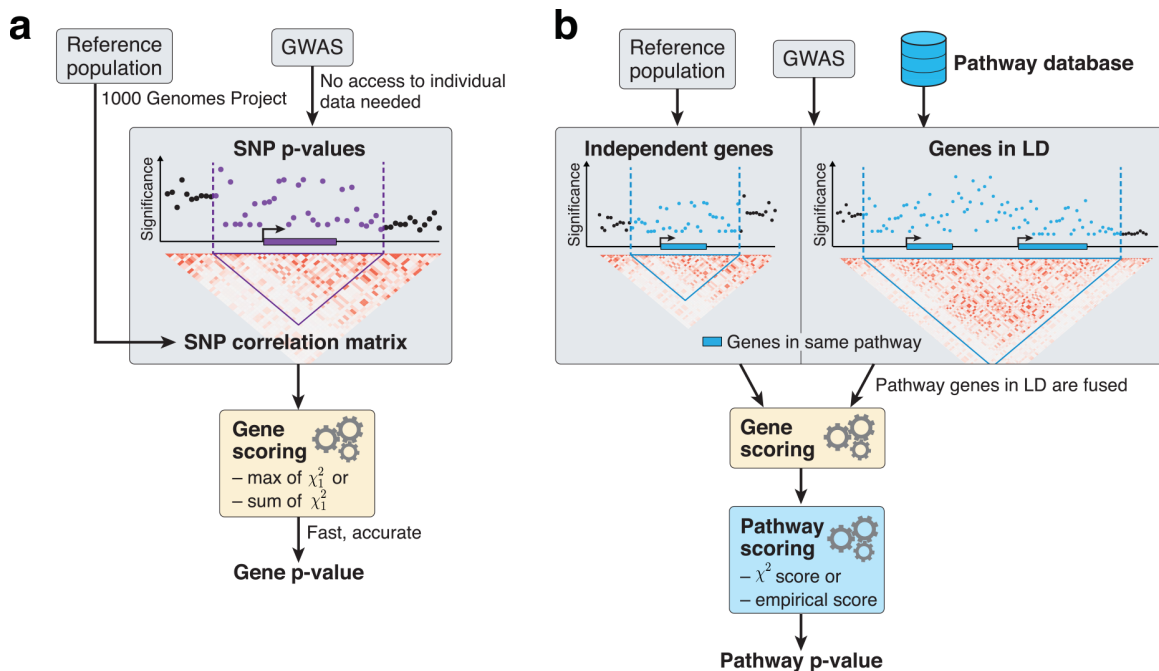
Pathway analysis aims to provide insight into the biological processes involved by aggregating the association signal observed for a collection of SNPs into a pathway level signal. This is generally carried out in two steps: first, individual SNPs are mapped to genes and their association p-values are combined into gene scores; second, genes are grouped into pathways and their gene scores are combined into pathway scores. Existing tools vary in the methods used for each step and the strategies employed to correct for correlation due to linkage disequilibrium.

SNPs are usually mapped to genes based on physical distance [3], linkage disequilibrium (LD) [4], or a combination of both [5]. Genes are commonly assigned to pathways using well-established databases (such as Gene Ontology [6], KEGG [7], PANTHER [8], REACTOME [9], BIOCARTA [10]) or in-house annotation (based on co-expression [4], for example).

Various methods have been developed to aggregate SNP summary statistics into gene scores [3,11,12]. A common aggregation method is to use only the most significant SNP within a window encompassing the gene of interest, for example by assigning the maximum-of-chi-squares (MOCS) as the gene score statistic [3,13] (the contributing chi-squared values can be obtained from SNP p-values by using the inverse chi-squared quantile transformation). Another method is to combine results for all SNPs in the gene region, for example by using the sum-of-chi-squares (SOCS) statistic [14]. Both the MOCS and SOCS statistics are confounded by several properties of the gene. Specifically, in both cases it is important to correct for gene size and LD structure to obtain a well-calibrated p-value for the statistic. In the remainder of this paper, we also refer to the p-values of the MOCS and the SOCS statistics as *max* and *sum gene scores*, respectively. P-values can be estimated by phenotype label permutation, but this method is both computationally intensive and requires access to genotype data of the actual study, which are rarely shared [15]. Thus, one often has access only to association summary statistics and not the individual genotypes. In this case, one method is to regress out confounding factors [3]. This approach is employed in the popular MAGENTA tool, but provides only a partial solution as substantial residual confounding still remains [3].

An alternative approach, which we take here, is to exploit the fact that the null distributions of the MOCS and SOCS statistics depend solely on the pairwise correlation matrix of the contributing genotypes. In the absence of the original genotypes, this correlation matrix can still be estimated from ethnicity-matched, publicly available genotypic data, as has been proposed by us and others for conditional multi-SNP analysis of GWAS results [16,17]. This approach has





**Fig 1. Overview of the methodology to compute gene and pathway scores.** a) We compute gene scores by aggregating SNP p-values from a GWAS meta-analysis (without the need for individual genotypes), while correcting for linkage disequilibrium (LD) structure. To this end, we use numerical and analytic solutions to compute gene p-values efficiently and accurately given LD information from a reference population (e.g. one provided by the 1000 Genomes Project[22]). Two options are available: the max and sum of chi-squared statistics, which are based on the most significant SNP and the average association signal across the region, respectively. b) We use external databases to define gene sets for each reported pathway. We then compute pathway scores by combining the scores of genes that belong to the same pathways, i.e. gene sets. The fast gene scoring method allows us to dynamically recalculate gene scores by aggregating SNP p-values across pathway genes that are in LD and thus cannot be treated independently. This amounts to fusing the genes and computing a new score that takes the full LD structure of the corresponding locus into account. We evaluate pathway enrichment of high-scoring (possibly fused) genes using one of two parameter-free procedures (chi-squared or empirical score), avoiding any p-value thresholds inherent to standard binary enrichment tests.

doi:10.1371/journal.pcbi.1004714.g001

been implemented in the *Versatile Gene-based Association Study* (VEGAS) software and yields results close to those from phenotype label permutation[11]. However, while VEGAS is faster than estimation via phenotype label permutation, it still relies on a Monte Carlo method for estimating the p-values. This limits its efficiency for highly significant gene scores.

Once gene scores have been computed, pathway analysis tools use various strategies to aggregate them across sets of related genes. The most common approach used for analysing GWAS meta-analysis results, as exemplified by the popular GWAS pathway analysis tool MAGENTA, is based on binary enrichment tests, which rely on a threshold parameter to define which genes are significantly associated with the trait[3,18]. However, with this strategy potential contributions of weakly associated genes that just missed the threshold are lost and there is no clear guidance on how the threshold parameter should be set. Indeed, it seems common practice to keep the default parameter without knowing whether other choices would produce better results[5].

In this work, we focus on improving two major aspects of pathway enrichment analysis (Fig 1). First, we incorporate numerical and analytic solutions for the p-value estimation of the MOCS and SOCS statistics. This removes the need for phenotype permutations or Monte Carlo simulations, thereby making the score computation faster. Second, we developed a rigorous type I error control strategy and implemented a modified Fisher method to compute parameter-free pathway scores[19]. While some elements of our algorithm have been proposed

in other fields of statistical genetics[20,21], the novelty of our method lies in the unique combination of sophisticated analytical methods employed for pathway analysis, which results in improved computational speed, precision, type I error control and power.

In the following, we first evaluate the performance of our tool, demonstrating its speed gains and robust control of type I error. Then, using precision-recall analyses, comparing small to large GWAS results for lipid traits and Crohn’s disease, we demonstrate that our pathway scoring approach exhibits a gain in power compared to binary enrichment. Finally, we apply our method to dozens of large meta-analysis studies and evaluate power by counting the number of pathways passing the Bonferroni-corrected p-value threshold.

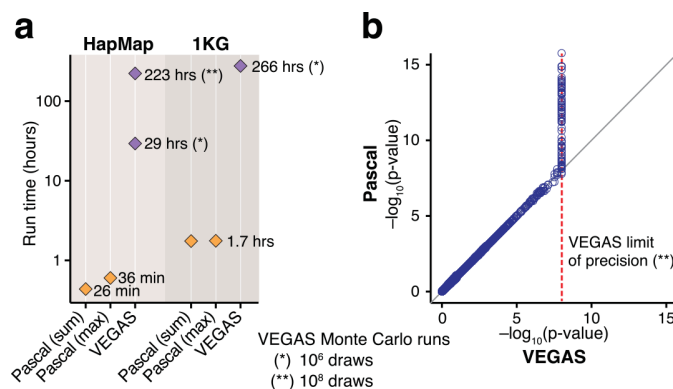
We provide this tool for gene and pathway scoring as a standalone, open-source software package called *Pascal*.

## Results

### *Pascal* computes genes scores rapidly and to very high precision

First, we compared the run time and precision of *Pascal* to those of VEGAS[11], one of the current state-of-the-art gene scoring tools. To this end, we applied both procedures to genome-wide p-values obtained from two large-scale GWAS meta-analyses: The first used about 2.5 million HapMap imputed SNPs[23,24] and the second was based on about 6.4 million SNPs imputed based on a common subset of 1000 Genomes Project (1KG) panel[22,25]. As benchmark we used the results from VEGAS for the former and VEGAS2 (a recent implementation of VEGAS that uses pre-computed LD matrices from 1KG[26]) for the latter. We observed a substantially smaller run time for our method in both cases (Fig 2A): for the HapMap imputed data, VEGAS took 29 hours to compute 18,132 gene scores, while *Pascal* was considerably faster, needing only about 30 minutes for either statistical test (sum or max) on a single core (Intel Xeon CPU, 2.8GHz). For the 1KG imputed data, *Pascal* finished the computation in under two hours for either statistic, whereas VEGAS2 took over ten days.

To compare the gene scores computed by the two methods, we increased the maximum number of Monte Carlo runs for VEGAS to  $10^8$ , at a high computational cost (about 9 days of runtime). We observed excellent concordance between the gene scores of *Pascal* and VEGAS,



**Fig 2. Comparing efficiency between VEGAS and *Pascal*.** a) Run times of VEGAS and *Pascal* (both options). Gene scores were computed on two GWAS (one HapMap imputed[23], one 1KG imputed[22,25]) for 18,132 genes on a single core. *Pascal* was compared to VEGAS for the HapMap imputed study and VEGAS2 for the 1KG-imputed study. For this plot, VEGAS and VEGAS2 were used with the default maximum number of Monte Carlo samples of  $10^6$  for both studies and additionally with  $10^8$  Monte Carlo samples for the HapMap imputed study. b) Scatter plot of  $-\log_{10}$ -transformed gene p-values for the sum gene scores obtained by VEGAS and *Pascal*, respectively. P-values above  $10^{-6}$  are in excellent concordance. Below this value VEGAS could not give precise estimates, since it was run with the maximal number of Monte Carlo samples set to  $10^6$ .

doi:10.1371/journal.pcbi.1004714.g002

except for scores below  $10^{-6}$ : since we restricted VEGAS to  $10^8$  Monte Carlo runs, it could not estimate p-values smaller than  $10^{-6}$  with good precision (Fig 2B). In contrast, *Pascal* can compute gene scores with high precision for p-values down to  $10^{-15}$ . In summary, the analytic solutions incorporated in the *Pascal* algorithm offer a dramatic increase in efficiency and precision. Direct comparison of the sum and max gene scores of *Pascal* revealed good concordance between the two scoring methods. In cases where the results of two methods disagree, max scores tend to be more significant (S3 Fig).

The results reported here are all based on GWAS of European cohorts, thus we used the European panel from 1KG as reference panel. To evaluate whether this panel approximates LD matrices derived from other European cohorts sufficiently well, we compared results when using genotypes taken from the *CoLaus* cohort as reference panel [27]. We saw good concordance between the different reference panels for both the sum and the max gene scores for the largest HDL blood lipid GWAS to-date [23] (S2 Fig).

### *Pascal* controls for inflation due to neighbouring genes

In general, methods that compute pathway scores from gene scores assume independence of these scores under the null hypothesis. However, neighbouring genes often have correlated scores due to LD, and are sometimes part of the same pathway. This results in a non-uniform pathway score p-value distribution under the null hypothesis. MAGENTA deals with this problem by pruning gene scores based on LD and using only the highest gene score in the region. However, this introduces a bias toward high gene scores into the calculation of pathway scores [3].

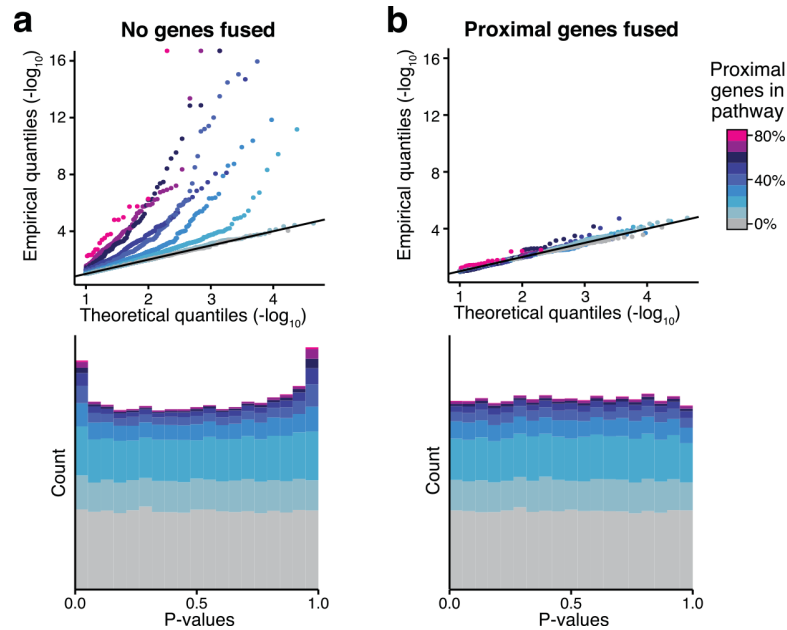
Our fast gene score calculation allows us to address this issue with a gene-fusion strategy. In brief, for each pathway harbouring correlated genes, gene scores are recomputed jointly for each correlated gene set (i.e. fusion-gene) using the same method as for individual genes (Fig 1B, Methods), thus taking the full LD structure of the corresponding region into account.

To see if our approach provides well-calibrated p-values, we simulated random phenotypes and calculated association p-values for all 1KG SNPs. We then employed our pathway analysis pipeline and checked if pathway p-values were uniformly distributed, as expected for random phenotypes. We found that without the gene-fusion strategy, pathway p-values are indeed inflated and, as expected, this inflation is stronger for pathways with many proximal genes (Fig 3A). In contrast, applying the gene-fusion strategy corrects the distribution of pathway score p-values to be uniform irrespective of the number of proximal genes (Fig 3B). Importantly, we did not see inflation for very small p-values with the gene-fusion strategy, which is essential for type I error control.

Going one step further, we also simulated *in-silico* phenotypes influenced by randomly selected causal SNPs. We explored two scenarios: one where 50 SNPs were randomly selected from the entire genome and another where random sampling was applied to gene regions only. The experiment was repeated 50 times and independent genetic data was used to generate the estimated pairwise correlation. Although in this case gene scores naturally deviate from the null distribution, we found that overall pathway p-values remain well calibrated (S14 and S15 Figs). Note that we explored only a limited set of simulation scenarios and cannot exclude that some settings might produce less well-calibrated results (see legend of S15 Fig).

### *Pascal* has higher sensitivity and specificity than hypergeometric pathway enrichment tests

A commonly used statistic to derive pathway scores from a ranked list of genes (or SNPs) is to first apply a fixed threshold in order to define a subset of elements that is considered to be significantly associated with the given trait. The pathway statistic is then computed using a



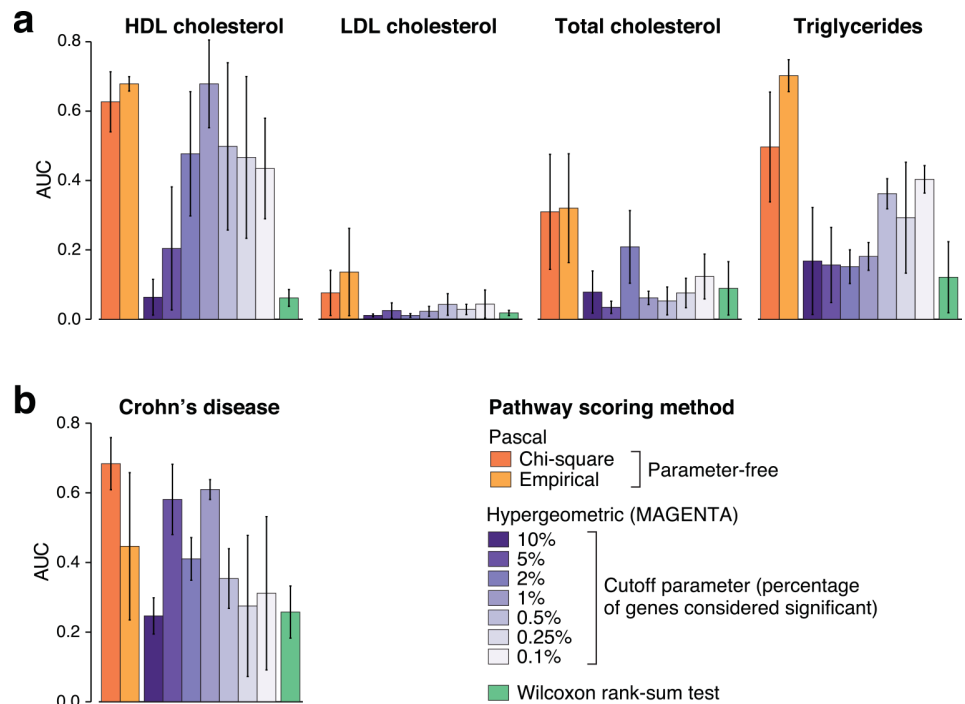
**Fig 3. Pathway scores for random phenotypes.** As input data we used 100 simulated instances of a random Gaussian phenotype and genotype data for 379 individuals from the EUR-1KG panel. Using the *Pascal* pipeline with sum gene scores and chi-squared pathway integration strategy we computed p-values for 1,077 pathways from our pathway library (results for max gene scores are similar, see S4 Fig). Panel (a) shows the p-value distributions without merging of neighbouring genes and (b) with merging of neighbouring genes (gene-fusion strategy). P-value distributions are represented by QQ-plots (upper panels) and histograms (lower panels). Results are colour-coded according to the fraction of genes in a given pathway that have a neighbouring gene in the same pathway, i.e. that are located nearby on the genome (distance <300kb). (a) P-values of pathways that contain genes in LD are strongly inflated without correction. (b) The gene fusion approach provides well-calibrated p-values independently of the number of pathway genes in LD.

doi:10.1371/journal.pcbi.1004714.g003

hypergeometric test evaluating whether the pathway contains more significant elements than expected. This approach is implemented, for example, in the tool MAGENTA[3]. Another common strategy is to use the rank-sum (Wilcoxon) test[3,28,29].

As described above, *Pascal* computes aggregate statistics without the need for defining a set of significant genes. We thus sought to compare this strategy with methods based on the hypergeometric or rank-sum tests. To this end, we tested performance on association results for four blood lipid traits obtained from of the *CoLaus* cohort[27]. We used a large meta-analysis of 188,577 individuals to define a reference set of associated pathways for each of the four lipid traits[23]. We then applied both pathway analysis methods to three non-overlapping, small subsets (1500 individuals) of the *CoLaus* study and compared how well the resulting pathways matched the reference set from the large study. We used the area under the precision-recall curve (AUC-PR) to quantify the performance of each method. Note that our choice was driven by the fact that precision-recall curves are preferred over receiver-operator-characteristic (ROC) curves when only a small fraction of tested pathways are in the reference set[30]. Our results show that *Pascal* outperforms both the hypergeometric and rank-sum based approaches (Fig 4A). Importantly, the better performance of *Pascal* is observed across a range of thresholds defining significant genes, including the optimal choice which is variable and *a priori* unknown across the different lipid phenotypes.

We applied the same evaluation strategy for GWAS data on Crohn's disease. We used the currently largest GWAS for Crohn's disease[31] to define a reference standard of associated



**Fig 4. Performance of pathway enrichment methods for blood lipid traits and Crohn's disease.** Displayed is the mean area under the precision-recall curve (AUC) for pathways identified using *Pascal*, a standard hypergeometric test at various gene score threshold levels, and a rank-sum test (vertical bars show the standard error). We show results for the max gene scores (sum gene score results are similar, see [S5 Fig](#)). a) Results for four blood lipid traits. The gold standard pathway list was defined as all pathways that show a significance level below  $5 \times 10^{-6}$  for any of the tested threshold parameters for hypergeometric tests in the largest study of lipid traits to date [23]. The significance level of  $5 \times 10^{-6}$  corresponds to the Bonferroni corrected, genome-wide significance threshold at the 0.5% level for a single method. For each phenotype, error bars denote the standard error computed from three independent subsamples of the *CoLaus* study (including 1500 individuals each). We see good overall performance of *Pascal* pathway scores, whereas results for discrete gene sets vary widely with the particular choice for the threshold parameter of hypergeometric test. b) Results for Crohn's disease using the same approach as in (a). A reference standard pathway list was defined as in (a) using the largest study of Crohn's disease traits to date [31]. We observe that the chi-squared strategy performs at least as well as all other strategies in this setting, whereas performance of the hypergeometric testing strategy varies.

doi:10.1371/journal.pcbi.1004714.g004

pathways. We then applied both pathway analysis methods to results from two individual cohorts participating in the meta-analysis that contained at least 1000 cases [31–33]. We observed that the chi-squared-method performed at least as well as all other strategies in this setting (Fig 4B). Overall, we saw similar results for both max and sum gene scores (S5 Fig).

### *Pascal* has higher power than hypergeometric test based pathway enrichment in a wide range of traits

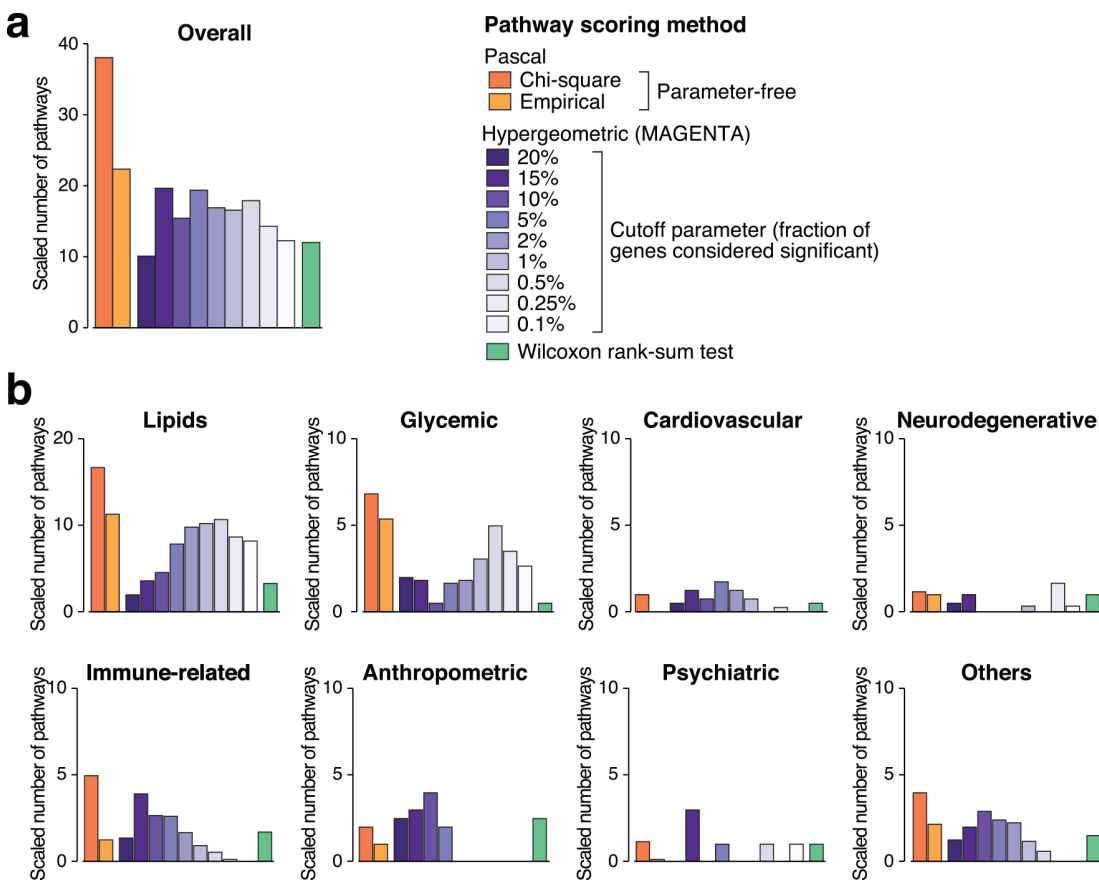
Having established that *Pascal* accurately controls type I error rate for simulated phenotypes and better recovers truly associated pathways for blood lipid traits as well as Crohn's disease, we next sought to evaluate its power when applied to large meta-analytic studies on a broad range of traits, where no ground truth can be defined.

To this end, we compared *Pascal* with the methods based on the hypergeometric test (using 9 different threshold values) and the rank-sum test proposed by Segrè *et al.* [3] for 118 GWAS

(S1 Table). All GWAS were derived from European populations justifying the use of the European 1KG genotypes as reference population. For a given GWAS, we asked how many tested pathways reached genome-wide significance at the Bonferroni-corrected p-value threshold of 0.05. Our results indicate that globally our approach has higher power than either the methods using the hypergeometric test (across all tested thresholds), or the rank-sum test (Figs 5A and S6). For individual traits (Fig 5B), specific choices of the threshold parameter of the hypergeometric test sometimes reveal more pathways, but again the value of the optimal threshold varies across traits and cannot be known *a priori*.

When splitting the GWAS into high powered (more than 50,000 individuals) and low powered studies (less than 50,000 individuals), we saw that in both cases we gain power by using *Pascal* although the effect was more pronounced for low powered GWAS (S7 Fig).

Hypergeometric enrichment testing is hampered by the fact that the optimal threshold is not known in advance. A strategy to overcome this could be to merge hypergeometric pathway scores coming from different sets of thresholds, further corrected for the effective size of the threshold sets. While such an aggregated hypergeometric testing improved performance, it was still outperformed by *Pascal* (S10 and S11 Figs).

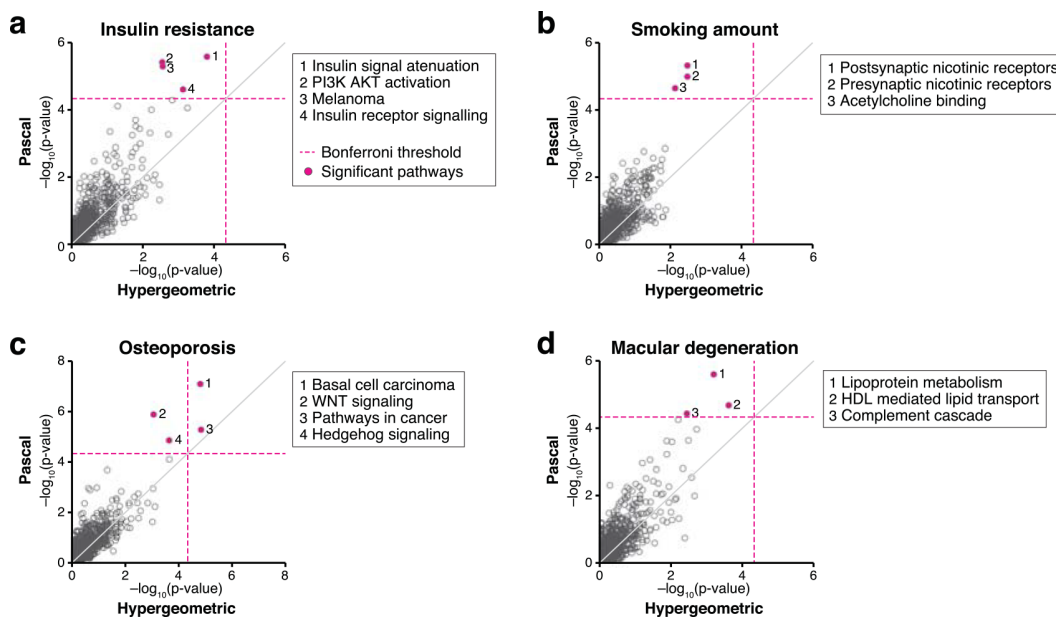


**Fig 5. Power of pathway scoring methods across diverse traits and diseases.** Bar heights represent the number of pathways found to be significant after Bonferroni-correction. Within a given trait group, results are aggregated for all tested GWAS studies. 65 GWAS had at least one significant pathway in one of the tested methods. For each GWAS, the raw number of significant pathways was divided by the number of pathways found by the best performing method. This was done in order to avoid that a few studies with many emerging pathways dominate. We show results for the MOCS gene scores (SOCS gene score results are similar, see S6 Fig). (a) Results are aggregated over all trait groups. (b) Results for different trait groups.

doi:10.1371/journal.pcbi.1004714.g005

One of the proposed pathway scoring methods transforms the ranked gene p-values such that they follow a chi-squared distribution. The chi-squared distribution is a special case of the Gamma distribution with shape parameter 0.5. Thus we also examined whether using other shape parameters of the Gamma distribution could improve performance (see [Methods](#), [S12](#) and [S13](#) Figs). This analysis suggested that the chi-squared pathway scoring method represents a good compromise for a wide range of genetic architectures.

We found numerous examples of biologically plausible pathways discovered by *Pascal* that were not found by a standard binary enrichment analysis ([Fig 6](#), [S2 Table](#)). For insulin resistance [[34](#)] we found the REACTOME pathway *insulin signal attenuation* to be genome-wide significant. Notably, none of the genes in this pathway was found to contain a genome-wide significant SNP in the original publication. Another example is bone mineral density in women (LS-BMD) [[35](#)]. We found the *Hedgehog* and *Wnt* pathways to be significant, both of which are known to be involved in osteoblast biology [[36](#)]. Again, standard binary enrichment did not reach genome-wide significance. For smoking behaviour (measured in cigarettes per day) [[37](#)], we found pathways related to *nicotinic acetylcholine receptors*. For macular degeneration, we found *lipoprotein* and *complement system* involvement, which both have support in the literature [[38,39](#)]. These examples illustrate that the improvements made by *Pascal* not only lead to better performance on benchmarks, but may also have a dramatic impact on the interpretation of GWAS results in practice.



**Fig 6. Examples of pathway enrichments comparing *Pascal* (chi-squared method) to the hypergeometric method.** Displayed are results for four phenotypes showing improvement when using *Pascal* instead of the hypergeometric (binary) enrichment strategy at the 5% threshold level. Underlying gene scores were calculated using the sum method. Dashed lines refer to the Bonferroni significance level when correcting for the number of pathways (1077). Besides from few cancer-related pathways, all pathways highlighted by this analysis have been implied by prior research (see main text). (a) For the trait *insulin resistance*, *Pascal* scored the pathway *insulin signal attenuation* first, followed by two other trait-relevant pathways (*PI3K AKT activation* and *insulin receptor signalling*), while the hypergeometric test did not find any significant pathways. (b) For *smoking amount* (number of cigarettes per day), *Pascal* revealed three significant pathways related to *nicotinic acetylcholine receptors*. (c) For *osteoporosis*, two cancer-related pathways scored significant using both *Pascal* and the hypergeometric test, but only *Pascal* revealed the *WNT* and *Hedgehog* signaling pathways, which are known to be involved in osteoblast biology. (d) For *macular degeneration*, *Pascal* found three significant, trait-relevant pathways related to lipoproteins and the complement system.

doi:10.1371/journal.pcbi.1004714.g006

## Discussion

In this work, we presented a new tool called *Pascal* (*Pathway scoring algorithm*) that specifically addresses both gene scoring and pathway enrichment, making significant advancement with respect to the state-of-the-art:

First, our gene score calculation combines analytical and numerical solutions to properly correct for multiple testing on correlated data[21]. While some of these approaches have already been applied within the rare variant field[20] (typically in a gene-wise fashion) we provide a streamlined implementation that can run genome-wide analyses without the need for any Monte Carlo simulations (making it about 100 times faster and more precise than the widely used software VEGAS).

Second, our pathway scoring integrates individual gene scores without the need for a tuneable threshold parameter to dichotomize gene scores for binary membership enrichment analysis (as done for example by MAGENTA). The choice of such a parameter is not straightforward and our method usually performs better, regardless of the chosen parameter.

Third, we show that the null distribution of enrichment p-values for pathways that contain genes in linkage disequilibrium is non-uniform due to an “over-counting” of gene association signals. This is a potential source of type I error underestimation and our method corrects for this phenomenon using a *gene fusion* approach, which considers genes in LD as single entities.

We have extensively evaluated the performance of *Pascal* for several real data sets. These comparisons demonstrated the rigorous control of type I error and superior predictive power in a wide range of trait and power settings in terms of enhanced precision-recall curves.

As an additional global measure of power, we considered the number of significantly enriched pathways for a large number of GWAS meta-analysis summary statistics. On average, our approach resulted in higher numbers of significant pathway scores than any binary enrichment strategy. Given its precise type I error control, this provides additional evidence of increased power for a wide range of traits. Indeed, the elevated rate of putatively involved pathways produced by our method not only reflects its higher sensitivity, but also already generates new hypotheses for further studies.

Taken together, our results demonstrate the superior performance of our approach compared to standard binary enrichment and rank-sum tests. Although methods with tuneable parameters might yield improved results in a particular setting, it is difficult to predict the optimal parameter choice. Indeed, the optimal parameter depends on sample size, as well as complexity and heritability of the phenotype. Another issue with binary enrichment tests is that the hypergeometric distribution is discrete, which leads to conservative p-values, especially if the expected number of successful draws is low. Our pathway scoring approach avoids this problem. Also, our approach lends itself to naturally extending pathway scoring in case genes have probabilistic membership in predefined pathways.

Users of our method will still have to make two choices: how to convert SNP p-values to gene scores (max or sum gene scores), and how to transform gene scores into pathway scores (empirical or chi-squared). We do not see evidence that one gene scoring method systematically outperforms the other in the context of our chi-squared pathway scoring method, while there seems to be a better performance for sum gene score when using the empirical approach (S8 Fig). To investigate this phenomenon we winsorized p-values (i.e. extreme p-values below  $10^{-12}$  were set to  $10^{-12}$ ) and saw that the max gene score combined with empirical sampling suffered far less performance loss (S9 Fig). We therefore conclude that the power loss is due to outlier gene scores. The max gene-scores can lead to very high gene scores for high-powered studies. In the extreme case one gene might reach scores so high that it precludes detection of pathways not containing that gene when *the empirical sampling* strategy is used.



Future work could attempt to enhance several other aspects of our pathway enrichment analysis. For example, here we mapped SNPs to genes only based on physical distance, while potential improvements could be attained by incorporating additional information, such as eQTL data [40] and functional annotations, to assign weights to different association signals within a locus. While our approach is amenable to such a weighting scheme, this would potentially require the introduction of tuneable parameters, which we avoided so far. Furthermore, one may attempt to redefine gene sets based on external unbiased large-scale molecular data, such as expression data, while so far we only used the established (but likely biased) pathway collections [4]. To this end, we already integrated *Pascal* into a pipeline to analyze the connectivity between trait-associated genes across over 400 tissue-specific regulatory, co-expression and protein-protein interaction networks, further demonstrating its value for network-based analysis of GWAS results (Marbach *et al.*, submitted).

As an additional caveat, we should mention that *Pascal* uses the European 1KG sample as reference population per default. This choice may not be appropriate if the studied sample is not of European origin. In this case the user is encouraged to supply *Pascal* with the appropriate reference panel. Also, SNPs with low MAF are by default excluded from the analysis, because the low number of individuals in the European 1KG sample limits the accuracy of the LD estimate for low frequency variants. If the user wishes to include lower frequency variants, the use of a reference sample containing more individuals is recommended.

To conclude, *Pascal* implements fast and rigorous analytical methods into a single analysis pipeline tailored for gene scoring and pathway enrichment analysis that can be run on a desktop computer. We thus hope that *Pascal* will be useful to the GWAS community in a range of applications and play a pivotal role in leveraging the rich information encoded in GWAS results both for single traits and—given its efficiency and power—in particular also for high-dimensional molecular traits.

Our tool is available as a single standalone executable java package containing all required additional data at: <http://www2.unil.ch/cbg/index.php?title=Pascal> (short URL: <http://goo.gl/t4U5z6>).

## Materials and Methods

### Gene scores

The *Pascal* gene scoring method consists of the following steps (Fig 1A). First, we assign SNPs to genes if they are located within a given window around the gene body. For the experiments reported in this paper, we used windows extending 50kb up and downstream from the gene. A reference population is required to estimate the correlation structure between Z-scores of SNP association values. Here, we used the European population of the 1000 Genomes Project (1KG) [22], which allows us to apply our approach flexibly to summary statistics from diverse panels (HapMap, 1KG imputed, metaboChip or ImmunoChip).

Under the null hypothesis, it can be shown that the Z-scores of  $n$  SNPs in our gene region as multivariate normal:

$$\mathbf{z} \sim \mathcal{N}_n(0, \Sigma)$$

where  $\Sigma$  is the pair-wise SNP-by-SNP correlation matrix (see Section ‘Derivation of the sum score’ for details).

We define our base statistics, the SOCS ( $T_{sum}$ ) and MOCS ( $T_{max}$ ), as:

$$T_{sum} = \sum_{i=1}^n z_i^2$$

and

$$T_{max} = \max(z_i^2),$$

respectively. It can be shown that  $T_{sum}$  is distributed according to the weighted sum of  $\chi_1^2$ -distributed random variables:

$$T_{sum} \sim \sum_i \lambda_i \chi_1^2$$

where  $\lambda_i$  is the  $i$ -th eigenvalue of  $\Sigma$ . Its distribution function can be evaluated numerically (see Section [Algorithmic details for gene-score calculations](#) for details). To estimate the null distribution of  $T_{max}$  we make use of the fact that

$$P[T_{max} \geq t] = P[\max(|z_i|) \geq t] = 1 - P(|z_i| < t, i = 1, 2, \dots, n).$$

This amounts to a rectangular integration over a multivariate normal, for which an efficient algorithm is available[41]. The current implementation of this integration is suitable to estimate p-values larger than  $10^{-15}$ . To approximate gene-wise p-values below this limit we multiply the minimum p-value of SNPs in the region with the effective number of tests within the gene (see Section [Algorithmic details for gene-score calculation](#)).

## Gene fusion

Pathway analysis methods typically assume that the gene scores used to define pathway enrichment are independent. However, functionally related genes often cluster on the genome and harbor SNPs in LD, leading to correlated gene scores that violate this assumption. To circumvent this problem, we check for a given pathway if any of its genes that cluster physically close on the chromosome are in LD. If so, for the calculation of the pathway score, we consider a single entity (a so-called *fusion-gene*) consisting of all the SNPs of the gene cluster. We then replace the genes in the cluster by this *fusion-gene* and calculate its gene score, but only for the calculation of the score for this particular pathway. The pathway score is then computed from the p-values of independent pathway genes and fusion genes that integrate the associational signals from dependent pathway genes (Fig 1B). In this way, the LD structure of neighbouring pathway genes is taken into account. Our gene scoring method facilitates this approach because it is sufficiently fast and scalable for recomputing the scores of all fusion genes.

## Pathway scores

For pathway analysis, we propose a parameter free enrichment strategy that does not require the specification of a gene score threshold, and thus allows weakly associated genes to contribute to pathway enrichment. The general approach consists of three steps: (1) gene scores are transformed so that they follow a target distribution, (2) a test statistic is computed by summing the transformed scores of pathway member genes and fusion-genes, and (3) analytic or empirical methods are used to evaluate whether the observed test statistic is higher than expected, i.e., the pathway is enriched for trait-associated genes. We considered two variants of this approach for pathway scoring (see overview in S1 Fig). The first variant is termed as the *chi-squared method*:

1. Gene score p-values are ranked such that the lowest p-value gets the highest rank. The rank value is then divided by the number of genes plus one to obtain a uniform distribution.
2. Uniform distribution values are transformed by the  $\chi_1^2$ -quantile function to obtain a  $\chi_1^2$ -distribution of gene scores.

- $\chi_1^2$ -gene scores of a given pathway of size  $m$  are summed and tested against a  $\chi_m^2$ -distribution.

The second variant is the *empirical sampling* method:

- Gene score p-values are directly transformed with the  $\chi_1^2$ -quantile function to obtain new gene scores:  $F_{\chi_1^2}^{-1}(1 - p)$ .
- A raw pathway score for a pathway of size  $m$  is computed by summing the transformed gene scores for all pathway genes.
- A Monte Carlo estimate of the p-value is obtained by sampling random gene sets of size  $m$  and calculating the fraction of sets reaching a higher score than gene set of the given pathway.

We also tested a generalization of the chi-squared method where the inverse  $\chi_1^2$ -quantile transformation of the p-value ranks was replaced by the inverse Gamma-quantile transformation with varying shape parameter. For shape parameter of 0.5, the results coincide with results from the chi-squared method.

For our benchmarking procedures we created a pathway library by combining the results from KEGG[7,42], REACTOME[9] and BIOCARTA[10] that we downloaded from MsigDB [43].

### Derivation of the sum-score

Let  $\mathbf{z}$  be the vector of Z-statistics coming from regressing the phenotype on each of the  $n$  SNPs within a gene-region. By construction, each Z-statistic has zero mean under the null. When both the outcome trait and the genotypes are standardized, the linear regression Z-statistics are essentially the scalar products of the genotype and the phenotype vectors. In other words, each Z-statistic in the region represents a weighted average of the same set of independent, identically distributed random variables. It can be shown that the correlation between two such mixtures, i.e. two Z-statistics, equals to the correlation between the weights, i.e. the correlation between the corresponding SNPs. Thus, the covariance matrix of  $\mathbf{z}$  is simply the pairwise SNP-by-SNP correlation matrix, denoted by  $\Sigma$ . Furthermore, the central limit theorem ensures that in case of sufficiently large sample size the Z-statistics are normally distributed. These facts put together yield that—under the null-hypothesis that no signal is present— $\mathbf{z}$  follows a multivariate normal distribution,  $\mathbf{z} \sim \mathcal{N}_n(0, \Sigma)$ . For a detailed derivation see supplementary material in Xu et al[44] for example. Note that the between SNP correlation matrix  $\Sigma$  can be estimated from external data[17,45].

The eigenvalue decomposition of  $\Sigma$  is

$$\Sigma = \mathbf{\Gamma} \mathbf{\Lambda} \mathbf{\Gamma}^T,$$

where  $\mathbf{\Gamma}$  and  $\mathbf{\Lambda}$  are the matrices of eigenvectors and eigenvalues, respectively. We see that multiplying  $\mathbf{z}$  with the inverse of the square-root of  $\Sigma$  leads to a vector of independent random variables. Let  $\mathbf{y}$  be defined as

$$\mathbf{y} = \mathbf{\Lambda}^{-1/2} \mathbf{\Gamma}^T \mathbf{z},$$

then

$$\mathbf{y} \sim \mathcal{N}_n(0, \mathbf{I}_n).$$

It follows that

$$\mathbf{z}^T \mathbf{z} = \mathbf{z}^T \mathbf{\Gamma} \mathbf{\Gamma}^T \mathbf{z} = \mathbf{y}^T \mathbf{\Lambda} \mathbf{y} \sim \sum_i \lambda_i \chi_1^2,$$

where  $\lambda_i$  is the  $i$ -th eigenvalue of  $\Sigma$  and  $\chi_1^2$  represents the chi-squared distribution.

### Parameter settings

If not stated otherwise, our tool was always used with the following settings. We extended gene regions by 50kb upstream and downstream for gene scoring. Only SNPs that reached a MAF of 0.05 in European 1KG sample were used. For pathway score calculation, we removed the HLA-region. The gene-fusion parameter was set to 1Mb, so that when calculating a particular pathway score, all pathway-member genes less than 1Mb apart were fused for the calculation. We also removed genes containing more than 3000 SNPs except during speed benchmarking (Fig 2) where all SNPs were used.

### Simulation settings for type I error control of the pathway scores

We used genotypes for 379 individuals from the EUR-1KG cohort[22]. Corresponding phenotype values were simulated as independent, standard normally distributed variables. Univariate Z-scores for each of the 2,692,429 tested SNPs were calculated using linear regression. Simulations were repeated 100 times. Since we investigated the impact of gene-fusion, the LD matrix was estimated from the same data set to avoid any influence that might come from out-of-sample LD estimation.

### Algorithmic details for gene-score calculations

**Max-score.** The algorithm first tries to use Monte Carlo simulation to derive p-values. Should the p-value be too small to be estimated within a few Monte Carlo draws, the procedure makes use of an algorithm for rectangular multivariate normal integration[41]. The implementation of the integration algorithm that is used is suitable to estimate p-values larger than  $10^{-15}$ . In addition, this implementation is limited to correlation matrices of size below 1000 due to numerical stability concerns. Therefore, SNPs that are in very high LD ( $r^2 > 0.98$ ) are pruned to lower the size of the correlation matrix. If more than 1000 SNPs fall into the gene or the gene-wise p-value is below  $10^{-15}$ , we approximate the gene score by multiplying the minimal SNP-wise p-value in the gene region by the effective number of tests. The effective number of tests is calculated as the minimum number of principal components needed to explain 99.5% of total variance[46].

**Sum-score.** The algorithm relies on the Davies algorithm to calculate distribution function values of weighted sums of independent  $\chi_1^2$ -distributed random variables[47]. In case of convergence problems the Farebrother algorithm is used as a backup[48,49].

**Web resources.** A stand-alone executable for *Pascal* can be found at <http://www2.unil.ch/cbg/index.php?title=Pascal>. The *Pascal* source code can be found at <https://github.com/dlampart/Pascal>.

### Supporting Information

**S1 Fig. Overview of pathway scoring strategies in *Pascal*.** Pathway scores are computed from gene scores. The upper panel shows a typical gene score distribution, where the pathway gene scores are indicated in black. In order to compute pathway scores, the original gene score p-values need to be transformed. To this end we use one of two strategies: in our *empirical strategy* (lower left panel), gene score p-values are directly transformed with the inverse  $\chi_1^2$ -quantile

function  $F_{\chi_1^2}^{-1}(1 - p)$  to obtain scores, which are then summed across all pathway genes. A Monte Carlo estimate of the p-value is then obtained by sampling random gene sets of the same size and calculating the fraction of sets reaching a higher score than that of the given pathway. In the *chi-squared method* (bottom right panel), the gene score p-values are first ranked such that the lowest p-value ranks highest. The rank values are then divided by the number of genes plus one to define new p-values ( $p_{\text{rank}}$ ) that are distributed uniformly by definition. From there, we proceed as for the empirical strategy just replacing  $p$  by  $p_{\text{rank}}$ . Also, since the scores are guaranteed to be chi-squared distributed, the computation of their corresponding p-value can be done analytically without any loss in precision.  
(PDF)

**S2 Fig. Comparison of results for different reference panels.** Comparing p-values computed using LD matrices from the European 1000 Genome reference panel and the *CoLaus* cohort. GWAS summary statistics were taken from a large-scale blood-HDL level meta-analysis. Results are compared for (a) max gene scores; (b) max gene scores excluding gene scores that were computed with the effective number of tests approximation; and (c) sum gene scores. There is good concordance in all cases.  
(PDF)

**S3 Fig. Comparison of max and sum gene scores.** We compared max and sum gene scores directly for a large-scale blood HDL level meta-analysis. Only gene scores up to  $10^{-15}$  are displayed, which truncated 6 genes with very large max scores.  $R^2$  between the  $-\log_{10}$ -transformed variables is 90%. Max scores tend to be larger when the two methods do not agree.  
(PDF)

**S4 Fig. Pathway scores for random phenotypes using max gene scores.** P-values for 1077 pathways from our pathway library were computed for 100 random phenotypes using the *Pascal* pipeline using max gene scores and *chi-squared* pathway integration strategy (a) without merging of neighbouring genes and (b) with merging of neighbouring genes (gene-fusion strategy). P-value distributions are represented by QQ-plots (upper panels) and histograms (lower panels). Results are colour-coded according to the fraction of genes in a given pathway that have a neighbouring gene in the same pathway, i.e. that are located nearby on the genome (distance <300kb). (a) P-values of pathways that contain genes in LD are strongly inflated without correction. (b) The gene fusion approach provides well-calibrated p-values independently of the number of pathway genes in LD.  
(PDF)

**S5 Fig. Performance of pathway enrichment methods for blood lipid traits and Crohn's disease using sum of squares (SOCS) statistics for defining gene scores.** Displayed is the mean area under the precision-recall curve (AUC) for pathways identified using *Pascal*, a standard hypergeometric test at various gene score thresholds, and a rank-sum test (vertical bars show the standard error). We show results for the SOCS gene scores (MOCS gene score results are similar, see Fig 4 in the main text). a) Results for four blood lipid traits. A reference standard pathway list was defined as all pathways that show a significance level below  $5 \times 10^{-6}$ , for any of the tested threshold parameters for hypergeometric tests in the largest study of lipid traits to date. The significance level of  $5 \times 10^{-6}$  corresponds to the Bonferroni corrected, genome-wide significance threshold at the 0.5% level for a single method. For each phenotype, error bars denote the standard error computed from three independent subsamples of the *CoLaus* study (including 1500 individuals each). We see good overall performance of *Pascal* pathway scores, whereas results for discrete gene sets vary widely with the particular choice for the threshold

parameter of hypergeometric test. b) Results for Crohn's disease using the same approach as in (a). A reference standard pathway list was defined as all pathways that show a significance level below  $5 \times 10^{-6}$  for *any* of the tested threshold parameters for hypergeometric tests in the largest study of Crohn's disease traits to date. We observe that the chi-squared strategy outperforms all other strategies in this setting, whereas performance of the hypergeometric testing strategy varies.

(PDF)

**S6 Fig. Power of pathway scoring methods across diverse traits and diseases using sum of squares (SOCS) statistics for defining gene scores.** Bar heights represent the number of pathways found to be significant after Bonferroni correction. Within a given trait group, results are aggregated for all tested GWAS studies. 65 GWAS had at least one significant pathway in one of the tested method. For each GWAS, the raw number of significant pathways was divided by the number of pathways found by the best performing method. This was done to avoid that a few studies with many emerging pathways dominate. We show results for the SOCS gene scores (MOCS gene score results are similar, see Fig 5). (a) Results are aggregated over all trait groups. (b) Results for different trait groups.

(PDF)

**S7 Fig. Power of pathway scoring methods stratified with respect to sample size.** Only GWAS studies for quantitative traits were used. Top panels (a,b) show results for max gene scores and bottom panels (c,d) show results for sum gene scores. (a,c) Results for all studies where the number of individuals was below 50,000. (b,d) Results for studies with sample sizes above 50,000. We see power gains in all cases. The improvements are particularly pronounced in lower powered GWAS.

(PDF)

**S8 Fig. Power comparison max and sum gene scores for pathway analysis.** Bar heights represent the number of pathways found to be significant after Bonferroni correction. Within a given trait group, results are aggregated for all tested GWAS studies. For each GWAS, the raw number of significant pathways was divided by the number of pathways found by the best performing method. Results for SOCS and MOCS as well as the chi-square and empirical pathway scores are displayed. We observe a drop in performance for the combination of MOCS gene scores with empirical pathway scores.

(PDF)

**S9 Fig. Power analysis for max gene scores with capped gene scores.** Bar heights represent the number of pathways found to be significant after Bonferroni correction. Within a given trait group, results are aggregated for all tested GWAS studies. For each GWAS, the raw number of significant pathways was divided by the number of pathways found by the best performing method. Max gene scores using empirical sampling pathway scores (emp) and chi-squared pathway scores (chi2) are compared to max gene scores combined with empirical sampling, where outlier gene scores ( $p\text{-value} < 10^{-12}$ ) are set to  $10^{-12}$  (empCapped). We chose the capping value such that the maximum  $-\log_{10}$  p-value was roughly in the middle between genome wide significance threshold (8) and the maximum value that can be calculated for the sum statistic (15).

(PDF)

**S10 Fig. Power of *Pascal* pathway scoring methods compared to aggregated hypergeometric scores (MOCS).** The same data as in Fig 5 is plotted here. However, instead of comparing *Pascal* pathway scoring methods with results for all hypergeometric threshold separately, we

defined a new aggregated pathway score that picks the optimal threshold for each pathway over a range of hypergeometric threshold and correcting for the multiple number of tests by Bonferroni correction. Results for different sets of thresholds are displayed. Set1 refers to the complete set of thresholds (i.e.: 25%, 15%, 10%, 5%, 2%, 1%, 0.25%, 0.1%). Set2 refers to a set with thresholds more 'spread out' (i.e.: 25%, 5%, 1%, 0.25). We see that *Pascal* has better performance, except when combining the 'empirical sampling' pathway scoring method with max gene scores.

(PDF)

**S11 Fig. Power of *Pascal* pathway scoring methods compared to aggregated hypergeometric scores (SOCS).** The same data as in Fig 5 is plotted here. However, instead of comparing *Pascal* pathway scoring methods with results for all hypergeometric threshold separately, we defined a new aggregated pathway score that picks the optimal threshold for each pathway over a range of hypergeometric threshold and correcting for the multiple number of tests by Bonferroni correction. Results for different sets of thresholds are displayed. Set1 refers to the complete set of thresholds (i.e.: 25%, 15%, 10%, 5%, 2%, 1%, 0.25%, 0.1%). Set2 refers to a set with thresholds more 'spread out' (i.e.: 25%, 5%, 1%, 0.25). We see that *Pascal* has better performance.

(PDF)

**S12 Fig. Power of gamma distribution for pathway analysis (MOCS).** Bar heights represent the number of pathways found to be significant after Bonferroni correction. Different bars signify results for a different gamma shape parameter value. For each GWAS, the raw number of significant pathways was divided by the number of pathways found by the best performing method. Upper left panel 'All' refers to all traits stacked. We present here MOCS gene score based results. 52 GWAS showed at least one significant pathway in one of the evaluated scenarios.

(PDF)

**S13 Fig. Power of gamma distribution for pathway analysis (SOCS).** Bar heights represent the number of pathways found to be significant after Bonferroni correction. Different bars signify results for a different gamma shape parameter value. For each GWAS, the raw number of significant pathways was divided by the number of pathways found by the best performing method. Upper left panel 'All' refers to all traits stacked. We present here MOCS gene score based results. 60 GWAS showed at least one significant pathway in one of the evaluated scenarios.

(PDF)

**S14 Fig. Distribution of pathway scores for simulated phenotypes influenced by causal SNPs in coding regions.** We first sampled 50 random SNPs assayed in *CoLaus* in or close to coding regions. Using the genotypes of the *CoLaus* study we then simulated phenotypes by adding up the sampled 50 SNPs with a normally distributed effect size with a variance of 0.04 plus Gaussian noise (with a variance of 1). We then ran GWAS for the simulated phenotype to obtain association summary statistics. The experiment was repeated 50 times. On average, this resulted in 18 independent, genome-wide significant gene score hits for each simulated GWAS (for the MOCS statistic). We applied *Pascal* to compute pathway scores for each of the 50 simulated GWAS. We found that the resulting pathway scores are well calibrated, i.e., they do not show inflation or deflation regardless of the setting used (max or sum gene score, chi2 or empirical enrichment test). The QQ-plots show the median value for each quantile across the 50 simulated GWAS. The shaded areas correspond to 95% confidence intervals for the median (estimated from 2000 bootstrap samples of size 50, with replacements). Similar results were

obtained when varying the type and number of simulated causal SNPs and their effect size. (PDF)

**S15 Fig. Distribution of pathway scores for simulated phenotype influenced by causal SNPs in coding and non-coding regions.** These QQ-plots correspond to an analysis equivalent to that of [S14 Fig](#) but with 50 SNPs chosen uniformly from all SNPs assayed in *CoLaus*, rather than from genic regions only. On average, this resulted in 12 independent, genome-wide significant gene score hits for each simulated GWAS (using the MOCS statistic). Note that this does not completely exclude the possibility of less well-calibrated scores in other settings. Deviations from perfectly calibrated scores may occur in the cases where true SNP associations are present, because the gene wise test statistic may have varying power for different genes depending on the genetic architecture of the associated phenotype and on certain gene properties (such as gene length, LD structure, SNP coverage, or SNP allele frequency). If a set of pathways contains many pathways enriched (or depleted) for genes with such confounding factors, inflation or deflation is possible.

(PDF)

**S1 Table. Details of GWAS used.** Details of the 118 GWAS that we used for comparing *Pascal* with other methods.

(TXT)

**S2 Table. Pathways found by *Pascal*.** Tables of pathways discovered by *Pascal* for the 118 GWAS.

(TXT)

## Author Contributions

Conceived and designed the experiments: DL DM ZK SB. Performed the experiments: DL DM. Analyzed the data: DL DM. Wrote the paper: DL DM RR ZK SB.

## References

1. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet.* 2012; 90: 7–24. doi: [10.1016/j.ajhg.2011.11.029](https://doi.org/10.1016/j.ajhg.2011.11.029) PMID: [22243964](https://pubmed.ncbi.nlm.nih.gov/22243964/)
2. Hou L, Zhao H. A review of post-GWAS prioritization approaches. *Front Genet.* 2013; 4: 280. doi: [10.3389/fgene.2013.00280](https://doi.org/10.3389/fgene.2013.00280) PMID: [24367376](https://pubmed.ncbi.nlm.nih.gov/24367376/)
3. Segrè A V, Groop L, Mootha VK, Daly MJ, Altshuler D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet.* 2010; 6: e1001058. doi: [10.1371/journal.pgen.1001058](https://doi.org/10.1371/journal.pgen.1001058) PMID: [20714348](https://pubmed.ncbi.nlm.nih.gov/20714348/)
4. Pers TH, Karjalainen JM, Chan Y, Westra H-J, Wood AR, Yang J, et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun.* Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2015; 6: 5890. Available: doi: <http://dx.doi.org/10.1038/ncomms6890> PMID: [25597830](https://pubmed.ncbi.nlm.nih.gov/25597830/)
5. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet.* 2014; 46: 1173–86. doi: [10.1038/ng.3097](https://doi.org/10.1038/ng.3097) PMID: [25282103](https://pubmed.ncbi.nlm.nih.gov/25282103/)
6. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000; 25: 25–29. doi: [10.1038/75556](https://doi.org/10.1038/75556) PMID: [10802651](https://pubmed.ncbi.nlm.nih.gov/10802651/)
7. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Res.* 2014; 42: D199–D205. doi: [10.1093/nar/gkt1076](https://doi.org/10.1093/nar/gkt1076) PMID: [24214961](https://pubmed.ncbi.nlm.nih.gov/24214961/)
8. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res.* 2003; 13: 2129–2141. doi: [10.1101/gr.772403](https://doi.org/10.1101/gr.772403) PMID: [12952881](https://pubmed.ncbi.nlm.nih.gov/12952881/)



9. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: A database of reactions, pathways and biological processes. *Nucleic Acids Res.* 2011; 39: D691–D697. doi: [10.1093/nar/gkq1018](https://doi.org/10.1093/nar/gkq1018) PMID: [21067998](https://pubmed.ncbi.nlm.nih.gov/21067998/)
10. Nishimura D. *BioCarta. Biotech Softw Internet Rep.* 2001; 2: 117–120. doi: [10.1089/152791601750294344](https://doi.org/10.1089/152791601750294344)
11. Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, et al. A versatile gene-based test for genome-wide association studies. *Am J Hum Genet. The American Society of Human Genetics;* 2010; 87: 139–145. doi: [10.1016/j.ajhg.2010.06.009](https://doi.org/10.1016/j.ajhg.2010.06.009) PMID: [20598278](https://pubmed.ncbi.nlm.nih.gov/20598278/)
12. Li MX, Gui HS, Kwan JSH, Sham PC. GATES: A rapid and powerful gene-based association test using extended Simes procedure. *Am J Hum Genet.* 2011; 88: 283–293. doi: [10.1016/j.ajhg.2011.01.019](https://doi.org/10.1016/j.ajhg.2011.01.019) PMID: [21397060](https://pubmed.ncbi.nlm.nih.gov/21397060/)
13. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* 2011; 21: 1109–21. doi: [10.1101/gr.118992.110](https://doi.org/10.1101/gr.118992.110) PMID: [21536720](https://pubmed.ncbi.nlm.nih.gov/21536720/)
14. Wang L, Jia P, Wolfinger RD, Chen X, Grayson BL, Aune TM, et al. An efficient hierarchical generalized linear mixed model for pathway analysis of genome-wide association studies. *Bioinformatics.* 2011; 27: 686–692. doi: [10.1093/bioinformatics/btq728](https://doi.org/10.1093/bioinformatics/btq728) PMID: [21266443](https://pubmed.ncbi.nlm.nih.gov/21266443/)
15. Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet.* 2007; 81: 1278–1283. doi: [10.1086/522374](https://doi.org/10.1086/522374) PMID: [17966091](https://pubmed.ncbi.nlm.nih.gov/17966091/)
16. Ehret GB, Lamparter D, Hoggart CJ, Whittaker JC, Beckmann JS, Kutalik Z. A multi-SNP locus-association method reveals a substantial fraction of the missing heritability. *Am J Hum Genet.* 2012; 91: 863–871. doi: [10.1016/j.ajhg.2012.09.013](https://doi.org/10.1016/j.ajhg.2012.09.013) PMID: [23122585](https://pubmed.ncbi.nlm.nih.gov/23122585/)
17. Yang J, Ferreira T, Morris AP, Medland SE, Madden PAF, Heath AC, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics.* 2012. pp. 369–375. doi: [10.1038/ng.2213](https://doi.org/10.1038/ng.2213) PMID: [22426310](https://pubmed.ncbi.nlm.nih.gov/22426310/)
18. Holmans P, Green EK, Pahwa JS, Ferreira M a R, Purcell SM, Sklar P, et al. Gene Ontology Analysis of GWA Study Data Sets Provides Insights into the Biology of Bipolar Disorder. *Am J Hum Genet.* 2009; 85: 13–24. doi: [10.1016/j.ajhg.2009.05.011](https://doi.org/10.1016/j.ajhg.2009.05.011) PMID: [19539887](https://pubmed.ncbi.nlm.nih.gov/19539887/)
19. Evangelou M, Smyth DJ, Fortune MD, Burren OS, Walker NM, Guo H, et al. A Method for Gene-Based Pathway Analysis Using Genomewide Association Study Summary Statistics Reveals Nine New Type 1 Diabetes Associations Genetic Epidemiology. *Genet Epidemiol.* 2014; 38: 661–670. doi: [10.1002/gepi.21853](https://doi.org/10.1002/gepi.21853) PMID: [25371288](https://pubmed.ncbi.nlm.nih.gov/25371288/)
20. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet.* 2011; 89: 82–93. doi: [10.1016/j.ajhg.2011.05.029](https://doi.org/10.1016/j.ajhg.2011.05.029) PMID: [21737059](https://pubmed.ncbi.nlm.nih.gov/21737059/)
21. Conneely KN, Boehnke M. So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. *Am J Hum Genet.* 2007; 81: 1158–1168. doi: [10.1086/522036](https://doi.org/10.1086/522036) PMID: [17966093](https://pubmed.ncbi.nlm.nih.gov/17966093/)
22. Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature.* 2012; 491: 56–65. doi: [10.1038/nature11632](https://doi.org/10.1038/nature11632) PMID: [23128226](https://pubmed.ncbi.nlm.nih.gov/23128226/)
23. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. *Nat Genet.* 2013; 45: 1274–83. doi: [10.1038/ng.2797](https://doi.org/10.1038/ng.2797) PMID: [24097068](https://pubmed.ncbi.nlm.nih.gov/24097068/)
24. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature.* 2010; 467: 52–8. doi: [10.1038/nature09298](https://doi.org/10.1038/nature09298) PMID: [20811451](https://pubmed.ncbi.nlm.nih.gov/20811451/)
25. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature.* 2014; 506: 376–81. doi: [10.1038/nature12873](https://doi.org/10.1038/nature12873) PMID: [24390342](https://pubmed.ncbi.nlm.nih.gov/24390342/)
26. Mishra A, Macgregor S. VEGAS2: Software for More Flexible Gene-Based Testing. *Twin Res Hum Genet.* 2015; 18: 86–91. doi: [10.1017/thg.2014.79](https://doi.org/10.1017/thg.2014.79) PMID: [25518859](https://pubmed.ncbi.nlm.nih.gov/25518859/)
27. Firmann M, Mayor V, Vidal P, Bochud M, Pécoud A, Hayoz D, et al. The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. *BMC Cardiovascular Disorders.* 2008. p. 6. doi: [10.1186/1471-2261-8-6](https://doi.org/10.1186/1471-2261-8-6) PMID: [18366642](https://pubmed.ncbi.nlm.nih.gov/18366642/)
28. Heinig M, Petretto E, Wallace C, Bottolo L, Rotival M, Lu H, et al. A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. *Nature.* 2010; 467: 460–464. doi: [10.1038/nature09386](https://doi.org/10.1038/nature09386) PMID: [20827270](https://pubmed.ncbi.nlm.nih.gov/20827270/)

29. Burren OS, Guo H, Wallace C. VSEAMS : A pipeline for variant set enrichment analysis using summary GWAS data identifies IKZF3, BATF and ESRRA as key transcription factors in type 1 diabetes. *2014*;30: 0–26. doi: [10.1093/bioinformatics/btu571](https://doi.org/10.1093/bioinformatics/btu571)
30. Davis J, Goadrich M. The Relationship Between Precision-Recall and ROC Curves. *Proc 23rd Int Conf Mach Learn—ICML'06*. 2006; 233–240. doi: [10.1145/1143844.1143874](https://doi.org/10.1145/1143844.1143874)
31. Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet*. Nature Publishing Group; 2010; 42: 1118–25. doi: [10.1038/ng.717](https://doi.org/10.1038/ng.717) PMID: [21102463](https://pubmed.ncbi.nlm.nih.gov/21102463/)
32. Imielinski M, Baldassano RN, Griffiths A, Russell RK, Annese V, Dubinsky M, et al. Common variants at five new loci associated with early-onset inflammatory bowel disease. *Nat Genet*. 2009; 41: 1335–1340. doi: [10.1038/ng.489](https://doi.org/10.1038/ng.489) PMID: [19915574](https://pubmed.ncbi.nlm.nih.gov/19915574/)
33. Wellcome T, Case T, Consortium C. Genome-wide association study of 14, 000 cases of seven common diseases and. *Nature*. 2007; 447: 661–78. doi: [10.1038/nature05911](https://doi.org/10.1038/nature05911) PMID: [17554300](https://pubmed.ncbi.nlm.nih.gov/17554300/)
34. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet*. 2010; 42: 105–116. doi: [10.1038/ng.520](https://doi.org/10.1038/ng.520) PMID: [20081858](https://pubmed.ncbi.nlm.nih.gov/20081858/)
35. Estrada K, Styrkarsdottir U, Evangelou E, Hsu YH, Duncan EL, Ntzani EE, et al. Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *2012*;44: 491–501. doi: [10.1038/ng.2249](https://doi.org/10.1038/ng.2249)
36. Day TF, Yang Y. Wnt and hedgehog signaling pathways in bone development. *J Bone Joint Surg Am*. 2008; 90 Suppl 1: 19–24. doi: [10.2106/JBJS.G.01174](https://doi.org/10.2106/JBJS.G.01174) PMID: [18292352](https://pubmed.ncbi.nlm.nih.gov/18292352/)
37. Tobacco T, Consortium G. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet*. 2010; 42: 441–7. doi: [10.1038/ng.571](https://doi.org/10.1038/ng.571) PMID: [20418890](https://pubmed.ncbi.nlm.nih.gov/20418890/)
38. Bradley DT, Zipfel PF, Hughes AE. Complement in age-related macular degeneration: a focus on function. *Eye (Lond)*. 2011; 25: 683–693. doi: [10.1038/eye.2011.37](https://doi.org/10.1038/eye.2011.37)
39. Ebrahimi KB, Handa JT. Lipids, lipoproteins, and age-related macular degeneration. *J Lipids*. 2011; 2011: 802059. doi: [10.1155/2011/802059](https://doi.org/10.1155/2011/802059) PMID: [21822496](https://pubmed.ncbi.nlm.nih.gov/21822496/)
40. Lee D, Williamson VS, Bigdeli TB, Riley BP, Fanous a. H, Vladimirov VI, et al. JEPeG: a summary statistics based tool for gene-level joint testing of functional variants. *Bioinformatics*. 2014; 31: 1176–1182. doi: [10.1093/bioinformatics/btu816](https://doi.org/10.1093/bioinformatics/btu816) PMID: [25505091](https://pubmed.ncbi.nlm.nih.gov/25505091/)
41. Genz A. Numerical Computation of Multivariate Normal Probabilities. *J Comput Graph Stat*. 1992; 1: 141–149. doi: [10.1080/10618600.1992.10477010](https://doi.org/10.1080/10618600.1992.10477010)
42. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*. 1999. pp. 29–34. doi: [10.1093/nar/27.1.29](https://doi.org/10.1093/nar/27.1.29) PMID: [9847135](https://pubmed.ncbi.nlm.nih.gov/9847135/)
43. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette M a, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005; 102: 15545–50. doi: [10.1073/pnas.0506580102](https://doi.org/10.1073/pnas.0506580102) PMID: [16199517](https://pubmed.ncbi.nlm.nih.gov/16199517/)
44. Xu Z, Duan Q, Yan S, Chen W, Li M, Lange E, et al. DISSCO: direct imputation of summary statistics allowing covariates. *Bioinformatics*. 2015; 31: 2434–2442. doi: [10.1093/bioinformatics/btv168](https://doi.org/10.1093/bioinformatics/btv168) PMID: [25810429](https://pubmed.ncbi.nlm.nih.gov/25810429/)
45. Ehret GB, Lamparter D, Hoggart CJ, Whittaker JC, Beckmann JS, Kutalik Z. A multi-SNP locus-association method reveals a substantial fraction of the missing heritability. *Am J Hum Genet*. 2012; 91: 863–871. doi: [10.1016/j.ajhg.2012.09.013](https://doi.org/10.1016/j.ajhg.2012.09.013) PMID: [23122585](https://pubmed.ncbi.nlm.nih.gov/23122585/)
46. Gao X, Starmer J, Martin ER. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet Epidemiol*. 2008; 32: 361–369. doi: [10.1002/gepi.20310](https://doi.org/10.1002/gepi.20310) PMID: [18271029](https://pubmed.ncbi.nlm.nih.gov/18271029/)
47. B DR. The Distribution of a Linear Combination of x2 Random Variables. *J R Stat Soc Ser C*. 1980; 29: 323–333.
48. Farebrother R. Algorithm AS 204: the distribution of a positive linear combination of chi2 random variables. *J R Stat Soc Ser C*. 1984; 33: 332–339. doi: [10.2307/2347721](https://doi.org/10.2307/2347721)
49. Duchesne P, Lafaye De Micheaux P. Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods. *Comput Stat Data Anal*. 2010; 54: 858–862. doi: [10.1016/j.csda.2009.11.025](https://doi.org/10.1016/j.csda.2009.11.025)

### 3. Identifying chromatin accessibility regulators

In prokaryotes, binding of transcription factors to some DNA region is driven mainly by two factors: The affinity of the DNA to the transcription factor and the abundance of the transcription factor. In eukaryotes additional layers of regulation occur that complicate this picture. Most of the DNA is compacted and therefore not accessible to TF binding. Consequently, just using the presence of binding motif instance in a genomic region is typically poor predictor of transcription factor binding even if the transcription factor is known to be expressed. As mentioned in the introduction, genome-wide DNase1 assays, allow to identify regions of open (not compacted) chromatin. These regions are typically accessible to TF binding. Together with transcription factor motif information, open chromatin information allows to build predictors of TF binding that are much better than using motif instances alone. However, this does not tell us how the chromatin was opened in the first place and what leads it to be close up again, i.e. it does not tell us whether transcription factor binding is the cause or the consequence of open chromatin. One popular model is that members of a certain class of TFs called pioneer factors are indeed capable of binding motif instances in closed chromatin and are instrumental in driving this transition. Consequently these factors are thought to play a major role in cell type transitioning. From the above, we see that in prokaryotes, TF binding and TF expression is tightly linked, whereas in eukaryotes this relationship is further regulated through chromatin compaction and potentially further regulatory mechanism. However, one can surmise that for a special class of TFs the relationship still holds and these factors are instrumental in transitioning between open and closed chromatin and also in turn driving cells into different states. In this chapter, we apply this rationale to datasets on open chromatin and gene expression provided by the ENCODE project to determine, which transcription factors drive transitions between open and closed states. The signature of such factor is the correlation between expression values and average open chromatin state at its motif across the same set of cell lines. Our method assesses this correlation while accounting for the fact that some tested cell lines are more related than others.

We find many transcription factors showing evidence of driving transitions and high proportion of these transcription factors are known pioneer factors, i.e. play a role in opening up closed chromatin.

RESEARCH ARTICLE

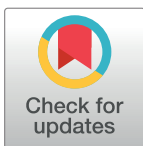
# Genome-Wide Association between Transcription Factor Expression and Chromatin Accessibility Reveals Regulators of Chromatin Accessibility

David Lamparter<sup>1,2</sup>, Daniel Marbach<sup>1,2</sup>, Rico Rueedi<sup>1,2</sup>, Sven Bergmann<sup>1,2,3</sup>\*, Zoltán Kutalik<sup>2,4</sup>\*

**1** Department of Computational Biology, University of Lausanne, Lausanne, Switzerland, **2** Swiss institute of Bioinformatics, Lausanne, Switzerland, **3** Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town, South Africa, **4** Institute of Social and Preventive Medicine (IUMSP), Lausanne University Hospital, Lausanne, Switzerland

\* These authors contributed equally to this work.

\* [zoltan.kutalik@unil.ch](mailto:zoltan.kutalik@unil.ch) (ZK); [sven.bergmann@unil.ch](mailto:sven.bergmann@unil.ch) (SB)



 OPEN ACCESS

**Citation:** Lamparter D, Marbach D, Rueedi R, Bergmann S, Kutalik Z (2017) Genome-Wide Association between Transcription Factor Expression and Chromatin Accessibility Reveals Regulators of Chromatin Accessibility. *PLoS Comput Biol* 13(1): e1005311. doi:10.1371/journal.pcbi.1005311

**Editor:** Ting Wang, Washington University in Saint Louis, UNITED STATES

**Received:** May 16, 2016

**Accepted:** December 15, 2016

**Published:** January 24, 2017

**Copyright:** © 2017 Lamparter et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** The code for reproduction (including scripts for data download) is available at: <https://github.com/dlampart/csrproject>

**Funding:** This work was supported by the Leenaards Foundation (<http://www.leenaards.ch>) [to ZK] the Swiss Institute of Bioinformatics (<http://www.isb-sib.ch/>) [to SB, to ZK], the Swiss National Science Foundation [31003A-143914 to ZK, FN 310030\_152724 / 1 to SB] and SystemsX.ch

## Abstract

To better understand genome regulation, it is important to uncover the role of transcription factors in the process of chromatin structure establishment and maintenance. Here we present a data-driven approach to systematically characterise transcription factors that are relevant for this process. Our method uses a linear mixed modelling approach to combine datasets of transcription factor binding motif enrichments in open chromatin and gene expression across the same set of cell lines. Applying this approach to the ENCODE dataset, we confirm already known and imply numerous novel transcription factors that play a role in the establishment or maintenance of open chromatin. In particular, our approach rediscovers many factors that have been annotated as pioneer factors.

## Author Summary

Transcription factor binding occurs mainly in regions of open chromatin. For many transcription factors, it is unclear whether binding is the cause or the consequence of open chromatin. Here, we used datasets on open chromatin and gene expression provided by the ENCODE project to predict which transcription factors drive transitions between open and closed states. A signature of such a factor is that its expression values are correlated to chromatin accessibility at its motif across the same set of cell lines. Our method assesses this correlation while accounting for the fact that some tested cell lines are more related than others. We find many transcription factors showing evidence of driving transitions and a high proportion of these transcription factors are known pioneer factors, i.e., they play a role in opening up closed chromatin.

(<http://www.systemsx.ch/>) [51RTP0\_151019 to ZK, via SysGenetiX to SB]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

## Introduction

In higher eukaryotes, certain sequence-specific transcription factors (TFs), which we will call *chromatin accessibility regulators* (CARs), are responsible for establishing and maintaining open chromatin configurations [1,2]. CARs therefore play a fundamental role in transcriptional regulation, because open chromatin configurations are necessary for additional TFs to bind and transcriptionally activate target genes.

CARs that can bind closed chromatin and open up chromatin are called pioneer TFs [3]. The comprehensive identification of pioneer TFs with high confidence still needs further research. While some pioneer TFs are well studied, others have only preliminary evidence, or are only computationally predicted. Some well studied examples include *FOXA1*, whose winged helix domains disrupt DNA–histone contacts, and *POU5F1*, *SOX2* and *KLF4*, which are used in production of induced pluripotent stem cells (iPSC) [4,5]. Further pioneer TFs such as *ASCL1*, *SPI1* and the *GATA* factors are used in transdifferentiation, and *PAX7* plays a role in pituitary melanotrope development [5–7]. However, not all pioneer TFs are involved in development and cell type conversions: the *CLOCK-BMAL1* heterodimer is part of the circadian clock and the tumour suppressor *TP53* is involved in the cell cycle, while its close homolog *TP63* is involved in skin development [8–10].

Recent studies suggest that maintaining open chromatin is a dynamic process with pioneer and other TFs binding and unbinding rapidly and continually recruiting additional chromatin remodelling factors that are not sequence specific [2,11,12]. TFs vary in their ability to recruit particular remodelling factors, for example the TFs *STAT5A/B* and *MYOG* motifs enrich in binding sites of the *SWI/SNF* remodelling complex but not in *ISWI* remodelling complex binding sites, whereas *YY1* motifs were found exclusively in *ISWI* complex binding sites [2]. A natural question then is which TFs are relevant to maintain open chromatin and can therefore be called CARs.

One approach to test whether a given TF is a CAR is to perform a knock-down of this TF followed by an open chromatin assay to see whether chromatin regions containing the respective motif preferentially change from open to closed [13]. However, this approach is very time consuming because it requires a separate knock-down experiment for each TF. To define pioneer TFs specifically, one can check if the TF has the ability to bind nucleosomal DNA *in vitro* and validate the results *in vivo* [14]. Recently, a computational method called *Protein Interaction Quantification (PIQ)* has been published that aims to recover pioneer TFs by estimating both TF binding and ensuing chromatin changes from the same DnaseI hypersensitivity (DHS) experiments [15]. However, *PIQ* did not predict some well known pioneer TFs such as *FOXA1*, *SOX2* and *POU5F1* showing that further improvements are possible [3].

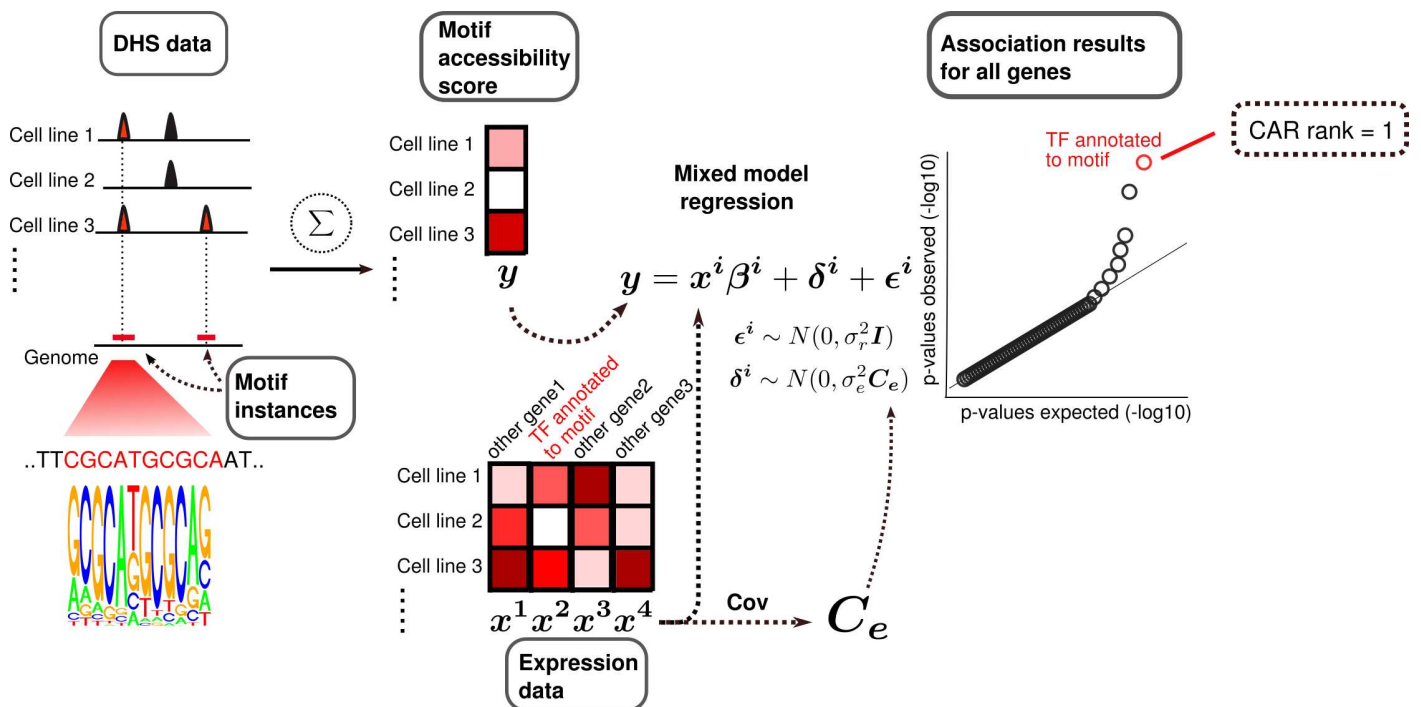
Here we introduce a data driven approach to predict CARs. Our approach relies on the joint analysis of a large collection of DHS and coordinated gene expression data to estimate TF activity independently of DHS data. We first define the *motif accessibility score* for a given TF for each cell line based on the enrichment of its binding motif in regions with open chromatin. We then associate these scores with gene expression values across all available cell lines. This should allow us to predict which factors have a role either in establishment or maintenance of open chromatin, although it will not reveal which mode predominates (to determine this, further experiments will be necessary).

We used our approach on data generated as part of the ENCODE project [16,17]. This uncovered numerous TFs whose motif accessibility is robustly associated with mRNA expression across 109 cell lines suggesting either a role in the establishment or maintenance of open chromatin. Also, we see that our uncovered TFs are strongly enriched for known pioneer TFs. This suggests that the TFs we identified are good candidates for CARs.

## Results

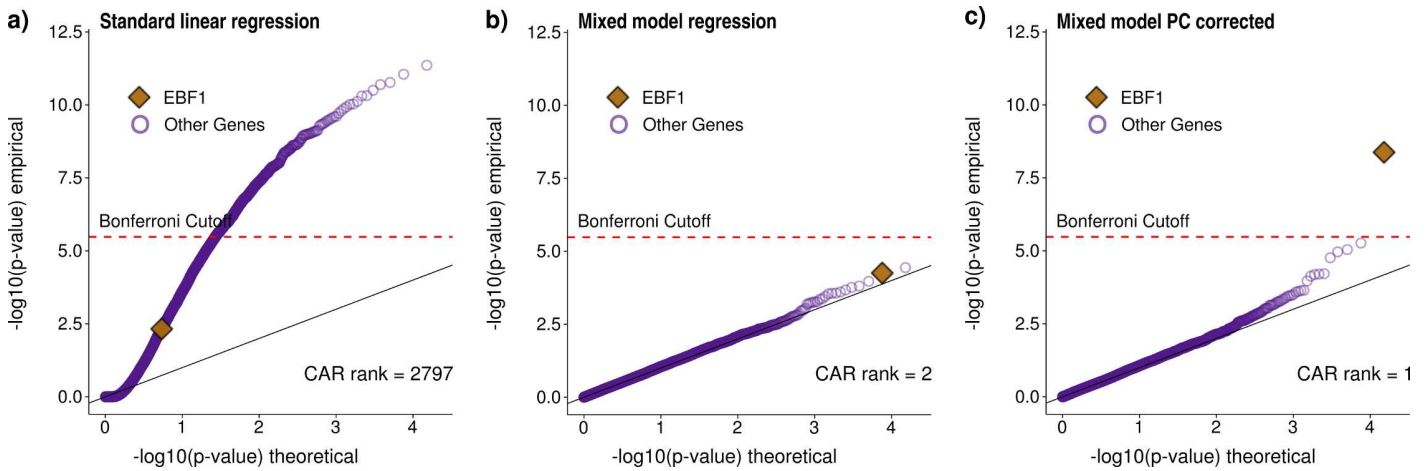
### A linear mixed model approach to predict chromatin accessibility regulators

Our approach rests on the assumption that the activity of a CAR is correlated with the amount of open chromatin in the vicinity of its potential binding sites. Both quantities can be estimated from genomic data: For the CAR activity we use its gene expression level as a proxy for the active protein concentration. The effect of this activity is approximated by the open chromatin fraction of the genome around its binding motif instances (Fig 1). Specifically, we count the number of instances of the binding motif of a given TF in the open chromatin fraction of the genome to define a motif accessibility score. A naive approach would be to use standard linear regression between the motif accessibility score and the expression level of a given TF to identify CAR candidates. Yet, this method has an elevated type I error rate, as it does not account for confounding due to cell line relatedness or batch effects. To overcome this limitation, we use here a linear mixed model (LMM) framework, where a random effect accounts for such confounding factors (which has been shown to work well in genetic association studies [18–20]). For a given motif, we use the linear mixed model framework to find the association p-value between its accessibility score and the measured expression of the TF gene. We then compare this p-value to the p-values calculated using the measured expression of each of the other genes as regressors. If confounding is controlled for, most association p-values should follow a uniform [0,1] distribution. Furthermore, if the TF is a CAR, its p-value should be low



**Fig 1. Mixed model approach for identification of chromatin accessibility regulators.** For a TF binding motif, we search for all its instances in the genome. For each cell line, we calculate the accessibility score by counting how many motif instances are found in the open chromatin fraction of the genome. After further normalization, these accessibility scores are compared to gene expression values for all genes via regression (Methods). To account for confounding, we use mixed model regression, where an additional random component is used with the same covariance structure as the gene expression matrix. To be considered a CAR candidate, motif accessibility of a TF must show strong association (low p-value) with the expression of the corresponding TF gene compared to other genes. The gene-level *CAR rank* of a TF is defined as the rank of its association p-value among the p-values for all genes.

doi:10.1371/journal.pcbi.1005311.g001



**Fig 2. Association between motif accessibility and mRNA expression for the putative chromatin accessibility regulator *EBF1*.** Three different regression models (a-c) were used to compute association p-values between the accessibility of a given TF motif (here *EBF1*) and mRNA expression for each of the assayed 15K protein-coding genes. Results are visualized as qq-plots showing the  $-\log_{10}$  transformed p-values. (a) Association p-values obtained using standard linear regression. Due to confounding, p-values are strongly inflated and *EBF1* motif accessibility does not show strong association with *EBF1* expression compared to other genes. (b) The linear mixed model (LMM) successfully corrects for confounding, with most p-values following the null distribution as expected. The association between *EBF1* motif accessibility and *EBF1* expression now ranks second among all genes and first among all TFs, although it does not pass the Bonferroni significance threshold. (c) Additionally controlling for the first principal component of the motif accessibility matrix corrects for a strong batch effect (Methods), which further improves the signal. Using this approach, *EBF1* motif accessibility showed the strongest association precisely with *EBF1* expression (i.e., the gene-level CAR rank equals one), suggesting that *EBF1* may be a CAR, in agreement with the literature [22]. As a further illustration for the improvements achieved using the mixed model approach S1 Fig shows the analogous plot for FOXA1, the first discovered pioneer factor [4,5].

doi:10.1371/journal.pcbi.1005311.g002

compared to other genes. We thus define the *CAR rank* of a TF as the rank of its association p-value among all genes (see example in Fig 1). Low CAR ranks indicate strong association between motif accessibility and TF expression, suggesting that the TF is a CAR.

Specifically, we used DHS data as well as mRNA expression data across 109 cell lines. To calculate motif accessibility scores we used 325 TF binding motifs from the HOCOMOCO database [21]. As expected, we observed severe confounding when using standard linear regression, which was controlled using linear mixed effect model regression (Fig 2).

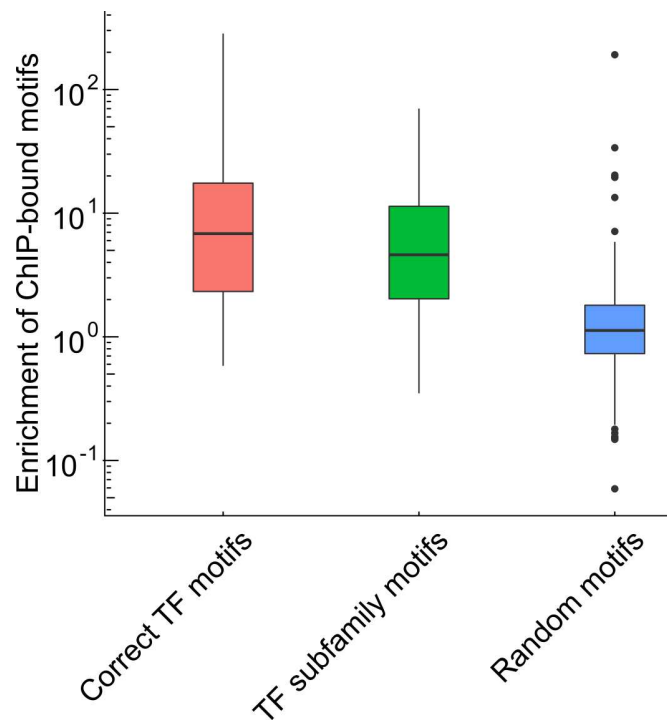
### ChIP-seq shows widespread binding of homologous TFs to each other's motifs

Our method relies on TF motif accessibility and expression data to predict CARs. However, evolutionarily related TFs have similar binding motifs [23]. Motif accessibility may therefore associate not only with the expression of the annotated TF, but also with the expression of a homologous TF with a similar motif. Therefore, we mapped TFs into subfamilies using the homology-based clustering TFClass [24]. The 1,557 TFs were grouped into 397 subfamilies. Using a collection of 329 ChIP-seq profiles from ENCODE, we saw strong enrichment of TF motifs in ChIP-seq peaks of the TF as well as its subfamily members (Fig 3). We therefore consider any strong association between a motif and a member of the subfamily of its TF as a signal for a CAR.

### Comprehensive prediction of chromatin accessibility regulators

Next, we used the linear mixed model strategy to predict CARs among TFs. We used 325 motifs from HOCOMOCO (after filtering motifs showing low overlap with DHS signal, see Methods). For each motif, we used a linear mixed effect model to compute its association with





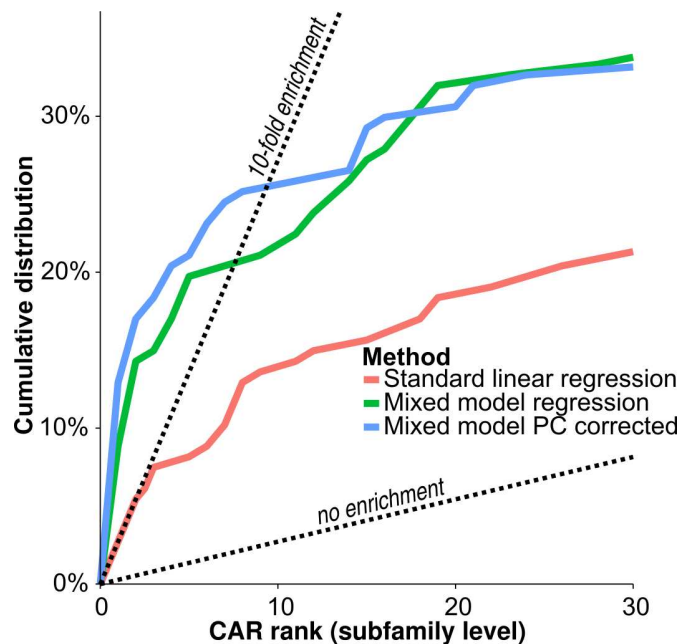
**Fig 3. Enrichment of bound motifs for a given TF and its subfamily members.** All TF ChIP-seq experiments from the Myers-lab released as part of the ENCODE project were downloaded. For each TF ChIP-seq experiment we also obtained the corresponding TF motif from the HOCOMOCO database [25]. For a given ChIP-seq experiment, we looked at the processed DHS peaks in the same cell line. We partitioned DHS peaks into two groups depending on whether they were bound by the TF (overlap with a ChIP-seq peak) or not. We then calculated both the fraction of bound and unbound DHS peaks containing a given motif. The enrichment of bound motifs was defined as the ratio of these two fractions. Results are shown from left to right for: the motifs of the TFs that were assayed in the corresponding ChIP-seq experiments (Correct TF motifs), motifs of other TFs from the same subfamily (TF subfamily motifs), and randomly sampled motifs (Random motifs). During sampling, each motif was sampled as often as the number of ChIP-seq experiment available for that motif. We see strong enrichment of TF motifs in ChIP-seq peaks of the TF as well as its subfamily members.

doi:10.1371/journal.pcbi.1005311.g003

mRNA expression for 1,188 known TFs. Due to the redundancy of motifs within the same TF subfamily (see preceding section), we also computed CAR ranks at the level of TF subfamilies. To this end, we retained the most significant association p-value within each subfamily corrected for subfamily size (see [Methods](#) and [S2 Fig](#)). Under the null model (when TFs are not CARs), CAR ranks should be uniformly distributed across all subfamilies, so that deviation from uniformity indicates presence of CARs.

We found strong enrichment of low CAR ranks at the subfamily level ([Fig 4, S1 Table](#)). The enrichment was stronger when using mixed modelling instead of standard linear regression, underlining again the importance of proper control for confounding factors. When looking at the threshold that leads to 10-fold enrichment of low CAR ranks compared to uniformity (i.e., 10% false discovery rate), we found that 25% of all subfamilies have a CAR rank that falls below that threshold. These results show that many TFs do have an impact on the open chromatin fraction and can be defined as CARs.

To validate our results based on the ENCODE dataset, we applied our CAR calling strategy to data from another large scale effort, the ROADMAP Epigenomics consortium [26]. Coordinated open chromatin and expression data have been released for 56 samples. For 29 of these samples, open chromatin was assayed directly. For the other samples, open chromatin



**Fig 4. Method comparison across all subfamilies.** Cumulative distribution of *CAR* ranks at the subfamily level for the 147 tested subfamilies using the three different modelling strategies: ‘standard linear regression’, ‘mixed model regression’ and ‘mixed model PC corrected’ (see legend of Fig 2 and Methods). We see strong enrichment of low ranks implying deviation from the null hypothesis. The linear mixed modelling increases enrichment of low *CAR* ranks.

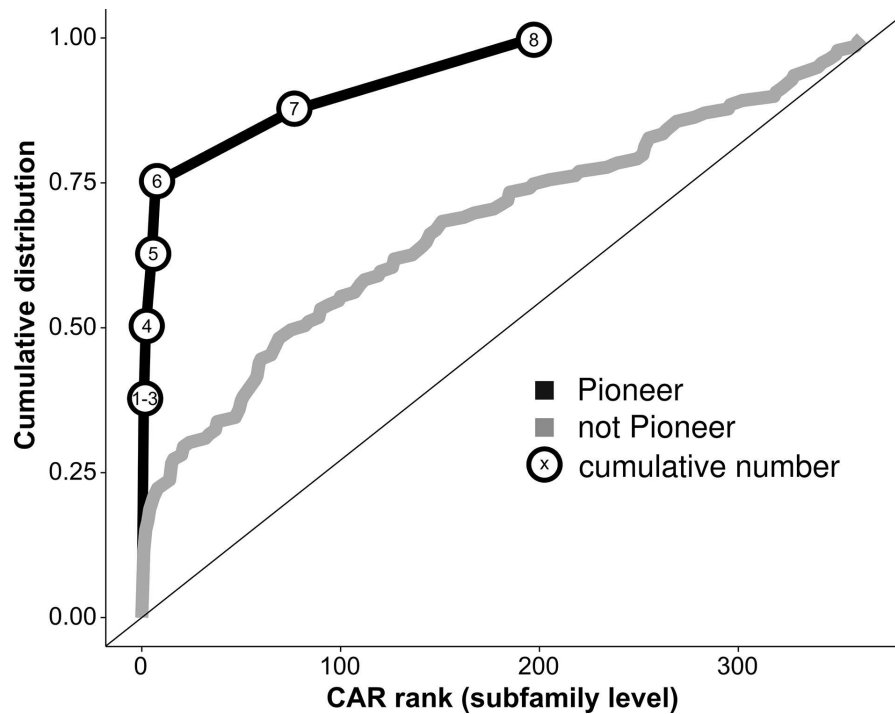
doi:10.1371/journal.pcbi.1005311.g004

information was imputed from other available epigenetic measurements. The ROADMAP collection is derived mainly from human tissue samples and primary cell lines (whereas ENCODE is biased towards immortalized cell lines). Further differences are that expression was measured using RNA sequencing. We applied our method to these datasets and compared results to the results derived in ENCODE. Most subfamilies predicted to be CARs in ROADMAP were recovered in ENCODE (see S3 Fig). Furthermore, while subfamilies predicted to be CARs in ENCODE showed enrichment for low *CAR* ranks in ROADMAP, subfamilies not predicted to be CARs in ENCODE did not show enrichment for low *CAR* ranks in ROADMAP (see S4 Fig). These results are concordant with both datasets, pointing toward the same factors being CARs and the higher power of the ENCODE data to detect CARs, potentially due to higher sample size, reliance on direct measurements of DHS and lower fraction of complex tissue samples.

To evaluate the impact of the motif search strategy, we investigated the robustness of the pipeline with respect to the motif search. Results were stable and power was only affected by varying motif cutoffs (S5 Fig, S6 Fig). Additionally, we investigated whether choosing the cutoff based on ChIP-seq data changed results. For each TF with available ChIP-seq data, we used an individual cutoff such that all called binding sites have fixed true positive rate (using the ChIP-seq data as the ground truth). Again, results were stable no matter how the cutoff was assigned (S7 Fig and S8 Fig).

### Pioneer TF subfamilies enrich in predicted chromatin accessibility regulators

As mentioned above, one well-defined class of CARs are pioneer TFs that can bind and open closed chromatin. Therefore, subfamilies annotated to known pioneer TFs should have low



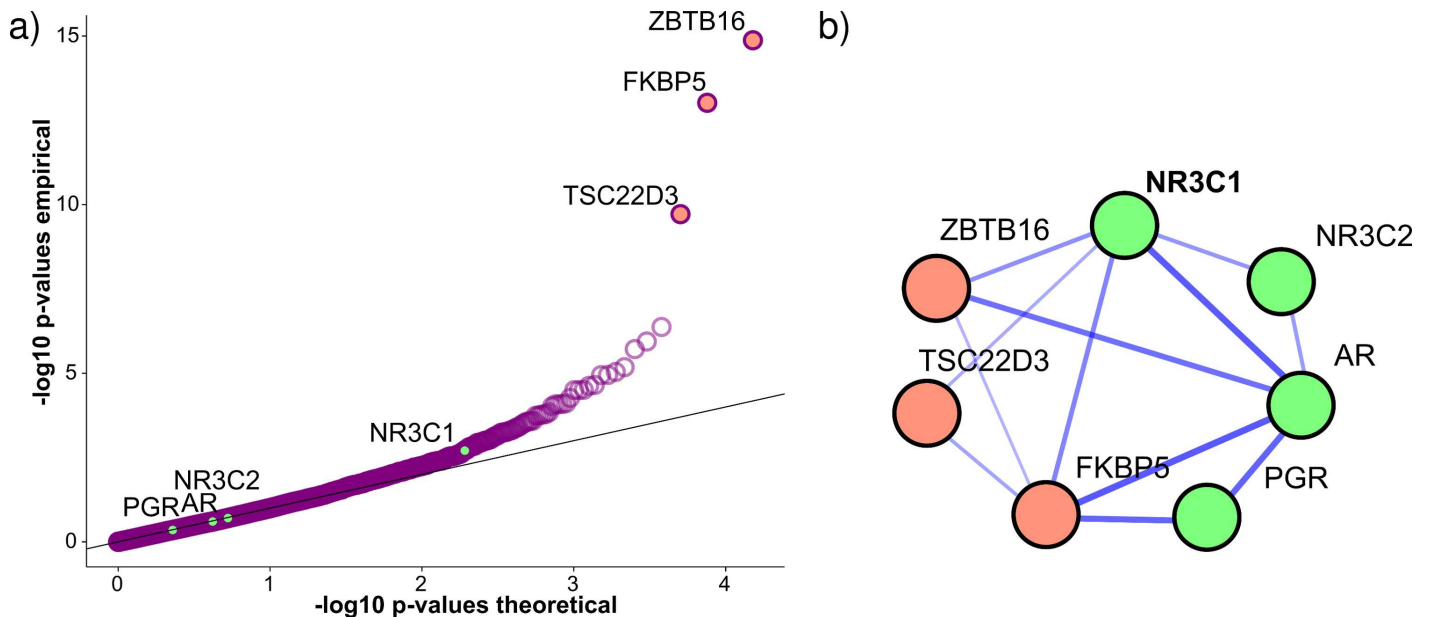
**Fig 5. Known pioneer TF subfamilies strongly enrich in predicted chromatin accessibility regulators.** Shown in grey is a scaled cumulative distribution plot for subfamily level CAR ranks of subfamilies not annotated as pioneers in Iwafuchi-Doi et al. [3]. In black, we see the cumulative number of pioneer subfamilies that reached at least a given CAR rank. Six out of eight subfamilies show a low CAR rank, which is more than three times as many as one would expect on average when sampling from non-pioneer subfamilies.

doi:10.1371/journal.pcbi.1005311.g005

CAR ranks. To test enrichment formally, we used a recently published list of established pioneer TF subfamilies (Methods) [3]. We asked whether these subfamilies were predicted as CARs using our methodology. For eight subfamilies in the list for which we had the motif, six showed at least ten-fold enrichment (i.e. having a CAR rank at the subfamily level below ten) (Fig 5). To assess significance, we used the Wilcoxon ranksum test leading to a p-value of 0.0087. When using the hypergeometric test with 10-fold enrichment cutoff (Fig 4), the p-value was even lower ( $P = 0.0016$ ). Because our approach to uncover CARs is biased towards TFs with large mRNA expression variability (S9 Fig), we sought to control for potential confounding introduced by the fact that the tested pioneer factors might also have large expression variability. Controlling for expression variability only slightly increased the p-values from 0.0087 to 0.024 and from 0.0016 to 0.0027, respectively.

### Downstream genes can show strong associations for activating chromatin accessibility regulators

It is known that the activity of some TFs is mainly regulated by the level of their cofactors rather than their own protein concentration [27]. These TFs are often present in their inactive form in the cell, which can then be quickly activated upon binding of the cofactor. This allows the cell to rapidly respond to environmental cues. An example of this phenomenon are steroid receptor TFs, which initiate transcriptional changes upon steroid hormone binding [28]. In such cases, one would not expect a strong association between the mRNA expression level of a



**Fig 6. Strong associations between GR-like receptor motif and glucocorticoid response genes.** a) Association results for motif accessibility of the TF *NR3C1*, which belongs to the GR-like receptor subfamily, and mRNA expression across all genes.  $-\log_{10}$  transformed p-values are shown in a QQ-plot. *NR3C1* motif accessibility shows strong association with mRNA expression of three glucocorticoid response genes (orange), but only weak association with expression of *NR3C1* and other GR-like receptor TFs (green). In this example, motif accessibility is strongly associated with downstream gene expression, but only weakly with expression of the TF itself. b) The network shows functional relationships among the GR-like receptor TFs (green) and the three most strongly associated genes (orange), which are all glucocorticoid response genes. The strength of links shows confidence in functional relationship given in the *STRING* database. We see numerous links between the downstream glucocorticoid response genes and the GR-like receptor TFs in the *STRING* database, confirming their functional relatedness, where *NR3C1* has the most links to associated genes.

doi:10.1371/journal.pcbi.1005311.g006

receptor TF and its motif accessibility because mRNA expression would rather be correlated to the amounts of inactive TF protein in the cell, while TF activity should depend on the strength of the environmental stimulus. However, if the TF strongly activates mRNA expression of other genes, it might be possible to predict whether the TF is a chromatin accessibility regulator by looking at associations between the motif accessibility of the TF and the expression of its downstream genes.

To explore this strategy, we looked at associations across all genes and motifs that were below the overall Bonferroni threshold ( $9.6 \times 10^{-9}$ ). For five out of 13 such motifs, members of the corresponding subfamily had top scores. In four further cases, a gene from a TF subfamily was ranked close to the top that was highly related (i.e. part of the same family [24]) to the motifs' corresponding subfamily but not identical with it. This suggests that the TF subfamily clustering was too fine-grained in these cases. Surprisingly, for one motif, the significant association had a negative effect size (the negative association was observed between *NUDT11* and the motif for *RARG*), which might reflect an indirect effect. The remaining three motifs were all annotated to the GR-like receptors, which encompass four TFs (*AR*, *NR3C1*, *NR3C2*, *PGR*). The accessibilities of these three motifs all associated strongly with the expression of three genes (*FKBP5*, *ZBTB1*, *TSC22D3*). When using the *STRING* database to check for functional links between these genes, all genes had links to a GR-like receptor (Fig 6) [29]. In fact, all three genes are known to be glucocorticoid response genes. These results suggest that some GR-like receptors might act as a CAR. For strongly activating factors, the power of the analysis can therefore be strengthened by incorporating results from downstream genes.

## Discussion

It is well known that TF binding correlates with open chromatin [17]. However, for many TFs, it is not clear whether their binding is the cause or the consequence of open chromatin. Here, we used datasets provided by ENCODE to predict chromatin accessibility regulator candidates, i.e., TFs that are able to establish or maintain open chromatin configurations. We devised an approach using linear mixed models to deal with the extensive confounding that one encounters in genome-wide data from heterogeneous sources. Our method uncovers a set of TFs whose expression is associated with their motif accessibility, suggesting a role in maintenance of an open chromatin configuration.

Potentially our methodology could be extended to histone modification data instead of DHS data. We applied our method to H3K4me3 data for cell-lines but did not see strong enrichment (S10 Fig).

Because pioneer TFs are by definition CARs, our predictions should be enriched for known pioneer TFs. We tested this formally for a list of pioneer TF subfamilies recently published by Iwafuchi-Doi et al. [3]. Six out of eight pioneer subfamilies were indeed predicted by our method to be CARs: *FOXA1*, *GATA6*, *KLF4*, *SOX2*, *SPI1* and *TP63* were the pioneer TFs driving these signals. The two subfamilies not predicted to be CARs were *POU5* and *CLOCK*. *SOX2* was the gene most strongly associated with *POU5F1* motif accessibility with a low p-value of  $5 \times 10^{-6}$  (S11 Fig). *POU5F1* acts together with *SOX2* to maintain undifferentiated states [30]. The two TFs also physically interact and a recent study proposed a model where *SOX2* guides *POU5F1* to target sites [31]. The *CLOCK* subfamily members have a role in the cell cycle, acting as TFs for the circadian pacemakers [32]. It is possible that average mRNA expression of these TFs in unsynchronized cell lines is not a meaningful measure for their activity. In addition to the eight aforementioned factors we found further factors discussed in the pioneer TF literature such as *TFAP2C*, *EBF1*, *CEBPD/B*, *OTX2*, *NFKB* and *STAT5* (Table 1) [22,33–37]. In addition, when combining our predictions with those from the PIQ method [15], we observed substantial performance improvement compared to either method alone (S12 Fig).

One limitation of our approach is that it cannot discern between open chromatin establishing TFs and open chromatin maintaining TFs. A way to discern the relative roles could be to

**Table 1. Predicted pioneer factors.** Shown are the CAR ranks of factor subfamilies that were discussed in the main text. These included subfamilies labelled pioneers in [3] and consequently used as a member of the true positive set used in Fig 5 (these subfamilies are set in bold face). Additionally, subfamilies are shown that are predicted to be CARs and for which there exist limited literature evidence for pioneer activity. For each subfamily, the top-scoring gene among all genes in the subfamily is mentioned. A complete table for all tested subfamilies is given in S1 Table.

Subfamily name	Top gene in subfamily	CAR rank (subfamily level)	Pioneer evidence
C/EBP	<i>CEBPD</i>	1	[34]
<b>AP-2</b>	<i>TFAP2C</i>	1	[3,33]
<b>Krüppel-like factors</b>	<i>KLF4</i>	1	[3,4]
<b>FOXA</b>	<i>FOXA1</i>	1	[3–5]
<b>Group B</b>	<i>SOX2</i>	1	[3,4]
NF-kappaB p65 subunit-like factors	<i>RELB</i>	1	[38]
Early B-Cell Factor-related factors	<i>EBF1</i>	1	[22]
<b>Two zinc-finger GATA factors</b>	<i>GATA6</i>	2	[3,5]
STAT factors	<i>STAT5B</i>	2	[37]
OTX	<i>OTX2</i>	3	[35]
<b>Spi-like factors</b>	<i>SPI1</i>	5	[1,7]
<b>Arnt-like factors</b>	<i>ARNTL2</i>	76	[3]
<b>POU5 (Oct-3/4-like factors)</b>	<i>POU5F1</i>	197	[3,4]

doi:10.1371/journal.pcbi.1005311.t001

perform overexpression and knock-down experiments followed by an open chromatin assay for the TFs found by our approach. While this is out of the scope for the current study, we hope that our method can help in prioritizing such experimental efforts.

Further, by its very nature, our methodology cannot with certainty resolve between TFs that belong to the same sub-family. It shares this weakness with almost any method relying on TF motifs. The procedure associates the expression values of each TF separately to the motif accessibilities and one strong association is enough to lead to low CAR ranks for the subfamily. The TF in the subfamily whose expression is the most strongly associated to one of the subfamily motif is naturally also the strongest candidate for CAR activity. (This information is given in [Table 1](#) as well as in [S1 Table](#)). However, if the expression values of the subfamily members are also strongly correlated, we cannot be sure which ones are driving the association.

It is also clear that multiple conditions have to be met for the approach to work. First and foremost, mRNA expression has to be correlated sufficiently with protein concentration of the CAR. Typically, only a fraction of the variation in protein concentration can be explained by variation in mRNA abundances [39]. Nevertheless, better power of our approach can always be achieved by increasing sample size, as long as there is at least some correlation. Further, it is reasonable to assume that our approach will perform better on TFs with a large dynamic range across cell types. This seems indeed to be the case, since most TFs predicted to be CARs tend to have large mRNA expression variance ([S9 Fig](#)). Sampling more and diverse cell lines could address this issue, because it should increase the dynamic range.

This restriction would also suggest that our approach is biased against cell type specific TFs. However, when looking at tissue expression patterns ([www.gtexportal.org](http://www.gtexportal.org) [40]) of the predicted CARs, we found both: TFs that showed expression in a large proportion of cell lines such as *EBF1* and *STAT5B* as well as quite specific TFs. Examples of specific CARs are *SPI1*, which only showed expression in whole blood, and *OTX2*, which only showed expression in some brain regions. It is possible that the use of immortalized cell lines leads to larger gene expression variability in the sample facilitating the detection of such tissue-specific CARs.

For some TFs, activity mainly depends on cofactors. For example, for steroid hormone receptors, hormone molecules activate a pool of inactive TF already present in the cell. In such cases measuring TF activity with gene expression measures can be misleading and one would not expect an association between the expression of a TF and the accessibility of its motif. For example, for the accessibility score of *NR3C1*, we saw much stronger associations with the expression levels of a small set of glucocorticoid response genes (*ZBTB16*, *FKBP5*, *TSC22D3*) than that of *NR3C1* itself [41–43]. This difference in signal strength is in line with the activity of *NR3C1* being mainly regulated by glucocorticoid binding and not *NR3C1* gene expression levels. Of note, *NR3C1* was reported to have pioneer activity [1].

In summary, we exploited the rich data source of ENCODE to find TFs whose mRNA expression levels are directly linked to the open chromatin fraction of the genome. Although our approach in its current form is able to find TFs with strong associations, it is also clear that increasing power by adding more cell lines would find more TFs with an association. From the current data, we would estimate that at least 25% of TF subfamilies show a low CAR rank at the subfamily level, suggesting that the regulation of chromatin accessibility is a pervasive phenomenon amongst TFs.

## Materials and Methods

### Motif accessibility score creation

Annotated open chromatin (FDR <0.01) peaks were downloaded from the EBI website (see URL section) and trimmed to the top 90,000 peaks for each cell line. 426 motifs were

downloaded from the HOCOMOCO website and aligned to the reference genome with FIMO [21,44]. Motif occurrences with a p-value below  $10^{-5}$  were kept for processing. For each motif, we counted the number of DHS peaks overlapping a motif instance in a given cell line using bedops [45]. Results were filtered to motifs that were present in at least 150 DHS peaks on average, leaving 344 motifs. For a given motif, we quantile-normalized the values to follow a normal distribution yielding the raw motif-activity matrix with rows corresponding to motifs and columns corresponding to cell lines. The resulting matrix was iteratively scaled to zero mean and unit standard deviation, first row-wise (across cell lines) then column-wise, until convergence [46,47]. Next, we saw that the cell-line wise covariance matrix had a very large first eigenvalue, with a corresponding eigenvector that did not track well the different tissue origins of the various cell lines. Assuming that this leading principal component largely captured batch effects, we chose to regress out the first eigenvector from each row of the matrix, leading to better agreement between expression and motif accessibility correlation matrices (S13 Fig). After this step, we quantile-normalized the data per motif to follow a normal distribution to ensure that the assumptions of the applied statistical model were met. To map motifs to TFs and TF subfamilies, we used the *TfClass* hierarchy [24]. Of the 344 tested motifs, we mapped 330 to a TF and its subfamily. Of these, 325 had expression data available for a subfamily member.

### Expression matrix creation

We downloaded raw expression microarray data from the GEO repository (GSE1909 and GSE15805). (ENCODE micro-array data was used instead of RNA-seq because to-date more cell lines with DHS information have also RNA expression measured by micro-array than RNA-seq). We background corrected and normalized using the RMA-algorithm implemented in the oligo package to process all arrays for which DHS data was also available [48,49]. Only the core set data was used. The data were summarized to gene level [50]. Only results that had a one-to-one mapping between genes and gene probesets were kept. 15,119 genes could be annotated in this fashion. Because for many cell lines more than one experiment was conducted, we summarized multiple plates by averaging gene results across experiments. The resulting matrix was iteratively scaled to zero mean and unit standard deviation, first row-wise (across cell lines) then column-wise, until convergence [46,47].

### Linear mixed effect model

The model proposed is

$$y = x_i \beta^i + \delta^i + \epsilon^i.$$

Where  $y$  is a vector of motif accessibility scores across  $n$  cell lines,  $x_i$  is the expression vector of gene,  $i$ ,  $\beta^i$  is the effect size of gene  $i$ :

$$\epsilon^i \sim N_n(0, \sigma_r^2 I_n)$$

and

$$\delta^i \sim N_n(0, \sigma_e^2 C_e).$$

$C_e$  is the covariance matrix of the  $n \times p$  expression matrix:

$$C_e = \frac{1}{p} \sum_{i=1}^p x_i x_i^T.$$

For each gene  $i$ ,  $\beta^i$ ,  $\sigma_r$ , and  $\sigma_e$  are estimated via maximum likelihood and the null hypothesis  $\beta^i = 0$  is tested via a likelihood ratio test [18,20]. More details on this procedure are given in [S1 Appendix](#).

### Calculating CAR ranks at the subfamily level

For each motif in HOCOMOCO, we used the mixed model association results across all 1,188 known TFs for which we had mRNA expression data [21]. This yielded a matrix of association p-values for all pairs of 325 motifs (belonging to 147 *TFClass* subfamilies) and 1188 TFs (belonging to 368 *TFClass* subfamilies). Due to the fact that homologous TFs have similar binding motifs, we sought to aggregate results into *CAR ranks* at the subfamily level ([S1 Fig](#)). To achieve this, we reduced the 325 x 1188 motif-TF association matrix to a 147 x 368 matrix of associations between motif subfamilies and TF subfamilies. In practice, for each motif subfamily-TF subfamily pair we collected the most significant p-value among all motif-TF pairs in these subfamilies and multiplied it with the total number of such motif-TF pairs to correct for subfamily size. Finally, for each motif *subfamily*, we ranked the adjusted p-values across all TF subfamilies and defined its *CAR rank* as the rank of its corresponding TF subfamily.

### Calculating pioneer subfamily enrichment

To get an external annotation of pioneer factors, we used a recently published list of established and predicted pioneer factors ([Table 1](#) in Iwafuchi-Doi et al. [3]). We used a hypergeometric test at the 10-fold enrichment cut-off ([Fig 4](#)), as well as a ranksum enrichment test. To derive a ranksum statistic, we summed the CAR ranks of the eight subfamilies annotated as pioneers. To assess significance of this statistic, we used permutation tests: For each of the 50,000 permutation samples, we picked eight CAR ranks from the set of subfamilies not annotated as pioneers and summed them to derive 50,000 permutation sample statistics. The p-value was approximated as the fraction of permutation sample statistics of greater or equal size as the statistic derived for the annotated pioneers. To control pioneer enrichment for mRNA expression variation, we first calculated the expression variance of each TF across all cell lines. The distribution of variance values was transformed to follow a standard normal distribution. We then used the maximal expression variance observed for any TF in each subfamily. To assess significance, we used permutation tests: we sampled eight non-pioneer subfamily level CAR ranks 50,000 times. However, subfamilies were not sampled uniformly: We sampled four non-pioneer subfamilies with maximal expression variance between the 0th and the 50th quantile of the eight pioneer subfamilies, and four non-pioneer subfamilies with maximal expression variance between the 50th and the 100th quantile of the eight pioneer subfamilies.

### Processing ROADMAP data

RNA-seq data were downloaded from the ROADMAP website (see section ‘URLS’) for 56 cell lines. We used only genes with average read count above 50, which removed 12% of genes. The number of reads plus a pseudo-count of one to were log-transformed. Samples were then quantile normalized to the average mean distribution [51]. The resulting matrix was iteratively scaled to zero mean and unit standard deviation, first row-wise (across cell lines) then column-wise, until convergence [46,47].

To derive motif accessibility scores, imputed DHS data were downloaded for 56 cell lines from the ROADMAP website (see section ‘URLs’). From these datasets motif accessibility scores were derived in the same fashion as for the ENCODE DHS data. To derive CAR ranks, the same strategy was employed as for the ENCODE dataset.



## Defining ChIP-seq guided motif cutoff

To compare fixed motif cutoffs to a variable motif cutoff guided by ChIP-seq, the following procedure was used. ChIP-seq data from the Myers and Snyder lab in the ENCODE collection for which dnase1 and expression data were available were downloaded and each ChIP-seq experiment was mapped to a dnase1 experiment based on cell line and to the motif of the TF, yielding mappings to 75 motifs (belonging to 50 subfamilies). For a given motif and cell line pair for which ChIP-seq data (as well as DHS data) was available, each DHS region was annotated with the p-value of its most significant motif instance (given that they contained a motif with p-value below  $5 \times 10^{-5}$ ) as well as whether it overlapped with a ChIP-seq peak. The motif p-value cutoff was defined such that a fixed fraction of peaks with motifs below that cutoff would validate in the ChIP-seq experiment. Three true positive rates were chosen for this comparison 0.3, 0.5 and 0.7 (see [S7 Fig](#), [S8 Fig](#)). Only experiments were used for which it was possible to choose a motif cutoff such that the highest validation rate (i.e. 0.7) could be reached. If multiple ChIP-seq experiments were available per motif, the median p-value cutoff was chosen for each validation rate. We compared these strategies using a fixed cutoff for all motifs of  $10^{-5}$ , which was used throughout the rest of the paper. Results obtained are similar when using ChIP-seq guided cutoffs or fixed cutoffs.

## URLs

Code for reproduction (including scripts for data download) is available at: <https://github.com/dlampart/csrproject>

ENCODE DHS peaks were downloaded from: [http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration\\_data\\_jan2011/byDataType/openchrom/jan2011/fdrPeaks/](http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/openchrom/jan2011/fdrPeaks/)

ROADMAP expression data were downloaded from: <http://egg2.wustl.edu/roadmap/data/byDataType/rna/expression/57epigenomes.N.pc.gz>

ROADMAP imputed DHS peaks were downloaded from: <http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidatedImputed/narrowPeak/>

ENCODE histone files were downloaded from: <ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwHistone/>

## Supporting Information

### S1 Appendix. Supporting methods.

(PDF)

**S1 Table. Results for comprehensive prediction of chromatin accessibility regulators.** The table shows CAR ranks at subfamily level for each motif in the HOCOMOCO library. Subfamily identifiers correspond to the identifier used in *TFClass*. Additionally, the gene in the annotated subfamily with the highest gene wise CAR rank is given.

(XLSX)

**S1 Fig. Association between motif accessibility and mRNA expression for the bona fide pioneer factor *FOXA1*.** Three different regression models (a-c) were used to compute association p-values between the accessibility of a given TF motif (here *FOXA1*) and mRNA expression for each of the assayed 15K protein-coding genes. Results are visualized as QQ-plots showing the  $-\log_{10}$  transformed p-values. (a) Association p-values obtained using standard linear regression. Due to confounding, p-values are strongly inflated and *FOXA1* motif accessibility shows only mild association with *FOXA1* expression compared to other genes. (b) The linear mixed model (LMM) successfully corrects for confounding, with most p-values following the null distribution as expected. The association between *FOXA1* motif accessibility and

*FOXA1* expression now ranks second among all genes and first among all TFs, although it does not pass the Bonferroni significance threshold. (c) Additionally controlling for the first principal component of the motif accessibility matrix corrects for a strong batch effect (Methods) and further lowers the CAR rank. Using this approach, *FOXA1* motif accessibility showed the strongest association precisely with *FOXA1* expression (i.e., the gene-level CAR rank equals one), in line with literature on *FOXA1* being a pioneer factor (Cirillo et al. 2002) (Cirillo et al. 2002; Soufi et al. 2015).

(PNG)

**S2 Fig. Overview of procedure to calculate CAR ranks on the subfamily level.** We cluster TFs and motifs according to subfamily definitions given in *TFClass*. For each bicluster, we define the bicluster score as the most significant p-value between any TF and motif members of the bicluster corrected for bicluster size. We then rank bicluster scores across the TF subfamilies. If the bicluster joining a TF cluster and its corresponding motifs is ranked low, this is an indication of CAR activity.

(PNG)

**S3 Fig. CARs predicted from ENCODE data enrich in subfamilies with low CAR ranks in the ROADMAP dataset.** DHS and expression data available for 56 samples (29 with assayed DHS and 27 with imputed DHS) as part of the ROADMAP data collection were used to predict CARs. Shown are CAR enrichment curves for ENCODE results stratified by CAR ranks derived from ROADMAP. Displayed are the following strata: ROADMAP CAR rank <10 (N = 9 observations in total), ROADMAP CAR rank <20 (N = 20 observations in total), ROADMAP CAR rank <30 (N = 25 observations in total), ROADMAP CAR rank <60 (N = 38 observations in total), ROADMAP CAR rank <100 (N = 58 observations in total), ROADMAP CAR rank >= 100 (N = 86 observations in total). We see that subfamilies with low ROADMAP CAR rank also tend to be predicted to be CARs when using the ENCODE data. This enrichment gets weaker for subfamilies with lower ROADMAP CAR ranking.

(PNG)

**S4 Fig. CARs ranks from ROADMAP data enrich only in subfamilies predicted to be CARs in ENCODE.** DHS and expression data, available as part of the ROADMAP data collection, were used to predict CARs. Shown are CAR enrichment curves for ROADMAP results stratified by CAR predictions derived from ENCODE. Displayed are the following strata: ENCODE CAR rank <10 (N = 37 observations in total), ENCODE CAR rank >= 10 (N = 107 observations in total). While we see enrichment for low ROADMAP CAR rank in subfamilies predicted to be CARs via the ENCODE data, we see no enrichment in low ROADMAP CAR ranks for other subfamilies.

(PNG)

**S5 Fig. CAR detection power is stable to changes in motif cutoffs.** Cumulative distribution of CAR ranks at the subfamily level using the three different motif cutoffs:  $10^{-5}$  (used throughout the paper) is compared to  $10^{-6}$  (yielding 9.3 fewer motifs on average [median]) and  $5 \cdot 10^{-5}$  (yielding 5.2 more motifs assigned on average). For each setting, we filtered motifs that did not overlap at least 150 DHS regions per cell line on average. Only subfamilies passing this filter in all settings were included (62 subfamilies in total). Power mildly increased at low CAR ranks for more stringent cutoffs at the cost of fewer motifs passing filtering. However, at false discovery rate of 10% power was nearly identical.

(PNG)

**S6 Fig. CAR prediction is stable with respect to changes in motif cutoffs.** Shown are pairwise comparisons of different motif cutoffs. For each cutoff we derived CAR ranks for all tested

subfamilies yielding one CAR rank list per cutoff. Pairwise comparisons of these lists were performed in the following manner: For each pair of rank lists, the first list was used to split the tested subfamilies into a 'CAR set' and its complement based on whether a subfamily had CAR rank below 10. For the second results list, two separate CAR enrichment curves were drawn, one curve for the 'CAR set' defined via the first list (black) and its complement (grey). Rows denote the cutoff used to derive the 'CAR set' and columns denote the cutoff used to draw the enrichment curves. For each setting, we filtered motifs that did not overlap at least 150 DHS regions per cell line on average. Only subfamilies passing this filter in all settings were included (62 subfamilies in total). We see that CARs predicted are stable with respect to varying motif cutoffs.

(PNG)

**S7 Fig. CAR detection power does not improve systematically when guiding motif cutoffs via ChIP-seq.** Shown are cumulative distribution of CAR ranks at the subfamily level comparing fixed motif cutoff of  $10^{-5}$  (used throughout the paper) is compared to variable motif cutoffs guided by ChIP-seq data, where motif cutoffs are adjusted such that called binding sites (i.e. DHS sites containing a motif instance) have a fixed validation rate compared to a gold standard defined by ChIP-seq. Chosen validation rates are 0.3, 0.5 and 0.7. For each setting, we filtered motifs that did not overlap at least 150 DHS regions per cell line on average. Only subfamilies passing this filter in all settings were included (32 subfamilies in total). While we see some variation in power, the variation is not systematic.

(PNG)

**S8 Fig. ChIP-seq data guiding motif cutoffs yields similar CAR predictions as regular motif cutoff.** Shown are pairwise comparisons of different motif cutoff methods. For each cutoff method we derived CAR ranks for all tested subfamilies yielding one CAR rank list per method. Pairwise comparisons of these lists were performed in the following manner: For each pair of rank lists, the first list was used to split the tested subfamilies into a 'CAR set' and its complement based on whether a subfamily had CAR rank below 10. For the second results list, two separate CAR enrichment curves were drawn, one curve for the 'CAR set' defined via the first list (black) and its complement (grey). Rows denote the cutoff method used to derive the 'CAR set' and columns denote the cutoff method used to draw the enrichment curves. A fixed motif cutoff of  $10^{-5}$  (also used throughout the paper) is compared to variable motif cutoffs guided by ChIP-seq data, where motif cutoffs are adjusted such that called binding sites (i.e. DHS sites containing a motif instance) have a fixed validation rate when compared to ChIP-seq. Chosen validation rates are 0.3, 0.5 and 0.7. For each setting, we filtered motifs that did not overlap at least 150 DHS regions per cell line on average. Only subfamilies passing this filter in all settings were included (32 subfamilies in total). We see that CARs predicted are stable with respect to varying motif cutoffs.

(PNG)

**S9 Fig. Predicted chromatin accessibility regulators tend to have higher expression variation.** We derived the variance of expression for all transcription factors across micro-arrays after RMA normalization and averaging expression values for experiments derived from the cell types. Displayed is a density distribution of the maximal expression variance observed in each subfamily. We partitioned TF subfamilies into two groups depending on whether they had family level CAR ranks of 1 or not. We observe that top ranked subfamilies do have substantially higher variance on average than other subfamilies (linear regression p-value  $< 10^{-3}$ ).

(PNG)

**S10 Fig. Histone-wise motif activities do not substantially associate with TF expression values.** H3K4me3 peak data for 51 cell lines were downloaded from ENCODE and histone-wise motif activity was computed and normalized analogously to for DHS data, regressing out the first principal component. We performed the mixed model regression where H3K4me3-based motif accessibility data are regressed on gene expression adding a random effect with the same covariance structure as the expression matrix (denoted ‘histone’). To assess the DHS-independent contribution of H3K4me3 histone activities, we added DHS-based motif accessibility as a covariate (denoted ‘DHS-adjusted histone’). We see that subfamily ranks for both of these strategies do not substantially enrich in low ranks. While ‘histone’ performs mildly better, this is likely due to correlation between the histone activity and DHS activity. In contrast, when DHS-based motif accessibility data was adjusted for H3K4me3-based motif accessibility, we see a still substantial enrichment (see “histone-adjusted DHS” curve). This experiment was performed by regressing DHS motif accessibility on gene expression while adding H3K4me3-based motif accessibilities as a covariate plus a random effect with the same covariance structure as the expression matrix. This shows that of the two activity measures, only DHS activity substantially associates with expression.

(PNG)

**S11 Fig. SOX2 expression associates strongly with POU5F1 motif accessibility.** The QQ-plot shows the p-value distribution obtained from the LMM associating the accessibility of the *POU5F1* motif to gene expression values across all genes. We see the strongest association to *SOX2* expression.

(PNG)

**S12 Fig. precision-recall curves of CAR ranks and PIQ pioneer scores and their combination.** Displayed are the precision-recall curves using annotation from Iwafuchi-Doi et al. (2014) as true set. Motif wise PIQ pioneer scores were extracted from Sherwood et al. (2014). For each subfamily, we defined its PIQ pioneer score as the maximal pioneer score for its subfamily members. For 77 subfamilies, data were available from both approaches of which 7 were in the true set. For both CAR ranks and PIQ pioneer scores, precision-recall curves were drawn (CAR rank precision-recall curve starts at 0.43 recall, because many subfamilies share CAR rank of one). Additionally, both scores were combined: For each scoring method, results were ranked (rank ties was replaced by the minimum). For each subfamily, its combined rank is the maximal rank across both methods. A low rank can therefore only be achieved when both methods yielded low ranks. We see that the combined strategy outperforms both base strategies.

(PNG)

**S13 Fig. Removing first principal component from motif accessibility matrix leads to similar correlation structures between motif accessibility and expression.** Displayed are pair-wise correlation matrices with squared entries across cell lines for motif accessibilities (a); motif accessibilities with the first principal component removed (b) and (c) for expression values. Further, the first 25 eigenvalues of these matrices are shown in (d). The motif accessibility matrix has a very dominant first principal component. After removal of the first principal component, the correlation structure of motif accessibility and expression show a similar structure.

(PNG)

## Author Contributions

**Conceived and designed the experiments:** DL ZK SB DM.

**Performed the experiments:** DL.

**Analyzed the data:** DL.

**Wrote the paper:** DL ZK SB DM RR.

## References

- Zaret KS, Carroll JS. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.* 2011; 25: 2227–2241. doi: [10.1101/gad.176826.111](https://doi.org/10.1101/gad.176826.111) PMID: [22056668](https://pubmed.ncbi.nlm.nih.gov/22056668/)
- Morris SA, Baek S, Sung M-H, John S, Wiench M, Johnson TA, et al. Overlapping chromatin-remodeling systems collaborate genome wide at dynamic chromatin transitions. *Nat Struct Mol Biol.* 2013; 21: 73–81. doi: [10.1038/nsmb.2718](https://doi.org/10.1038/nsmb.2718) PMID: [24317492](https://pubmed.ncbi.nlm.nih.gov/24317492/)
- Iwafuchi-Doi M, Zaret KS. Pioneer transcription factors in cell reprogramming. *Genes and Development.* 2014. pp. 2679–2692. doi: [10.1101/gad.253443.114](https://doi.org/10.1101/gad.253443.114) PMID: [25512556](https://pubmed.ncbi.nlm.nih.gov/25512556/)
- Soufi A, Garcia MF, Jaroszewicz A, Osman N, Pellegrini M, Zaret KS. Pioneer Transcription Factors Target Partial DNA Motifs on Nucleosomes to Initiate Reprogramming. *Cell.* 2014.
- Cirillo LA, Lin FR, Cuesta I, Friedman D, Jarnik M, Zaret KS. Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol Cell.* 2002; 9: 279–289. PMID: [11864602](https://pubmed.ncbi.nlm.nih.gov/11864602/)
- Vierbuchen T, Ostermeier A, Pang ZP, Kokubu Y, Südhof TC, Wernig M. Direct conversion of fibroblasts to functional neurons by defined factors. *Nature.* 2010; 463: 1035–1041. doi: [10.1038/nature08797](https://doi.org/10.1038/nature08797) PMID: [20107439](https://pubmed.ncbi.nlm.nih.gov/20107439/)
- Feng R, Desbordes SC, Xie H, Tillo ES, Pixley F, Stanley ER, et al. PU.1 and C/EBPalpha/beta convert fibroblasts into macrophage-like cells. *Proc Natl Acad Sci U S A.* 2008; 105: 6057–6062. doi: [10.1073/pnas.0711961105](https://doi.org/10.1073/pnas.0711961105) PMID: [18424555](https://pubmed.ncbi.nlm.nih.gov/18424555/)
- Menet JS, Pescatore S, Rosbash M. CLOCK:BMAL1 is a pioneer-like transcription factor. *Genes Dev.* 2014; 28: 8–13. doi: [10.1101/gad.228536.113](https://doi.org/10.1101/gad.228536.113) PMID: [24395244](https://pubmed.ncbi.nlm.nih.gov/24395244/)
- Sammons MA, Zhu J, Drake AM, Berger SL. TP53 engagement with the genome occurs in distinct local chromatin environments via pioneer factor activity. *Genome Res.* 2015; 25: 179–188. doi: [10.1101/gr.181883.114](https://doi.org/10.1101/gr.181883.114) PMID: [25391375](https://pubmed.ncbi.nlm.nih.gov/25391375/)
- Koster MI. P63 in Skin Development and Ectodermal Dysplasias. *J Invest Dermatol.* 2010; 130: 2352–8. doi: [10.1038/jid.2010.119](https://doi.org/10.1038/jid.2010.119) PMID: [20445549](https://pubmed.ncbi.nlm.nih.gov/20445549/)
- Voss TC, Schiltz RL, Sung MH, Yen PM, Stamatoyannopoulos JA, Biddie SC, et al. Dynamic exchange at regulatory elements during chromatin remodeling underlies assisted loading mechanism. *Cell.* 2011; 146: 544–554. doi: [10.1016/j.cell.2011.07.006](https://doi.org/10.1016/j.cell.2011.07.006) PMID: [21835447](https://pubmed.ncbi.nlm.nih.gov/21835447/)
- Voss TC, Hager GL. Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nature reviews. Genetics.* 2014. pp. 69–81.
- Schulz KN, Bondra ER, Moshe A, Villalta JE, Lieb JD, Kaplan T, et al. Zelda is differentially required for chromatin accessibility, transcription-factor binding and gene expression in the early *Drosophila* embryo. *Genome Res.* 2015;
- Cirillo LA, Lin FR, Cuesta I, Friedman D, Jarnik M, Zaret KS. Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol Cell.* 2002; 9: 279–289. PMID: [11864602](https://pubmed.ncbi.nlm.nih.gov/11864602/)
- Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal A a, van Hoff JP, et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotechnol.* 2014; 32: 171–8. doi: [10.1038/nbt.2798](https://doi.org/10.1038/nbt.2798) PMID: [24441470](https://pubmed.ncbi.nlm.nih.gov/24441470/)
- Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012; 489: 57–74. doi: [10.1038/nature11247](https://doi.org/10.1038/nature11247) PMID: [22955616](https://pubmed.ncbi.nlm.nih.gov/22955616/)
- Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. *Nature.* Nature Publishing Group; 2012; 488: 75–82.
- Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient control of population structure in model organism association mapping. *Genetics.* 2008; 178: 1709–1723. doi: [10.1534/genetics.107.080101](https://doi.org/10.1534/genetics.107.080101) PMID: [18385116](https://pubmed.ncbi.nlm.nih.gov/18385116/)
- Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet.* 2010; 42: 348–54. doi: [10.1038/ng.548](https://doi.org/10.1038/ng.548) PMID: [20208533](https://pubmed.ncbi.nlm.nih.gov/20208533/)

20. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. *Nature Methods*. 2011. pp. 833–835. doi: [10.1038/nmeth.1681](https://doi.org/10.1038/nmeth.1681) PMID: [21892150](https://pubmed.ncbi.nlm.nih.gov/21892150/)
21. Kulakovskiy I V., Medvedeva YA, Schaefer U, Kasianov AS, Vorontsov IE, Bajic VB, et al. HOCO-MOCO: A comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res*. 2013; 41.
22. Treiber T, Mandel EM, Pott S, Gy??ry I, Firner S, Liu ET, et al. Early B cell factor 1 regulates B cell gene networks by activation, repression, and transcription-independent poising of chromatin. *Immunity*. 2010; 32: 714–725. doi: [10.1016/j.immuni.2010.04.013](https://doi.org/10.1016/j.immuni.2010.04.013) PMID: [20451411](https://pubmed.ncbi.nlm.nih.gov/20451411/)
23. Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res*. 2010; 20: 861–873. doi: [10.1101/gr.100552.109](https://doi.org/10.1101/gr.100552.109) PMID: [20378718](https://pubmed.ncbi.nlm.nih.gov/20378718/)
24. Wingender E, Schoeps T, Dönitz J. TFClass: An expandable hierarchical classification of human transcription factors. *Nucleic Acids Res*. 2013; 41.
25. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, et al. Architecture of the human regulatory network derived from ENCODE data. *Nature*. 2012; 489: 91–100. doi: [10.1038/nature11245](https://doi.org/10.1038/nature11245) PMID: [22955619](https://pubmed.ncbi.nlm.nih.gov/22955619/)
26. Kundaje, Anshul Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015; 518: 317–330. doi: [10.1038/nature14248](https://doi.org/10.1038/nature14248) PMID: [25693563](https://pubmed.ncbi.nlm.nih.gov/25693563/)
27. Evans RM, Mangelsdorf DJ. Nuclear receptors, RXR, and the big bang. *Cell*. 2014. pp. 255–266. doi: [10.1016/j.cell.2014.03.012](https://doi.org/10.1016/j.cell.2014.03.012) PMID: [24679540](https://pubmed.ncbi.nlm.nih.gov/24679540/)
28. Lu NZ, Wardell SE, Burnstein KL, Defranco D, Fuller PJ, Giguere V, et al. International Union of Pharmacology. LXV. The Pharmacology and Classification of the Nuclear Receptor Superfamily: Glucocorticoid, Mineralocorticoid, Progesterone, and Androgen Receptors. *Pharmacol Rev*. 2006; 58: 782–97. doi: [10.1124/pr.58.4.9](https://doi.org/10.1124/pr.58.4.9) PMID: [17132855](https://pubmed.ncbi.nlm.nih.gov/17132855/)
29. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*. 2015; 43: D447–D452. doi: [10.1093/nar/gku1003](https://doi.org/10.1093/nar/gku1003) PMID: [25352553](https://pubmed.ncbi.nlm.nih.gov/25352553/)
30. Buganim Y, Faddah DA, Jaenisch R. Mechanisms and models of somatic cell reprogramming. *Nat Rev Genet*. 2013; 14: 427–439. doi: [10.1038/nrg3473](https://doi.org/10.1038/nrg3473) PMID: [23681063](https://pubmed.ncbi.nlm.nih.gov/23681063/)
31. Chen J, Zhang Z, Li L, Chen BC, Revyakin A, Hajj B, et al. Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. *Cell*. 2014; 156: 1274–1285. doi: [10.1016/j.cell.2014.01.062](https://doi.org/10.1016/j.cell.2014.01.062) PMID: [24630727](https://pubmed.ncbi.nlm.nih.gov/24630727/)
32. Fu L, Lee CC. The circadian clock: pacemaker and tumour suppressor. *Nat Rev Cancer*. 2003; 3: 350–361. doi: [10.1038/nrc1072](https://doi.org/10.1038/nrc1072) PMID: [12724733](https://pubmed.ncbi.nlm.nih.gov/12724733/)
33. Tan SK, Lin ZH, Chang CW, Varang V, Chng KR, Pan YF, et al. AP-2γ regulates oestrogen receptor-mediated long-range chromatin interaction and gene transcription. *EMBO J*. 2011; 30: 2569–2581. doi: [10.1038/emboj.2011.151](https://doi.org/10.1038/emboj.2011.151) PMID: [21572391](https://pubmed.ncbi.nlm.nih.gov/21572391/)
34. Plachetka A, Chayka O, Wilczek C, Melnik S, Bonifer C, Klempnauer K-H. C/EBPβ induces chromatin opening at a cell-type-specific enhancer. *Mol Cell Biol*. 2008; 28: 2102–2112. doi: [10.1128/MCB.01943-07](https://doi.org/10.1128/MCB.01943-07) PMID: [18195047](https://pubmed.ncbi.nlm.nih.gov/18195047/)
35. Buecker C, Srinivasan R, Wu Z, Calo E, Acampora D, Faial T, et al. Reorganization of enhancer patterns in transition from naive to primed pluripotency. *Cell Stem Cell*. 2014; 14: 838–853. doi: [10.1016/j.stem.2014.04.003](https://doi.org/10.1016/j.stem.2014.04.003) PMID: [24905168](https://pubmed.ncbi.nlm.nih.gov/24905168/)
36. Hayden MS, Ghosh S. NF-κB, the first quarter-century: Remarkable progress and outstanding questions. *Genes Dev*. 2012; 26: 203–234. doi: [10.1101/gad.183434.111](https://doi.org/10.1101/gad.183434.111) PMID: [22302935](https://pubmed.ncbi.nlm.nih.gov/22302935/)
37. Hagan CR, Knutson TP, Lange CA. A common docking domain in progesterone receptor-B links DUSP6 and CK2 signaling to proliferative transcriptional programs in breast cancer cells. *Nucleic Acids Res*. 2013; 41: 8926–8942. doi: [10.1093/nar/gkt706](https://doi.org/10.1093/nar/gkt706) PMID: [23921636](https://pubmed.ncbi.nlm.nih.gov/23921636/)
38. Hori S. c-Rel: A pioneer in directing regulatory T-cell lineage commitment? *European Journal of Immunology*. 2010. pp. 664–667. doi: [10.1002/eji.201040372](https://doi.org/10.1002/eji.201040372) PMID: [20162555](https://pubmed.ncbi.nlm.nih.gov/20162555/)
39. Maier T, Güell M, Serrano L. Correlation of mRNA and protein in complex biological samples. *FEBS Letters*. 2009. pp. 3966–3973. doi: [10.1016/j.febslet.2009.10.036](https://doi.org/10.1016/j.febslet.2009.10.036) PMID: [19850042](https://pubmed.ncbi.nlm.nih.gov/19850042/)
40. Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. The human transcriptome across tissues and individuals. *Science (80-)*. 2015; 348: 660–665.
41. Wasim M, Carlet M, Mansha M, Greil R, Ploner C, Trockenbacher A, et al. PLZF/ZBTB16, a glucocorticoid response gene in acute lymphoblastic leukemia, interferes with glucocorticoid-induced apoptosis. *J Steroid Biochem Mol Biol*. 2010; 120: 218–227. doi: [10.1016/j.jsmb.2010.04.019](https://doi.org/10.1016/j.jsmb.2010.04.019) PMID: [20435142](https://pubmed.ncbi.nlm.nih.gov/20435142/)

42. Galigniana NM, Ballmer LT, Toneatto J, Erlejman AG, Lagadari M, Galigniana MD. Regulation of the glucocorticoid response to stress-related disorders by the Hsp90-binding immunophilin FKBP51. *Journal of Neurochemistry*. 2012. pp. 4–18.
43. Rog-Zielinska E a, Craig M, Manning JR, Richardson R V, Gowans GJ, Dunbar DR, et al. Glucocorticoids promote structural and functional maturation of foetal cardiomyocytes: a role for PGC-1 $\alpha$ . *Cell Death Differ*. 2015; 22: 1–11.
44. Grant CE, Bailey TL, Noble WS. FIMO: Scanning for occurrences of a given motif. *Bioinformatics*. 2011; 27: 1017–1018. doi: [10.1093/bioinformatics/btr064](https://doi.org/10.1093/bioinformatics/btr064) PMID: [21330290](https://pubmed.ncbi.nlm.nih.gov/21330290/)
45. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, et al. BEDOPS: High-performance genomic feature operations. *Bioinformatics*. 2012; 28: 1919–1920. doi: [10.1093/bioinformatics/bts277](https://doi.org/10.1093/bioinformatics/bts277) PMID: [22576172](https://pubmed.ncbi.nlm.nih.gov/22576172/)
46. Efron B. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction* (Institute of Mathematical Statistics Monographs). Reprint. Cambridge University Press; 2013.
47. Olshen RA, Rajaratnam B. Successive normalization of rectangular arrays. *Ann Stat*. 2010; 38: 1638–1664. doi: [10.1214/09-AOS743](https://doi.org/10.1214/09-AOS743) PMID: [20473354](https://pubmed.ncbi.nlm.nih.gov/20473354/)
48. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003; 4: 249–264. doi: [10.1093/biostatistics/4.2.249](https://doi.org/10.1093/biostatistics/4.2.249) PMID: [12925520](https://pubmed.ncbi.nlm.nih.gov/12925520/)
49. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. *Bioinformatics*. 2010; 26: 2363–2367. doi: [10.1093/bioinformatics/btq431](https://doi.org/10.1093/bioinformatics/btq431) PMID: [20688976](https://pubmed.ncbi.nlm.nih.gov/20688976/)
50. Wells CA, Mosbergen R, Korn O, Choi J, Seidenman N, Matigian NA, et al. Stemformatics: Visualisation and sharing of stem cell gene expression. *Stem Cell Res*. 2013; 10: 387–395. doi: [10.1016/j.scr.2012.12.003](https://doi.org/10.1016/j.scr.2012.12.003) PMID: [23466562](https://pubmed.ncbi.nlm.nih.gov/23466562/)
51. Bolstad BM, Irizarry R., Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 2003; 19: 185–193. PMID: [12538238](https://pubmed.ncbi.nlm.nih.gov/12538238/)

## 4. Investigation of the genetic control of *Drosophila* body size

Animal models have long been a dominant object of study in genetics, long before cheap genotyping made unbiased investigation of the human genetics feasible. The reason is mainly the ability to perform experimental manipulations. Instead of relying on the natural variation in outbred population, in animal models, one can also modify the genetic variation. Examples are random knockdown screens or experimental setups where inbred strains are cross-bred to perform linkage studies [57, 58]. Not only can the genetics be controlled but also the environment leading to studies where genetics almost completely explains the phenotypic variation. Furthermore, results obtained can be followed up with additional experiments in the same model organism (and when using inbred species even the same genotype) with relative ease. It is clear however, that studies using artificially controlled genetics will have limited information on the genetics of the natural out-bred population. The following study is an investigation of the genetics of *Drosophila* body size measures. In contrast to the bulk of genetic studies investigating *Drosophila* growth control, a natural population of outbreeding *Drosophila* was sampled randomly to allow for better inference of the genetics of a natural *drosophila* population. The study is noteworthy for being the first GWAS of *Drosophila* on body size traits. It made use of extensive environmental control and confounding correction, with the use of mixed models to better model the phenotypes. It further is notable for the fact that follow-up experiments were performed on candidate genes.

The study was spearheaded by Dr. Sybille Vonesch, who gathered the data and had conceived the experiment together with her supervisor Prof. Ernst Hafen. I contributed various computational and statistical analyses of the data. Specifically, I helped establish and perform the statistical modeling of the phenotypes and established a GWAS analysis strategy. Further, I performed gene-wise analysis as well as epistatic analysis and devised and implemented a computational strategy to detect gene set enrichment in epistatic analysis results. The results were published in PLOS Genetics in a paper written by Vonesch. To the text, I contributed sections on the



technical details of the analysis performed. This paper is reproduced below.

RESEARCH ARTICLE

# Genome-Wide Analysis Reveals Novel Regulators of Growth in *Drosophila melanogaster*

Sibylle Chantal Vonesch<sup>1‡</sup>, David Lamparter<sup>2</sup>, Trudy F. C. Mackay<sup>3</sup>, Sven Bergmann<sup>2</sup>, Ernst Hafen<sup>1\*</sup>

**1** Institute of Molecular Systems Biology, ETH Zürich, Zürich, Switzerland, **2** Department of Medical Genetics, University of Lausanne, Lausanne, Switzerland, **3** Department of Biological Sciences, Program in Genetics, W. M. Keck Center for Behavioral Biology, North Carolina State University, Raleigh, North Carolina, United States of America

‡ Current address: Genome Biology Unit, European Molecular Biology Laboratories, Heidelberg, Germany

\* [hafen@imsb.biol.ethz.ch](mailto:hafen@imsb.biol.ethz.ch)



CrossMark  
click for updates

 OPEN ACCESS

**Citation:** Vonesch SC, Lamparter D, Mackay TFC, Bergmann S, Hafen E (2016) Genome-Wide Analysis Reveals Novel Regulators of Growth in *Drosophila melanogaster*. PLoS Genet 12(1): e1005616. doi:10.1371/journal.pgen.1005616

**Editor:** Gregory S. Barsh, Stanford University School of Medicine, UNITED STATES

**Received:** May 2, 2015

**Accepted:** September 28, 2015

**Published:** January 11, 2016

**Copyright:** © 2016 Vonesch et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

**Funding:** This work was funded by grant SXRTX0-123851 from SystemsX.ch, the Swiss National Science Foundation grant 31003AB\_135699 and financial support from ETH Zurich to EH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

## Abstract

Organismal size depends on the interplay between genetic and environmental factors. Genome-wide association (GWA) analyses in humans have implied many genes in the control of height but suffer from the inability to control the environment. Genetic analyses in *Drosophila* have identified conserved signaling pathways controlling size; however, how these pathways control phenotypic diversity is unclear. We performed GWA of size traits using the *Drosophila* Genetic Reference Panel of inbred, sequenced lines. We find that the top associated variants differ between traits and sexes; do not map to canonical growth pathway genes, but can be linked to these by epistasis analysis; and are enriched for genes and putative enhancers. Performing GWA on well-studied developmental traits under controlled conditions expands our understanding of developmental processes underlying phenotypic diversity.

## Author Summary

Genetic studies in *Drosophila* have elucidated conserved signaling pathways and environmental factors that together control organismal size. In humans, hundreds of genes are associated with height variation, but these associations have not been performed in a controlled environment. As a result we are still lacking an understanding of the mechanisms creating size variability within a species. Here, under carefully controlled environmental conditions, we identify naturally occurring genetic variants that are associated with size diversity in *Drosophila*. We identify a cluster of associations close to the *kek1* locus, a well-characterized growth regulator, but otherwise find that most variants are located in or close to genes that do not belong to the conserved pathways but may interact with these in a biological network. We validate 33 novel growth regulatory genes that participate in diverse cellular processes, most notably cellular metabolism and cell polarity. This study is

the first genome-wide association analysis of natural variants underlying size in *Drosophila* and our results complement the knowledge we have accumulated on this trait from mutational studies of single genes.

## Introduction

How animals control and coordinate growth among tissues is a fundamental question in developmental biology. A detailed mechanistic but global understanding of the processes taking place during normal physiological development is furthermore relevant for understanding pathological growth in cancers. Classical genetic studies in *Drosophila* have revealed core molecular mechanisms governing growth control and have shed light on the role of humoral factors and the environment on adult size [1–4]. Two major pathways regulate size, the Insulin/TOR pathway, which couples systemic growth to nutrient availability; and the Hippo tumor suppressor pathway, which controls cell survival and proliferation in developing organs [5–7]. However, growth control is complex [8–10], and the interactions between components of these pathways with each other and with unknown molecules and extrinsic factors remain poorly understood. Studies focusing on single or a few genes can only capture individual aspects of the entire system of networks underlying this trait, which is especially problematic when individual alleles have subtle and context-dependent effects [11, 12]. Therefore, global genome-wide approaches are needed for a better understanding of the genetic control of size. One genome-wide approach is to study multifactorial natural genetic perturbations as they occur in a segregating, phenotypically diverse population.

Artificial selection experiments have revealed that naturally occurring populations of *Drosophila melanogaster* show abundant genetic variation for size, with heritabilities approaching 50% [13]. Usually selection for size results in correlated responses in the same direction for all body parts and overall weight, indicating a common genetic architecture [14]. Selection responses differ between populations but not between sexes [15]. Body size is an important component of fitness in *D. melanogaster* since there are parallel clines in body size and correlated traits clines across different continents [16, 17]. Loci on chromosome 3R and 2R are, respectively, associated with body size and wing area [16, 17]; interestingly, the majority of the 3R loci seem to be located within the polymorphic chromosomal inversion *In(3R)Payne* [18, 19]. Candidate genes and variants associated with size within *In(3R)Payne* include *hsr-omega*, the microsatellite loci DMU25686 and AC008193, and genes in the Insulin signaling pathway (*InR*, *Tsc1*, *Akt1*) [20, 21]. Similarly, the frequency of the polymorphic inversion *In(2L)t* is associated with a body size cline across several continents; genes in the IIS/TOR pathway (*chico*, *Pten*, *Tor*) are located in the inversion region and *Pi3K21B* and *Idgfs 1–3* are located immediately proximal to it [21]. Naturally segregating alleles in *smp-30* (*Dca*) and *InR* have been causally associated with body weight [22, 23]. Recently, a long-term selection experiment identified hundreds of loci with allele frequency differences between large and small populations [24], indicating that the genetic basis of naturally occurring variation in size is highly polygenic. Candidate loci were enriched for genes implicated in post-embryonic development, metamorphosis and cell morphogenesis. The genes included components of the EGFR, Hippo and many other growth pathways, as well as canonical IIS/TOR signaling genes. Therefore, dissecting the genetic basis of naturally occurring variation in body size has the potential to uncover novel variants in known loci affecting body size as well as identify novel genes.

The advent of next-generation sequencing technology has enabled the rapid and relatively cheap acquisition of complete genome sequences, and thereby the generation of very dense

genotype information that enables genome-wide association (GWA) mapping with a much higher resolution than previously possible. GWA studies aim to link variation in quantitative traits to underlying genetic loci in populations of unrelated individuals genome-wide [25, 26]. GWAS have been pioneered [27, 28] and widely applied in humans and are now a routinely used tool in model organisms such as *Arabidopsis* [29, 30], *Drosophila* [31–33] and mouse [34] as well as in various crop [35, 36] and domestic animal species [37–41], where they have substantially broadened our understanding of the genetics of complex traits. To date there are no GWA analyses of size in *Drosophila*, but GWA studies of height have revealed that many loci with small effect contribute to size variation in human populations [9, 10, 42, 43], which contrasts with a much simpler genetic architecture of size in domestic animals, where as a consequence of breeding few loci have relatively large effect sizes that jointly explain a large proportion of size variation [38, 44]. Although many loci affecting human height have been identified by GWA analyses, deducing the underlying molecular mechanisms by which they affect size is challenging. Larger genome regions and not single genes are mapped; uncontrolled environmental variability makes it difficult to identify causal links between genotype and phenotype; and functional validation cannot be performed in humans [12, 28, 45–47].

In contrast to human studies, GWA studies in model organisms benefit from the feasibility of functional validation, more stringent environmental control and, when using inbred strains, the possibility of measuring many genetically identical individuals to obtain an accurate estimate of the phenotype for a given trait and genotype. All three factors can substantially improve the power of a GWA analysis. The establishment of the inbred, sequenced lines of the *Drosophila* Genetic Reference Panel (DGRP) [48, 49] has made GWA analysis in *Drosophila* widely applicable. The DGRP lines harbor the substantial natural genetic variation present in the original wild population and show copious phenotypic variation for all traits assayed to date [31–33, 48, 50, 51].

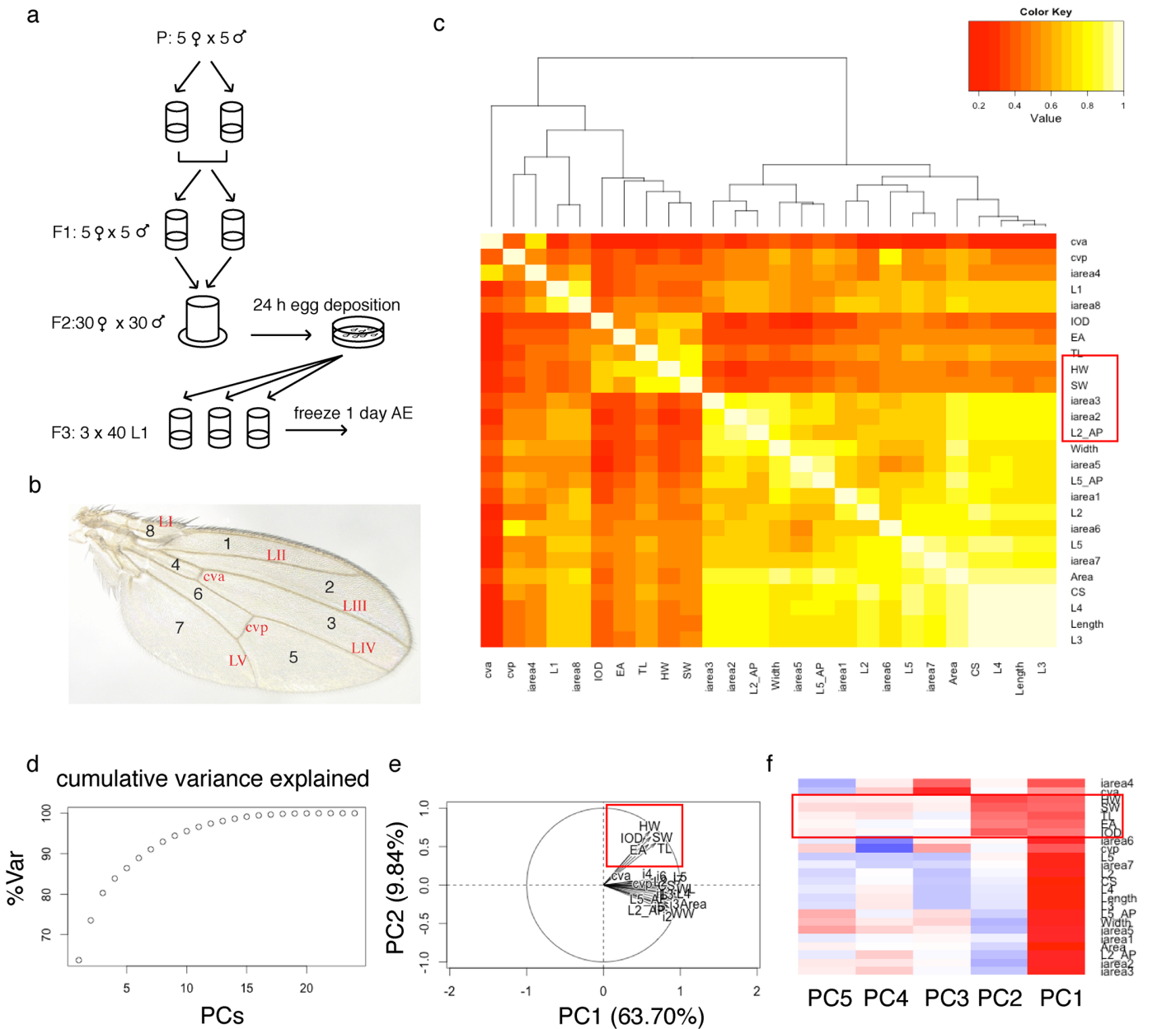
Here, we used the DGRP to perform single- and two-locus associations for size-related developmental traits in *Drosophila*. We find pervasive trait and sex-specificity of top variants, validate a substantial number of novel growth regulators, and extend our knowledge of the genetic control of size beyond existing growth regulatory networks.

## Results

### Quantitative genetic analysis of size

We cultured 143 DGRP lines under conditions we had previously shown to reduce environmental influences on size (S1 Table, Fig 1A) and measured five body and 21 wing traits (S1 Table, Fig 1B). The cross trait genetic correlations were positive and generally high among all features except small veins and areas that were difficult to quantify accurately, indicating shared genetic architecture of the various size measures. We observed two modules of higher correlation, one formed by wing traits and the second by head/thorax traits (Figs 1C and S1), indicating that the genetic architecture is more similar among wing features and among head/thorax features than between traits of the wing and head/thorax. Principal component analysis (PCA) of 23 of the 26 size traits (L1, L6 and iarea8 were excluded since measuring these traits accurately was very difficult) revealed that the first two PCs explained nearly 75% of the observed phenotypic variation. The first component reflected an overall size element and the second component separated wing from head/thorax traits (Figs 1D–1F and S1).

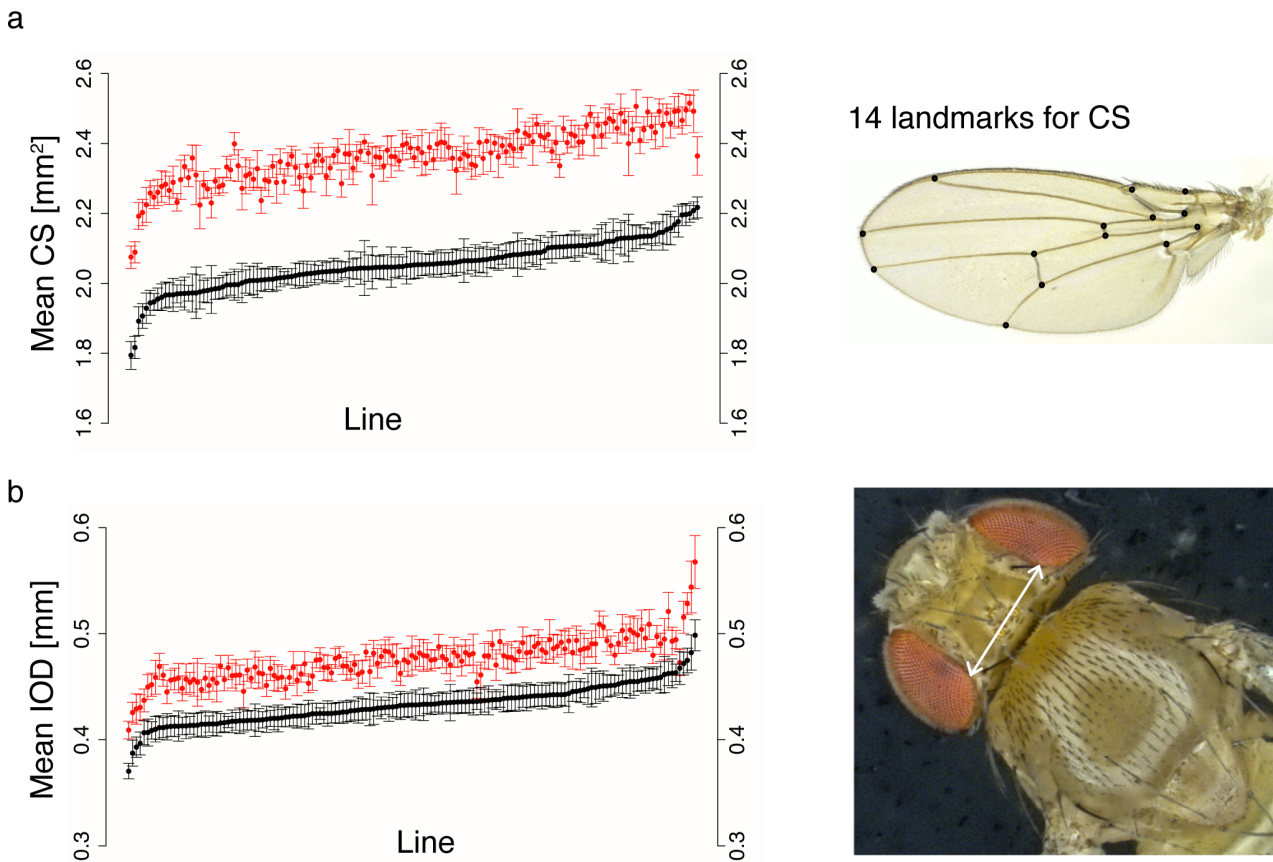
Given the observed redundancy of the phenotypes, we chose only one trait from each high-correlation module for further in-depth analysis: centroid size (CS, reflecting growth processes in the wing disc), and interocular distance (IOD, reflecting eye disc growth), respectively (Fig 2A and 2B). IOD showed the lowest genetic correlation with CS of all head/thorax traits



**Fig 1. Analysis of 26 size traits in the DGRP.** (a) Standardized *Drosophila* culture conditions for the quantification of morphometric traits. The protocol extends over three generations and efficiently controls known covariates of size, such as temperature, humidity, day-night-cycle and crowding. Additionally, effects of other environmental covariates, such as intra-vial environment, light intensity and incubator position, are randomized. (b) Illustration of the wing features. L2\_AP and L5\_AP are not illustrated; they comprise the area between the AP boundary and L2 or L5, respectively, and serve as measures for the size of the anterior and posterior part of the wing. (c) Genetic correlation between morphometric traits in females. Two modules of higher correlation are clearly visible (bright yellow): one encompassing almost all wing features and one comprising all head/thorax traits. (d) Cumulative variance explained in female data by increasing number of principal components. (e) Variables factor map. PC1 and PC2 separate the data into two groups. (f) Correlation between PCs and traits. PC1 reflects a general size component and PC2 is highly correlated with head/thorax traits, effectively splitting the data into two groups.

doi:10.1371/journal.pgen.1005616.g001

(0.46 in females and 0.51 in males). Interestingly, the allometric coefficient  $b$  describing the relationship  $CS = a \cdot X^b$  (where  $X = IOD$  or  $TL$ ) varied substantially between lines, from near independence ( $b = 0$ ) to hyperallometry (positive allometry  $b > 1$ ) (S2 Fig, S1 Table). We



**Fig 2. Phenotypic variation in the DGRP for two size traits.** Plots show mean phenotypic values for (a) centroid size and (b) interocular distance. Each dot represents the mean phenotype per line of males (black) and corresponding females (red), with error bars denoting one standard deviation. Lines are ordered on the x-axis according to male trait value, from lowest to highest: consequently, the order of lines is different for each plot. Raw phenotypes and line means are listed in [S2 Table](#). To the right are illustrations of both measures.

doi:10.1371/journal.pgen.1005616.g002

observed extensive phenotypic and genetic variation in both phenotypes ([Fig 2A and 2B](#), [S1](#) and [S2 Tables](#)), which was reflected in the substantial broad-sense heritabilities ( $H^2_{CS} = 0.63$ ,  $H^2_{IOD} = 0.69$ ). Furthermore, both traits showed significant genetic variation in sex dimorphism but similar heritability estimates for males and females and high cross-sex genetic ( $r_{MF}$ ) and phenotypic correlation ([S1 Table](#)). 15% of phenotypic variance in centroid size could be attributed to raising flies on different food batches, which only differed by the day on which they were prepared (according to the same protocol) ([S1 Table](#)). Though nutrition is a well-studied size-determining factor [[52](#)], we were surprised at the substantial phenotypic effects elicited by even such a small nutritional variation. Although the environmental effect of food batch was markedly lower for IOD (3%), we used batch-mean corrected phenotypes in all subsequent analyses to remove this effect.

### Single-marker and gene based GWAS identify novel loci associated with size variation

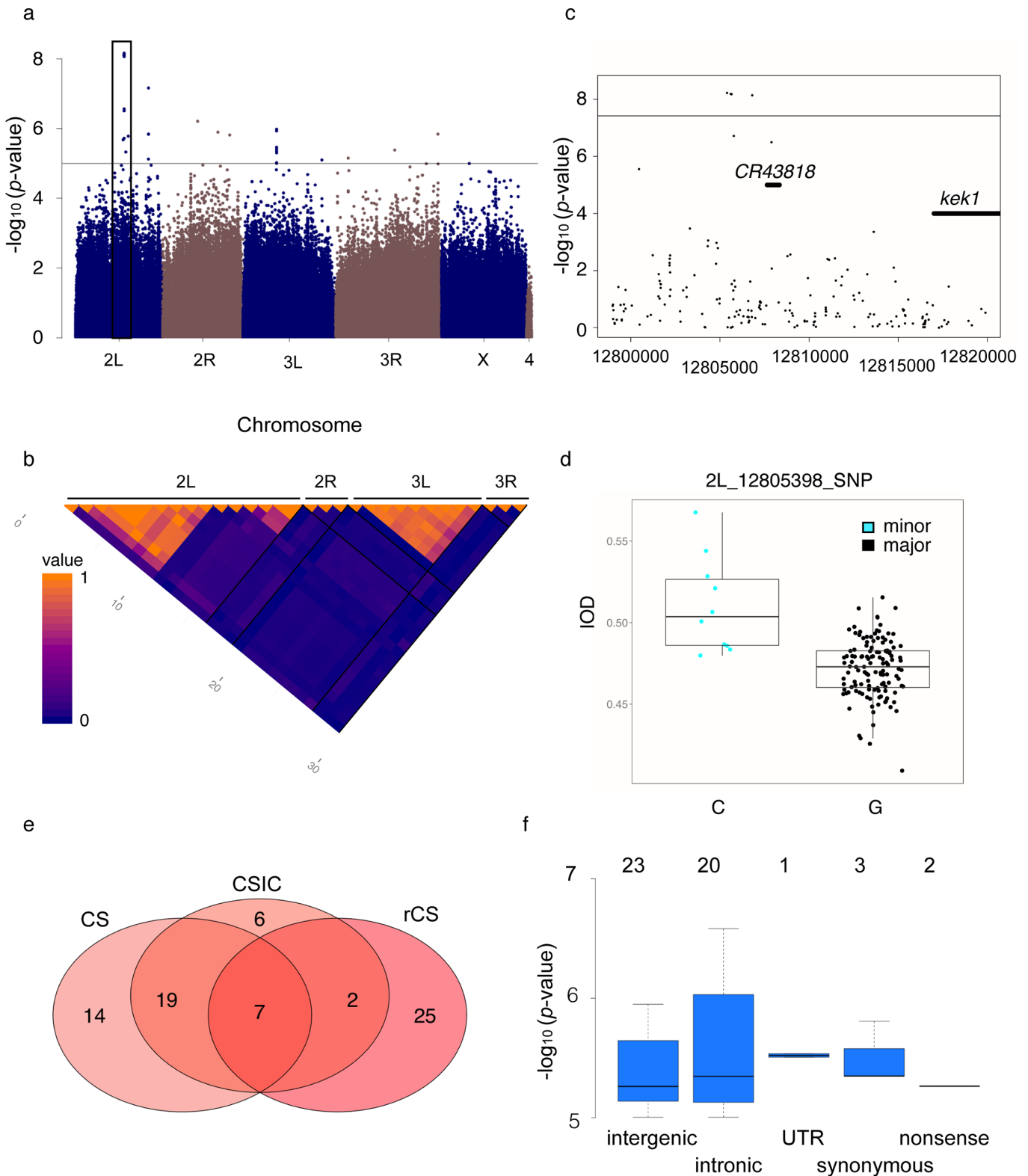
To identify common loci contributing to size variation in *Drosophila*, we performed single marker GWA analyses for 1,319,937 SNPs for a wing disc derived (CS) and an eye disc derived (IOD) size measure using Fast-LMM [[53](#)]. This association method uses a linear mixed model

to capture confounders such as population structure and cryptic relatedness. As the genetic correlation between CS and IOD was moderate (0.46 and 0.51 for females and males, respectively), we expected to map both shared and trait-specific SNPs. To find loci that specifically affect variation in wing size unrelated to the overall organismal size variation we constructed an additional phenotype (rCS) that had the effect of IOD on CS removed via regression. In addition to the effect of the food batch, two cosmopolitan inversions, *In(2L)t* and *In(3R)Mo*, were correlated with both CS and IOD and we addressed their effect on size by modeling their presence in the homozygous state (S2 Fig), yielding the inversion-corrected phenotypes CS<sub>IC</sub> and IOD<sub>IC</sub>. *In(3R)P*, which is known to be correlated with *Drosophila* size [19], was present in the homozygous state in only one line; therefore, we could not estimate its effect on size.

Only for one trait (IOD in females) did we observed significantly associated SNPs when applying a stringent Bonferroni corrected  $p$ -value threshold of  $3.8 \times 10^{-08}$ . However, the significance of these six SNPs dropped below the genome-wide level when we applied GWAS on rank-normalized IOD, which was probably due to an outlier line ( $>4SD$ ) in the minor allele class of all six SNPs. Overall, the  $p$ -values between normalized and non-normalized GWAS showed good correlation and the locus clearly segregates with size, as 75% of the major allele class lines had a smaller IOD than the lines of the minor allele class (S3 Fig). The six SNPs were all located in a cluster on chromosome 2L (2L: 12'805'398–12'806'812), 12–13kb upstream of the gene encoding the EGFR pathway regulator *kek1* (Fig 3A and 3C). Three more SNPs in this locus were annotated to *kek1*, but did not survive Bonferroni correction. All nine SNPs formed a haplotype, with lines having either all minor or all major alleles of these SNPs, and the minor allele haplotype was associated with an increased IOD (Figs 3B, 3D and S3). In total, 198 SNPs are located in the 20 kb genomic region upstream of the *kek1* transcript start site. This region showed high conservation between species (DGRP Freeze 2 genome browser, <http://genome.ucsc.edu>) and several blocks of higher LD are formed across it (Figs 3B and S3), which could be attributable to its proximity to *In(2L)t* (2L: 2'225'744–13'154'180) [54]. However, none of the lines with the minor allele haplotype was either homo- or heterozygous for this inversion, and they were distributed across all four food batches (S3 Fig). Interestingly, a noncoding RNA, *CR43818*, was located in the 20kb region upstream of *kek1*, and the region was spanned by the intron of *CG9932*, a poorly characterized gene that interacts genetically with *Bx* and *Chi* during wing development [55]. Clearly there are signs for functionality of this locus, and several good candidates for causal variants. Further experiments are required to elucidate the molecular mechanism of this association and its potential connection to *In(2L)t*.

As QQ-plots showed a departure from uniformity for  $p$ -values below  $10^{-05}$  (S4 and S5 Figs) we picked candidate loci using this nominal significance threshold for hypothesis generation and functional validation. The corresponding  $q$ -values for each SNP are listed in S3 Table. This yielded between 31 and 51 SNPs for females and between 17 and 36 SNPs for males, with little overlap between top associations and moderate correlation of overall SNP ranks between sexes (S3 and S4 Tables; S6, S7 and S8 Figs), consistent with significant sex by line variances and departure of the cross-sex genetic correlations from unity in the quantitative genetic analyses.

Correcting for the segregation of polymorphic inversions generally enhanced the power of the GWA analyses, as was evident by more loci reaching nominal significance. Nevertheless, the majority (65–86%) of SNPs from the GWA analysis with uncorrected trait values remained candidates in the GWA analysis with corrected phenotypes. Somewhat surprisingly, despite the significant genetic correlation between CS and IOD, no candidate SNPs were shared between these phenotypes (Fig 3E, S4 Table). In both sexes, approximately one-third of top SNPs was shared between the absolute and relative CS GWA analyses, suggesting variation in relative versus absolute organ size may be achieved through genetic variation at both shared and private loci.



**Fig 3. Genome-wide association of size traits.** (a) Manhattan plot of the SNP  $p$ -values from the IOD GWAS in females shows that nominally associated SNPs are distributed over all chromosomes. Negative  $\log_{10} p$ -values are plotted against genomic position, the black horizontal line denotes the nominal



significance threshold of  $10^{-05}$  and the black box marks the location of the cluster of Bonferroni-significant SNPs upstream of *kek1* on 2L. (b) Correlation between SNPs nominally associated with female IOD. The cluster of Bonferroni-significant SNPs on 2L shows high correlation among individual SNPs over a larger region, whereas most other SNPs except a few in a narrow region on 3L represent individual associations. Blue = No correlation, orange = complete correlation. Pixels represent individual SNPs and black lines divide chromosomes. (c) Locus zoom plot of the region 20kb upstream of *kek1* that harbors the genome-wide significant associations. The black horizontal line denotes the genome-wide significance threshold ( $p = 3.8 \times 10^{-08}$ ) and the locations of *kek1* and the ncRNA *CR43818* are marked by broad black lines. (d) Lines with the minor allele genotype at the most significantly associated locus have a larger IOD than lines with the major allele. (e) Overlap in the number of nominally associated SNPs for different wing traits in females. The overlap is bigger between the absolute wing size phenotypes and only a few SNPs are candidates for all traits. (f) Nominally associated SNPs are most abundant in the intergenic space and in regulatory regions. Boxes show the distribution of negative  $\log_{10}$   $p$ -values of the SNPs nominally associated to rCS in females among site classes. Numbers of SNPs belonging to each site class are denoted above the boxes. As a SNP can fall into multiple classes, the sum of SNPs from all site classes is higher than the total number of nominally associated SNPs.

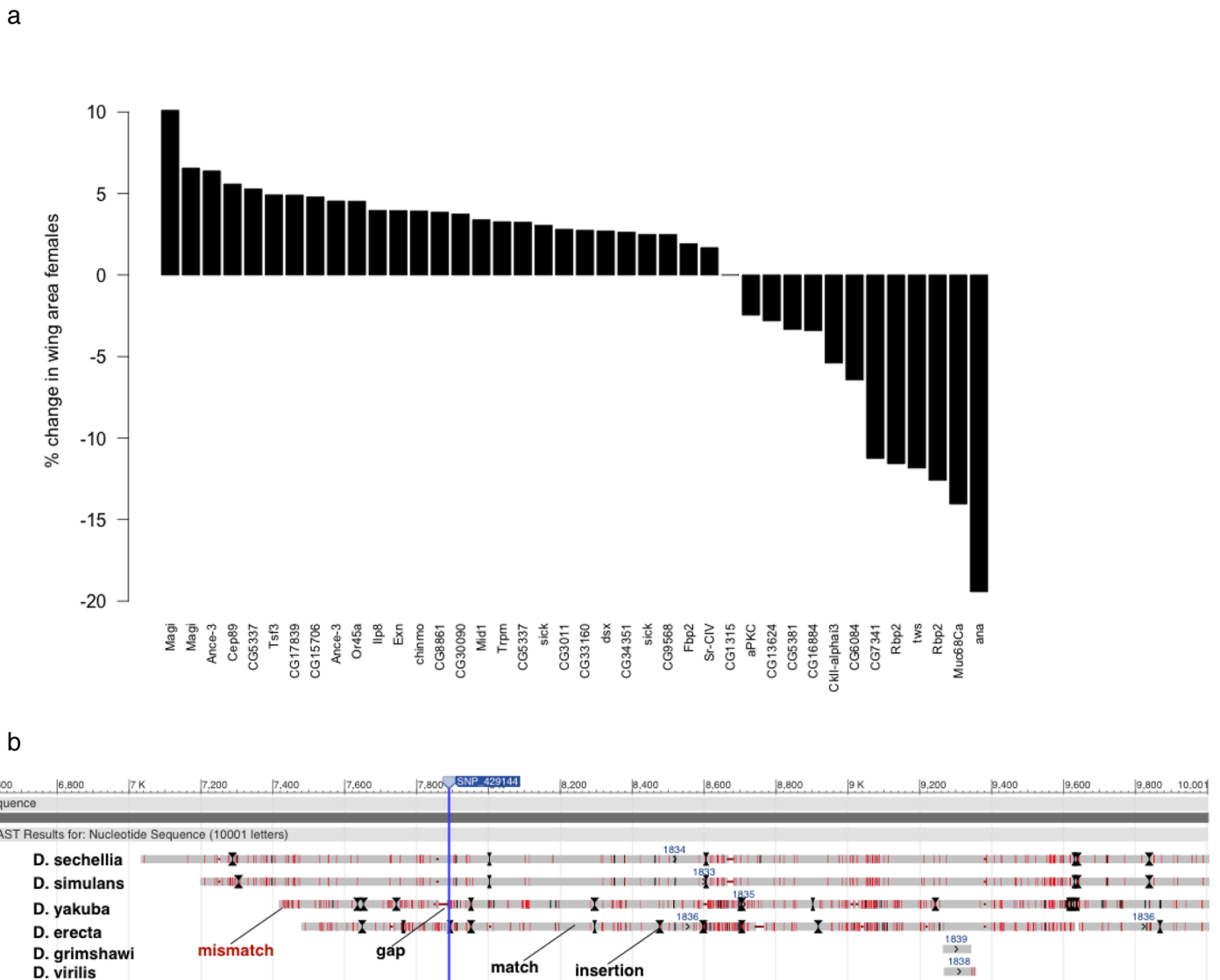
doi:10.1371/journal.pgen.1005616.g003

Nominally associated variants predominantly mapped to intergenic regions, but were nevertheless enriched in gene regions ( $p < 0.001$ , hypergeometric test) (Fig 3F, S5 Table), demonstrating that associations were not randomly distributed across the genome. For gene-level analyses we determined candidates for each phenotype as genes having a nominally significant SNP in or within 1kb of their transcribed region, yielding a total of 107 genes over all phenotypes. Only the candidate gene sets for rCS were enriched for STRING curated interactions and only the candidate list for CSF was enriched for functional categories (positive regulation of Rho signal transduction and melanotic encapsulation of foreign target), though growth was among the top categories for CSM<sub>IC</sub> (FDR corrected  $p = 0.08$ ) [56]. Given the large number of genes already known to play a role in growth control we were surprised that only few canonical growth genes contained or were close to nominally associated SNPs. Exceptions included several SNPs near or in the genes coding for *Ilp8*, TOR and EGFR pathway components and regulators of tissue polarity and patterning. However, some SNPs that narrowly missed the candidate reporting threshold localized to further growth regulatory genes, such as the Hippo pathway components *ex* and *wts*.

The small number of canonical growth pathway genes detected might be explained by the lack of SNPs with large effects in these genes, which is plausible considering the essential role of many growth regulators. We therefore wanted to test whether the combined signal of SNPs with small effects (each too small to reach significance on its own) across known growth genes might be significant. To this end we determined gene-based statistics using the sum of chi-squares VEGAS method [57], which computes a  $p$ -value for each gene considering all SNPs within a gene while correcting for gene length and linkage disequilibrium between SNPs. None of the genes reached genome-wide significance ( $p < 3.75 \times 10^{-06}$ ) (S6 Table). The overlap between the 20 top scoring genes from this analysis with our GWA candidate genes was small for each individual phenotype and even when combining the VEGAS analyses from all phenotypes only 11 of our 97 VEGAS top scoring genes contained a SNP that reached significance on its own in one of our GWA analyses. We did not find GO or interaction enrichment [56] and as in the individual GWA analyses, top candidates were largely novel with respect to growth control.

### Functional validation of candidate genes reveals novel regulators of size

We selected a subset (41% to 69%) of candidates identified by each of our six wing size GWAS (CS, CS<sub>IC</sub> and rCS in both sexes) for functional validation by tissue-specific RNAi. A total of 64% to 79% of tested genes had significant effects on wing area ( $p < 0.001$ , Wilcoxon rank sum test, S7 Table, Figs 4A and S9). We achieved similar validation rates for gene-based candidates. In contrast, only 42% of a set of 24 randomly selected genes had significant effects on wing size in females (S7 Table). The overall proportion of validated candidates versus random genes was significantly different ( $p = 0.02$ , Fisher's exact test) and Wilcoxon test  $p$ -values showed different distributions between candidate and random knockdowns ( $p = 0.02$ , Wilcoxon test, S10 Fig).



**Fig 4. Associated SNPs overlap 33 functionally diverse novel candidate genes for wing size determination and localize within putative enhancer elements.** (a) Validated genes in females. Bars show the percent change in median wing area compared to *CG1315* RNAi upon wing-specific knockdown of candidate genes. Only the lines yielding a significant wing size change ( $p < 0.001$ , Wilcoxon rank sum test) are depicted. (b) Alignment of the 2kb region on chromosome arm 2L upstream of the *D. melanogaster* *ex* locus that shows sequence conservation across *Drosophila* species. The position of the SNP is indicated by the vertical blue line. The *D. melanogaster* sequence is represented by the dark grey bar at the top ("Sequence"). The respective sequences of each compared species are represented below. Light grey regions are matches to the *D. melanogaster* sequence, red regions are mismatches, gaps in the alignment are denoted by horizontal red lines and insertions by black lines and arrows.

doi:10.1371/journal.pgen.1005616.g004

This combined evidence suggests an advantage in power for identifying growth regulators by GWA over randomly testing genes. The validated candidates constitute 33 functionally diverse novel growth regulators (S11 Fig).

Knockdown of genes in the whole eye disc often affects eye area, which in turn leads to a bigger or smaller head area between the eyes. As we used exactly this area for determining IOD, we were concerned that effects of knockdown on compound eye development could not be discerned from effects on IOD specifically. For this reason we chose not to perform validation for IOD candidates. However, since we observed similar coefficients of variation, effect

sizes, and q-values for IOD and CS (S1 and S3 Tables) we would expect the proportion of validated genes to be similar for IOD.

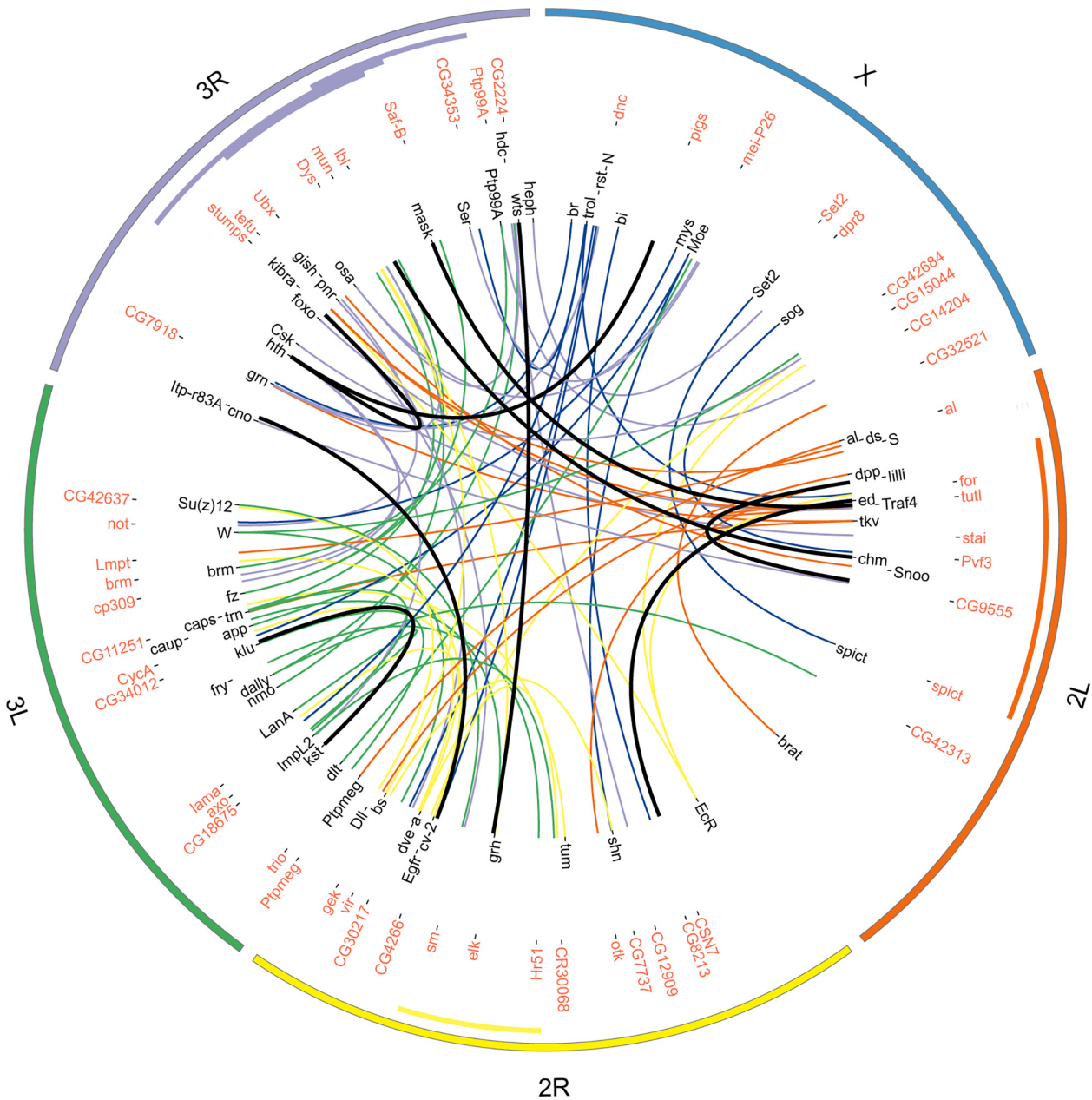
## Two-locus association reveals novel interactions

To place novel genes within the network of known growth pathways, we next performed tests for two-locus associations [58] to  $CS_{IC}$ ,  $IOD_{IC}$  and  $rCS$  in both sexes with SNPs in 306 growth genes as focal SNPs (S8 Table). This gene list was combined from genes listed as influencing wing development (The Interactive Fly, <http://www.sdbonline.org/sites/fly/aimorph/wing.htm>), commonly known growth genes from the IIS/TOR, EGFR and Hippo pathways, and growth regulators identified in screens by our group. This list is not comprehensive but should serve as a rough framework for the most relevant growth pathways. Overall, 15 interactions reached Bonferroni-corrected significance ( $p < 7.9 \times 10^{-13}$ ), but we observed none of our GWA candidates among the significant epistasis partners. Generally, more interactions reached genome wide significance in males than in females. The most significant interaction ( $CSM_{IC}$ ,  $p = 5.79 \times 10^{-15}$ ) occurred between *mask*, a positive regulator of JAK/STAT signaling [59] and *tutl*, a JAK/STAT target gene during optic lobe development [60] (Fig 5). Furthermore, among the top five interactions we found one between *nkd*, a downstream target of *Dpp* [60, 61], and the tyrosine phosphatase *Ptp99A* ( $CSF_{IC}$ ,  $p = 8.79 \times 10^{-14}$ ), which has been shown to interact with InR and the Ras signaling pathway [62, 63]. Furthermore, though we detected none of the significant interactions on DroID [64, 65] *mask* and *tutl*, and *Ptp99A* and *nkd* shared more DroID interactors than 95% of all possible pairs of the 32 genes involved in the significant interactions. Due to their already known growth-related functions we consider the interactions between these genes as prime candidates for future functional validation.

To investigate whether our GWA candidates or genes from the ‘previously known’ catalog would be enriched further down the list, we lowered the stringency for reporting interactions to a discovery threshold of  $p < 10^{-09}$ . Counting only those interactions where the interacting SNP lay in or within 1kb of a gene (S8 Table, total 1,353 interactors across all phenotypes) we found enrichment for development, morphogenesis and signaling categories (Bonferroni corrected  $p < 0.001$ ) [56], which supports a role of these genes in growth control. Notably, the  $rCSM$  list (Fig 5) was additionally enriched for genes involved in regulation of metabolic processes. Among them were 73 of the 306 genes in the ‘previously known’ catalog and 35 of our 107 overall GWAS candidates. However, these overlaps did not reach significance ( $p$ -value of 0.46 and 0.22, respectively). We next asked whether the candidate gene sets identified by normal GWAS and the epistasis approach were nevertheless biologically related to each other. To this end we used the STRING database [56], which revealed that the number of observed curated interactions between the two gene sets was much larger than expected by chance ( $p < < 0.001$ ). Analyzing pairwise interactions may thus help to place genes into pre-established networks.

## Intergenic SNPs are preferentially located in regions with enhancer signatures and overlap lincRNA loci

Intergenic SNPs may be functional by changing the sequence of more distant regulatory elements or noncoding RNAs. We therefore tested whether intergenic GWAS candidate SNPs located to putative functional regions. We found enrichment ( $p < 0.01$ , hypergeometric test) of SNPs lying in regions with H3K4Me1 or H3K27Ac, epigenetic signatures of active enhancers (S5 Table) [66], and in lincRNA loci [67], which have been implied in developmental regulation and are often enriched for trait-associated loci [68]. Though only loci associated with IOD



**Fig 5. Pairwise interactions between focal genes and DGRP SNPs for male wing size (rCSM).** The plot shows the focal genes annotated in black and the interactors in red. Interaction lines are colored according to the chromosome the focal gene is located on and the thick black lines denote Bonferroni-significant interactions. The outer circle marks the chromosome arms (2L = orange, 2R = yellow, 3L = green, 3R = purple, X = blue). The colored bars inside the inner circle mark the locations of cosmopolitan inversions (orange: *In(2L)t*; yellow: *In(2R)NS*; purple: *In(3R)K*, *In(3R)P*, *In(3R)Mo*).

doi:10.1371/journal.pgen.1005616.g005

in females were enriched for SNPs localizing to lincRNA loci, we found one SNP lying in a lincRNA among the top variants for rCSF and IODF<sub>IC</sub> (S5 Table).

A SNP 2kb upstream (position 2L: 429144) of the Hippo pathway regulator *ex* narrowly missed the reporting threshold ( $p = 1.7 \times 10^{-5}$ , CSM). However, its genomic location suggests this variant could affect a novel regulatory region for this gene. The region surrounding it was

annotated with the enhancer methylation signatures H3K4Me1 and H3K27Ac and had assigned state 4 of the 9 state chromatin model suggestive of a strong enhancer [66, 69]. Further annotations included H3K9Ac, a mark of transcriptional start sites, histone deacetylase binding sites and an origin of replication. To further assess functionality, we investigated whether the sequence around this SNP was conserved across taxa by performing multiple sequence alignment using BLAST [70] (S9 Table). Indeed, the region immediately upstream of the *D. melanogaster ex* gene showed high similarity to ~3kb regions slightly more upstream of *expanded* orthologs in the genomes of *D. sechellia*, *D. yakuba* and *D. erecta* (S9 Table, Fig 4B). This combined evidence suggests a functional region immediately upstream of the *D. melanogaster ex* gene, but additional experiments are required to corroborate functionality and to establish an involvement in growth control and a mechanism for influencing size.

### Human orthologs of candidate genes are associated with height, obesity and a variety of other traits

To investigate conservation to humans and further elucidate putative functions of candidate genes, we searched for orthologous proteins in humans. We found human orthologs [71] for 62 of our 107 GWA candidate genes, of which seven had a good confidence ortholog (score  $\geq 3$ ) associated with height, pubertal anthropometrics or growth defects (S10 Table). Of the 423 loci involved in human height mapped in a more recent meta analysis [10], five contained a gene that was orthologous to one of our GWAS candidates. Using all human-*Drosophila* ortholog relationships reported by DIOPT-DIST [71] as background, this results in an enrichment *p*-value of 0.001. However, given the large number of genes implicated in height in humans, which is likely to further increase with the number of individuals used for association and the small overlap of five genes with our loci, the reported link is tenuous and needs more support from better-powered studies. Nevertheless, the evidence for an involvement in growth control from GWAS in both organisms and experimental support from validation in *Drosophila* corroborates a biological function of these genes in the determination of body size.

### Discussion

We applied several GWA methods to developmental traits that have been extensively studied by single gene analyses in *Drosophila* as a complementary approach for identifying loci underlying size variation. Our single-marker GWAS revealed only one SNP cluster close to the known growth gene *kek1* to be significantly associated with body size when using a conservative Bonferroni correction. Yet, in contrast to human GWA analyses, which require independent replication, we exploited the fact that our model organism is amenable to direct validation strategies and tested candidates corresponding to a much lower significance threshold of  $10^{-05}$  for an involvement in size determination. Using tissue-specific RNAi, we validated 33 novel genes affecting *Drosophila* wing size. Nominally significant intergenic associations were preferentially located in regions with an enhancer signature and overlapped lincRNA loci. A SNP upstream of the *expanded* locus was in an evolutionarily conserved region, indicating the presence of a putatively functional element. A two-locus epistasis screen identified several genome-wide significant interactions between known growth genes and novel loci, showing that targeted epistasis analysis can be used to extend existing networks. Our study shows that despite limited statistical power, insights into the genetic basis of trait variation can be gained from analyzing nominal associations through functional and enrichment analyses and performing targeted locus-locus interaction studies.

## Single-marker and two-locus GWA of size

Our study adds 33 novel growth genes and 15 genomic loci that may interact with known growth genes to the extensive number of loci already implied in size regulation from single gene studies. That only a few *bona fide* growth genes were among the nominally significant candidates could be due to selection against functional variation in natural populations and/or during subsequent inbreeding and/or that the effect sizes of SNPs in these genes are too small to be detected in the DGRP. Though it has been shown for other phenotypes that the identified loci seldom overlap between mutational and GWA approaches, we expected a higher overlap for size as this trait has been exceptionally intensively studied in *Drosophila* and as a result we have extensive prior knowledge on the underlying genes. Our validation of a substantial number of novel genes underscores the complementarity of the GWA approach to classical genetics and highlights the importance of probing natural variants. However, future studies would benefit from utilizing bigger population sizes in order to improve statistical power and from investigating populations with different geographic origins, to address population specificity of associated variants.

With the exception of the *kek1* cluster, all the most significant wing size associations mapped to putative novel growth genes: *CG6091*, a de-ubiquitinating enzyme whose human ortholog has a role in innate immunity; *CG34370*, which was recently identified in a GWA analysis of lifespan and lifetime fecundity in *Drosophila* [72]; and, surprisingly, *dsx*, a gene well characterized for its involvement in sex determination, fecundity and courtship behavior. *dsx* showed a significant effect on wing size in both sexes, *CG6091* was only validated in males despite reaching a smaller *p*-value in the female GWAS, and *CG34370* did not show a significant wing size change in either sex. Due to the obvious limitations of RNAi as a validation approach for SNPs we think it important to investigate the roles of *CG6091* and *CG34370* in growth control by other approaches before discarding them as false associations. As genes affecting growth also impact on general and reproductive fitness of organisms it is not surprising that most of the candidate variants in or close to genes lie in regulatory regions, potentially modulating splicing, RNA turnover or RNA/protein abundance. Our data support the general notion that intergenic SNPs can impact phenotypes, either by affecting transcript abundance of protein coding genes (e.g. through distal enhancer elements) or via noncoding RNAs, which have been shown to regulate many biological functions including cellular processes underlying growth [73–75].

There are few nonsense and missense SNPs among our candidates (one and six, respectively); these variants are prime contenders for effects on protein function. However, confirming such effects requires testing the SNP in an isogenic background. Knockdown of most candidate genes resulted in a small change in median wing size (-19.4% to 10.1%), indicating a redundant or mildly growth enhancing or suppressing role in this tissue, which may explain why they were not discovered by classical mutagenesis screens. However, larger effects might be observed upon ubiquitous knockout, knockdown or overexpression.

Epistasis analysis revealed 15 loci showing Bonferroni-significant interactions with SNPs in previously known growth genes, demonstrating the usefulness of this approach for extending existing biological networks. In addition to the interactions described in the main text, we note there is an interesting interaction between *eIF2D* (*ligatin*) and *SNF4agamma* [76]. Furthermore, we found putative biological interactors for several GWA candidates among the top interactions that did not reach genome-wide significance e.g. *Lar* with *InR*. *Lar* can phosphorylate *InR* [62], so polymorphisms at these two loci could act synergistically to modulate *InR* activity. The enrichment of annotated interactions between our GWA candidates and epistasis partners shows that different analyses yielding different top associations uncover common

underlying genetic networks. A similar combinatorial approach has been successful in study using the DGRP [33], underscoring that combinatorial approaches can help placing candidates from different analyses into a joint biological network, and provide a basis for further hypothesis driven investigation of the roles and connectivity of novel and known genes.

### The *kek1* locus

The region upstream of *kek1* is the only locus where a larger genome region shows association. The minor allele frequency of this haplotype was between 4.7% and 6.2% (present in 7–10 lines). We explain the large effect of this locus by the fact that most lines with the minor alleles of the significant SNPs have an IOD that exceeds the 75<sup>th</sup> percentile of the IOD distribution in lines with the major allele. Obviously the effect may be overestimated due to relatively few lines having the minor allele, and due to effects of the genetic background. Nevertheless, the region contains some prime contenders for an effect on size: the SNPs lie in a region that could serve as a regulatory element for *kek1* or the poorly characterized CG9932, which has been implied in wing disc development in another study [55]. Furthermore, an uncharacterized noncoding RNA lies close by and could be linked to the SNP haplotype, due to generally higher LD in this region. Also, the region is spanned by *In(2L)t*, which itself shows association to size, and the observed effect could be due to this inversion. None of the lines with the minor allele haplotype had any copy of the inversion, which agrees with the observation that *In(2L)t* homozygous flies are relatively smaller than lines without the arrangement. On the other hand, only two lines in our dataset were homozygous for *In(2L)t* (DGRP\_350 and DGRP\_358), leaving many of the lines with the major allele with none or only one copy (nine lines) of the inversion, and we corrected for its presence, so we would not expect this to cause the association.

### Biological roles of novel growth genes

Apart from expected processes like signaling, transcription, translation and morphogenesis, we validated genes involved in transmembrane transport, planar cell polarity (PCP), metabolism and immunity. A total of 24 single marker or two way interaction candidate genes from our GWA analyses were discovered to be enhancers or suppressors of major growth pathways in another study [77] (S11 Table) and 15 were associated with nutritional variation in *Drosophila* [78], supporting their role in growth control. We did not see a large overlap between our candidates and the candidates identified by artificial selection on body size by Turner *et al.* [24], who identified many classical growth genes. We explain this discrepancy by the different methods to recover underlying genetic loci. Selection can enrich for rare alleles, while these cannot be probed in GWAS. If these also had large effects, they would be strongly differentially selected for in the Turner study. In contrast, we would expect large effect alleles of canonical growth genes to be mostly rare in the DGRP lines due to pleiotropic effects of these genes on fitness traits, and thus not evaluated in the GWA. Furthermore, Turner *et al.* used a population from California, which likely has a different allele composition from our North Carolina population. A minor factor may be the rather coarse and general size measure used by Turner *et al.* Sieving selected for generally bigger flies, with no distinction between flies with bigger wings, heads, thoraxes, or legs that obstructed passage through the sieve. As growth of individual body parts is controlled by both systemic and organ intrinsic factors [1–4], and measuring overall size likely identifies the systemic, general factors, this could explain some of the discrepancy. Likewise, we do not identify two previously reported large effect alleles of *InR* or *smp-30* [23, 24]. The *smp-30* allele was identified in a population of different geographic origin and could thus simply not be present in the DGRP lines. The *InR* allele is an indel, whereas we only analyzed single nucleotide polymorphisms in this study. Although both genes likely contain further

polymorphisms that are present in the DGRP, their effects may be far smaller and thus not detectable given the restricted size of the DGRP. In terms of chromosomal loci, we do identify SNPs on 3R, several of them located in the region spanned by *In(3R)P* (3R: 12,257,931–20,569,732) [54]. Furthermore, our data show a trend of *In(3R)P* correlating with size. However, the inversion was present in too few lines of our dataset to reliably estimate its effect in a model.

Most of the loci we validated have not been previously linked with growth in *Drosophila*. The yeast ortholog of *Mid1* (validated only in females but stronger association in males), is a stretch activated  $Ca^{2+}$  channel with a role in the polarized growth of mating projections [79, 80]). As mechanical tension plays a role during growth of imaginal discs, this channel could act in translating such signals to intracellular signaling pathways via the second messenger  $Ca^{2+}$ . The human ortholog of another candidate, the transmembrane channel *Trpm* (associated and validated only in females), showed association with anthropometric traits during puberty, indicating a role during the postnatal growth phase. The mucin *Muc68Ca*, identified in the top 20 of the gene-based association to rCSF, showed one of the largest knockdown effects (14 and 15% reduction in size in females and males, respectively). Mucins form a protective layer around vital organs, and the expression pattern of *Muc68Ca* in the larval midgut concurs with a putative effect on growth via the control of intestinal integrity [81].

A dual role in PCP, the establishment of cell polarity within a plane in an epithelium, and growth control has been shown for many genes, which regulate these two processes via distinct but coordinated downstream cascades [82, 83]. *Lar* (no significant effect on size in validation), *aPKC*, the Fz target Kermit [84] and the motor proteins Dhc64C (not included in validation but contained a nonsense SNP reaching nominal significance) and Khc-73 (validated in males but stronger association in females, though with positive effect size), whose human ortholog is significantly associated with height, are implied in PCP establishment. Kermit and motor proteins act downstream in the PCP cascade and likely have specialized roles for this process, but PCP can itself impact on growth, as proper establishment of polarity provides the orientation of cell division, and loss of a PCP component in zebrafish causes a reduction in body length [85]. Interestingly, *kermit* was a candidate interactor of EGFR, which acts in a combinatorial manner with Fz signaling in PCP [86], providing a biological basis for this interaction.

Metabolic genes are prime candidates for improving our understanding of growth, which depends on the amount of energy and precursors available for biosynthesis, and thus to metabolic coordination. The recent findings that the growth and PCP regulator Fat can couple growth and metabolism and mitochondrial proteins can causally affect growth pathway activity [87] underscore the importance of metabolic coordination. A missense SNP in the validated candidate *Cep89*, a gene involved in mitochondrial metabolism and growth in *Drosophila* and humans [88] was associated with most wing phenotypes. Elucidating the function of *Cep89* and other validated candidates with putative roles in metabolism, e.g. *CG3011*, *CG6084* and *Fbp2*, whose human ortholog has been linked to growth defects and cancer [e.g. 89], may provide further insight into this coordination.

## Sexual dimorphism of size

Of the top 100 SNPs for each trait only 25% - 43% are shared between the sexes, a surprisingly small overlap given the high genetic and phenotypic correlations between sexes. In some cases the SNPs still lie in the same gene, implying that this gene differentially affects size in both sexes, but the responsible SNP is different. In other cases, we only detect associated SNPs for a gene in one sex. Here, the gene may affect size in both sexes but genetic variation in this gene affects size differentially in only one sex. As we have a low powered study this is only a hint and



these results need to be further analyzed in a bigger population or by allele replacements using e.g. the CRISPR/Cas9 system to be corroborated. Unfortunately we cannot conclude anything about sex-specificity of our variants from the knockdown results. A knockdown is a very different perturbation from the effects of, for example, a regulatory variant. The knockdowns are performed in a different background than the one the association was discovered in, and they have much larger effect on the levels of a gene than a regulatory variant. So even though RNAi on a gene might show effects in males and females it does not exclude that different alleles of a SNP in this gene only affect wing size differentially in one sex.

Interestingly, an intronic SNP in the sex determination gene *dsx* had the lowest *p*-value in the female relative wing size GWAS but had a smaller effect size in males. *Dsx* is a transcription factor with sex-specific isoforms, and has many targets with sex- and tissue specific effects [90]. We also observed sex-specificity for the genome-wide significant two-locus interactions.

Considering that males use their wings to produce a courtship song that is instrumental for mating success, it may well be possible that selection pressure is different for male and female wing size or wing size in relation to other body parts. Indeed, the selection response for wing length seems to be more constrained in males [91]. In our dataset, wing length is highly correlated with CS, our main wing size measure. Menezes *et al.* observed males with more elongated wings but also smaller males had the highest mating successes [92]. These studies and our data suggest there may be subtle differences in the genetic networks underlying size determination in males and females in natural populations, a possibility that is neglected in single gene studies and thus would be worthwhile exploring.

## Conclusions and future perspectives

Growth control has been well studied, particularly in *Drosophila*, where many genes and pathways affecting growth have been documented by mutational analyses. However, such screens are far from saturation and do not scale well to investigating effects of combinations of mutations. Here we took advantage of naturally occurring, multifactorial perturbations genome-wide to identify novel genes affecting growth and to place them in genetic interaction networks. Rather than deepening our understanding of growth control, the identification of ever more growth regulators raises new questions about how all these loci interact to govern growth. The challenge for the future will be to shift our focus from studying genes in isolation towards investigating them in the context of developmental networks, and to assess the effects of network perturbations on intermediate molecular phenotypes of transcript, protein and metabolite levels.

## Materials and Methods

### *Drosophila* medium and strains

Fly food was prepared according to the following recipe: 100 g fresh yeast, 55 g cornmeal, 10 g wheat flour, 75 g sugar, 8 g bacto-agar and 1 liter tap water. Experiments were performed with 149 of the DGRP lines. RNAi lines used are listed in [S7 Table](#).

### Standardized culture conditions

Lines were set up in duplicate vials, with five males and five females per vial. After seven days, the parental flies were removed. From the F<sub>1</sub>, five males and five females were put together in duplicate vials and discarded after seven days of egg laying. From the F<sub>2</sub>, thirty males and thirty females were mated in a laying cage with an apple juice agar plate plus a yeast drop as food source and allowed to acclimatize for 24 hours. A fresh plate of apple juice agar plus yeast drop

was then applied and flies were left to lay eggs for another 24 hours. From this plate, F<sub>3</sub> L1 larvae were picked with forceps and distributed into three replicate vials, with 40 larvae per vial. The food surface in the vials was scratched and 100 μl of ddH<sub>2</sub>O added prior to larvae transfer. The adult F<sub>3</sub> flies were pooled from the three vials and frozen at -20°C approximately 1–2 days after eclosion. The whole experiment was performed in a dedicated incubator (DR-36VL, CLF Plant Climatics GmbH) with a 12-hour day-night cycle, constant humidity of 65–68% and constant temperature of 25.5°C +/- 1°C. Vials were shuffled every two days during the first and second round of mating but left at a fixed position in the incubator for the duration of the development of the F<sub>3</sub> generation.

For the parental generation, lines were all set up on the same day on the same food batch. For the F<sub>1</sub> matings, different food batches had to be used due to different developmental timing of the lines. F<sub>2</sub> matings were set up using the same batch of apple agar plates and yeast for all lines. F<sub>3</sub> larvae were distributed on four different food batches and the batch number was recorded for each line.

The control experiment (S1 Table) was performed using the same procedure as above, except that the same food batch was used for all flies of a generation. We used the DGRP lines DGRP\_303, DGRP\_732, DGRP\_721 and DGRP\_908 for this experiment because they had comparable generation times and set up ten replicates of each of these lines according to the standardized culture conditions.

## Phenotyping and morphometric measurements

Depending on the number of flies available, between five and twenty-five flies per sex and line were measured for the dataset (median 25 flies per sex and line, mean 23 (CS<sub>females</sub>, CS<sub>males</sub>, IOD<sub>males</sub>) and 24 for IOD<sub>females</sub>; exact numbers are given in S2 Table). For the experimental generation we distributed a total of 19,200 larvae in four batches spaced throughout 1.5 weeks according to developmental timing, and the final dataset consisted of morphometric data of 6,978 flies, 3,500 females and 3,478 males. For the control experiment we phenotyped 25 flies per replicate, sex and line, resulting in a total of 2,000 flies (1,000 males, 1,000 females). Flies were positioned on a black apple agar plate and photographed using a VHX-1000 digital light microscope (KEYENCE). Morphometric body traits were measured manually using the VHX-1000 dedicated measurement software. If intact the right and otherwise the left wing was removed and mounted in water on a glass slide for wing image acquisition. Morphometric measurements were extracted from the wing images using WINGMACHINE [93] and MATLAB (MATLAB version R2010b Natick, Massachusetts: The MathWorks Inc.)

Centroid size was measured as the square root of the summed squared distances of 14 landmarks from the center of the wing (Fig 1). Interocular distance was measured from eye edge to eye edge along the anterior edge of the posterior ocelli and parallel to the base of the head.

## Quantitative genetic analysis

All analyses were performed in R Studio using the R statistical language version 2.15 (<http://www.R-project.org>). PCA was performed on data of individual flies using the package *FactoMineR*. Allometric coefficients (b) were determined for each line and sex from the model  $\log(y) = \log(a) + b * \log(x)$ , where  $y = CS$  and  $x = IOD$  or  $TL$ , using the *lm()* function in the *stats* package. 95% confidence intervals for the parameter b were computed using the *confint()* function in the *stats* package. The total phenotypic variance in the control experiment was partitioned using the mixed model  $Y = S + L + SxL + R(L) + \epsilon$ , where S is the fixed effect of sex, L is the random effect of line (genotype), SxL is the random effect of line by sex interaction, R is the random effect of replicate and  $\epsilon$  is the within line variance. The brackets represent that replicate is

nested within line. The total phenotypic variance in the dataset was partitioned using the mixed model  $Y = S + L(F) + SxL(F) + F + \varepsilon$ , where  $S$  is the fixed effect of sex,  $L$  is the random effect of line (genotype),  $SxL$  is the random effect of line by sex interaction,  $F$  is the random effect of food batch and  $\varepsilon$  is the within line variance. The random effects of line and line by sex are nested within food batch, as each line was raised only on one of the four food batches. Models of this form were fitted using the *lmer()* function in the *lme4* package in R. We also ran reduced models separately for males and females. The *rand()* function in the *lmerTest* package was used to assess significance of the random effects terms in the dataset.

Relative contributions of the variance components to total phenotypic variance ( $\sigma^2_P$ ) was calculated as  $\sigma^2_i / \sigma^2_P$  where  $\sigma^2_i$  represents any of  $\sigma^2_L$ ,  $\sigma^2_{LxS}$ ,  $\sigma^2_F$ ,  $\sigma^2_R$ ,  $\sigma^2_E$ , and  $\sigma^2_P = \sigma^2_L + \sigma^2_{LxS} + \sigma^2_C + \sigma^2_E$ .  $\sigma^2_C$  stands for  $\sigma^2_R$  in the control dataset and for  $\sigma^2_F$  in the analysis of the GWAS dataset.  $\sigma^2_L$  = variance due to genotype,  $\sigma^2_{LxS}$  = variance due to genotype by sex interactions,  $\sigma^2_F$  = variance due to food,  $\sigma^2_R$  = variance due to replicate and  $\sigma^2_E$  = residual (intra-line) variance. The broad sense heritability for each trait was estimated as

$H^2 = \sigma^2_G / \sigma^2_P = (\sigma^2_L + \sigma^2_{LxS}) / (\sigma^2_L + \sigma^2_{LxS} + \sigma^2_C + \sigma^2_E)$ . The cross-sex genetic correlation was calculated as  $r_{MF} = \sigma^2_L / (\sigma_{LF} \sigma_{LM})$  where  $\sigma^2_L$  is the variance among lines from the analysis pooled across sexes, and,  $\sigma_{LF}$  and  $\sigma_{LM}$  are, respectively, the square roots of the among line variance from the reduced models of females and males. Similarly, cross-trait genetic correlations were calculated as  $r_{AB} = \sigma^2_{G(AB)} / (\sigma_{GA} \sigma_{GB})$  where  $\sigma^2_{G(AB)}$  is the genetic covariance between traits A and B, and  $\sigma_{GA}$  and  $\sigma_{GB}$  are the square roots of the genetic variance for traits A and B, respectively. The phenotypic correlation between sexes was determined using the *cor()* function with method = "spearman" in R.

## Phenotypes for GWAS

We found a large effect of food batch on CS, and inversions *In(2L)t* and *In(3R)Mo* were associated with IOD and to a lesser extent CS. We modeled these covariates using a mixed model. The food batch was modeled by a random effect and the rearrangements were coded as (0,1,2) depending on whether both, one or no inversion was present in the homozygous state. We did not observe correlation between *Wolbachia* infection status and any trait and thus did not include this as a covariate in the model. Specifically, the models used were:  $CS_{raw} = \alpha + X_1\beta_1 + X_2\beta_2 + Fu + \varepsilon$ , where  $X_1$  refers to the sex covariate,  $X_2$  refers to the inversion covariate,  $\varepsilon \sim N_n(0, \sigma_\varepsilon^2 I_n^2)$  with  $n$  being the number of lines,  $u \sim N_k(0, \sigma_u^2 I_k)$  with  $k$  being the number of food batches and  $Fu$  an  $(n,k)$ -indicator matrix, associating each line to its respective food batch. The GWA analyses were performed using the estimated residual of this model ( $CS = \varepsilon$ ).

To find loci that specifically affected variation in wing size unrelated to the overall body size variation we constructed an additional phenotype (rCS) that had the effect of IOD on CS removed via regression:  $CS_{raw} = \alpha + IOD + X_1\beta_1 + Fu + \varepsilon$ , where  $X_1$  and  $Fu$  refer again to the sex-effect and the food batch effect. We did not model the inversions because the residuals of this model were not correlated with the inversions. The residuals  $\varepsilon$  from this regression were used as relative size phenotypes. All phenotypes were rank normalized before GWAS.

## Association analyses

We performed GWA analyses using male and female line means. Genotypes for 143 of the 149 lines were obtained from the DGRP Freeze 2 website (<http://dgrp2.gnets.ncsu.edu>). Only SNPs that were missing in a maximum of ten lines and occurred in at least ten lines (7% of the measured lines, 1,319,937 SNPs in total) were used. GWA was performed using FaST-LMM [53] for separate sexes. This association method uses a linear mixed model to capture confounders such as population structure and cryptic relatedness. Association results were visualized using

the *manhattan()* function in the R package *qqman* [94]. To determine correlation between SNPs for a given phenotype we extracted the genotype of the top  $n$  SNPs ( $p < 10^{-05}$ ) and calculated the correlation between genotypes at these loci across all DGRP lines used in the GWA analyses. We used the FaST-LMM SNP  $p$ -values to apply the sum of chi-squares VEGAS method [57] to calculate gene wise statistics. Gene boundaries were defined using annotation from popDrowser (<http://popdrowser.uab.cat/gb2/gbrowse/dgrp/>), but we included also SNPs lying within 1,000 bp up- or downstream of these margins. The correlation matrix was calculated from the genotypes themselves.

## GO annotation and interaction enrichment

To determine enrichment of functional classes, annotate genes with functions and curated interactions among our candidate genes, we used the functional annotation and protein interaction enrichment tools from STRING [56].

## RNAi validation

SNPs with an association  $p$ -value  $< 10^{-05}$  lying in a gene region or  $\pm 1$  kb from a gene were mapped to that gene. From the gene based VEGAS analysis, we chose the top 20 genes from each list as candidates. RNAi lines for a subset of candidate genes for each wing phenotype were ordered from VDRC [95]. For one gene, *chinmo*, there was no appropriate line available from VDRC and we instead tested two Bloomington lines (26777 (y[1] v[1]; P{y[+t7.7] v[+t1.8]} = TRiP.JF02341}attP2) and 33638 (y[1] v[1]; P{y[+t7.7] v[+t1.8]} = TRiP.HMS00036}attP2/TM3, Sb[1]), indicated in S7 Table with (BL). For the random control knockdowns we tested a set of 24 genes that did not contain a significant SNP in or within 1 kb of their transcribed region. We did the random knockdowns only in females to more effectively assess more genes for the same labor. As we wanted to address the controls like an additional phenotype (random) we chose a number of genes comparable to the numbers of candidates for other phenotypes. We chose females because we generally had more candidate genes in females than in males. All RNAi lines used are listed in S7 Table. For wing size candidates, validation was performed by crossing males of the respective RNAi line to virgin females carrying the *GAL4* transcriptional activator under the control of the *nubbin* (*nub*) promoter. The VDRC line containing a *UAS*-RNAi construct against the *CG1315* (GD library, transformant ID 47097) gene served as a negative control for the knockdowns. We decided to use this line as reference because it was in the same background as most of our tester lines, an essential factor to consider when assessing genes that presumably only have a small effect on size upon knockdown. The *CG1315* knockdown had never shown an effect in any setting and it allowed us to evaluate unspecific effects of RNAi knockdown on wing size. Prior to the experiment, driver lines were bred under controlled density to eliminate cross-generational effects of crowding on size. Wings were phenotyped as described above and wing area used as a phenotypic readout. Change in median wing area relative to the control was tested with a Wilcoxon rank sum test (function *wilcoxon.test()* in R) for each line and for separate sexes. If possible, 25 flies per cross and sex were phenotyped for statistical analysis, however sometimes the number of progeny was lower. The number of phenotyped flies per cross and sex is given in S7 Table. We used the *fisher.test()* function in R to determine if the proportion of validated genes was different among candidates and random lines, and the *wilcoxon.test()* function to test for a difference in median  $p$ -value between candidates and random lines. The comparison between candidate and random knockdowns was done for females exclusively as only this sex was measured for the random lines. Only genes not previously implied in wing development or growth control were included

in the analysis, which excluded *chinmo*, *aPKC*, *tw5* and *Ilp8* from the candidates and *EloA* and *spz5* from the random list.

### Epistatic analyses

We explored epistatic interactions between SNPs lying within and 1 kb around genes that were previously found to be involved in growth or wing development in *Drosophila* against all DGRP SNPs with missingness <11 and present in at least 10% of the lines. We compiled a list of SNPs within and 1 kb up- or down-stream of genes that were previously known to play a role in growth control (14,137 SNPs) or wing development (43,498 SNPs) and used these as focal SNPs (S8 Table). All phenotypes were normalized to follow a standard normal distribution for this analysis to make sure that no severely non-normal distributions occurred within any of the four marker classes per locus. We used FasT-Epistasis [58] calculating interactions for all pairs between the focal SNPs and the set of all SNPs satisfying the above criteria (1,100,811 SNPs). Bonferroni corrected significance would thus require  $p < 7.9 \times 10^{-13}$ . Interactions were visualized using Circos [96]. To calculate significance for the overlap between genes found via epistasis and a given gene list, we first positionally indexed all  $n$  SNPs that were used in the epistasis analysis. We recorded the set of indices of SNPs with  $p < 10^{-09}$  yielding set  $K$ :  $K = \{k: \text{SNP}_k \text{ is an epistasis hit}\}$ . We then generated random samples.

For random sample  $j$ , do:

For all elements in  $K$ , add a random integer  $r_j$  between 0 and  $n-1$ . Define new index as the modulo  $n$ :  $k_i^j = \text{mod}(k_i + r_j, n)$ , which yields  $K^j = \{k_i^j; j = 1, \dots, m\}$ . Given the shifted positions  $K^j$ , we look up the SNP positions  $P_{K^j}$ . For a given gene list, we record the number  $x_j$  of gene regions that overlap a position in  $P_{K^j}$ . Let  $x$  be number of gene regions overlapping an epistasis hit. Our  $p$ -value estimate is then  $P_{\text{approx}} \approx 1/m \sum 1_{\{x_j \geq x\}}$ .

### Intergenic element enrichment analysis

We determined the number of SNPs from each GWA candidate list and the overall number of SNPs that located within modENCODE [66] elements annotated with Histone 3 lysine 4 mono-methylation (H3K4Me1) or Histone 3 lysine 27 acetylation (H3K27Ac) or lincRNA loci. For the H3K4Me1/H3K27Ac enrichments we restricted ourselves to three developmental stages (L2, L3, pupae), which we considered to be the most relevant interval for gene activity affecting growth of imaginal discs. We obtained a table with lincRNAs in the *Drosophila* genome from the study of Young *et al.* [67] and searched for enrichment of SNPs located in those lincRNA loci. Enrichment was tested using a hypergeometric test (function *phyper()* in R).

### BLAST alignment

We downloaded the sequence of the region 10 kb upstream of the annotated transcription start site of the *expanded* locus (2L: 421227..431227) from FlyBase [97], as well as the sequence of the same relative region for seven of the twelve *Drosophila* species [98], which contained the ortholog of the *expanded* gene in the same orientation in the genome. We performed multiple sequence alignment using the discontinuous megablast option on NCBI BLAST [70].

### Annotation with human orthologs

We combined candidate genes from GWA analyses of all phenotypes and searched for orthologs in humans using DIOPT-DIST [71]. Enrichment of GWA candidates for genes with human orthologs associated with height [10] was determined with a hypergeometric test (function *phyper()* in R). We determined *Drosophila* orthologs of gene annotations of all associated

SNPs in Wood *et al.* (total 697), resulting in 374 ortholog pairs supported by at least 3 prediction tools, and searched for overlap of these orthologs with the 62 of our GWAS candidate genes that had a human ortholog supported by at least 3 prediction tools, which resulted in 12 matches. Of those, only five matches were supported by three or more prediction tools (score  $\geq 3$ ) and we used only those for enrichment calculation. As background we used the total number of *Drosophila*-Human ortholog relationships (= 28,605) [71].

## Supporting Information

**S1 Fig. Analysis of the male dataset.** a) Genetic correlation between morphometric traits in males. The two modules of higher correlation observed in females are still visible (bright yellow in the upper left and lower right corners) but the overall clustering is more influenced by the more inaccurately measured smaller veins and areas. b) Cumulative variance explained in male data by increasing number of principal components. As in the female dataset, the first two PCs explain almost 75% of the variance in the data. c) Factor map for the variables. PCs 1 and 2 split the data into two groups. d) Correlation between PCs and traits. PC1 reflects a general size component and PC2 is highly correlated with head/thorax traits, effectively splitting the data in two groups. (PDF)

**S2 Fig. Allometry and inversions.** Histograms of the estimates for the allometric coefficient  $b$  for the relationship between CS and IOD in females (a), in males (b) and between CS and TL in females (c) and males (d). e) Boxplot and individual datapoints of the data in a-d. Red = females and black = males. 95% confidence intervals for  $b$  (S1 Table) are very broad for some lines due to few datapoints used for fitting, so these are just very rough estimates for the allometric relationship. Nevertheless there is variation among lines for all evaluated relationships. f) The effect of cosmopolitan inversions on wing size. Lines are plotted according to the number of homozygous inversion arrangements they have: 0 (red) = neither *In(2L)t* nor *In(3R)Mo* present, 1 (green) = homozygous for either *In(2L)t* or *In(3R)Mo*, 2 (blue) = homozygous for both *In(2L)t* and *In(3R)Mo*. Datapoints are individual flies. (PDF)

**S3 Fig. The minor and major haplotype of genome-wide significant SNPs show differential association with female IOD.** a) The minor allele haplotype of the genome-wide significant cluster is associated with an increased IOD in females. Boxplots of female IOD by genotype at the nine SNPs annotated to *kek1*. SNPs marked by a star pass Bonferroni correction. Grey = major allele, white = minor allele. b) Lines with the minor haplotype are distributed across all four foodbatches. Black dots = major allele, blue dots = minor allele. The IOD distribution for each foodbatch is plotted for females for the most significant SNP. The distribution is the same for all other SNPs of the cluster as all minor alleles form a haplotype. c) Correlation between  $p$ -values from GWAS with normalized IOD ( $y$ -axis) and non-normalized iod ( $x$ -axis) in females. Axes are on the  $-\log_{10}$  scale. d) Several blocks of higher LD are visible in the region 20kb upstream of *kek1*. Blue = no correlation, orange = complete correlation. (PDF)

**S4 Fig. QQ-plots from GWA in females for all traits show a departure from uniformity of top associations.** Observed association  $p$ -values are  $-\log_{10}$  transformed ( $y$ -axis) and plotted against the  $-\log_{10}$  transformed theoretically expected  $p$ -values under the assumption of no association (uniform distribution,  $x$ -axis). Centroid size (a), inversion corrected centroid size (b), interocular distance (c), inversion corrected interocular distance (d) and relative centroid size (e). (PDF)

**S5 Fig. QQ-plots from GWAS in males for all traits show a departure from uniformity of top associations.** Observed association  $p$ -values are  $-\log_{10}$  transformed (y-axis) and plotted against the  $-\log_{10}$  transformed theoretically expected  $p$ -values under the assumption of no association (uniform distribution, x-axis). Centroid size (a), inversion corrected centroid size (b), interocular distance (c), inversion corrected interocular distance (d) and relative centroid size (e).

(PDF)

**S6 Fig. Correlation between associated ( $p < 10^{-05}$ ) SNPs in females.** The SNPs are ordered according to chromosome arm (2L, 2R, 3L, 3R, X) and black dividers separate chromosomes. Within one chromosome arm SNPs are ordered according to their position on that chromosome with each tile representing one SNP. The color code is depicted on the right: orange = complete correlation (1) and blue = no correlation (0). Centroid size (a), inversion corrected centroid size (b), interocular distance (c), inversion corrected interocular distance (d) and relative centroid size (e).

(PDF)

**S7 Fig. Correlation between associated ( $p < 10^{-05}$ ) SNPs in males.** The SNPs are ordered according to chromosome arm (2L, 2R, 3L, 3R, X) and black dividers separate chromosomes. Within one chromosome arm SNPs are ordered according to their position on that chromosome with each tile representing one SNP. The color code is depicted on the right: orange = complete correlation (1) and blue = no correlation (0). Centroid size (a), inversion corrected centroid size (b), interocular distance (c), inversion corrected interocular distance (d) and relative centroid size (e).

(PDF)

**S8 Fig. Correlation of SNP  $p$ -values between the sexes.** SNP  $p$ -values in females (x-axis) are plotted against their respective  $p$ -values in males (y-axis). The Spearman rank correlation is given for each trait and the red lines denote the significance cutoff. a = CS, b =  $CS_{IC}$ , c = IOD, d =  $IOD_{IC}$ , e = rCS.

(PDF)

**S9 Fig. RNAi knockdown results males.** Percent change in median wing area compared to CG1315 RNAi upon wing-specific knockdown of the validated candidate genes in males. Only the lines yielding a significant wing size change ( $p < 0.001$ , Wilcoxon rank sum test) are depicted. Median, 25<sup>th</sup> and 75<sup>th</sup> percentile for each are given in [S7 Table](#).

(PDF)

**S10 Fig. Comparison of  $p$ -values and effect sizes between candidate and control RNAi.** a:  $-\log_{10}$  transformed  $p$ -value densities of the candidate (black) and combined control (red) data sets. The two  $p$ -value distributions differ by a location shift that is not zero (i.e. are not the same); specifically, the  $-\log_{10}$  transformed control  $p$ -value distribution (red) is shifted towards the left of the  $-\log_{10}$  transformed candidate  $p$ -value distribution (black) (one sided Wilcoxon rank sum test  $p = 0.02$ ). b: The distribution of candidate effect sizes (percent change in wing size upon knockdown) is shifted towards positive effect sizes (white boxes), whereas the control knockdown effect size distribution (red) is more centered on 0. The two exceptions at -28% (CG17646) and -42% (CG3704) are lines whose wings not only show a size reduction but also considerable morphological defects (c).  $N = 43$  candidates (white),  $N = 22$  control (red); only data from females was used for these analyses.

(PDF)

**S11 Fig. Functional annotation of the 33 validated candidate genes based on DAVID GO annotation.**

(PDF)

**S1 Table. Quantitative genetic analysis.** Control experiment: Only 2% of total population variance in CS and IOD was due to flies coming from replicate vials, a negligible fraction compared to the 78% and 71% attributable to differences in genotype. This indicates that the standardized culture protocol sufficiently deals with confounding effects on size phenotypes.  $N = 2000$ , 25 flies/sex of 10 replicates of four DGRP lines. Allometry: Allometric coefficients ( $b$ ) calculated from the equation  $\log(\text{CS}) = \log(a) + b^* \log(\text{trait})$  and their 95% confidence interval (CI) are given for the allometric relationships between CS and IOD and CS and TL for each line and sex. Though the CI varies substantially for some lines due to few data points used for fitting the models, the upper CI boundary is close to 0 for some (e.g. 28157 females (RAL228) for the CS-IOD relationship). QGA Dataset: Quantitative genetic analysis of CS and IOD in the dataset consisting of  $N = 6978$  flies from 149 DGRP lines.  $p_{\text{Sex}}$  = significance of fixed effect of sex,  $p_{\text{Line}}$  = significance of random effect of Line,  $p_{\text{Sex} \times \text{Line}}$  = significance of random effect of Line by sex interaction,  $p_{\text{Replicate}}$  = significance of random effect of replicate,  $p_{\text{Food}}$  = significance of random effect of foodbatch. Estimated parameters are variance due to genotype ( $V_G$ ), genotype by sex interaction ( $V_{G \times S}$ ), food ( $V_F$ ), replicate ( $V_R$ ) and intra-line variance ( $V_E$ ), as well as the cross-sex genetic ( $r_{MF}$ ) and phenotypic correlation between sexes.  $H^2$  = broad-sense heritability and  $V_P$  = total phenotypic variance. Phenotypic variation: Smallest and largest trait values and the percent difference are given per sex for each phenotype. Population means (Mean) and standard deviations (STD) are used to calculate the coefficient of variation (CV). (XLSX)

**S2 Table. Phenotypic data.** Raw data for all traits and line means (Mean), standard deviations (STD) and number of phenotyped flies (N) listed by sex for centroid size (CS) and interocular distance (IOD).

(XLS)

**S3 Table. GWAS results.** Nominally significant SNPs ( $p < 10^{-05}$ ) from GWAS for Centroid size (CS), inversion modeled centroid size ( $CS_{IC}$ ), interocular distance (IOD), inversion modeled interocular distance ( $IOD_{IC}$ ) and relative centroid size (rCS) in both sexes. (XLS)

**S4 Table. Between-sex and -phenotype overlap of nominally significant SNPs.** GWAS SNPs: Number of nominally significant SNPs ( $p < 10^{-05}$ ) identified in each of the GWAS and percent overlap between sexes. SNPs F = number of SNPs nominally significant in females, SNPs M = number of SNPs nominally significant in males, Common MF = number of SNPs nominally significant in both sexes, % M in F = percent nominally significant male SNPs also nominally significant in females, % F in M = percent nominally significant female SNPs also nominally significant in males. Between sex overlap: SNP and gene level overlap between sexes for different thresholds. Percent overlap of SNPs (exact associated positions), annotated SNPs (including annotation of SNPs located in intergenic regions where the annotation is often not reliable, as the next genes are frequently more than 20kb away), and genes (only includes SNPs that locate within or 1kb around a gene). A consistent difference in associated loci in terms of the exact SNP that is associated (top block) is detectable between sexes, though the loci apparently are located more or less in a similar region (middle block). For SNPs in or close to genes, the differences between sexes are more pronounced: of the top 100 associated genes about 50% overlap, while the other 50% are private to one sex but this percentage increases with inclusion of more genes (bottom block, top 10,000 is >90% overlap, in total there are around 14,000



genes in these lists). Between phenotype overlap: Proportion of nominally significant SNPs that are shared between phenotypes.

(XLS)

**S5 Table. Enrichment analysis.** Enrichment: Enrichment  $p$ -values for SNPs localizing to different genomic regions. LincRNA: LincRNA loci that overlap significant SNPs and their expression during different developmental stages. LincRNA data from Young *et al.* [67].

(XLS)

**S6 Table. Top 20 genes for each trait identified by the VEGAS method.** Last column indicates the percent change in median wing size of genes tested by RNAi, and asterisks (\*\*\*) indicate significant change ( $p < 0.001$ ).

(XLS)

**S7 Table. Validation results candidates and random genes.** The lines are ordered according to decreasing significance of the change in median wing area upon knockdown (Wilcoxon rank sum test). Crosses in columns pigmentation, bristles, veins indicate a slightly abnormal corresponding phenotype. N = number of individuals tested, Median = median wing area, Q25 and Q75 = first and third quartile of wing area distribution. MAF 3 or MAF 5 in brackets means a SNP in or near this gene was found among the top associated genes in a GWAS for wing size with a lower MAF cut-off (SNPs present in min. 3% or 5% of lines). This gene was tested for wing size since a SNP in it showed association to body size with the used MAF cutoff and thus a corresponding RNAi line was available. Overview: Overview validation results. Number of SNPs that were tested and the number and percentage that were validated for each trait. We tested the known genes *chinmo*, *aPKC*, *tws* and *Ilp8* as positive controls, but did not include them in the calculation of these percentages. Fisher's exact test: We performed a two-sided Fisher's exact test to determine if the proportions of validated genes was different between candidates and random genes. The results are shown for different Wilcoxon test  $p$ -value validation thresholds. Only not previously known candidates and random lines were used ( $N_{\text{candidates}} = 43$ ,  $N_{\text{random}} = 22$ )

(XLS)

**S8 Table. Epistasis results.** Focal genes: Genes previously implied in growth regulation or wing development that were used as focal genes for epistasis. Top interactions ( $p < 10^{-09}$ ) for female absolute inversion corrected wing size (CSF<sub>IC</sub>), male absolute inversion corrected wing size (CSM<sub>IC</sub>), female absolute inversion corrected body size (IODF<sub>IC</sub>), male absolute inversion corrected body size (IODM<sub>IC</sub>), female relative wing size (rCSF) and male relative wing size (rCSM). X is the interactor locus and Y the focal (= previously known) locus.

(XLS)

**S9 Table. Multiple sequence alignment (MSA) results.** Expanded orthologs: Genomic location of *ex* orthologs in 12 *Drosophila* species [98]. Name of the ortholog, its genomic location and orientation are shown. For the MSA we only used species that contained the gene in the same orientation as *D. melanogaster* (+). MSA details: Details of MSA of the 10kb region upstream of *D. melanogaster ex* gene. Sequence = genomic region in each species used in the MSA, Identity = percent identical nucleotides, Aligned Length = length of alignment, Query Cover = percent of the query sequence (*D. melanogaster* sequence) aligned to the sequence in the respective species, E-value = significance of alignment (expected number of high scoring pairs with score at least as high as the score of the current alignment), Score = strength of alignment.

(XLS)

**S10 Table. Human orthologs of putative and validated *Drosophila* growth regulators and their association to human complex traits.**

(XLS)

**S11 Table. GWAS/epistasis candidates reported by other studies.** Candidates found as suppressors or enhancers of major growth pathways by Schertel *et al.* [77]. Candidates associated with nutritional indices in the study of Unckless *et al.* [78].

(XLS)

## Acknowledgments

We thank Anna Troller, Benjamin Schlager and Anni Strässle for manual support during experiments. Stocks obtained from the Bloomington *Drosophila* Stock Center (NIH P40OD018537) and from the Vienna *Drosophila* RNAi Center were used in this study.

## Author Contributions

Conceived and designed the experiments: EH SCV. Performed the experiments: SCV. Analyzed the data: SCV DL. Contributed reagents/materials/analysis tools: TFCM DL SB. Wrote the paper: SCV. Edited the manuscript: SCV EH TFCM SB. Conceived the approach: EH TFCM SB.

## References

1. Oldham S, Bohni R, Stocker H, Brogiolo W, Hafen E. Genetic control of size in *Drosophila*. *Phil. Trans. R. Soc. B: Biological Sciences* 2000; 355: 945–952. PMID: [11128988](#)
2. Johnston LA, Gallant P. Control of growth and organ size in *Drosophila*. *Bioessays* 2002; 24: 54–64. PMID: [11782950](#)
3. Mirth CK, Riddiford LM. Size assessment and growth control: how adult size is determined in insects. *Bioessays* 2007; 29: 344–355. PMID: [17373657](#)
4. Shingleton AW. The regulation of organ size in *Drosophila*: physiology, plasticity, patterning and physical force. *Organogenesis* 2010; 6: 76–87. PMID: [20885854](#)
5. Oldham S, Hafen E. Insulin/IGF and target of rapamycin signaling: a TOR de force in growth control. *Trends Cell Biol.* 2003; 13: 79–85. PMID: [12559758](#)
6. Pan D. Hippo signaling in organ size control. *Genes Dev.* 2007; 21: 886–897. PMID: [17437995](#)
7. Tumaneng K, Russell RC, Guan KL. Organ size control by Hippo and TOR pathways. *Curr. Biol.* 2012; 22: R368–R379. doi: [10.1016/j.cub.2012.03.003](#) PMID: [22575479](#)
8. Gockel J, Robinson SJW, Kennington WJ, Goldstein DB, Partridge L. Quantitative genetic analysis of natural variation in body size in *Drosophila melanogaster*. *Heredity* 2002; 89: 145–153. PMID: [12136418](#)
9. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 2010; 467: 832–838 (2010). doi: [10.1038/nature09410](#) PMID: [20881960](#)
10. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* 2014; 46: 1173–1186. doi: [10.1038/ng.3097](#) PMID: [25282103](#)
11. Falconer DS, Mackay TFC. *Introduction to Quantitative Genetics*. 4th ed. Harlow, Essex, UK: Longmans Green; 1996.
12. Lynch M, Walsh B. *Genetics and Analysis of Quantitative Traits*. Sunderland MA, USA: Sinauer Associates, Inc; 1998.
13. Robertson FW, Reeve E. Studies in quantitative inheritance I. The effects of selection of wing and thorax length in *Drosophila melanogaster*. *J. Genet.* 1952; 50: 414–448.
14. Partridge L, Langelan R, Fowler K, Zwaan B, French V. Correlated responses to selection on body size in *Drosophila melanogaster*. *Genetics Research* 1999; 74: 43–54.

15. Trotta V, Calboli FCF, Ziosi M, Cavicchi S. Fitness variation in response to artificial selection for reduced cell area, cell number and wing area in natural populations of *Drosophila melanogaster*. *BMC Evol. Biol.* 2007; 7: Suppl 2, S10. PMID: [17767726](#)
16. Gockel J, Robinson SJW, Kennington WJ, Goldstein DB, Partridge L. Quantitative genetic analysis of natural variation in body size in *Drosophila melanogaster*. *Heredity* 2002; 89: 145–153. PMID: [12136418](#)
17. Calboli FCF, Kennington WJ, Partridge L. QTL mapping reveals a striking coincidence in the positions of genomic regions associated with adaptive variation in body size in parallel clines of *Drosophila melanogaster* on different continents. *Evolution* 2003; 57: 2653–2658. PMID: [14686541](#)
18. Rako L, Anderson AR, Sgro CM, Stocker AJ, Hoffmann AA. The association between inversion In(3R) Payne and clinally varying traits in *Drosophila melanogaster*. *Genetica* 2006; 128: 373–384. PMID: [17028965](#)
19. Kennington WJ, Hoffmann AA, Partridge L. Mapping regions within cosmopolitan inversion In(3R) Payne associated with natural variation in body size in *Drosophila melanogaster*. *Genetics* 2007; 177: 549–556. PMID: [17603103](#)
20. Weeks AR, McKechnie SW, Hoffmann AA. Dissecting adaptive clinal variation: markers, inversions and size/stress associations in *Drosophila melanogaster* from a central field population. *Ecology Letters* 2002; 5: 756–763.
21. DeJong G, Bochdanovits Z. Latitudinal clines in *Drosophila melanogaster*: body size, allozyme frequencies, inversion frequencies, and the insulin-signalling pathway. *J. Genet.* 2003; 82: 207–223. PMID: [15133196](#)
22. McKechnie SW, Blacket MJ, Song SV, Rako L, Carroll X, Johnson TK, Jensen LT, Lee SF, Wee CW, Hoffmann AA. A clinally varying promoter polymorphism associated with adaptive variation in wing size in *Drosophila*. *Mol. Ecol.* 2010; 19: 775–784. doi: [10.1111/j.1365-294X.2009.04509.x](#) PMID: [20074315](#)
23. Paaby AB, Bergland AO, Behrman EL, Schmidt PS. A highly pleiotropic amino acid polymorphism in the *Drosophila* insulin receptor contributes to life-history adaptation. *Evolution* 2014; 68: 3395–3409. doi: [10.1111/evo.12546](#) PMID: [25319083](#)
24. Turner TL, Stewart AD, Fields AT, Rice WR, Tarone AM. Population-Based Resequencing of Experimentally Evolved Populations Reveals the Genetic Basis of Body Size Variation in *Drosophila melanogaster*. *PLoS Genet.* 2011; 7: e1001336. doi: [10.1371/journal.pgen.1001336](#) PMID: [21437274](#)
25. Womack JE, Jang HJ, Lee MO. Genomics of complex traits. *Ann. N. Y. Acad. Sci.* 2012; 1271: 33–36. doi: [10.1111/j.1749-6632.2012.06733.x](#) PMID: [23050961](#)
26. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS- a review. *Plant Methods* 2013; 9: p. 29. doi: [10.1186/1746-4811-9-29](#) PMID: [23876160](#)
27. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* 2005; 6: 95–108. PMID: [15716906](#)
28. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 2008; 9: 356–369. doi: [10.1038/nrg2344](#) PMID: [18398418](#)
29. Bergelson J, Roux F. Towards identifying genes underlying ecologically relevant traits in *Arabidopsis thaliana*. *Nat. Rev. Genet.* 2010; 11: 867–879. doi: [10.1038/nrg2896](#) PMID: [21085205](#)
30. Meijón M, Satbhai SB, Tsuchimatsu T, Busch W. Genome-wide association study using cellular traits identifies a new regulator of root development in *Arabidopsis*. *Nat. Genet.* 2013; 46: 77–81. doi: [10.1038/ng.2824](#) PMID: [24212884](#)
31. Jumbo-Lucioni P, Ayroles JF, Chambers M, Jordan KW, Leips J, Mackay TFC, et al. Systems genetics analysis of body weight and energy metabolism traits in *Drosophila melanogaster*. *BMC Genomics* 2010; 11: 297. doi: [10.1186/1471-2164-11-297](#) PMID: [20459830](#)
32. Jumbo-Lucioni P, Bu S, Harbison ST, Slaughter JC, Mackay TFC, Moellering DR, et al. Nuclear genomic control of naturally occurring variation in mitochondrial function in *Drosophila melanogaster*. *BMC Genomics* 2012; 13: p. 659. doi: [10.1186/1471-2164-13-659](#) PMID: [23171078](#)
33. Swarup S, Huang W, Mackay TFC, Anholt RRRH. Analysis of natural variation reveals neurogenetic networks for *Drosophila* olfactory behavior. *Proc. Natl. Acad. Sci. USA* 2013; 110: 1017–1022. doi: [10.1073/pnas.1220168110](#) PMID: [23277560](#)
34. Flint J, Eskin E. Genome-wide association studies in mice. *Nat. Rev. Genet.* 2012; 13: 807–817. doi: [10.1038/nrg3335](#) PMID: [23044826](#)
35. Huang X, Zhao Y, Wei X, Li C, Wang A, Zhao Q, et al. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.* 2011; 44: 32–39. doi: [10.1038/ng.1018](#) PMID: [22138690](#)

36. Lipka AE, Gore MA, Magallanes-Lundback M, Mesberg A, Lin H, Tiede T, et al. Genome-wide association study and pathway-level analysis of tocochromanol levels in maize grain. *G3 (Bethesda)* 2013; 3: 1287–1299.
37. García-Gámez E, Gutiérrez-Gil B, Sahana G, Sánchez JP, Bayón Y, Arranz JJ. GWA Analysis for Milk Production Traits in Dairy Sheep and Genetic Support for a QTN Influencing Milk Protein Percentage in the LALBA Gene. *PLoS ONE* 2012; 7: e47782. doi: [10.1371/journal.pone.0047782](https://doi.org/10.1371/journal.pone.0047782) PMID: [23094085](https://pubmed.ncbi.nlm.nih.gov/23094085/)
38. Makvandi-Nejad S, Hoffman GE, Allen JJ, Chu E, Gu E, Chandler AM, et al. Four Loci Explain 83% of Size Variation in the Horse. *PLoS ONE* 2012; 7: e39929. doi: [10.1371/journal.pone.0039929](https://doi.org/10.1371/journal.pone.0039929) PMID: [22808074](https://pubmed.ncbi.nlm.nih.gov/22808074/)
39. Maxa J, Neuditschko M, Russ I, Förster M, Medugorac I. Genome-wide association mapping of milk production traits in Braunvieh cattle. *Journal of Dairy Science* 2012; 95: 5357–5364. doi: [10.3168/jds.2011-4673](https://doi.org/10.3168/jds.2011-4673) PMID: [22916942](https://pubmed.ncbi.nlm.nih.gov/22916942/)
40. Lee SH, Choi BH, Lim D, Gondro C, Cho YM, Dang CG, et al. Genome-Wide Association Study Identifies Major Loci for Carcass Weight on BTA14 in Hanwoo (Korean Cattle). *PLoS ONE* 2013; 8: e74677. doi: [10.1371/journal.pone.0074677](https://doi.org/10.1371/journal.pone.0074677) PMID: [24116007](https://pubmed.ncbi.nlm.nih.gov/24116007/)
41. Minozzi G, Nicolazzi EL, Stella A, Biffani S, Negrini R, Lazzari B, et al. Genome Wide Analysis of Fertility and Production Traits in Italian Holstein Cattle. *PLoS ONE* 2013; 8: e80219. doi: [10.1371/journal.pone.0080219](https://doi.org/10.1371/journal.pone.0080219) PMID: [24265800](https://pubmed.ncbi.nlm.nih.gov/24265800/)
42. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 2010; 42: 565–569. doi: [10.1038/ng.608](https://doi.org/10.1038/ng.608) PMID: [20562875](https://pubmed.ncbi.nlm.nih.gov/20562875/)
43. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JN, et al. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* 2011; 43: 519–525. doi: [10.1038/ng.823](https://doi.org/10.1038/ng.823) PMID: [21552263](https://pubmed.ncbi.nlm.nih.gov/21552263/)
44. Sutter NB, Bustamante CD, Chase K, Gray MM, Zhao K, Zhu L, et al. A Single IGF1 Allele Is a Major Determinant of Small Size in Dogs. *Science* 2007; 316: 112–115. PMID: [17412960](https://pubmed.ncbi.nlm.nih.gov/17412960/)
45. Oksenberg JR, Baranzini SE, Sawcer S, Hauser SL. The genetics of multiple sclerosis: SNPs to pathways to pathogenesis. *Nat. Rev. Genet.* 2008; 9: 516–526. doi: [10.1038/nrg2395](https://doi.org/10.1038/nrg2395) PMID: [18542080](https://pubmed.ncbi.nlm.nih.gov/18542080/)
46. Thomas D. Gene–environment-wide association studies: emerging approaches. *Nat. Rev. Genet.* 2010; 11: 259–272. doi: [10.1038/nrg2764](https://doi.org/10.1038/nrg2764) PMID: [20212493](https://pubmed.ncbi.nlm.nih.gov/20212493/)
47. Vilhjálmsson BJ, Nordborg M. The nature of confounding in genome-wide association studies. *Nat. Rev. Genet.* 2013; 14: 1–2. doi: [10.1038/nrg3382](https://doi.org/10.1038/nrg3382) PMID: [23165185](https://pubmed.ncbi.nlm.nih.gov/23165185/)
48. Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, et al. The *Drosophila melanogaster* Genetic Reference Panel. *Nature* 2012; 482: 173–178. doi: [10.1038/nature10811](https://doi.org/10.1038/nature10811) PMID: [22318601](https://pubmed.ncbi.nlm.nih.gov/22318601/)
49. Huang W, Massouras A, Inoue Y, Pfeiffer J, Ramia M, Tarone AM, et al. Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. *Genome Res.* 2014; 24: 1193–1208. doi: [10.1101/gr.171546.113](https://doi.org/10.1101/gr.171546.113) PMID: [24714809](https://pubmed.ncbi.nlm.nih.gov/24714809/)
50. Ayroles JF, Carbone MA, Stone EA, Jordan KW, Lyman RF, Magwire MM, et al. Systems genetics of complex traits in *Drosophila melanogaster*. *Nat. Genet.* 2009; 41: 299–307. doi: [10.1038/ng.332](https://doi.org/10.1038/ng.332) PMID: [19234471](https://pubmed.ncbi.nlm.nih.gov/19234471/)
51. Massouras A, Waszak SM, Albarca-Aguilera M, Hens K, Holcombe W, Ayroles JF, et al. Genomic Variation and Its Impact on Gene Expression in *Drosophila melanogaster*. *PLoS Genet.* 2012; 8: p. e1003055. doi: [10.1371/journal.pgen.1003055](https://doi.org/10.1371/journal.pgen.1003055) PMID: [23189034](https://pubmed.ncbi.nlm.nih.gov/23189034/)
52. Nijhout HF, Riddiford LM, Mirth C, Shingleton AW, Suzuki Y, Callier V. The developmental control of size in insects. *WIREs Dev. Biol.* 2014; 3: 113–134.
53. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. *Nat. Methods* 2011; 8: 833–835. doi: [10.1038/nmeth.1681](https://doi.org/10.1038/nmeth.1681) PMID: [21892150](https://pubmed.ncbi.nlm.nih.gov/21892150/)
54. Corbett-Detig RB, Hartl DL. Population genomics of inversion polymorphisms in *Drosophila melanogaster*. *PLoS Genet.* 2012; 8: e1003056. doi: [10.1371/journal.pgen.1003056](https://doi.org/10.1371/journal.pgen.1003056) PMID: [23284285](https://pubmed.ncbi.nlm.nih.gov/23284285/)
55. Bronstein R, Levkovitz L, Yosef N, Yanku M, Ruppin E, Sharan R, et al. Transcriptional regulation by CHIP/LDB complexes. *PLoS Genet.* 2010; 6: e1001063. doi: [10.1371/journal.pgen.1001063](https://doi.org/10.1371/journal.pgen.1001063) PMID: [20730086](https://pubmed.ncbi.nlm.nih.gov/20730086/)
56. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 2012; 41: D808–D815. doi: [10.1093/nar/gks1094](https://doi.org/10.1093/nar/gks1094) PMID: [23203871](https://pubmed.ncbi.nlm.nih.gov/23203871/)

57. Liu JZ, Mcrae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, et al. A Versatile Gene-Based Test for Genome-wide Association Studies. *Am. J. Hum. Genet.* 2010; 87: 139–145. doi: [10.1016/j.ajhg.2010.06.009](https://doi.org/10.1016/j.ajhg.2010.06.009) PMID: [20598278](https://pubmed.ncbi.nlm.nih.gov/20598278/)
58. Schüpbach T, Xenarios I, Bergmann S, Kapur K. FastEpistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics* 2010; 26: 1468–1469. doi: [10.1093/bioinformatics/btq147](https://doi.org/10.1093/bioinformatics/btq147) PMID: [20375113](https://pubmed.ncbi.nlm.nih.gov/20375113/)
59. Müller P, Kutenkeuler D, Gesellchen V, Zeidler MP, Boutros M. Identification of JAK/STAT signaling components by genome-wide RNA interference. *Nature* 2005; 436: 871–875. PMID: [16094372](https://pubmed.ncbi.nlm.nih.gov/16094372/)
60. Wang H, Chen X, He T, Zhou Y, Luo H. Evidence for tissue-specific Jak/STAT target genes in *Drosophila* optic lobe development. *Genetics* 2013; 195: 1291–1306. doi: [10.1534/genetics.113.155945](https://doi.org/10.1534/genetics.113.155945) PMID: [24077308](https://pubmed.ncbi.nlm.nih.gov/24077308/)
61. Yang L, Meng F, Ma D, Xie W, Fang M. Bridging Decapentaplegic and Wingless signaling in *Drosophila* wings through repression of naked cuticle by Brinker. *Development* 2012; 140, 413–422.
62. Madan LL, Veeranna S, Shameer K, Reddy CCS, Sowdhamini R, Gopal B. Modulation of Catalytic Activity in Multi-Domain Protein Tyrosine Phosphatases. *PLoS ONE* 2011; 6: e24766. doi: [10.1371/journal.pone.0024766](https://doi.org/10.1371/journal.pone.0024766) PMID: [21931847](https://pubmed.ncbi.nlm.nih.gov/21931847/)
63. Carter GW. Inferring gene function and network organization in *Drosophila* signaling by combined analysis of pleiotropy and epistasis. *G3* 2013; 3: 807–814. doi: [10.1534/g3.113.005710](https://doi.org/10.1534/g3.113.005710) PMID: [23550134](https://pubmed.ncbi.nlm.nih.gov/23550134/)
64. Murali T, Pacifico S, Yu J, Guest S, Roberts GG 3rd, Finley RL Jr. Droid 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for *Drosophila*. *Nucleic Acids Res.* 2011; 39: D736–D743. doi: [10.1093/nar/gkq1092](https://doi.org/10.1093/nar/gkq1092) PMID: [21036869](https://pubmed.ncbi.nlm.nih.gov/21036869/)
65. Yu J, Pacifico S, Liu G, Finley RL Jr. Droid: the *Drosophila* Interactions Database, a comprehensive resource for annotated gene and protein interactions. *BMC Genomics.* 2008; 9: p. 461. doi: [10.1186/1471-2164-9-461](https://doi.org/10.1186/1471-2164-9-461) PMID: [18840285](https://pubmed.ncbi.nlm.nih.gov/18840285/)
66. Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, et al. Unlocking the secrets of the genome. *Nature* 2009; 459: 927–930. doi: [10.1038/459927a](https://doi.org/10.1038/459927a) PMID: [19536255](https://pubmed.ncbi.nlm.nih.gov/19536255/)
67. Young RS, Marques AC, Tibbit C, Haerty W, Basset AR, Liu JL, et al. Identification and Properties of 1,119 Candidate LincRNA Loci in the *Drosophila melanogaster* Genome. *Genome Biol. Evol.* 2012; 4: 427–442. doi: [10.1093/gbe/evs020](https://doi.org/10.1093/gbe/evs020) PMID: [22403033](https://pubmed.ncbi.nlm.nih.gov/22403033/)
68. Hangauer MJ, Vaughn IW, McManus MT. Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. *PLoS Genetics* 2013; 9: e1003569. doi: [10.1371/journal.pgen.1003569](https://doi.org/10.1371/journal.pgen.1003569) PMID: [23818866](https://pubmed.ncbi.nlm.nih.gov/23818866/)
69. Ernst J, Kheradpour P, Mikkelsen TS, Shores N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 2011; 473: 43–49. doi: [10.1038/nature09906](https://doi.org/10.1038/nature09906) PMID: [21441907](https://pubmed.ncbi.nlm.nih.gov/21441907/)
70. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J. Mol. Biol.* 1990; 215: 403–410. PMID: [2231712](https://pubmed.ncbi.nlm.nih.gov/2231712/)
71. Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, et al. An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics* 2011; 12: 357. doi: [10.1186/1471-2105-12-357](https://doi.org/10.1186/1471-2105-12-357) PMID: [21880147](https://pubmed.ncbi.nlm.nih.gov/21880147/)
72. Durham MF, Magwire MM, Stone EA, Leips J. Genome-wide analysis in *Drosophila* reveals age-specific effects of SNPs on fitness traits. *Nat. Commun.* 2014; 5: p. 4338. doi: [10.1038/ncomms5338](https://doi.org/10.1038/ncomms5338) PMID: [25000897](https://pubmed.ncbi.nlm.nih.gov/25000897/)
73. Carrington JC, Ambros V. Role of microRNAs in plant and animal development. *Science* 2003; 301: 336–338. PMID: [12869753](https://pubmed.ncbi.nlm.nih.gov/12869753/)
74. Inui M, Martello G, Piccolo S. MicroRNA control of signal transduction. *Nat. Rev. Mol. Cell Biol.* 2010; 11: 252–263. doi: [10.1038/nrm2868](https://doi.org/10.1038/nrm2868) PMID: [20216554](https://pubmed.ncbi.nlm.nih.gov/20216554/)
75. Fatica A, Bozzoni I. Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.* 2014; 15: 7–21. doi: [10.1038/nrg3606](https://doi.org/10.1038/nrg3606) PMID: [24296535](https://pubmed.ncbi.nlm.nih.gov/24296535/)
76. Schleich S, Strassburger K, Janiesch PC, Koledachkina T, Miller KK, Haneke K et al. DENR-MCT1 promotes translation re-initiation downstream of uORFs to control tissue-growth. *Nature* 2014; 512: 208–212. doi: [10.1038/nature13401](https://doi.org/10.1038/nature13401) PMID: [25043021](https://pubmed.ncbi.nlm.nih.gov/25043021/)
77. Schertel C, Huang D, Björklund M, Bischof J, Yin D, Li R, et al. Systematic Screening of a *Drosophila* ORF Library in vivo Uncovers Wnt/Wg Pathway Components. *Dev. Cell* 2013; 25: 207–219. doi: [10.1016/j.devcel.2013.02.019](https://doi.org/10.1016/j.devcel.2013.02.019) PMID: [23583758](https://pubmed.ncbi.nlm.nih.gov/23583758/)
78. Unckless RL, Rottschaefer SM, Lazzaro BP. A Genome-Wide Association Study for Nutritional Indices in *Drosophila*. *G3* 2015; 5: 417–425. doi: [10.1534/g3.114.016477](https://doi.org/10.1534/g3.114.016477) PMID: [25583649](https://pubmed.ncbi.nlm.nih.gov/25583649/)

79. Iida H, Nakamura H, Ono T, Okumura MS, Anraku Y. MID1, a novel *Saccharomyces cerevisiae* gene encoding a plasma membrane protein, is required for Ca<sup>2+</sup> influx and mating. *Mol Cell Biol* 1994; 14: 8259–8271. PMID: [7526155](#)
80. Levin DE, Errede B. The proliferation of MAP kinase signaling pathways in yeast. *Curr. Opin. Cell Biol.* 1995; 7: 197–202. PMID: [7612271](#)
81. Syed ZA, Hård T, Uf A, van Dijk-Hård IF. A potential role for *Drosophila* mucins in development and physiology. *PLoS ONE* 2008; 3: e3041. doi: [10.1371/journal.pone.0003041](#) PMID: [18725942](#)
82. Povelones M, Howes R, Fish M, Nusse R. Genetic Evidence That *Drosophila* frizzled Controls Planar Cell Polarity and Armadillo Signaling by a Common Mechanism. *Genetics* 2005; 171: 1643–1654. PMID: [16085697](#)
83. Parsons LM, Grzeschik NA, Allott M, Richardson H. Lgl/aPKC and Crb regulate the Salvador/Warts/Hippo pathway. *Fly* 2010; 4: 288–293. PMID: [20798605](#)
84. Lin C, Katanaev VL. Kermit Interacts with Gαo, Vang, and Motor Proteins in *Drosophila* Planar Cell Polarity. *PLoS ONE* 2013; 8: e76885. doi: [10.1371/journal.pone.0076885](#) PMID: [24204696](#)
85. Hatakeyama J, Wald JH, Printsev I, Ho HYH, Carraway KL. Vangl1 and Vangl2: planar cell polarity components with a developing role in cancer. *Endocrine Related Cancer* 2014; 21: R345–R356. doi: [10.1530/ERC-14-0141](#) PMID: [24981109](#)
86. Weber U, Pataki C, Mihaly J, Mlodzik M. Combinatorial signaling by the Frizzled/PCP and Egfr pathways during planar cell polarity establishment in the *Drosophila* eye. *Dev. Biol.* 2008; 316: 110–123. doi: [10.1016/j.ydbio.2008.01.016](#) PMID: [18291359](#)
87. Sing A, Tsatskis Y, Fabian L, Hester I, Rosenfeld R, Serrichio M, et al. The Atypical Cadherin Fat Directly Regulates Mitochondrial Function and Metabolic State. *Cell* 2014; 158: 1293–1308. doi: [10.1016/j.cell.2014.07.036](#) PMID: [25215488](#)
88. van Bon BWM, Oortveld MAW, Nijtmans LG, Fenckova M, Nijhof B, Besseling J, et al. CEP89 is required for mitochondrial metabolism and neuronal function in man and fly. *Hum. Mol. Genet.* 2013; 22: 3138–3151. doi: [10.1093/hmg/ddt170](#) PMID: [23575228](#)
89. Pereira C, Queirós S, Galaghar A, Sousa H, Pimentel-Nunes P, Brandão C, et al. Genetic variability in key genes in prostaglandin E2 pathway (COX-2, HPGD, ABCC4 and SLCO2A1) and their involvement in colorectal cancer development. *PLoS ONE* 2014; 9: e92000. doi: [10.1371/journal.pone.0092000](#) PMID: [24694755](#)
90. Clough E, Jimenez E, Kim YA, Whitworth C, Neville MC, Hempel LU, et al. Sex- and Tissue-Specific Functions of *Drosophila* Doublesex Transcription Factor Target Genes. *Dev. Cell* 2014; 31: 761–773. doi: [10.1016/j.devcel.2014.11.021](#) PMID: [25535918](#)
91. Palenzona DL, Alicchio R. Differential response to selection on the two sexes in *Drosophila melanogaster*. *Genetics* 1973; 74: 533–542. PMID: [4200687](#)
92. Menezes BF, Vigoder FM, Peixoto AA, Varaldi J. The influence of male wing shape on mating success in *Drosophila melanogaster*. *Animal Behaviour* 2013; 85: 1217–1223.
93. Houle D, Mezey J, Galpern P, Carter A. Automated measurement of *Drosophila* wings. *BMC Evol. Biol.* 2003; 3: p.25. PMID: [14670094](#)
94. Turner SD. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv* doi: [10.1101/005165](#) <http://cran.r-project.org/web/packages/qqman/index.html> accessed 19.09.2014.
95. Dietzl G, Chen D, Schnorrer F, Su KC, Barinova Y, Fellner M et al. A genome-wide transgenic RNAi library for conditional gene inactivation in *Drosophila*. *Nature* 2007; 448: 151–156. PMID: [17625558](#)
96. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. *Genome Res.* 2009; 19: 1639–1645. doi: [10.1101/gr.092759.109](#) PMID: [19541911](#)
97. St Pierre SE, Ponting L, Stefancsik R, McQuilton P, FlyBase Consortium. FlyBase 102—advanced approaches to interrogating FlyBase. *Nucleic Acids Res.* 2014; 42: D780–8.
98. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, et al. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 2007; 450: 203–218. PMID: [17994087](#)

## 5. Discussion

### 5.1. How it all fits together: Interpretation of GWAS findings

The aim of genetics is to understand the genetic basis of traits by linking genetic variability to phenotypic variability. The genetic variability studied can be either artificially introduced (such as in knock-out studies) or the naturally occurring variability can be exploited. While many studied traits have a clear medical implication, some are of interest as models of traits with a particular genetic architecture because they are easy to study. In the field of genetics of complex phenotypes, one model trait of considerable interest is human height: It is easy to measure, allowing to assemble very large cohorts making in depth study feasible. Furthermore, it is highly heritable, extremely polygenic and driven to a large extent by common variants [13]. Therefore, lessons learned from studying human height can inform us about what will be needed to fully elucidate the genetics of complex diseases with similar architectures. These investigation, while very successful at uncovering new variants involved in the genetics of height, have also revealed that only a fraction of all contributing genetic variants have been uncovered so far [5].

One can expect that sample size will continue to increase leading to ever more annotated variants. However, even if all genetic variants influencing a complex trait such as human height were discovered, it is currently unclear how to best tackle the challenge of interpreting variants in the context of the biology involved. Compared to the problem of uncovering all variants that contribute to the heritability of a trait, the problem of biological interpretation is very multi-faceted and much less clearly defined, but is nevertheless crucial to reap the full benefit from the revolution in genetics currently underway. One strategy for interpretation is pathway analysis where prior biological knowledge is formalized into sets of genes with annotated functions and results from genetics studies are searched for enrichments, which in turn connects the biological function to the investigated trait. Chapter 2 showed work that follows this general reasoning.

A particularly attractive feature of genetics studies is that experimental perturbations are not necessarily required, making them a powerful tool to study human biology. However, to investigate the molecular mechanistic details of the genetic impact of a variant, experiments are often performed in model organisms because of the need for perturbation. A crucial question therefore is, how well a model organisms' phenotype may serve as a model for the human phenotype of interest. This question is of great practical importance. For example, to develop a pharmacological manipulation strategy of a phenotype, one typically needs an animal model of said phenotype on which to test potential treatments. The study of the genetics of animal model is therefore relevant not only for its own sake but also to see how generalizable results between different organisms are. The effort described in Chapter 4, an investigation into the genetics of drosophila growth control, can be seen under this aspect.

Another strategy to interpret genetic variants affecting complex phenotypes is to find simpler phenotypes that are affected by the variant and see the former as a consequence of the latter. In particular, gene expression regulation has emerged as a less complex phenotype in terms of which complex effects such as changes in body size can be investigated. But even though gene regulation as a phenotype is simpler, it is not fully understood. However, joint modelling of different molecular data types allows to investigate this process. The classical example is the expression quantitative trait locus (eQTL) where expression and genotypes are modeled together. eQTL mapping has been successful at finding SNPs in cis to gene loci affecting the expression levels for a substantial fraction of genes in a wide variety of tissues and cell types [59, 60]. While discovering SNPs that have an impact on gene expression in trans has been more challenging due to generally smaller effect sizes and multiple testing burden, studies with larger sample sizes did discover numerous trans-eQTLs suggesting that comprehensive regulatory maps might be constructed in the near future using this approach [7, 60]. However, there is a clear need for further integrative data analysis strategies to deepen the understanding of gene regulation. Chapter 3 showed such a method to elucidate effects of expression of TFs on



chromatin accessibility, an important layer of gene regulation.

These approaches can therefore be seen as attempts of providing context for regular GWAS analysis; by investigating how GWAS results cluster into functional categories, how well the genetics of a model organism generalizes to human genetics, or by deepening our understanding of the regulatory processes on the chromatin level which is a prerequisite for a full understanding of how genetic variation leads to variation in gene regulation.

## 5.2. Methodology overview of each project

While none of the three approaches are classical human GWAS, they drew inspiration from it, as in the case of the investigation of the genetics of drosophila body measures and the method developed to investigate chromatin state regulators, or were using GWAS results as their starting point.

**Drosophila growth control** Of the three projects, the investigation of the genetics of drosophila body measures was perhaps the closest to a traditional GWAS setup: A naturally occurring population was genotyped and phenotyped, and it was investigated which SNPs would associate to the phenotype. Still, there is a number of idiosyncrasies to the analysis that are specific to this experimental setup. The first is that, although a naturally occurring population is sampled, stable in-bred lines are then produced. The reason is experimental logistics: In-bred lines can be maintained in perpetuity allowing for additional phenotyping being performed later by other scientists. The downside is that it alters somewhat the genetic composition of the lines compared to the natural lines: The dominant genetic component will be amplified and will not be separable from the additive component. However, it is still much closer to the natural genetic variability than more traditional knock-out screens. Another advantage of using model organisms is that one can impose strict environmental control. This demands additional efforts in experimental

setup and statistical modeling as can be seen in the related chapter. Follow up studies to validate variants or genes can be performed in the actual model system whereas in human GWAS, one has to rely on proxy systems for validation. Again, our study provides an example of this principle. Further, animal studies are not plagued by privacy concerns as human studies are. This implies that genotypes and phenotypes can be made available for re-analysis and meta-analysis taking account of more complex modeling. This also applies to our study, as the phenotypes are freely available.

**Chromatin state regulators** In the second investigation we tried to uncover chromatin state regulators: transcription factors that are direct drivers between open and closed chromatin states, by developing novel methodology and applying it to repositories of public data [53, 54]. While parts of the pipeline have resemblance with a GWAS-type setup, the data set used, is of a very different nature. Genotypes and Phenotypes are replaced by two kinds of functional genomics measurements: gene expression and transcription factor motif accessibility (approximated by motif enrichment in open chromatin). Furthermore, instead of individuals from a population, the sampling units are various cell lines. To minimize confounding, we made use of multiple techniques discussed below in detail. While measuring genome-wide expression is well established, the motif accessibility measures in open chromatin relies on assays that have a much shorter history, making the claim that motif enrichment in open chromatin can function as a proxy for motif accessibility in terms of total binding events more controversial [46]. It seems clear from the literature and our own analysis, that for most TFs chromatin binding occurs mainly in open chromatin regions [49], though the literature also suggests that open chromatin and motif strength alone are not the only determinants of binding [52]. However, averaging across all TF motifs in the genome should yield an adequate motif accessibility measure, a conclusion that is also supported by our investigation using ChiP-seq data.

**GWAS pathway analysis** Further, we developed an approach to help analyse the results derived from GWAS while making sure that on the one hand the lessons learnt from pathway analysis for gene expression data are incorporated and on the other hand that the special properties of GWAS summary results data are accounted for. We used gene-wise  $p$ -values as basis of our pathway enrichment, making use of external genotype data to estimate the correlation structure between SNPs necessary to calculate said  $p$ -values. Independence between gene scores is a prerequisite of typical enrichment strategies, as enrichment tests assume that gene scores in a pathway are just an independent sample from all available gene scores. While GWAS summary statistics are correlated, their correlation (bar confounding) is only genomically local allowing for a correction strategy where neighbouring genes in the same pathway are fused for the duration of pathway gene score calculation. While this step ensures independence, it incurs further computational cost. We therefore made sure that all developed  $p$ -value calculation strategies were sufficiently fast and accurate to limit the analysis time even in settings where many genes are fused. In addition to deriving and implementing the methodology, we performed extensive testing on real data to make sure that the method performed comparatively well in high and low power settings alike, as usability in different settings was a major concern.

### **5.3. Confounding: A problem revisited in each project**

While statistical significance testing methodologies can vary a lot in the type of questions asked and the modeling strategies employed to answer them, there are some fundamental challenges that most have to deal with. Perhaps the biggest is the issue of confounding. Confounding happens, when a statistical association between two variables is observed due to a third unmeasured variable. In GWAS, the most common form of confounding variable is population structure, where the third unmeasured variable, confounding genotype and phenotype, is origin. Fortunately, the amount of confounding, as well as the confounding variable, can be estimated to a certain extent from the data itself, a boon made possible by the genome-wide era [32, 61]. All the covered

investigations had to tackle the problem of confounding in some way. The drosophila body measure GWAS used an approach of modeling general population structure via mixed modeling and additionally model the occurrence of large inversions separately, after evaluating their impact on the phenotype. In the investigation on chromatin accessibility regulators, confounding was probably most severe. Both functional genomics measurements are heavily driven by cell line origin leading to very strong confounding (as can be assessed by  $p$ -value inflation). Surprisingly, the mixed model and data cleaning strategy were remarkably successful at controlling confounding as evaluated by  $p$ -value inflation. While the mixed model strategy works well to correct for confounding variables that affect most tested genes, this confounding control strategy does not work, if there is a small number of confounded genes because they will not contribute substantially to the spectral distribution of the relationship matrix. Unfortunately, this type of confounding is also hard to diagnose, because it does not impact the bulk of the  $p$ -value distribution<sup>1</sup>. We circumvented this problem by leveraging substantial prior biological knowledge. Per investigated motif, we are only interested in one hypothesis: whether some TF having this motif shows a strong association. That particular hypothesis, compared to all other hypothesis, cannot be influenced systematically by confounding. Therefore we can be confident that the bulk of our results are not driven by spurious confounding.

In the case of the enrichment strategies for pathway analysis, confounding can be subtle and has a connection to comparability and biased pathway collections. Comparability refers to the fact that gene scores should not be biased with regards to variables other than their actual involvement in the trait of interest [64]. For instance, taking the raw minimum  $p$ -value of all

---

<sup>1</sup>Interestingly, a similar problem occurs in GWAS confounding control of low frequency variants, where very local population structure affects a limited number of low frequency variants but is not sufficiently represented in the spectral distribution [62]. In this case, the solution offered by Listgarten et al. consisted in constructing a relationship matrix using only variants with strong associations to the phenotype, thereby biasing the genetic relationship matrix toward affected low frequency variants [63]. One also needs to exclude all SNPs in the locus that is currently tested from genetic relationship matrix construction as they would contribute substantially to the much lower dimensional genetic relationship matrix. However, this solution is difficult to apply to our situation because for expression data, it is much more difficult to know which genes can be regarded as statistically independent in the absence of confounding, as this would necessitate detailed knowledge of the causal gene regulatory network.

SNPs in gene region for each gene as its gene score would lead to a bias, where longer genes had lower gene scores, which would bias any pathway analysis towards pathways containing many long genes. This is particularly problematic if the pathway collection is biased in terms of that same variable. For instance in the above example, if a certain pathway contains mainly short genes while another is mainly composed of long genes their enrichment scores would not be comparable. In the case of only null genes one can ensure comparability by using  $p$ -values as the basis for the underlying gene scores. If there are non-null genes, comparability is very hard to ensure because one would need to show that the pathway collection is unbiased in terms of factors affecting power. Continuing our example above, even if we are using genewise  $p$ -values as the basis for our enrichments, if there are many non-null genes, and gene size affects power to detect a gene as significant, again the two pathways would not be comparable. While this is a problem in theory, in our exploration of that issue via simulation did not show a bias for the genetic architectures and the pathway collections that we tested. However, our results cannot be readily generalized to other pathway collections and other genetic architectures.

#### **5.4. Causality**

A common problem in statistical data analysis is to decide how much can be learned from a statistical association about causality. This is a question where domain knowledge can help, but a conservative statistician prefers perturbation data in a randomized trial. Among the sciences, genetics is in the enviable position of being able to find causal relationships without the need for perturbation data because all genetic variants are fixed at birth and their segregation in the germ line can be regarded as statistically independent of other factors that could influence the phenotype. However, for common variants, this is not true for SNPs within the same locus. They can be strongly correlated, albeit in a predictable fashion, making it difficult to isolate the causal variant. On the gene level, the problem can persist and can be further complicated by the fact that causal SNPs might fall in regulatory regions with unclear connections to the gene of action.

In the investigation of drosophila growth control, this problem is less acute as the effective population size is much larger and LD much lower. In our pathway investigation, the use of human data and the need for gene level scores makes the annotation of SNPs to genes a relevant question. We opted to err on the side of low false negatives, allowing for the opportunity of the pathway enrichments to prioritize genes in loci with multiple genes.

The problem of causality is of particular interest in the chromatin accessibility study. Because we do not use genotype data in this case, causality statements have to be justified much more carefully. One way to do so is domain knowledge: A model, where the transcription factor expression level affects transcription factor binding measured across the genome is the most parsimonious one for the observed associations. However, the final test will be carefully designed perturbation experiments.

## **5.5. Future work**

Our investigation into the genetics of drosophila growth control can be seen as a pilot study into the genetics of natural drosophila populations, which hopefully, will be built upon in the future with further samples to reap the full benefits of the advantages of working with a model system. One attractive feature of working with wild populations of model organism that has been potentially underexplored, is the opportunity for detailed regulatory studies. As mentioned above, LD blocks are much smaller, which would make it particularly easy to pinpoint causal SNPs. Finding variants affecting molecular phenotypes such as expression and epigenetic marks could be of particular interest, because it would allow to precisely model what properties a SNP must have to lead to changes in expression.

With regards to the chromatin accessibility investigation, the work would clearly benefit from further experimental validation by experimental biologists as is common for purely computational data analysis. Further, the questions asked and potentially answered by our approach could

be refined if further genomics assays would be added. For example one could investigate the question which transcription factors change neighbouring histone methylation states upon binding in open chromatin regions by associating average histone methylation state in motif-carrying open chromatin regions with TF expression. While we attempted such an analysis, the current collections available did not yield positive results, probably due to low power. Further, it could be of interest to study whether changes in CAR expression levels have medical consequences. The fact that expression changes impact chromatin accessibility globally, could suggest that changes in CAR levels might have an increased likelihood of affecting health outcomes compared to changes in non-CAR TFs. Collocalisation, of SNPs affecting CAR expression and SNPs being disease relevant could be a test of this hypothesis. However, one has to bear in mind, that the variation in expression changes seen in eQTL studies are much smaller, than between different cell lines and extrapolation between the different regimes might not be straight forward.

More generally, the study uses a novel strategy for integrative data analysis of functional genomics data, where different types of functional genomics data is collected across multiple cell lines of various origins. The strategy for confounding control in particular could potentially be used to associate various functional elements across cell lines, depending on the interest of the researcher. For example, associating specific enhancer regions to gene expression could be of interest, because the mapping of enhancers to promoters is still a challenging problem [65].

With regards to *Pascal*, the method was featured in other genetics investigations, be it the analysis and evaluation of regulatory networks [66], for the analysis of a large exome study on human height (Marouli et al. Nature, accepted), and the forthcoming analysis and evaluation of module calling algorithms in a DREAM challenge. The results from this challenge might inform further avenues. One can, for example, envision that modules called from contestants might help refine predefined pathway sets in a cell type specific manner. Other extensions and use cases could be of interest. On the methodological side, one interesting feature would be to move from statistical testing to statistical estimation: Currently, pathways can only be tested

for significance. A question of interest might be how much heritability can be attributed to the pathway of interest. As is shown in Appendix C, one can derive the maximum likelihood of the heritability captured by SNPs within a pathway from the summary statistics alone.

On the side of use cases, plans exist to build a web portal for allowing users to supply gene lists to test for enrichment across a large list of GWAS. This setup would mesh well with a typical work-flow of experimental biologists generating gene lists of interest via literature review or high throughput experiments and has the potential to open the results derived from GWAS to a much wider community helping to fulfill the promise of GWAS to provide a framework to link medical traits of interest to fundamental molecular biology in a methodologically systematic way.



## A. Estimating the contribution of pairwise interactions to the genetic variance

### A.1. A computationally efficient estimation strategy

In the following, we will develop a simple strategy to estimate the heritability contribution of all SNP-SNP interactions. We will then investigate the power of the approach in a simplified setting and compare it to the power of estimating the additive heritability. These considerations will suggest that the number of individuals needs to be of the order of  $10^6$  but that lower numbers are needed if only a subset of SNP-SNP pairs are investigated.

We assume we are given a  $n \times 1$  phenotype vector  $\mathbf{y}$  normalized such that  $\sum_i y_i = 0$  and  $\mathbf{y}^T \mathbf{y} = n$ . Further, we have  $m \times n \times 1$  SNP genotype vectors, all of which are normalized in the same fashion as the phenotype vector.  $m$  might be the number of all SNPs in a genome-wide panel. Define  $S$  as the set of all ordered pairs of SNP indices (i.e.  $S$  contains  $m^2$  elements). We will try to fit the following random effects model:

$$\mathbf{y} = \sum_i (\mathbf{x}_i) b_i + \sum_{(i,j) \in S} (\mathbf{x}_i \cdot \mathbf{x}_j) b_{ij} + \boldsymbol{\epsilon},$$

where  $b_i \sim N(0, \sigma_g)$ ,  $b_{ij} \sim N(0, \sigma_h)$ ,  $\boldsymbol{\epsilon} \sim N(0, \sigma_e)$  and  $\cdot$  refers to the point-wise or Hadamard product. This is equivalent to estimating two random components additionally to the noise term with covariance matrices

$$\sigma_g^2 \mathbf{C}^g = \sigma_g^2 \sum_i (\mathbf{x}_i \mathbf{x}_i^T)$$

and

$$\sigma_h^2 \mathbf{C}^h = \sigma_h^2 \sum_{(i,j) \in S} (\mathbf{x}_i \cdot \mathbf{x}_j) (\mathbf{x}_i \cdot \mathbf{x}_j)^T.$$

Direct computation of  $\mathbf{C}^h$  might seem prohibitive because  $S$  might contain on the order of  $10^{12}$  elements. However, the computation can be easily reduced to the computation of  $\mathbf{C}^g$ .

**Proposition 1.** Define  $\mathbf{C}^h$  and  $\mathbf{C}^g$  as above, then

$$\mathbf{C}^h = \mathbf{C}^g \cdot \mathbf{C}^g.$$

*Proof.* For any pair of genotype vectors  $\mathbf{x}_i, \mathbf{x}_j$  it holds that

$$(\mathbf{x}_i \cdot \mathbf{x}_j)(\mathbf{x}_i \cdot \mathbf{x}_j)^T = (\mathbf{x}_i \mathbf{x}_i^T) \cdot (\mathbf{x}_j \mathbf{x}_j^T).$$

Plugging this relationship into the definition of  $\mathbf{C}^h$ , we get

$$\begin{aligned} \sum_{(i,j) \in S} (\mathbf{x}_i \cdot \mathbf{x}_j)(\mathbf{x}_i \cdot \mathbf{x}_j)^T &= \sum_{(i,j) \in S} (\mathbf{x}_i \mathbf{x}_i^T) \cdot (\mathbf{x}_j \mathbf{x}_j^T) \\ &= \sum_i (\mathbf{x}_i \mathbf{x}_i^T) \cdot \sum_j (\mathbf{x}_j \mathbf{x}_j^T) \\ &= \mathbf{C}^g \cdot \mathbf{C}^g. \end{aligned}$$

□

One can also remove from  $S$  all ordered pairs of SNP indices for SNP pairs in LD. This would remove dominance effects. The factor to remove is easily calculated by applying the same trick as above to consecutive blocks of SNPs. Define  $B_k$  as the set of SNP indices in the  $k$ th block and  $\mathbf{C}_k^g$  as

$$\mathbf{C}_k^g = \sum_{i \in B_k} (\mathbf{x}_i \mathbf{x}_i^T)$$

then we can remove all dominance effects by calculating

$$\mathbf{C}^d = \sum_k \mathbf{C}_k^g \cdot \mathbf{C}_k^g \cdot + 2 \sum_k \mathbf{C}_k^g \cdot \mathbf{C}_{k+1}^g.$$

and removing it from  $\mathbf{C}^h$ :

$$\mathbf{C}^{h*} = \mathbf{C}^h - \mathbf{C}^d$$

To estimate  $\sigma_g^2$  and  $\sigma_h^2$ , we can employ a variant of Haseman-Elston regression which is computa-

tionally very efficient. Note that

$$E[Y_i Y_j] = \sigma_g^2 \mathbf{C}_{ij}^g + \sigma_h^2 \mathbf{C}_{ij}^h + \sigma_e^2 \delta_{ij}$$

where  $\delta_{ij} = 1$  if  $i = j$  and 0 else. We then perform a regression across all individual pairs. In short: the time complexity to estimate the genetic variance of pairwise interactions, if we are prepared to use Haseman-Elston regression, is the same as calculating the regular genetic relationship matrix.<sup>2</sup>

## A.2. Power considerations

We will try to estimate the power of the maximum-likelihood estimator of the random effects model.

$$\mathbf{y} \sim N_n(\sigma_g^2 \mathbf{C} + \sigma_e^2 \mathbf{I}_n),$$

We will use this model to investigate how the power to estimate  $\boldsymbol{\theta} = (\sigma_g^2, \sigma_e^2)^T$  depends on  $\mathbf{C}$ . To simplify the situation, we assume that  $\sigma_g^2 \rightarrow 0$  and that  $\sigma_e^2 \rightarrow 1$ . This assumption while not reasonable for the marginal genetic component should be reasonable for the genetic components capturing interactions as the total contribution to heritability should be low. Because the covariance matrix of the maximum-likelihood estimate converges to the inverse of the information matrix, we can investigate the power by looking at the fisher information matrix. Low values on the diagonal of the inverse of the fisher information matrix imply a low variance estimator and high values imply a high variance estimator. We can show the following:

---

<sup>2</sup>Simulation was used to test the interaction variant of the Haseman-Elston regression. 100 independent SNPs (normally distributed for ease of simulation) and 2'000 observations were sampled. The parameters were set to  $\sigma_e^2 = 2, \sigma_g^2 = 0.02$  and  $\sigma_h^2 = 0.0002$ , and observations were simulated from the model 50 times. The Haseman-Elston approach outlined above (only using upper triangular values of the covariance matrices) was used to estimate the parameters. The mean of  $\sigma_g^2$  estimates was 0.01909 and their standard deviation 0.00267. The mean of  $\sigma_h^2$  estimates was 0.000202 and their standard deviation 0.000108.

**Proposition 2.** For the model above, assuming  $\sigma_g^2 \rightarrow 0$  and  $\sigma_e^2 \rightarrow 1$ , then

$$I_{1,1}^{-1}(\boldsymbol{\theta}) \rightarrow \frac{2}{\sum_{ij} c_{ij}^2 - (\sum_i c_{ii})^2/n},$$

where  $c_{ij}$  is the  $ij$ -th element of  $\mathbf{C}$ .

*Proof.* Since the Fisher information matrix of the multivariate normal distribution  $N_n(\mathbf{0}, \boldsymbol{\Sigma})$  is

$$I_{m,n} = \frac{1}{2} \text{tr}(\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_m} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_n})$$

We have

$$I_{1,1}(\boldsymbol{\theta}) = \frac{1}{2} \sum_i \frac{\lambda_i^2}{(\sigma_g^2 \lambda_i + \sigma_e^2)^2},$$

$$I_{2,2}(\boldsymbol{\theta}) = \frac{1}{2} \sum_i \frac{1}{(\sigma_g^2 \lambda_i + \sigma_e^2)^2},$$

and

$$I_{1,2}(\boldsymbol{\theta}) = \frac{1}{2} \sum_i \frac{\lambda_i}{(\sigma_g^2 \lambda_i + \sigma_e^2)^2},$$

where  $\lambda_i$  is the  $i$ -th eigenvalue of  $\boldsymbol{\Sigma}$ . From

$$I_{1,1}^{-1}(\boldsymbol{\theta}) = \frac{I_{2,2}(\boldsymbol{\theta})}{I_{2,2}(\boldsymbol{\theta})I_{1,1}(\boldsymbol{\theta}) - I_{1,2}^2(\boldsymbol{\theta})}$$

follows that when  $\sigma_g^2 \rightarrow 0$  and  $\sigma_e^2 \rightarrow 1$

$$I_{1,1}^{-1}(\boldsymbol{\theta}) \rightarrow \frac{2}{\sum_i \lambda_i^2 - (\sum_i \lambda_i)^2/n}$$

We have  $\sum_i \lambda_i = \text{tr}(\mathbf{C})$  and because  $\sum_i \lambda_i^2 = \text{tr}(\mathbf{C}\mathbf{C})$ , we can sum the squares of all entries of  $\mathbf{C}$  to get  $\sum_i \lambda_i^2$ . □

### A.3. Power of the pairwise interaction covariance matrix

When applying this reasoning to the pairwise interaction covariance matrix, we see that the variance of the estimator is approximately inversely proportional to the sum of the entries of  $\mathbf{C}$

each raised to the fourth power. We will investigate how large the expectation of this quantity is assuming  $m$  independent markers. (This is obviously an unreasonable assumption, However, since dependence between markers will lead to slower convergence of off-diagonal elements in  $\mathbf{C}$  to 0 this will yield an lower bound for  $\sum_{ij} c_{ij}^2$ . Therefore,  $m$  can be thought of as the effective number of independent markers.)

**Proposition 3.** *Assume that genotypes of different SNPs are independent e.g.  $E[x_{ki}x_{kj}] = E[x_{ki}]E[x_{kj}]$  for all  $k$  if  $i \neq j$  and that there is no population structure or cryptic relatedness e.g.  $E[x_{ki}x_{li}] = E[x_{ki}]E[x_{li}]$  for all  $i$  if  $k \neq l$ , then*

$$E[\sum_{i<j} c_{ij}^2] = \frac{(n^2 - n)}{2m},$$

where  $c_{ij}$  is the  $ij$ -th element in  $\mathbf{C}_{ij}^g$

*Proof.* We have

$$\begin{aligned} E[\sum_{i<j} c_{ij}^2] &= \sum_{i<j} E[c_{ij}^2] \\ &= \sum_{i<j} E[(\frac{1}{m} \sum_k (x_{ki}x_{kj}))^2] \\ &= \sum_{i<j} (\frac{1}{m^2} \sum_{kl} E[x_{ki}x_{kj}x_{li}x_{lj}]) \\ &= \sum_{i<j} (\frac{1}{m^2} \sum_k E[x_{ki}^2x_{kj}^2]) \end{aligned}$$

because

$$E[x_{ki}x_{kj}x_{li}x_{lj}] = E[x_{ki}]E[x_{kj}]E[x_{li}]E[x_{lj}] = 0$$

if  $k \neq l$ . Further

$$\begin{aligned}
E\left[\sum_{i < j} c_{ij}^2\right] &= \sum_{i < j} \left(\frac{1}{m^2} \sum_k E[x_{ki}^2] E[x_{kj}^2]\right) \\
&= \sum_{i < j} \left(\frac{1}{m} E[x^2]^2\right) \\
&= \frac{(n^2 - n)}{2m}
\end{aligned}$$

□

Using a similar strategy for  $\mathbf{C}^h$  we get,

**Proposition 4.** *Assume that genotypes of different SNPs are independent e.g.  $E[x_{ki}x_{kj}] = E[x_{ki}]E[x_{kj}]$  for all  $k$  if  $i \neq j$  and that there is no population structure or cryptic relatedness e.g.  $E[x_{ki}x_{li}] = E[x_{ki}]E[x_{li}]$  for all  $i$  if  $k \neq l$ , further assume that  $\exists C_1 < \infty$  s.t.  $E[x_{ki}^4] < C_1 \forall k, i$ . Also assume  $m \rightarrow \infty$  and  $n \rightarrow \infty$  s.t.  $\frac{m}{n} \rightarrow C_2$  with  $0 < C_2 < \infty$ . Then*

$$E\left[\sum_{i < j} c_{ij}^2\right] \rightarrow \frac{3n^2}{2m^2},$$

where  $c_{ij}$  is the  $ij$ -th element in  $\mathbf{C}_{ij}^h$ .

*Proof.*

$$\begin{aligned}
E[\sum_{i<j} c_{ij}^4] &= \sum_{i<j} E[c_{ij}^4] \\
&= \sum_{i<j} E[(\frac{1}{m} \sum_k (x_{ki}x_{kj}))^4] \\
&= \sum_{i<j} \frac{1}{m^4} \sum_{klvw} E[x_{ki}x_{kj}x_{li}x_{lj}x_{vi}x_{vj}x_{wi}x_{wj}] \\
&= \sum_{i<j} (\frac{1}{m^4} (\sum_{kl} 3E[x_{ki}^2x_{kj}^2x_{li}^2x_{lj}^2] - 2 \sum_k E[x_{ki}^4x_{kj}^4])) \\
&= \sum_{i<j} \frac{3}{m^2} (E[x^2]^4 - \frac{1}{m} E[x^2]^4 + \sum_k \frac{E[x_{ki}^4x_{kj}^4]}{3m^2}) \\
&= \sum_{i<j} (\frac{3}{m^2} - \frac{3}{m^3} + \sum_k \frac{E[x_{ki}^4x_{kj}^4]}{m^4}) \\
&\leq \frac{3(n^2 - n)}{2m^2} + \frac{(C_1 - 3)(n^2 - n)}{2m^3},
\end{aligned}$$

where

$$\frac{(C_1 - 3)(n^2 - n)}{2m^3} \rightarrow 0.$$

Also

$$E[\sum_{i<j} c_{ij}^4] \geq \frac{3(n^2 - n)}{2m^2} - \frac{3(n^2 - n)}{2m^3},$$

where

$$\frac{(3)(n^2 - n)}{2m^3} \rightarrow 0.$$

Furthermore

$$\frac{3(n^2 - n)}{2m^2} \rightarrow \frac{3n^2}{2m^2}.$$

□

We see that to achieve equivalent power to estimate the genetic variance of pairwise interactions

compared to the marginal case, we need a factor increase in sample size  $\alpha$  s.t.

$$\alpha \approx \sqrt{\frac{m}{3}}$$

Assuming  $3E5$  independent markers, we would need about 316 times(!) as many individuals to get equivalent power as in the marginal case. While this number seems large, it is not inconceivable that biobank cohorts large enough will be assembled in the near future that would allow to perform this analysis. One suggestion to remedy this is to restrict the number of interactions tested. As an example, it would be possible to estimate interactions between SNPs prioritized using regular GWAS where the number of SNPs to include can be guided by the considerations above. The analysis above ignores elements on the diagonal. The deviation of diagonal elements of  $C \cdot C$  from one are very close to 2 times the elements of  $C$ . Additionally diagonal elements contain  $G \times E$  contributions, and might be fitted separately. While we investigated the situation in the setting where  $\sigma_g^2 \rightarrow 0$  in an analytic fashion, a full treatment of the subject would require to investigate settings where this condition does not hold, potentially via simulation, which might be the subject of future work.



## B. Connection between random effects score test and the *Pascal* sum statistic

In the following, we will try to elucidate the connection between the random effects model score test and the *Pascal* sum statistic. To set up the notation, we first review the score test statistic. Let  $L(\boldsymbol{\theta})$  be the likelihood of some statistical model parametrized by a vector of parameters  $\boldsymbol{\theta}$ . We are interested to test some null hypothesis defined as a particular linear subspace of the parameter space covered by  $\boldsymbol{\theta}$ . Let  $\hat{\boldsymbol{\theta}}_0$  be the maximum likelihood estimate under the null. We define the score of the model at  $\hat{\boldsymbol{\theta}}_0$  as

$$U(\hat{\boldsymbol{\theta}}_0) = \frac{\partial \log(L(\hat{\boldsymbol{\theta}}_0))}{\partial \boldsymbol{\theta}}.$$

Also, we define the Information matrix as

$$I(\hat{\boldsymbol{\theta}}_0) = -E\left(\frac{\partial^2 \log(L(\hat{\boldsymbol{\theta}}_0))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right).$$

From this we can define the score test statistic for the null

$$S(\hat{\boldsymbol{\theta}}_0) = U^T(\hat{\boldsymbol{\theta}}_0)I^{-1}(\hat{\boldsymbol{\theta}}_0)U(\hat{\boldsymbol{\theta}}_0).$$

One can show that

$$S(\hat{\boldsymbol{\theta}}_0) \sim \chi_k^2,$$

where  $k$  is the difference between the rank of the full parameter space and the rank of subspace defined by the null.

Returning to the particular problem we want to investigate, let's assume that we are given a  $n \times 1$  phenotype vector  $\mathbf{y}$  normalized such that  $\sum_i y_i = 0$  and  $\mathbf{y}^T \mathbf{y} = n$ . Further, we have  $m$   $n \times 1$  genotype vectors for SNPs within a gene region of interest all of which are normalized in

the same fashion as the phenotype vector. We use the following random effects model:

$$\mathbf{y} = \sum_{i=1}^m \mathbf{x}_i b_i + \boldsymbol{\epsilon},$$

where  $b_i \sim N(0, \sigma_g)$  and  $\boldsymbol{\epsilon} \sim N_n(0, \sigma_e^2 \mathbf{I}_n)$ . We assume that we do not have access to the original phenotype and genotype data. Rather, we have access to the  $z$ -scores  $z_i = \mathbf{y}^T \mathbf{x}_i / \sqrt{n}$  and estimates of SNP-SNP correlations  $\rho_{ij} = \mathbf{x}_j^T \mathbf{x}_i / n$  for all  $j$  and  $i$  in  $1, \dots, m$ . We want to test whether  $\sigma_g^2 = 0$  using a score test. We first define  $\boldsymbol{\theta} = (\sigma_g^2, \sigma_e^2)^T$ . We need to calculate the score  $U(\hat{\boldsymbol{\theta}}_0)$  for the null that  $\sigma_g^2 = 0$  as well as the inverse of  $I(\hat{\boldsymbol{\theta}}_0)$ . We will show the following connection between the score test and the *Pascal* sum score.

**Proposition 5.** *Given the model and assumption above, let  $\sqrt{S(\hat{\boldsymbol{\theta}}_0)}$  be the square root of the score test statistic for the null hypothesis that  $\sigma_g^2 = 0$ . Then*

$$\sqrt{S(\hat{\boldsymbol{\theta}}_0)} = \frac{\sum_i z_i^2 - m}{\sqrt{\frac{4}{n^2} I^{-1}(\hat{\boldsymbol{\theta}}_0)_{1,1}}}, \quad (1)$$

where  $I^{-1}(\hat{\boldsymbol{\theta}}_0)_{1,1}$  is the first element of the inverse of the fisher information matrix and therefore independent of  $\mathbf{y}$ .

*Proof.* The random effects model is equivalent to

$$\mathbf{y} \sim N_n(\sigma_g^2 \mathbf{C} + \sigma_e^2 \mathbf{I}_n),$$

with

$$\sigma_g^2 \mathbf{C} = \sigma_g^2 \sum_i (\mathbf{x}_i \mathbf{x}_i^T).$$

The log-likelihood of the model is

$$l(\boldsymbol{\theta}) \propto -\frac{1}{2} \log(|\mathbf{C} \sigma_g^2 + \mathbf{I}_n \sigma_e^2|) - \frac{1}{2} \mathbf{y}^T (\mathbf{C} \sigma_g^2 + \mathbf{I}_n \sigma_e^2)^{-1} \mathbf{y}.$$

Let  $\mathbf{C} = \mathbf{\Gamma} \mathbf{\Lambda} \mathbf{\Gamma}^T$  be the spectral decomposition of  $\mathbf{C}$  and define  $\mathbf{y}' = \mathbf{\Gamma}^T \mathbf{y}$ . Then the log-likelihood

becomes

$$l(\boldsymbol{\theta}) \propto -\frac{1}{2} \sum_i \log(\lambda_i \sigma_g^2 + \sigma_e^2) - \frac{1}{2} \sum_i \frac{y_i'^2}{\lambda_i \sigma_g^2 + \sigma_e^2}.$$

Therefore, the first term of the score vector is

$$U(\hat{\boldsymbol{\theta}}_0)_1 = -\frac{1}{2\hat{\sigma}_e^2} \sum_i \lambda_i + \frac{1}{2\hat{\sigma}_e^4} \sum_i y_i'^2 \lambda_i.$$

With

$$\sum_i y_i'^2 \lambda_i = \mathbf{y}^T \mathbf{C} \mathbf{y} = n \sum_i z_i^2$$

and

$$\sum_i \lambda_i = \text{Tr}(\mathbf{C}) = \text{Tr}(\mathbf{X} \mathbf{X}^T) = \text{Tr}(\mathbf{X}^T \mathbf{X}) = nm,$$

we have

$$U(\hat{\boldsymbol{\theta}}_0)_1 = \frac{n}{2\hat{\sigma}_e^4} \sum_i z_i^2 - \frac{nm}{2\hat{\sigma}_e^2}.$$

The second term of the score statistic is

$$U(\hat{\boldsymbol{\theta}}_0)_2 = -\frac{n}{2\hat{\sigma}_e^2} + \frac{1}{2\hat{\sigma}_e^4} \sum_i y_i'^2 = \frac{n}{2\hat{\sigma}_e^2} \left( \frac{1}{\hat{\sigma}_e^2} - 1 \right).$$

because  $\sum_i y_i'^2 = \sum_i y_i^2 = n$ . From the same fact follows that the maximum likelihood estimate of  $\sigma_e^2$  under the null ( $\sigma_g^2 = 0$ ) is 1. Therefore,

$$U(\hat{\boldsymbol{\theta}}_0)_2 = 0.$$

We can now calculate the score test statistic:

$$\begin{aligned} S(\hat{\boldsymbol{\theta}}_0) &= U(\hat{\boldsymbol{\theta}}_0)_1^2 I_{1,1}^{-1}(\hat{\boldsymbol{\theta}}_0) \\ &= \left( \sum_i z_i^2 - m \right)^2 I_{1,1}^{-1}(\hat{\boldsymbol{\theta}}_0) \frac{n^2}{4} \end{aligned}$$

□

To actually calculate  $I_{1,1}^{-1}(\hat{\boldsymbol{\theta}}_0)$ , we note that for a normal distribution of the form  $N_n(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$  is

$$I_{m,n} = \frac{1}{2} \text{Tr}(\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_m} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_n}).$$

Therefore

$$\begin{aligned} I_{1,1}(\hat{\boldsymbol{\theta}}_0) &= \frac{1}{2} \sum_i \lambda_i^2, \\ I_{2,2}(\hat{\boldsymbol{\theta}}_0) &= \frac{n}{2} \\ I_{1,2}(\hat{\boldsymbol{\theta}}_0) &= \frac{1}{2} \sum_i \lambda_i = \frac{nm}{2}. \end{aligned}$$

Therefore

$$I_{1,1}^{-1}(\hat{\boldsymbol{\theta}}_0) = \frac{2}{(\sum_i \lambda_i^2 - nm^2)}.$$

$\sum_i \lambda_i^2$  is not directly available to us because we do not have genotype level data. But we can estimate it from a reference population. Note that

$$\sum_i \lambda_i^2 = \text{Tr}(CC^T) = \text{Tr}(XX^T XX^T) = \text{Tr}((X^T X)(X^T X)) = n^2 \sum_{ij} \rho_{ij}^2.$$

Because we have estimates for  $\rho_{ij}$  from reference populations, we can estimate  $I_{1,1}^{-1}(\hat{\boldsymbol{\theta}}_0)$  efficiently.

## C. Estimating heritability within a genic region via maximum likelihood

In the Appendix B, we have shown a connection between the *Pascal* sum statistic and a score test for a particular random effects model. When instead of testing, we try to estimate the parameters of this random effect model, we essentially estimate the contribution of this genic region to the additive genetic variance. In the following, we will present a procedure to do so. Again, let's make the same assumptions as in Appendix B. However, now we will try to estimate  $\boldsymbol{\theta}$  via maximum likelihood. As before the log-likelihood of the model is

$$l(\boldsymbol{\theta}) \propto -\frac{1}{2} \log(|\mathbf{C}\sigma_g^2 + \mathbf{I}_n\sigma_e^2|) - \frac{1}{2} \mathbf{y}^T (\mathbf{C}\sigma_g^2 + \mathbf{I}_n\sigma_e^2)^{-1} \mathbf{y}.$$

We will further assume that  $m < n$  (a reasonable assumption as the length of  $\mathbf{y}$  is quite typically larger than the number of SNPs in a given genic region). Under this assumption,  $\mathbf{C}$  not full rank. Again, we neither have access to  $\mathbf{C}$  nor  $\mathbf{y}$ . Rather, we have access to an estimate of  $\Sigma$  and  $\mathbf{z}$  where

$$\begin{aligned} \Sigma &= \mathbf{X}^T \mathbf{X} \\ \mathbf{z} &= \frac{1}{\sqrt{n}} \mathbf{X}^T \mathbf{y} \end{aligned}$$

We will try to define the maximum likelihood from these quantity accessible to us.

**Proposition 6.** *Let  $\Sigma = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$  be the eigenvalue decomposition of  $\Sigma$ . Then we can formulate the log-likelihood as*

$$l(\boldsymbol{\theta}) \propto -\frac{1}{2} \sum_{i=1}^n \log(\lambda_i \sigma_g^2 + \sigma_e^2) - \frac{n}{2\sigma_e^2} \left(1 - \sum_{i=1}^m \frac{z'_i{}^2}{\lambda_i}\right) - \frac{n}{2\sigma_e^2} \sum_{i=1}^m \frac{z'_i{}^2}{\lambda_i(\lambda_i \sigma_g^2 + \sigma_e^2)},$$

where  $\mathbf{z}' = \mathbf{V}^T \mathbf{z}$  with  $z'_i$  being the  $i$ -th element of  $\mathbf{z}'$  and  $\lambda_i$  being the  $i$ -th diagonal element of  $\boldsymbol{\Lambda}$  if  $i \leq m$  and 0 otherwise.

*Proof.* Define  $\mathbf{X} = \boldsymbol{\Gamma}\mathbf{D}\mathbf{V}^T$  as is the singular value decomposition of the genotype matrix, where

$\Gamma$  is a  $n \times m$ -matrix. We have

$$\sqrt{n}\mathbf{D}^{-1}\mathbf{z}' = \Gamma^T\mathbf{y}.$$

Set  $\mathbf{U}^\perp$  as some  $(n - m) \times m$  matrix composed of orthonormal vectors orthogonal to  $\mathbf{U}$ . Since

$$\mathbf{y}^T(\mathbf{C}\sigma_g^2 + \mathbf{I}_n\sigma_e^2)^{-1}\mathbf{y} = \mathbf{y}^T\Gamma(\mathbf{D}^2\sigma_g^2 + \mathbf{I}_m\sigma_e^2)^{-1}\Gamma^T\mathbf{y} + \frac{1}{\sigma_e^2}\mathbf{y}^T(\mathbf{U}^\perp)(\mathbf{U}^\perp)^T\mathbf{y},$$

with

$$\begin{aligned} \mathbf{y}^T\Gamma(\mathbf{D}^2\sigma_g^2 + \mathbf{I}_m\sigma_e^2)^{-1}\Gamma^T\mathbf{y} &= n\mathbf{z}'^T\mathbf{D}^{-1}(\mathbf{D}^2\sigma_g^2 + \mathbf{I}_m\sigma_e^2)^{-1}\mathbf{D}^{-1}\mathbf{z}' \\ &= n\sum_{i=1}^m \frac{z_i'^2}{\lambda_i(\lambda_i\sigma_g^2 + \sigma_e^2)}, \end{aligned}$$

and

$$\begin{aligned} \mathbf{y}^T(\mathbf{U}^\perp)(\mathbf{U}^\perp)^T\mathbf{y} &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{U}\mathbf{U}^T\mathbf{y} \\ &= n - n\mathbf{z}'^T\mathbf{D}^{-2}\mathbf{z}' \\ &= n\left(1 - \sum_{i=1}^m \frac{z_i'^2}{\lambda_i}\right). \end{aligned}$$

Further

$$\log(|\mathbf{C}\sigma_g^2 + \mathbf{I}_n\sigma_e^2|) = \sum_{i=1}^n \log(\lambda_i\sigma_g^2 + \sigma_e^2).$$

Assembling all terms we get the expression for the likelihood above. □

Because we can express the likelihood in terms of known or estimable quantities, we can now optimize the likelihood function easily via some numerical optimization routine such as newton-raphson. Alternatively, one could approximate the maximum likelihood by a second degree taylor polynomial around  $\sigma_g^2 = 0$  and maximize the resulting equation analytically.

## **D. Supplementary Information for Chapter 2**

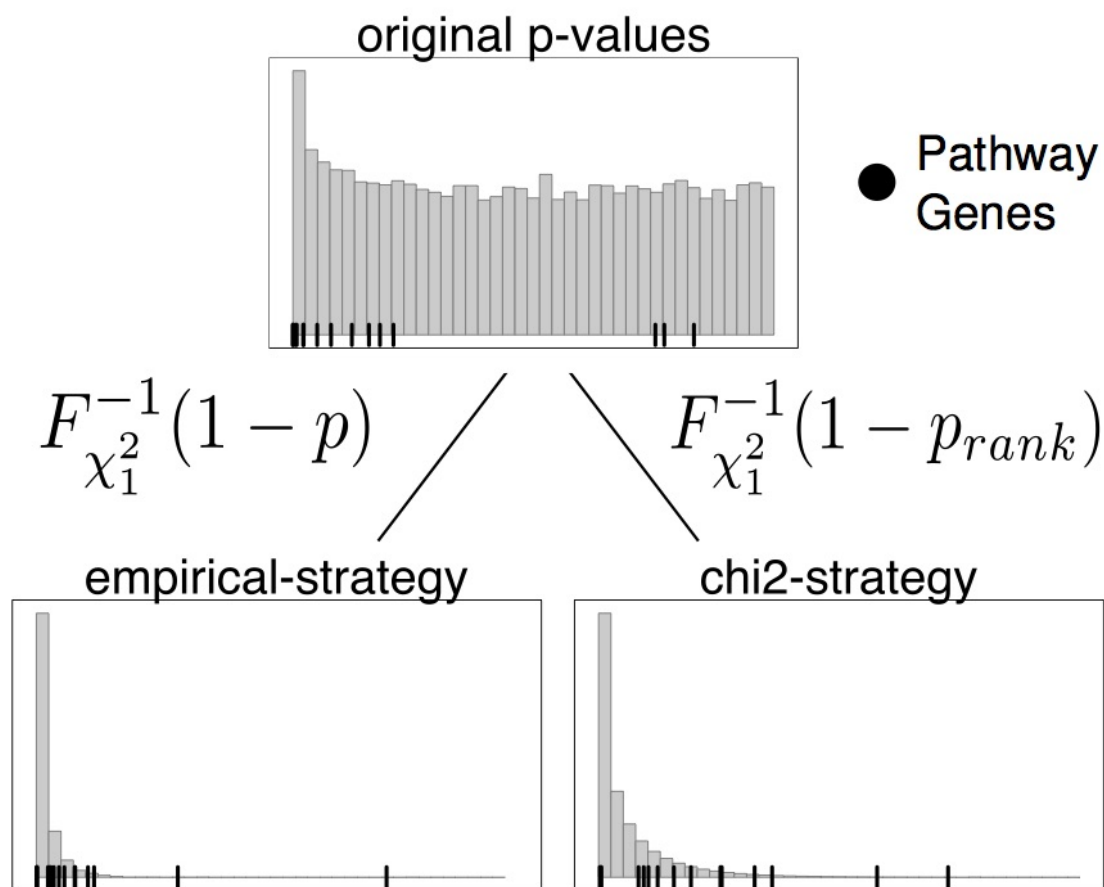


Fig S1: Overview of pathway scoring strategies in *Pascal*.

Pathway scores are computed from gene scores. The upper panel shows a typical gene score distribution, where the pathway gene scores are indicated in black. In order to compute pathway scores, the original gene score  $p$ -values need to be transformed. To this end we use one of two strategies: in our ‘empirical strategy’ (lower left panel), gene score  $p$ -values are directly transformed with the inverse -quantile function to obtain chi-square scores, which are then summed across all pathway genes. A Monte Carlo estimate of the  $p$ -value is then obtained by sampling random gene sets of the same size and calculating the fraction of sets reaching a higher score than that of the given pathway. In the ‘chi-square method’ (bottom right panel), the gene score  $p$ -values are first ranked such that the lowest  $p$ -value ranks highest. The rank values are then divided by the number of genes plus one to define new  $p$ -values ( $p$ -rank) that are distributed uniformly by definition. From there, we proceed as for the empirical strategy just replacing  $p$  by  $p$ -rank. Also, since the scores are guaranteed to be chi-square distributed, the computation of their corresponding  $p$ -value can be done analytically without any loss in precision.



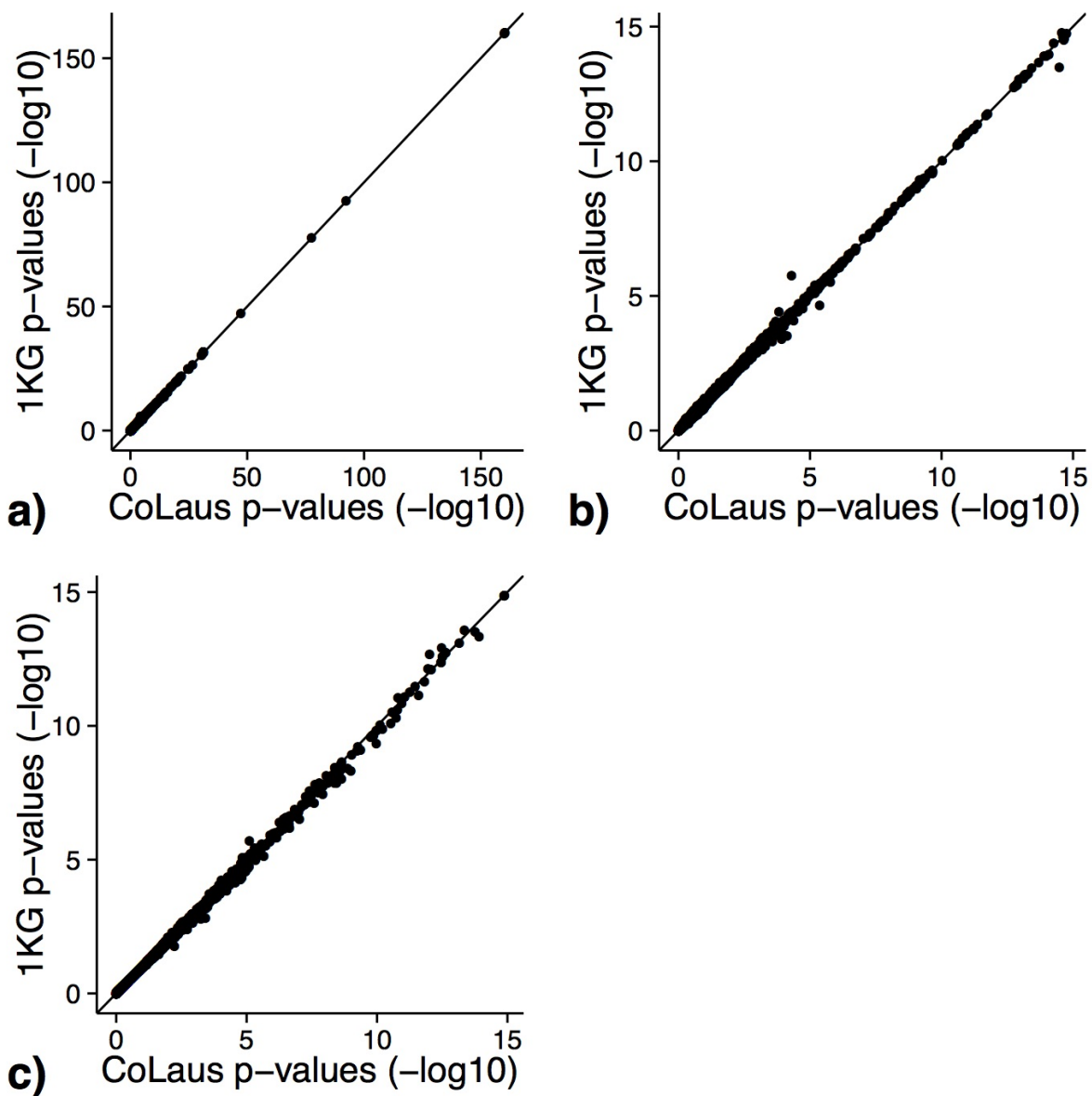


Fig S2: Comparison of results for different reference panels.

Comparing  $p$ -values using of European 1000 Genomes project reference panel to calculate LD-matrix versus those using the measured from CoLaus. GWAS summary statistics were taken from a large-scale blood-HDL level meta-analysis. a) comparing results for max gene scores. b) comparing results for max gene scores removing gene scores that were computed with the effective number of tests approximation. c) comparing results for sum gene scores. There is good concordance in all cases.

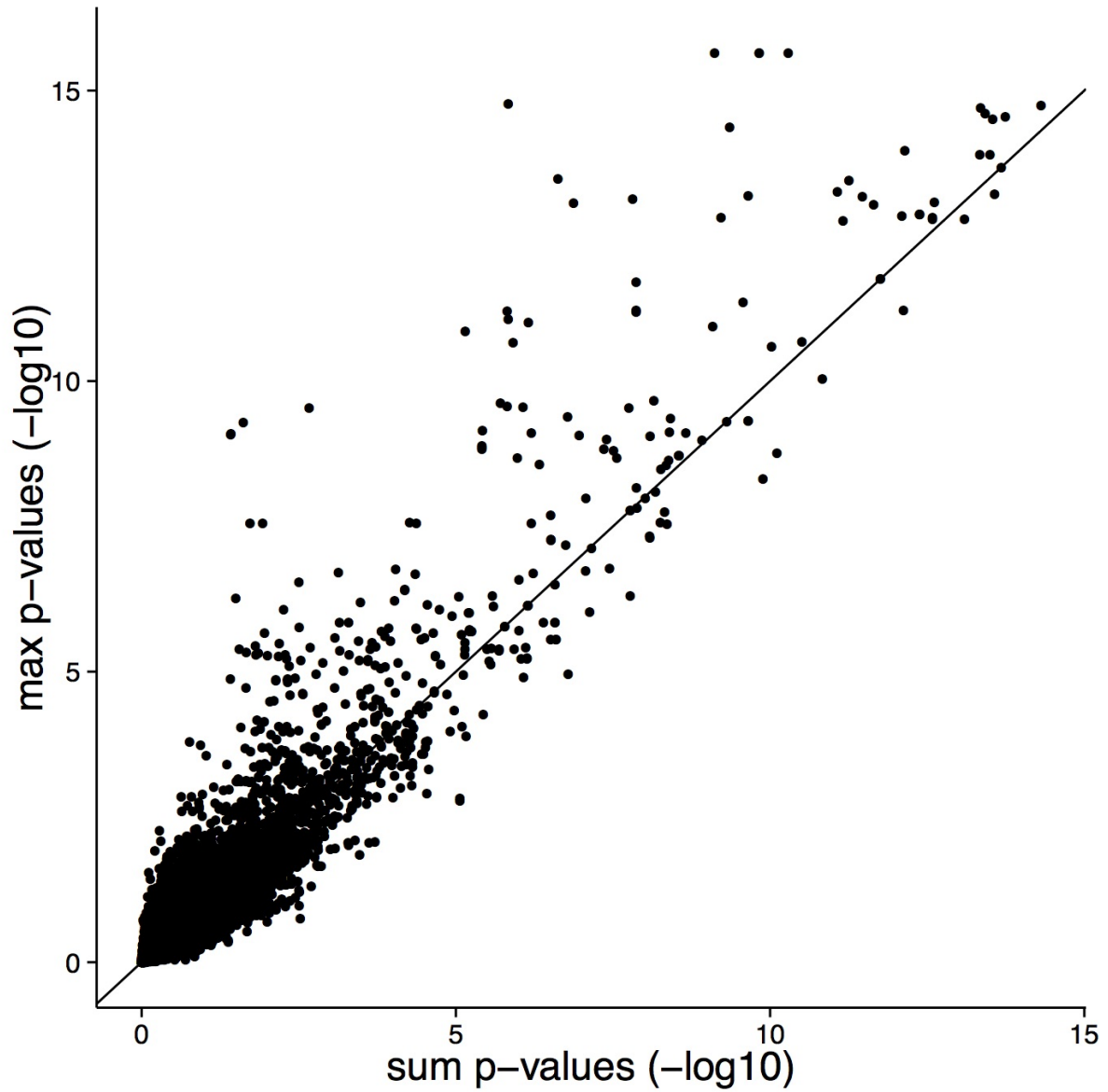


Fig S3: Comparison of max and sum gene scores.

We compared max and sum gene scores directly for a large-scale blood-HDL level meta-analysis. Only gene scores up to  $10^{-15}$  are displayed, which truncated 6 genes with very large max scores.  $R^2$  between the  $-\log_{10}$ -transformed variables is 90%. We see max scores tend to be larger when the two methods do not agree.

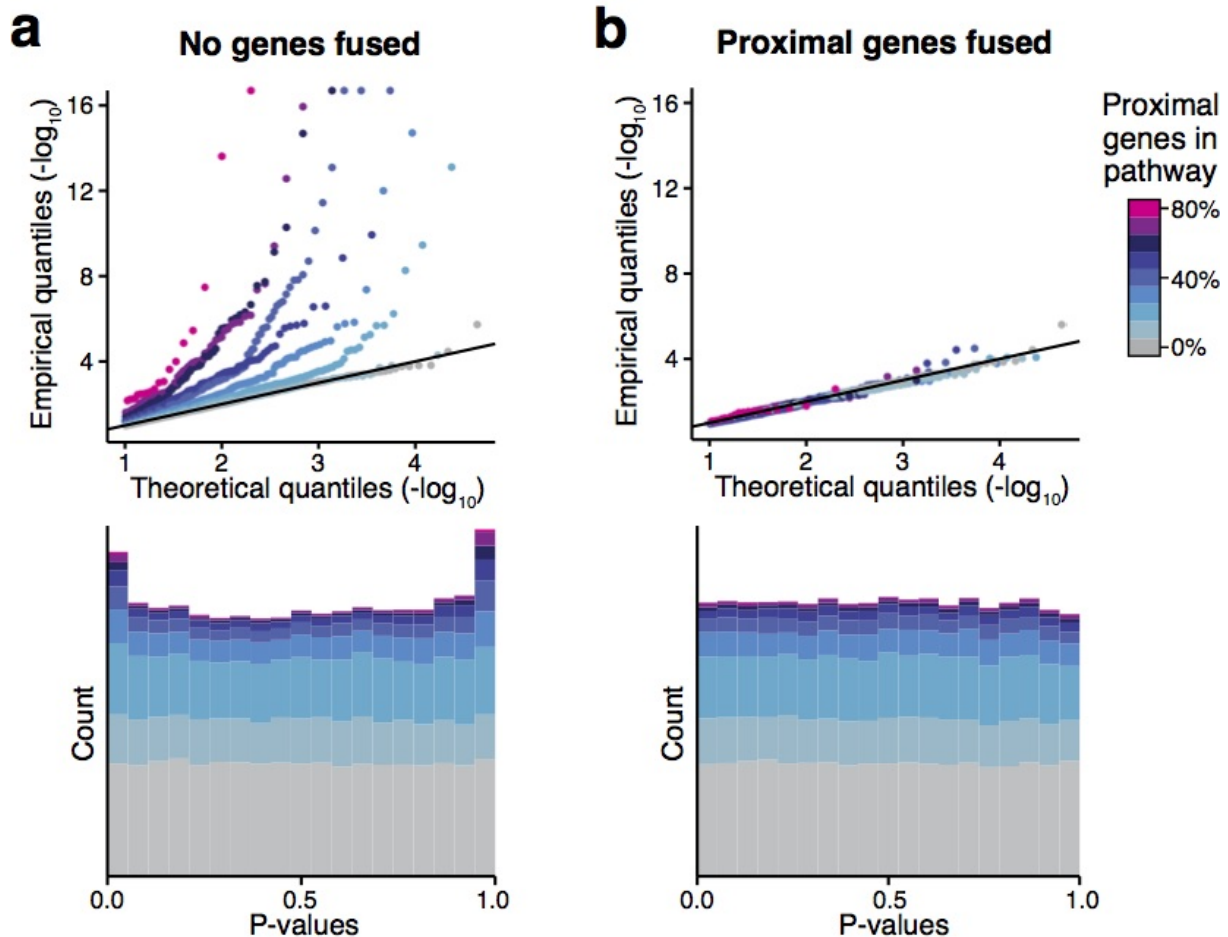


Fig S4: Pathway scores for random phenotypes using max gene scores.  $p$ -values for 1077 pathways from our pathway library were computed for 100 random phenotypes using the *Pascal* pipeline using max gene scores and chi2-pathway integration strategy. a) Without merging of neighbouring genes and (b) with merging of neighbouring genes (gene-fusion strategy).  $p$ -value distributions are represented by QQ-plots (upper panels) and histograms (lower panels). Results are colour-coded according to the fraction of genes in a given pathway that have a neighbouring gene in the same pathway, i.e. that are located nearby on the genome (distance  $<300\text{kb}$ ). a)  $p$ -values of pathways that contain genes in LD are strongly inflated without correction. b) The gene fusion approach provides well-calibrated  $p$ -values independently of the number of pathway genes in LD.

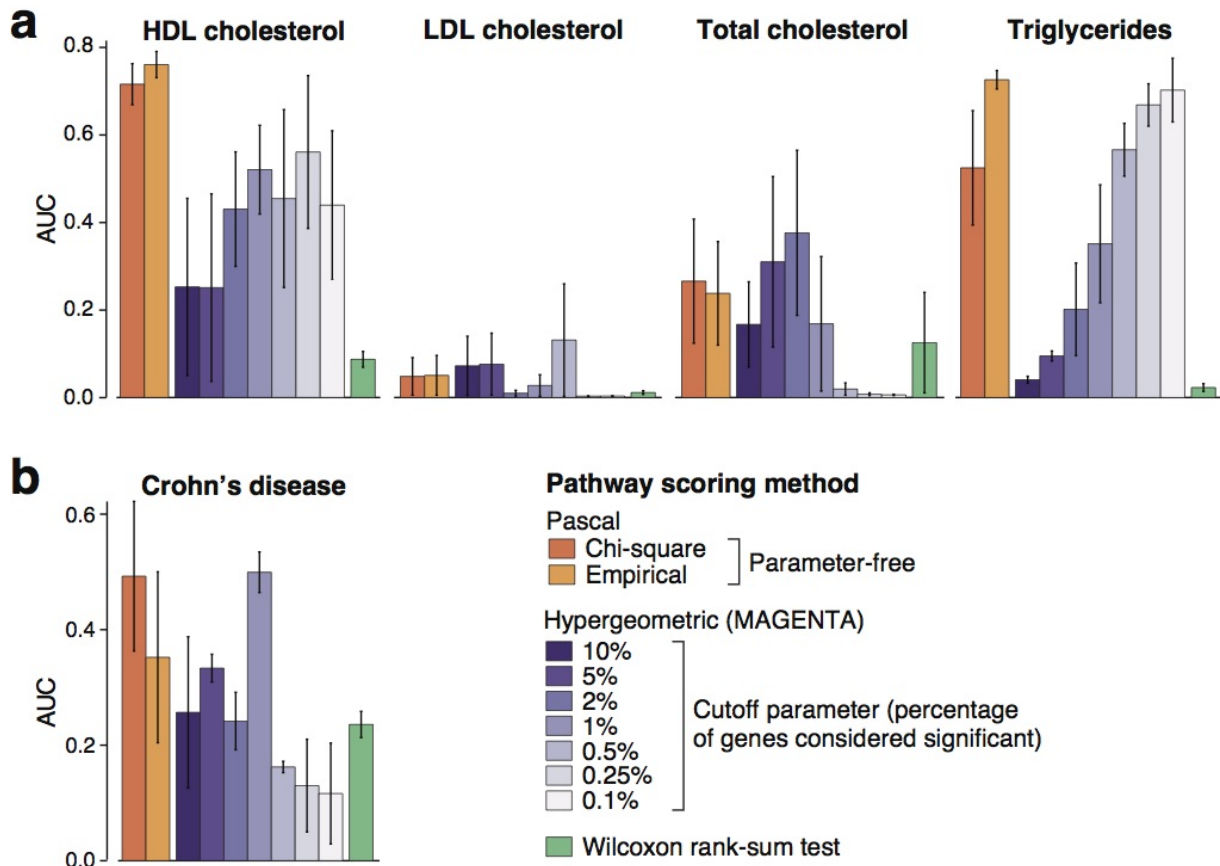


Fig S5: Performance of pathway enrichment methods for blood lipid traits and Crohn's disease using sum of squares (SOCS) statistics for defining gene scores.

Displayed is the mean area under the precision-recall curve (AUC) for pathways identified using Pascal, a standard hypergeometric test at various gene score thresholds, and a rank-sum test (vertical bars show the standard error). We show results for the SOCS gene scores (MOCS gene score results are similar, see Figure 4 in the main text). a) Results for four blood lipid traits. A reference standard pathway list was defined as all pathways that show a significance level below  $5 \cdot 10^{-06}$ , for any of the tested threshold parameters for hypergeometric tests in the largest study of lipid traits to date. The significance level of  $5 \cdot 10^{-06}$  corresponds to the Bonferroni corrected, genome-wide significance threshold at the 0.5% level for a single method. For each phenotype, error bars denote the standard error computed from three independent subsamples of the CoLaus study (including 1500 individuals each). We see good overall performance of Pascal pathway scores, whereas results for discrete gene sets vary widely with the particular choice for the threshold parameter of hypergeometric test. b) Results for Crohn's disease using the same approach as in (a). A reference standard pathway list was defined as all pathways that show a significance level below  $5 \cdot 10^{-06}$  for any of the tested threshold parameters for hypergeometric tests in the largest study of Crohn's disease traits to date. We observe that the chi-squared strategy outperforms all other strategies in this setting, whereas performance of the hypergeometric testing strategy varies.

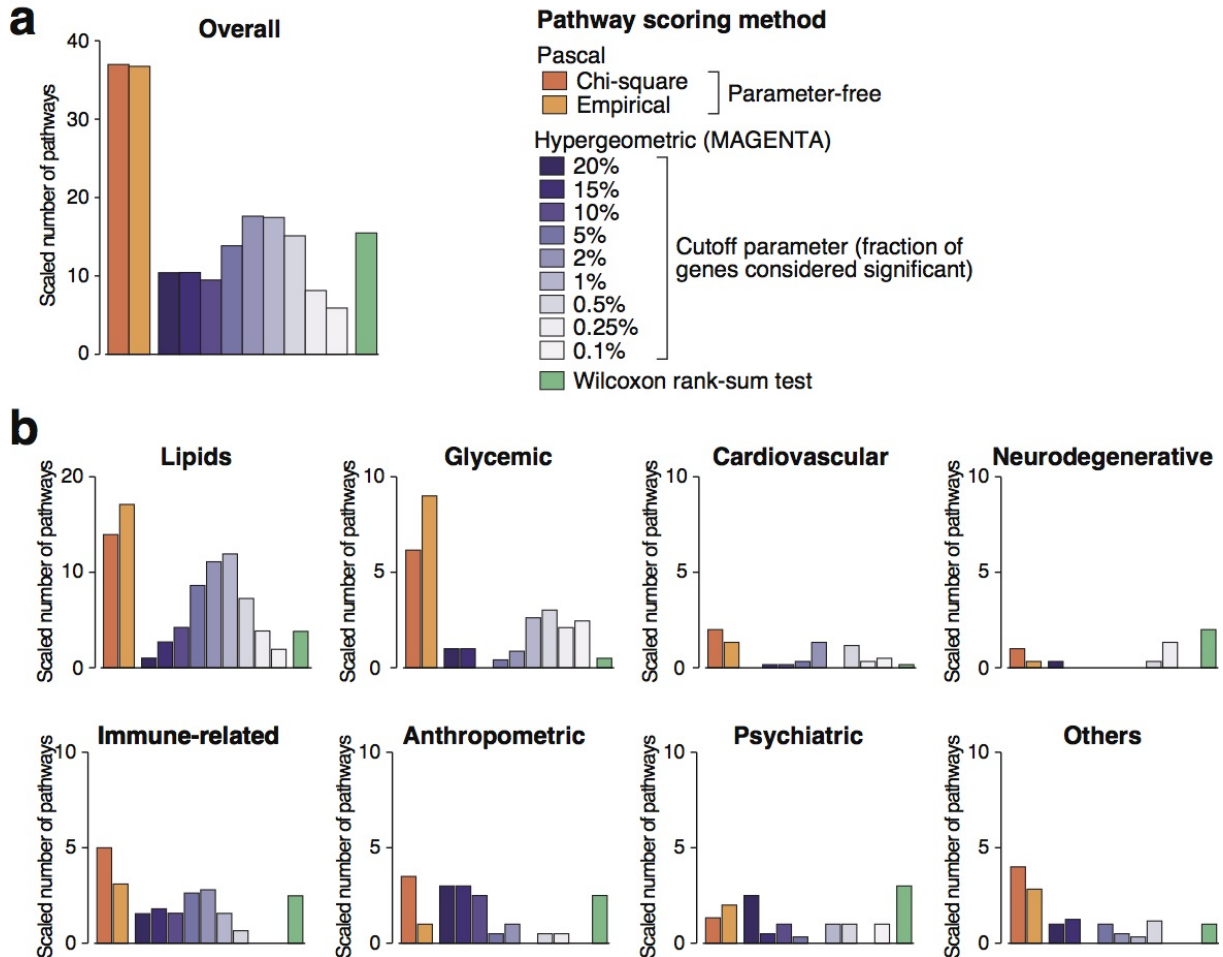


Fig S6: Power of pathway scoring methods across diverse traits and diseases using sum of squares (SOCS) statistics for defining gene score.

Bar heights represent the number of pathways found to be significant after Bonferroni correction. Within a given trait group, results are aggregated for all tested GWAS studies. 65 GWASs had at least one significant pathway in one of the tested method. For each GWAS, the raw number of significant pathways was divided by the number of pathways found by the best performing method. This was done to avoid that a few studies with many emerging pathways dominate. We show results for the SOCS gene scores (MOCS gene score results are similar, see Figure 5). a) Results are aggregated over all trait groups. b) Results for different trait groups.

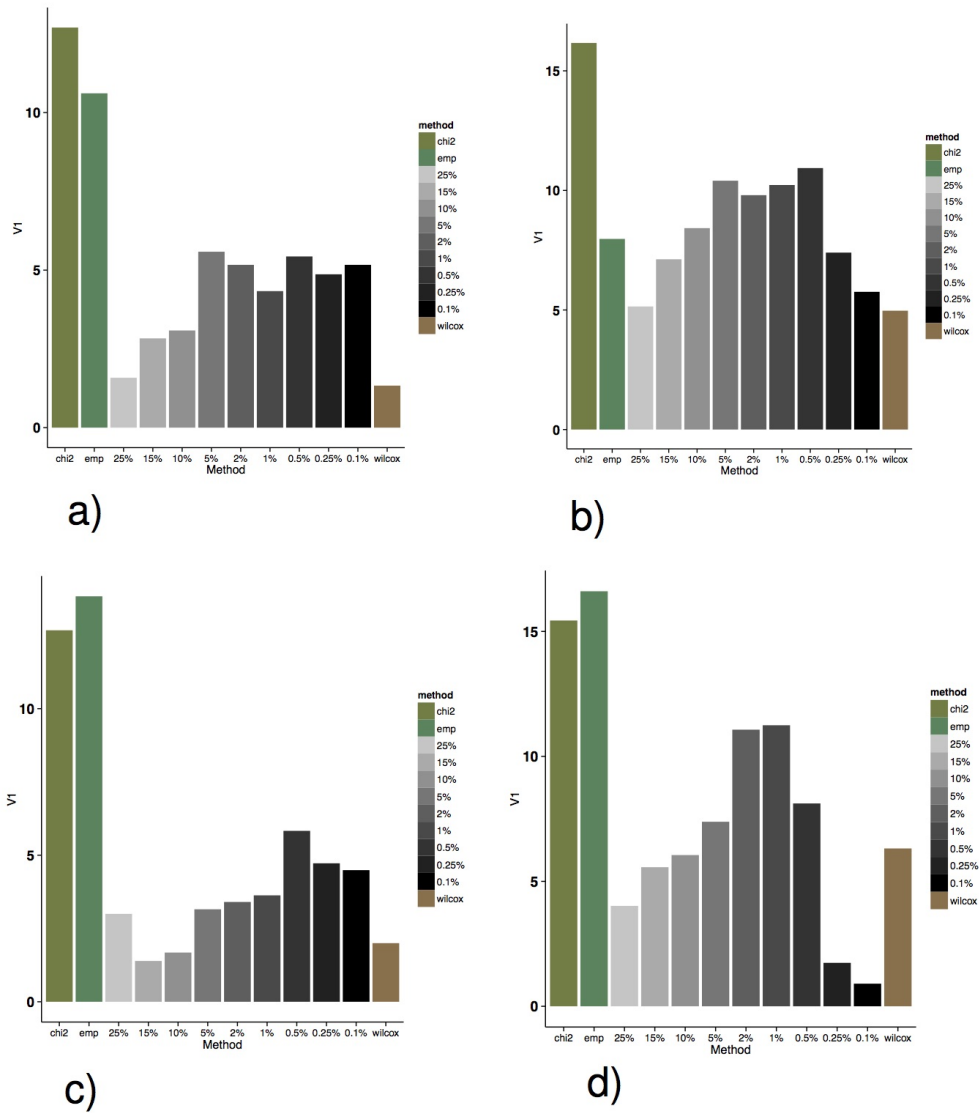


Fig S7: Power of pathway scoring methods stratified with respect to sample size. Only GWAS studies for quantitative traits were used. Top panels (a,b) show results for max gene scores and bottom panels (c,d) show results for sum gene scores. Left and side panels (a,c) show results for all studies where the number of individuals was below 50'000. Right hand side panels (b,d) show results for studies with sample sizes above 50'000. We see power gains in all cases. The improvements are particularly pronounced in lower powered GWAS.

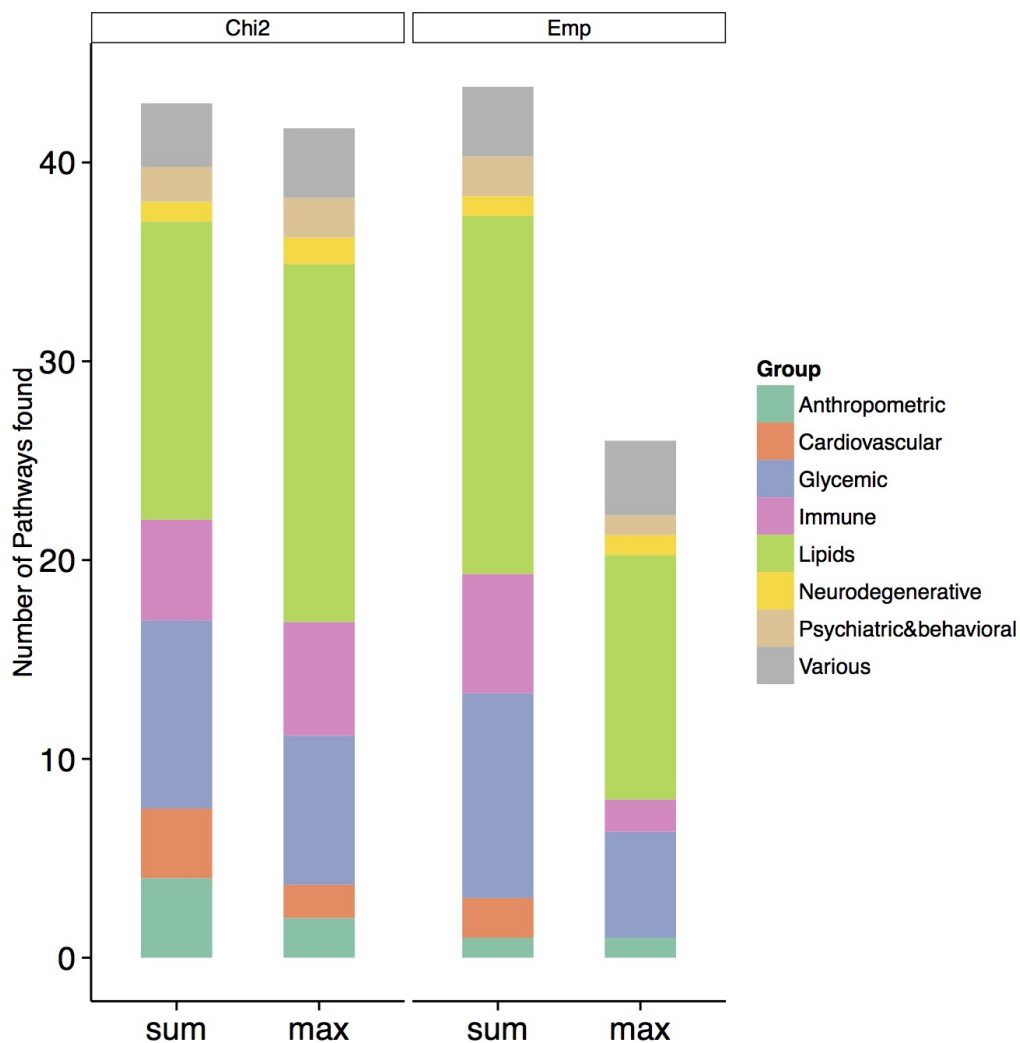


Fig S8: Power comparison max and sum gene scores for pathway analysis.

Bar heights represent the number of pathways found to be significant after Bonferroni correction. Within a given trait group, results are aggregated for all tested GWAS studies. For each GWAS, the raw number of significant pathways was divided by the number of pathways found by the best performing method. Results for SOCS and MOCS as well as the chi-square and empirical pathway scores are displayed. We observe a drop in performance for the combination of MOCS gene scores with empirical pathway scores.

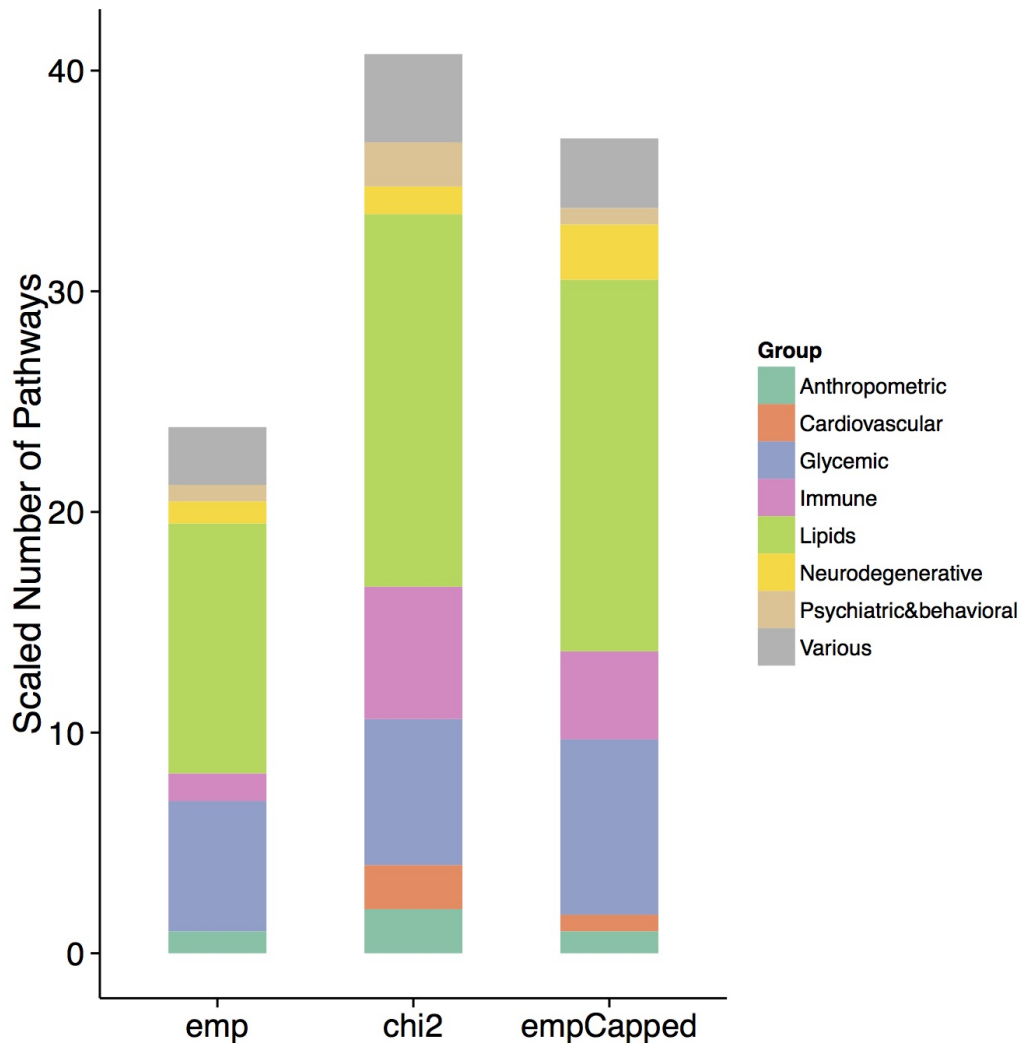
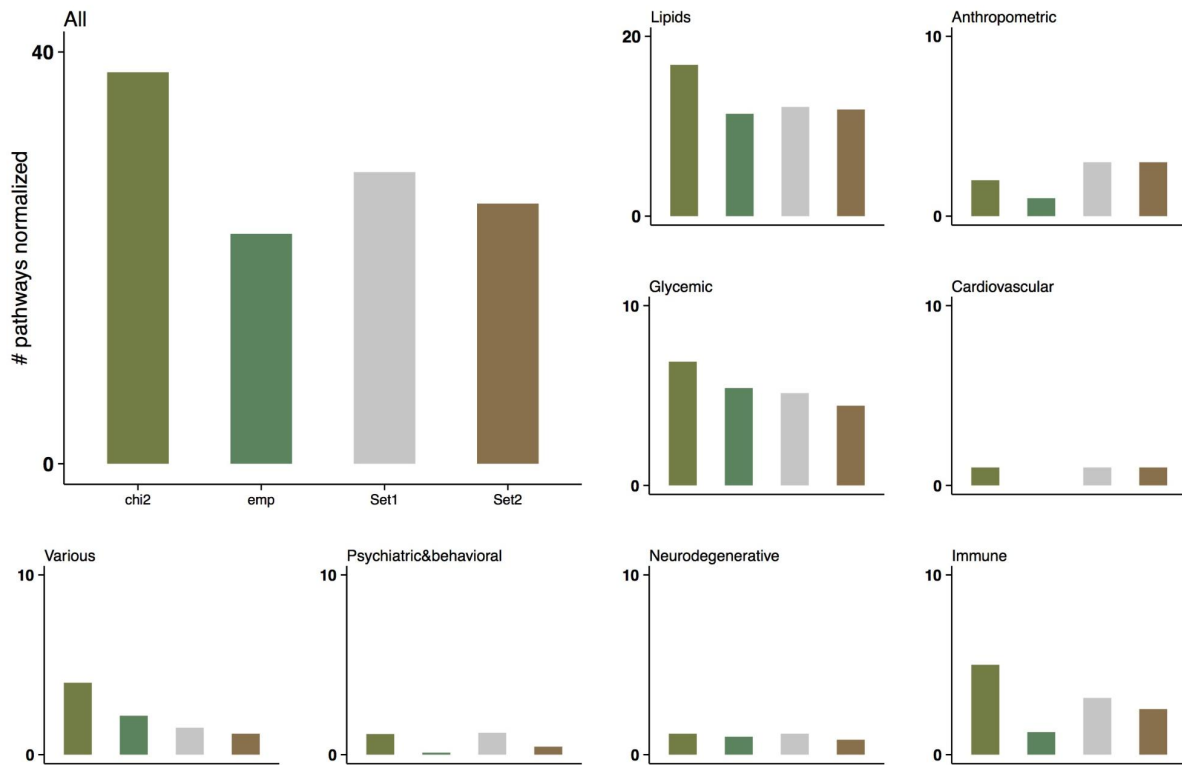


Fig S9: Power analysis for max gene scores with capped gene scores.

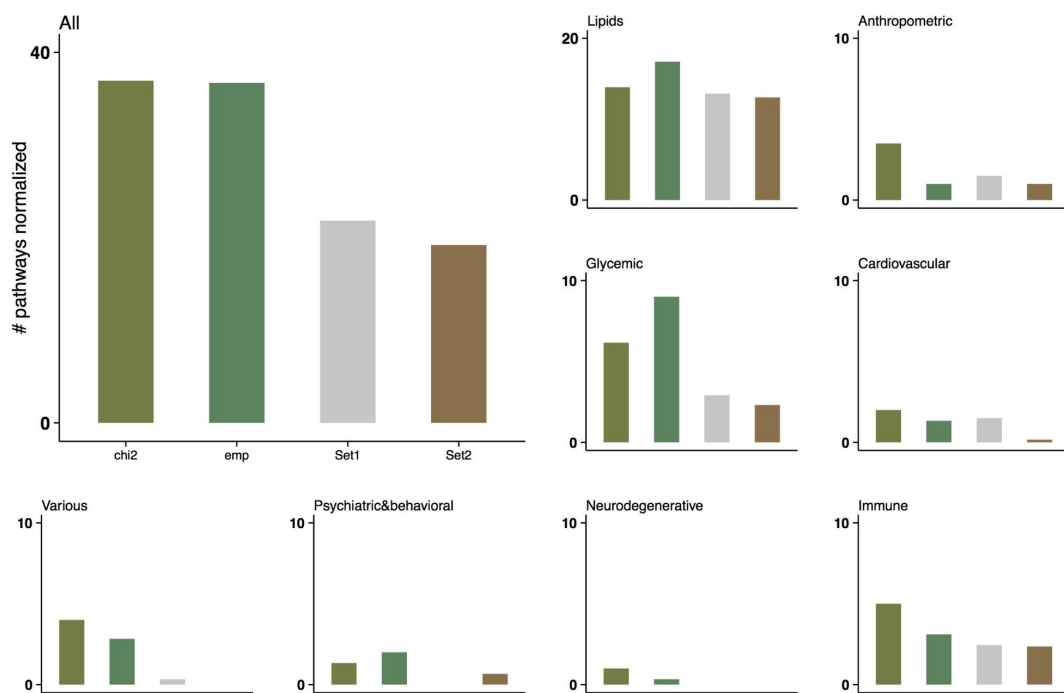
Bar heights represent the number of pathways found to be significant after Bonferroni correction. Within a given trait group, results are aggregated for all tested GWAS studies. For each GWAS, the raw number of significant pathways was divided by the number of pathways found by the best performing method. Max gene scores using empirical sampling pathway scores (emp) and chi-squared pathway scores (chi2) are compared to max gene scores combined with empirical sampling, where outlier gene scores ( $p$ -value  $< 10^{-12}$ ) are set to  $10^{-12}$  (empCapped). We chose the capping value such that the maximum  $-\log_{10}$   $p$ -value was roughly in the middle between genome wide significance threshold (8) and the maximum value that can be calculated for the sum statistic (15).





S10 Fig. Power of *Pascal* pathway scoring methods compared to aggregated hypergeometric scores (MOCS)

The same data as in Figure 5 is plotted here. However, instead comparing *Pascal* pathway scoring methods with results for all hypergeometric threshold separately, we defined a new aggregated pathway score that picks the optimal threshold for each pathway over a range of hypergeometric threshold and correcting for the multiple number of tests by Bonferroni correction. Results for different sets of thresholds are displayed. Set1 refers to the complete set of thresholds (i.e.: 25%, 15%, 10%, 5%, 2%, 1%, 0.25%, 0.1%). Set2 refers to a set with thresholds more ‘spread out’ (i.e.: 25%, 5%, 1%, 0.25). We see that *Pascal* has better performance, except when combining the ‘empirical sampling’ pathway scoring method with max gene scores.



S11 Fig. Power of *Pascal* pathway scoring methods compared to aggregated hypergeometric scores (SOCS)

The same data as in Figure 5 is plotted here. However, instead comparing *Pascal* pathway scoring methods with results for all hypergeometric threshold separately, we defined a new aggregated pathway score that picks the optimal threshold for each pathway over a range of hypergeometric threshold and correcting for the multiple number of tests by Bonferroni correction. Results for different sets of thresholds are displayed. Set1 refers to the complete set of thresholds (i.e.: 25%, 15%, 10%, 5%, 2%, 1%, 0.25%, 0.1%). Set2 refers to a set with thresholds more ‘spread out’ (i.e.: 25%, 5%, 1%, 0.25). We see that Pascal has better performance.

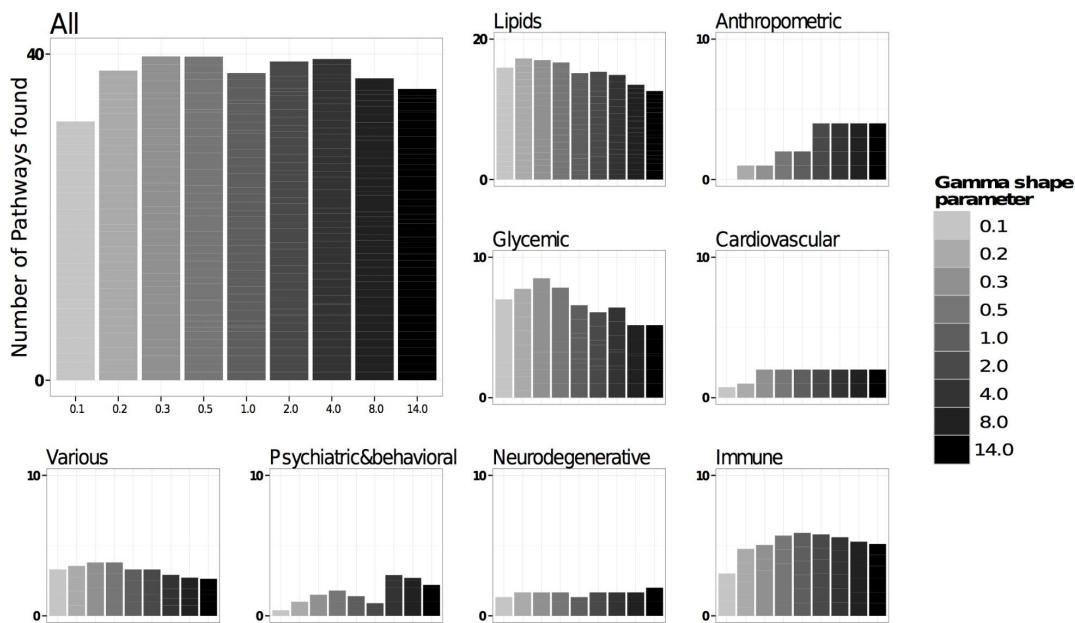


Fig S12: Power of gamma distribution for pathway analysis (MOCS).

Bar heights represent the number of pathways found to be significant after Bonferroni correction. Different bars signify results for a different gamma shape parameter value. For each GWAS, the raw number of significant pathways was divided by the number of pathways found by the best performing method. Upper left panel 'All' refers to all traits stacked. We present here MOCS gene score based results. 52 GWA studies showed at least one significant pathway in one of the evaluated scenarios.

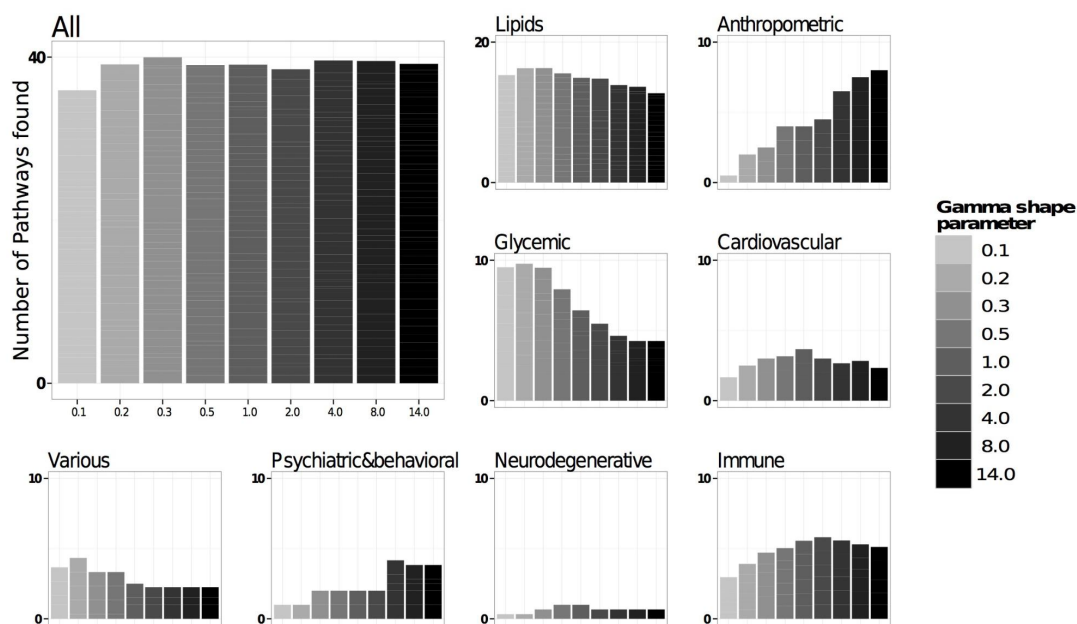


Fig S13: Power of gamma distribution for pathway analysis (SOCS).

Bar heights represent the number of pathways found to be significant after Bonferroni correction. Different bars signify results for a different gamma shape parameter value. For each GWAS, the raw number of significant pathways was divided by the number of pathways found by the best performing method. Upper left panel 'All' refers to all traits stacked. We present here SOCS gene score based results. 60 GWA studies showed at least one significant pathway in one of the evaluated scenarios.

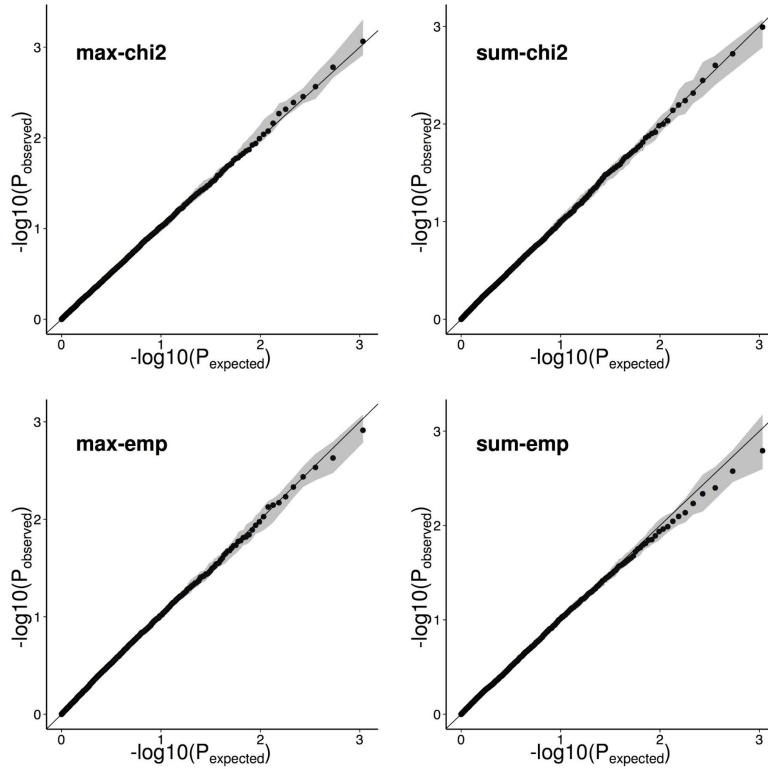


Fig S14: Distribution of pathway scores for simulated phenotypes influenced by causal SNPs in coding regions.

We first sampled 50 random SNPs assayed in CoLaus in or close to coding regions. Using the genotypes of the *CoLaus* study we then simulated phenotypes by adding up the sampled 50 SNPs with a normally distributed effect size with a variance of 0.04 plus Gaussian noise (with a variance of 1). We then ran GWAS for the simulated phenotype to obtain association summary statistics. The experiment was repeated 50 times. On average, this resulted in 18 independent, genome-wide significant gene score hits for each simulated GWAS (for the MOCS statistic). We applied *Pascal* to compute pathway scores for each of the 50 simulated GWASs. We found that the resulting pathway scores are well calibrated, i.e., they do not show inflation or deflation regardless of the setting used (max or sum gene score, chi2 or empirical enrichment test). The QQ-plots show the median value for each quantile across the 50 simulated GWASs. The shaded areas correspond to 95% confidence intervals for the median (estimated from 2000 bootstrap samples of size 50, with replacements). Similar results were obtained when varying the type and number of simulated causal SNPs and their effect size.

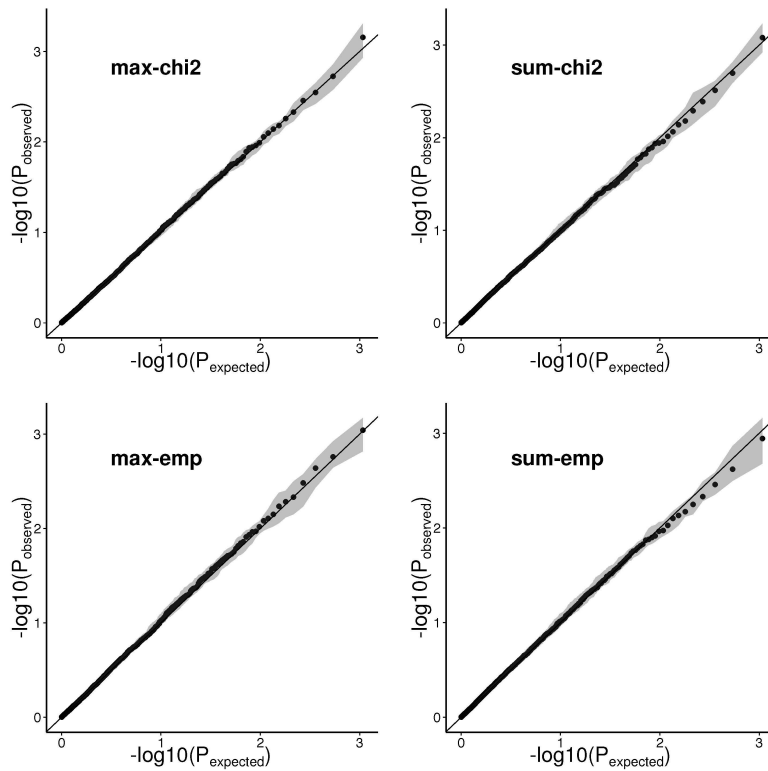


Fig S15: Distribution of pathway scores for simulated phenotype influenced by causal SNPs in coding and non-coding regions.

These QQ-plots correspond to an analysis equivalent to that of Fig S14 but with 50 SNPs chosen uniformly from all SNPs assayed in CoLaus, rather than from genic regions only. On average, this resulted in 12 independent, genome-wide significant gene score hits for each simulated GWAS (using the MOCS statistic). Note that this does not completely exclude the possibility of less well-calibrated scores in other settings. Deviations from perfectly calibrated scores may occur in the cases where true SNP associations are present, because the gene wise test statistic may have varying power for different genes depending on the genetic architecture of the associated phenotype and on certain gene properties (such as gene length, LD structure, SNP coverage, or SNP allele frequency). If a set of pathways contains many pathways enriched (or depleted) for genes with such confounding factors, inflation or deflation is possible.

## E. Supplementary Information for Chapter 3

### E.1. Supplementary Figures

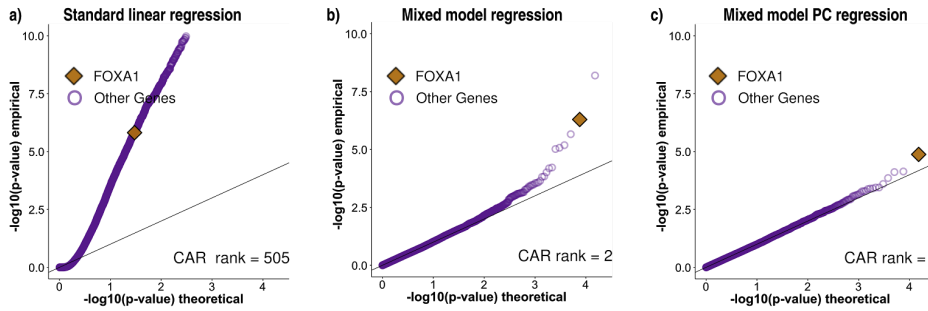


Fig S1: Association between motif accessibility and mRNA expression for the bona fide pioneer factor *FOXA1*.

Three different regression models (a-c) were used to compute association  $p$ -values between the accessibility of a given TF motif (here *FOXA1*) and mRNA expression for each of the assayed 15K protein-coding genes. Results are visualized as qq-plots showing the  $-\log_{10}$  transformed  $p$ -values. a) Association  $p$ -values obtained using standard linear regression. Due to confounding,  $p$ -values are strongly inflated and *FOXA1* motif accessibility shows only mild association with *FOXA1* expression compared to other genes. b) The linear mixed model (LMM) successfully corrects for confounding, with most  $p$ -values following the null distribution as expected. The association between *FOXA1* motif accessibility and *FOXA1* expression now ranks second among all genes and first among all TFs, although it does not pass the Bonferroni significance threshold. c) Additionally controlling for the first principal component of the motif accessibility matrix corrects for a strong batch effect (Methods), further lowers the CAR rank. Using this approach, *FOXA1* motif accessibility showed the strongest association precisely with *FOXA1* expression (i.e., the gene-level CAR rank equals one), in line with literature on *FOXA1* being a pioneer factor [67, 68]

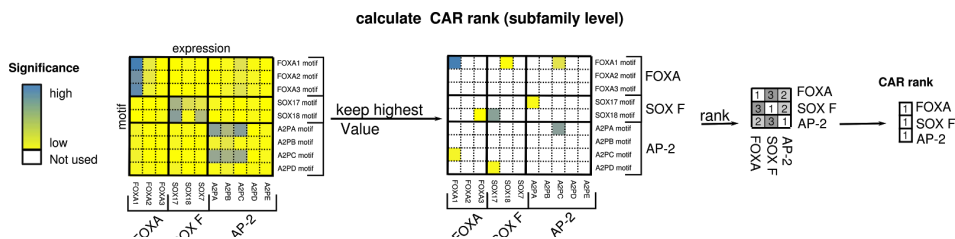


Fig S2: Overview of procedure to calculate CAR ranks on the subfamily level.

We cluster TFs and motifs according to subfamily definitions given in TFClass. For each bicluster, we define the bicluster score as the most significant  $p$ -value between any TF and motif members of the bicluster corrected for bicluster size. We then rank bicluster scores across the TF subfamilies. If the bicluster joining a TF cluster and its corresponding motifs is ranked low, this is an indication of CAR activity.

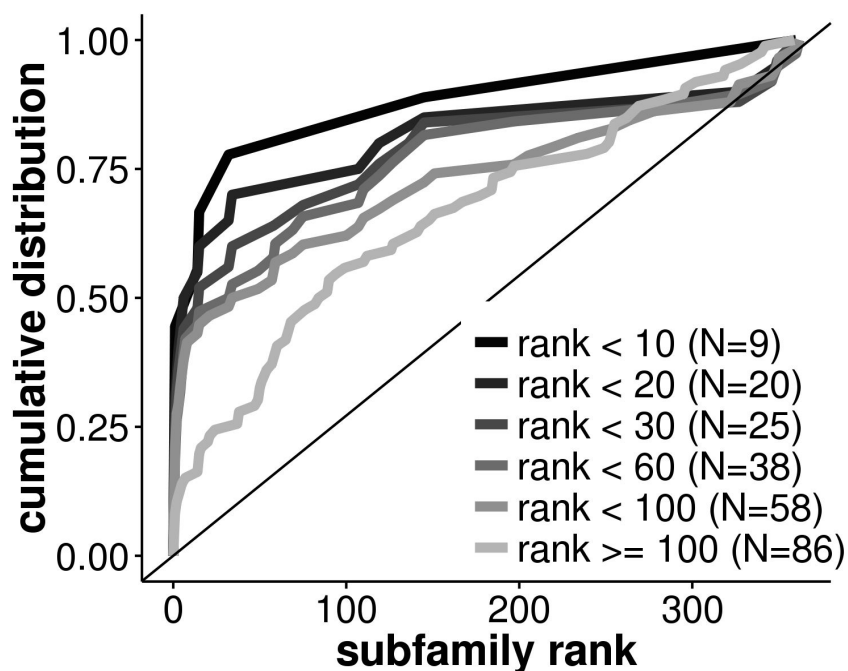


Fig S3: CARs predicted from ENCODE Data set enrich in subfamilies with low CAR ranks in the ROADMAP dataset.

DHS and expression data available for 56 samples (29 with assayed DHS and 27 with imputed DHS) as part of the ROADMAP data collection were used to predict CARs. Shown are CAR enrichment curves for ENCODE results stratified by CAR ranks derived from ROADMAP. Displayed are the following strata: ROADMAP CAR rank <10 (N=9 observations in total), ROADMAP CAR rank <20 (N=20 observations in total), ROADMAP CAR rank <30 (N=25 observations in total), ROADMAP CAR rank <60 (N=38 observations in total), ROADMAP CAR rank <100 (N=58 observations in total), ROADMAP CAR rank  $\geq$ 100 (N=86 observations in total). We see that subfamilies with low ROADMAP CAR rank are also tend to be predicted to be CARs when using the ENCODE data. This enrichment get weaker for subfamilies with lower ROADMAP CAR ranking.



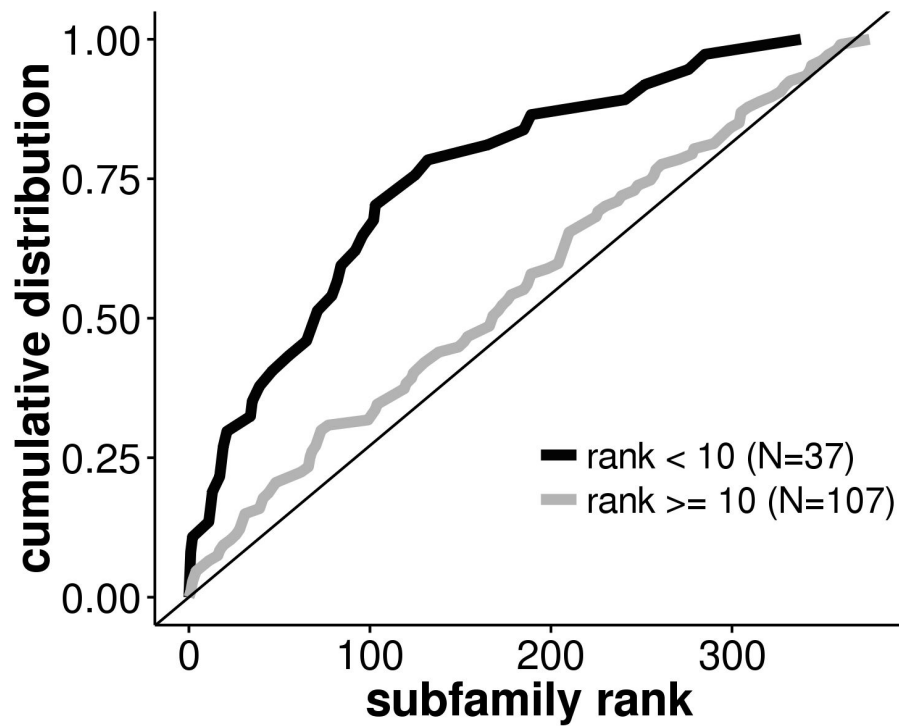


Fig S4: CARs ranks from ROADMAP data set enrich only in subfamilies predicted to be CARs in ENCODE.

DHS and expression data available as part of the ROADMAP data collection were used to predict CARs. Shown are CAR enrichment curves for ROADMAP results stratified by CAR predictions derived from ENCODE. Displayed are the following strata: ENCODE CAR rank  $< 10$  ( $N=37$  observations in total), ENCODE CAR rank  $\geq 10$  ( $N=107$  observations in total). While we see enrichment for low ROADMAP CAR rank in subfamilies predicted to be CARs via the ENCODE data, we see no enrichment in low ROADMAP CAR ranks for other subfamilies.

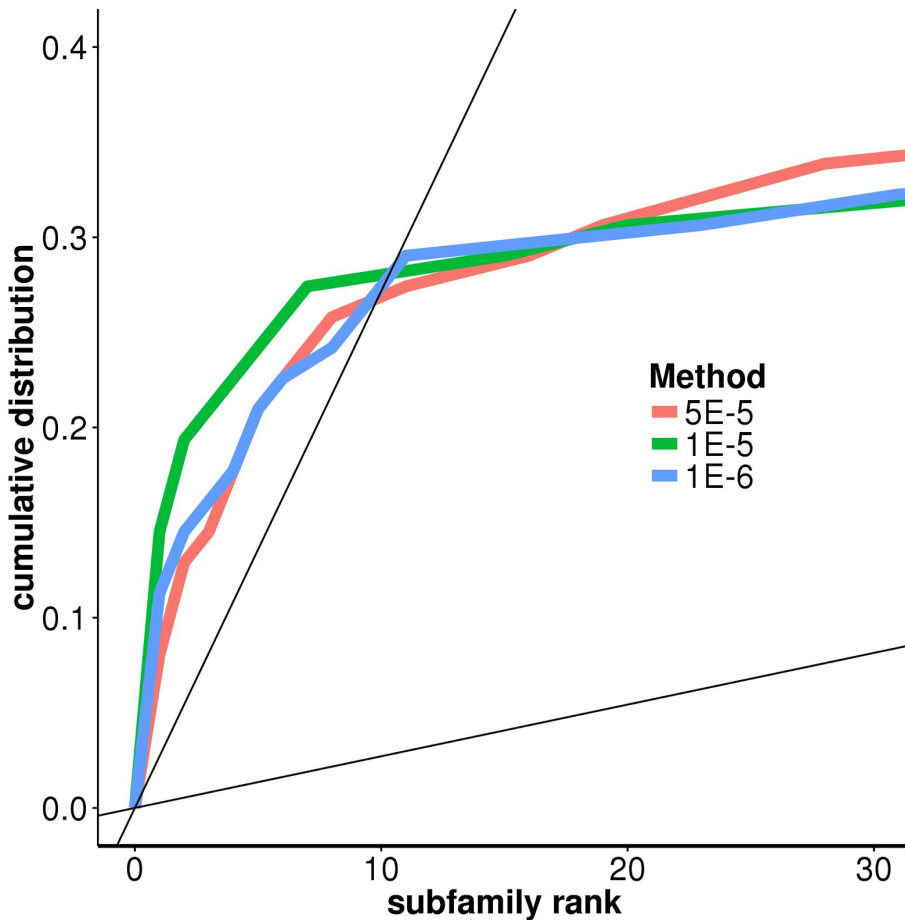


Fig S5: CAR detection power is stable to changes in motif cutoffs.

Cumulative distribution of CAR ranks at the subfamily level using the three different motif cutoffs:  $10^{-5}$  (used throughout the paper) is compared to  $10^{-6}$  (yielding 9.3 fewer motifs on average [median]) and  $5 \cdot 10^{-5}$  (yielding 5.2 more motifs assigned on average). For each setting, we filtered motifs that did not overlap at least 150 DHS regions per cell line on average. Only subfamilies passing this filter in all settings were included (62 subfamilies in total). Power mildly increased at low CAR ranks for more stringent cutoffs at the cost of fewer motifs passing filtering. However, at false discovery rate of 10% power was nearly identical.

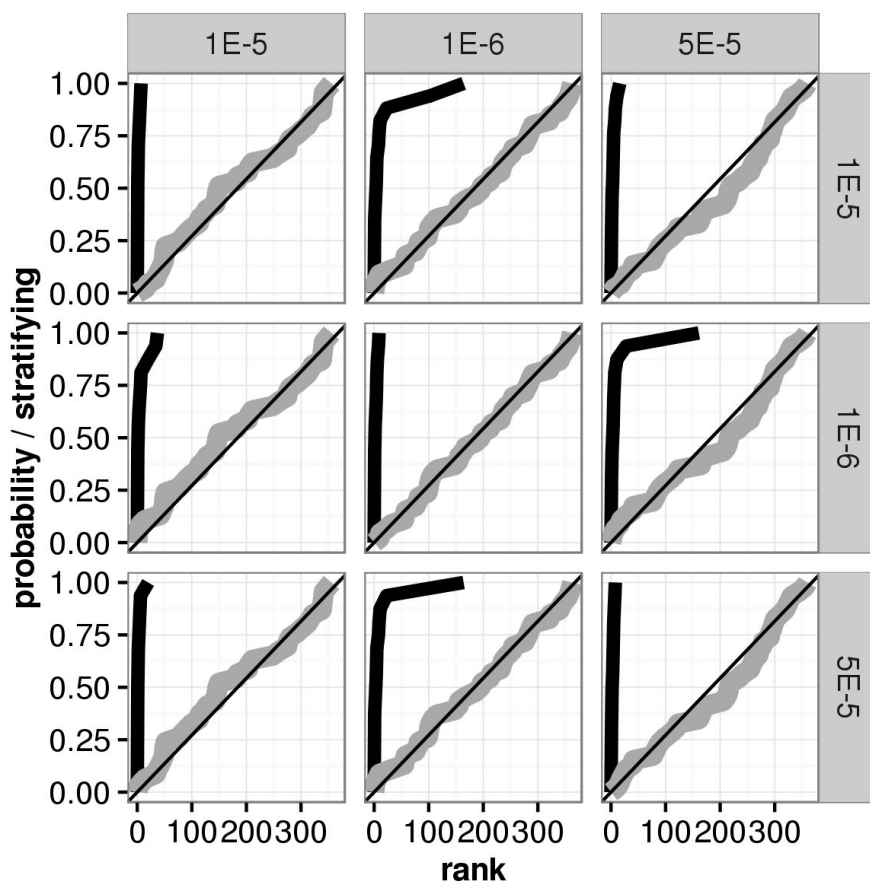


Fig S6: CAR prediction is stable w.r.t. changes in motif thresholds.

Shown are pairwise comparisons of different motif thresholds. For each threshold, we derived CAR ranks for all tested subfamilies yielding one CAR rank list per threshold. Pairwise comparisons of these lists were performed in the following manner: For each pair of rank lists, the first list was used to split the tested subfamilies into a 'CAR set' and its complement based on whether a subfamily had CAR rank below 10. For the second results list, two separate CAR enrichment curves were drawn, one curve for the 'CAR set' defined via the first list (black) and its complement (grey). Rows denote the threshold used to derive the 'CAR set' and columns denote the threshold used to draw the enrichment curves. For each setting, we filtered motifs that did not overlap at least 150 dhs regions per cell line on average. Only subfamilies passing this filter in all settings were included (62 subfamilies in total). We see that CARs predicted are stable with respect to varying motif cutoffs.

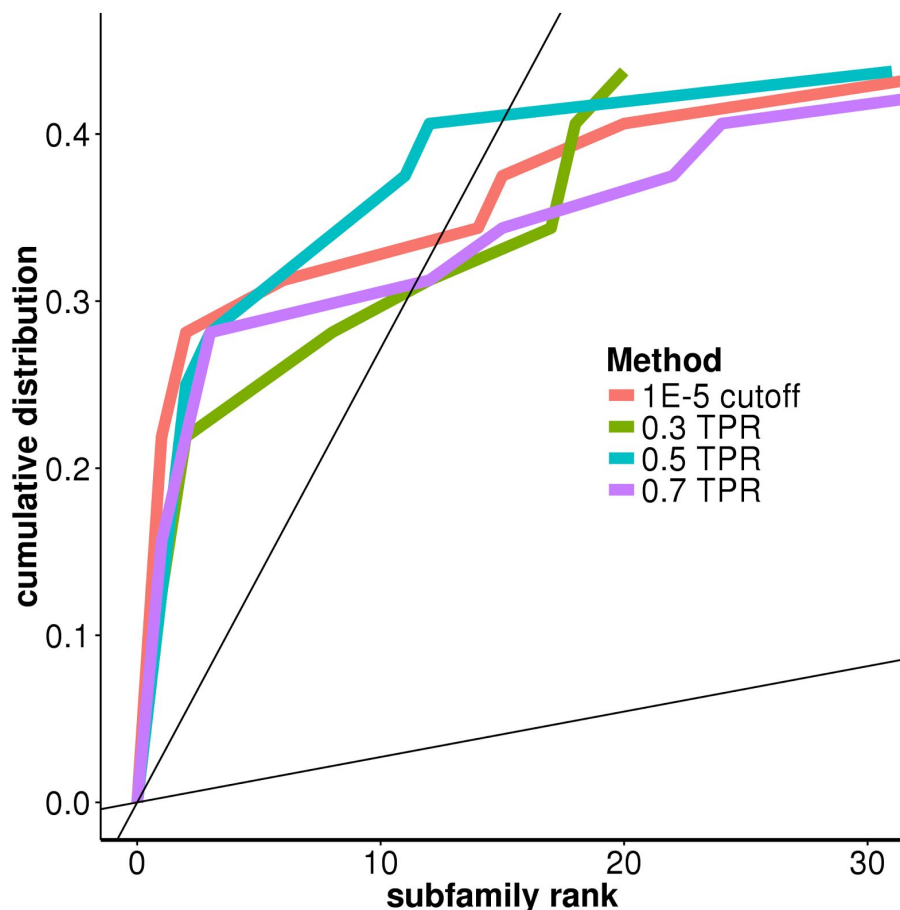


Fig S7: CAR detection power does not improve systematically when guiding motif cutoffs via ChIP-seq.

Shown are cumulative distribution of CAR ranks at the subfamily level comparing fixed motif cutoff of  $10^{-5}$  (used throughout the paper) is compared to variable motif cutoffs guided by ChIP-seq data, where motif cutoffs are adjusted such that called binding sites (i.e. DHS sites containing a motif instance) have a fixed validation rate compared to a gold standard defined by ChIP-seq. Chosen validation rates are 0.3, 0.5 and 0.7. For each setting, we filtered motifs that did not overlap at least 150 DHS regions per celline on average. Only subfamilies passing this filter in all settings were included (32 subfamilies in total). While we see some variation in power, the variation is not systematic.

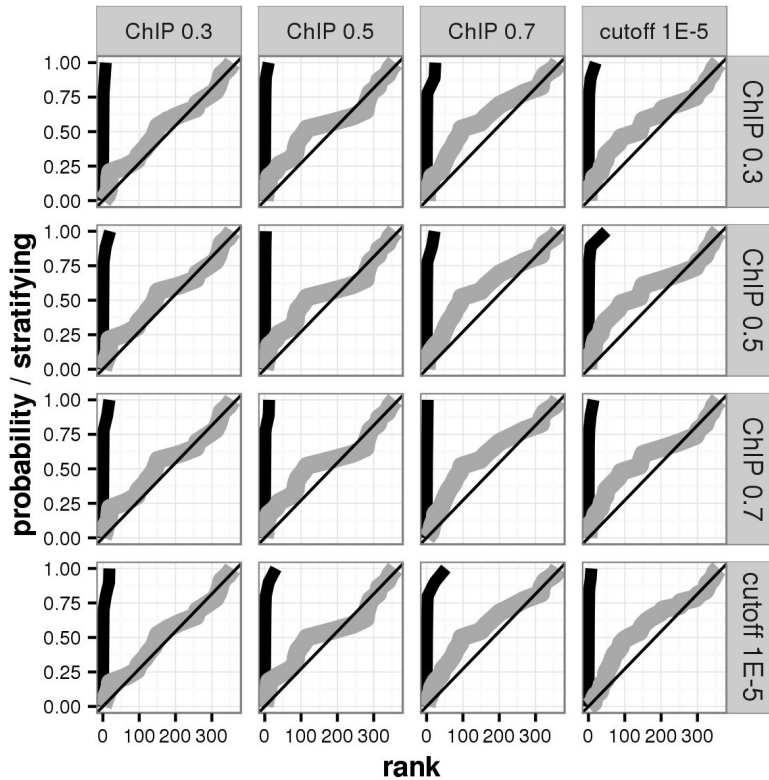


Fig S8: ChIP-seq data guiding motif thresholding yields similar CAR predictions as regular motif thresholding.

Shown are pairwise comparisons of different motif thresholding methods. For each thresholding method we derived CAR ranks for all tested subfamilies yielding one CAR rank list per method. Pairwise comparisons of these lists were performed in the following manner: For each pair of rank lists, the first list was used to split the tested subfamilies into a ‘CAR set’ and its complement based on whether a subfamily had CAR rank below 10. For the second results list, two separate CAR enrichment curves were drawn, one curve for the ‘CAR set’ defined via the first list (black) and its complement (grey). Rows denote the thresholding method used to derive the ‘CAR set’ and columns denote the thresholding method used to draw the enrichment curves. A fixed motif cutoff of  $10^{-5}$  (also used throughout the paper) is compared to variable motif cutoffs guided by ChIP-seq data, where motif cutoffs are adjusted such that called binding sites (i.e. DHS sites containing a motif instance) have a fixed validation rate when compared to ChIP-seq. Chosen validation rates are 0.3, 0.5 and 0.7. For each setting, we filtered motifs that did not overlap at least 150 DHS regions per celline on average. Only subfamilies passing this filter in all settings were included (32 subfamilies in total). We see that CARs predicted are stable with respect to varying motif cutoffs.

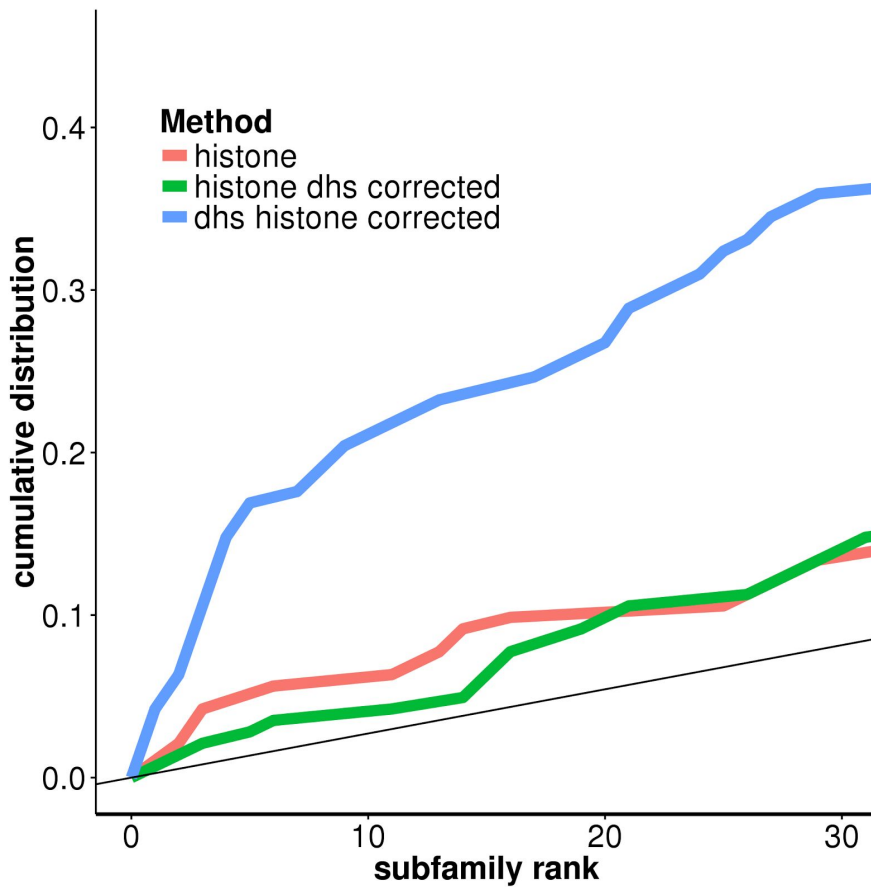


Fig S9: Histone-wise motif activities do not substantially associate with TF expression values. H3K4me3 peak data for 51 cell lines were downloaded from ENCODE and histone-wise motif activity was computed and normalized analogously to for DHS data, regressing out the first principal component. We performed the mixed model regression where H3K4me3-based motif accessibility data are regressed on gene expression adding a random effect with the same covariance structure as the expression matrix (denoted ‘histone’). To assess the DHS-independent contribution of H3K4me3 histone activities, we added DHS-based motif accessibility as a covariate (denoted ‘DHS-adjusted histone’). We see that subfamily ranks for both of these strategies do not substantially enrich in low ranks. While ‘histone’ performs mildly better, this is likely due to correlation between the histone activity and DHS activity. In contrast, the when DHS-based motif accessibility data was adjusted for H3K4me3-based motif accessibility, we see a still substantial enrichment (see “histone-adjusted DHS” curve). This experiment was performed by regressing DHS motif accessibility on gene expression while adding H3K4me3-based motif accessibilities as a covariate plus a random effect with the same covariance structure as the expression matrix. This shows that of the two activity measures, only DHS activity substantially associates with expression.

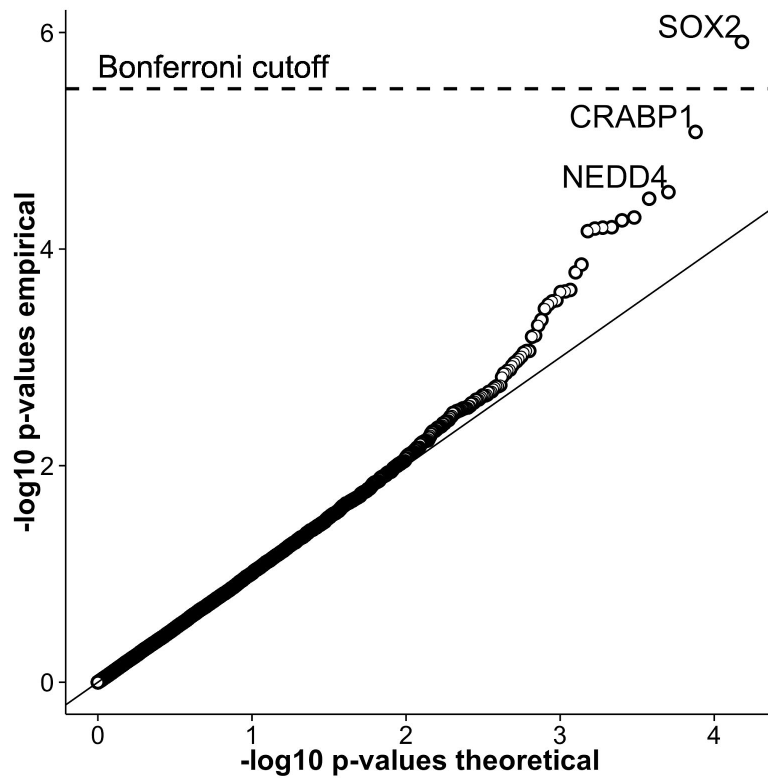


Fig S10: *SOX2* expression associates strongly with *POU5F1* motif accessibility. The QQ-plot shows the  $p$ -value distribution obtained from the LMM associating the accessibility of the *POU5F1* motif to gene expression values across all genes. We see the strongest association to *SOX2* expression.

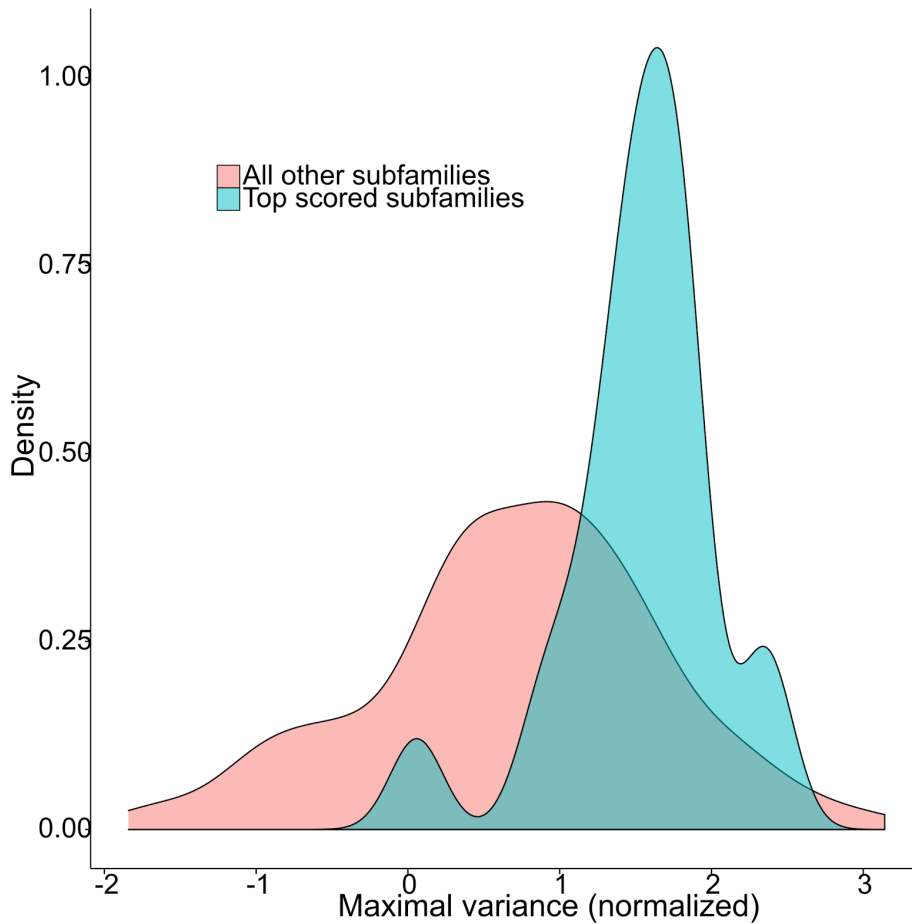


Fig S11: Predicted chromatin accessibility regulators tend to have higher expression variation. We derived the variance of expression for all transcription factors across micro-arrays after RMA normalization and averaging expression values for experiments derived from the cell types. Displayed is a density distribution of the maximal expression variance observed in each subfamily. We partitioned TF subfamilies into two groups depending on whether they had family level CAR ranks of 1 or not. We observe that top ranked subfamilies do have substantially higher variance on average than other subfamilies (linear regression  $p$ -value  $<10^{-03}$ ).



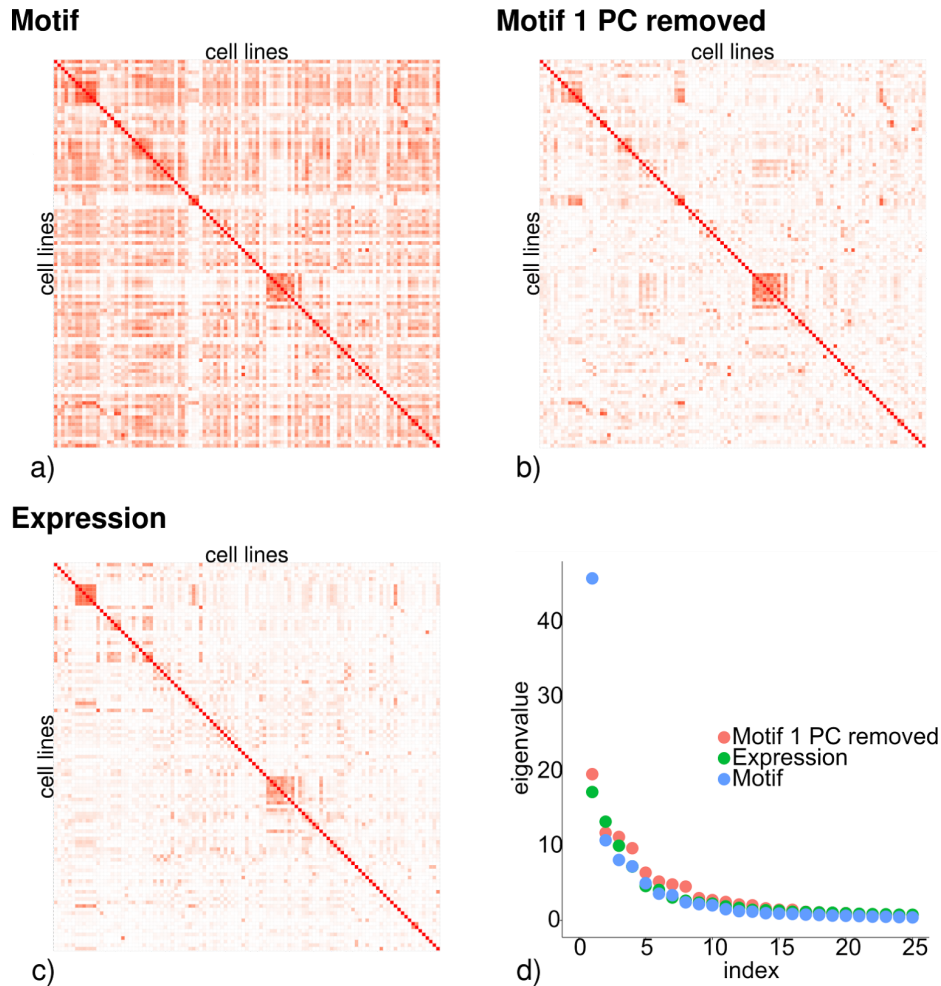


Fig S12: Removing first principal component from motif accessibility matrix leads to similar correlation structures between motif accessibility and expression.

Displayed are pair-wise correlation matrices with squared entries across cell lines for motif accessibilities (a); motif accessibilities with the first principal component removed (b) and (c) for expression values. Further, the first 25 eigenvalues of these matrices are shown in (d). The motif accessibility matrix has a very dominant first principal component. After removal of the first principal component, the correlation structure of motif accessibility and expression show a similar structure.

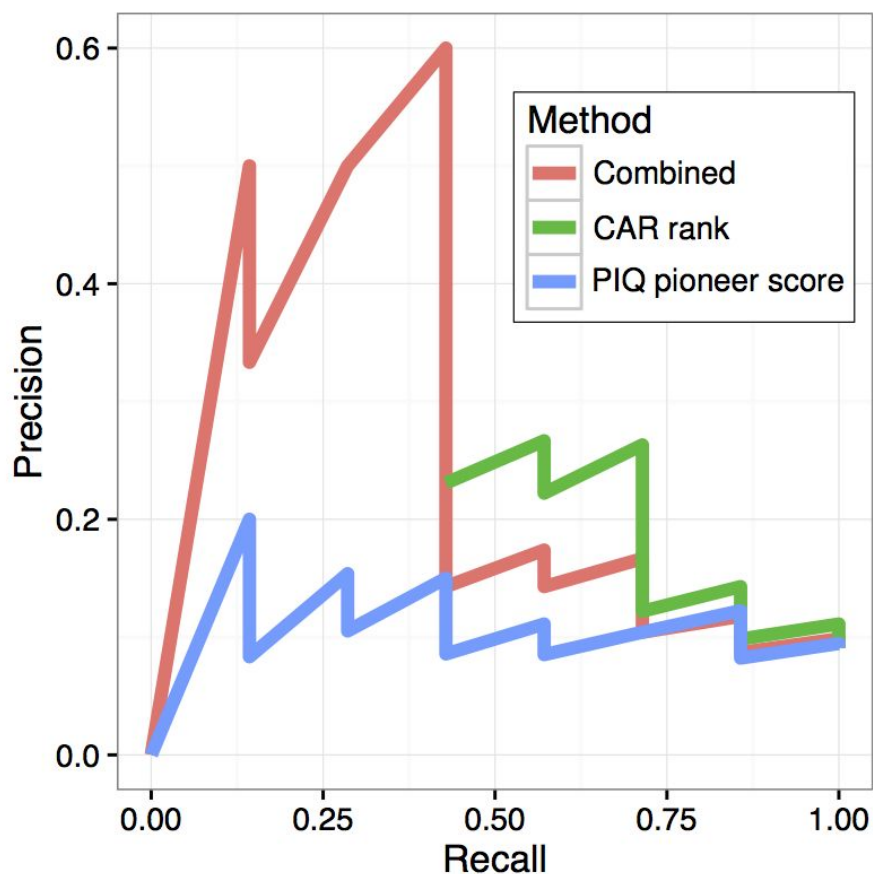


Fig S13: precision-recall curves of CAR ranks and PIQ pioneer scores and their combination. Displayed are the precision-recall curves using annotation from Iwafuchi-Doi et al. [69] as true set. Motif wise PIQ pioneer scores were extracted from Sherwood et al. [70]. For each subfamily, we defined its PIQ pioneer score as the maximal pioneer score for its subfamily members. For 77 subfamilies, data were available from both approaches of which 7 were in the true set. For both CAR ranks and PIQ pioneer scores, precision-recall curves were drawn (CAR rank precision-recall curve starts at 0.43 recall, because many subfamilies share CAR rank of one). Additionally, both scores were combined: For each scoring method, results were ranked (rank ties was replaced by the minimum). For each subfamily, its combined rank is the maximal rank across both methods. A low rank can therefore only be achieved when both methods yielded low ranks. We see that the combined strategy outperforms both base strategies.

## E.2. Supplementary Methods

As mentioned in the main text, we use the following linear mixed model,

$$\mathbf{y} = \mathbf{x}_i \beta^i + \boldsymbol{\delta}^i + \boldsymbol{\epsilon}^i,$$

where  $\mathbf{y}$  is a vector of motif accessibility scores across  $n$  cell lines,  $\mathbf{x}_i$  is the expression vector of gene  $i$ ,  $\beta^i$  is the effect size of gene  $i$ :

$$\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma_r^2 \mathbf{I}_n),$$

and

$$\boldsymbol{\delta} \sim N_n(\mathbf{0}, \sigma_e^2 \mathbf{C}_e),$$

with

$$\mathbf{C}_e = \frac{1}{p} \sum_{i=1}^p \mathbf{x}_i \mathbf{x}_i^T.$$

The likelihood function of is

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{n/2} (\sigma_e^2 \mathbf{C}_e + \sigma_r^2 \mathbf{I}_n)^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{y} - \mathbf{x}_i \beta^i)^T (\sigma_e^2 \mathbf{C}_e + \sigma_r^2 \mathbf{I}_n)^{-1} (\mathbf{y} - \mathbf{x}_i \beta^i)\right).$$

We define the spectral decomposition of  $\mathbf{C}_e$  as:

$$\mathbf{C}_e = \mathbf{\Gamma} \mathbf{\Lambda} \mathbf{\Gamma}^T.$$

For any values of  $\sigma_e^2$  and  $\sigma_r^2$  we have

$$(\sigma_e^2 \mathbf{C}_e + \sigma_r^2 \mathbf{I}_n) = \mathbf{\Gamma} (\sigma_e^2 \mathbf{\Lambda}_e + \sigma_r^2 \mathbf{I}_n) \mathbf{\Gamma}^T,$$

i.e.: the eigenvectors of the mixture matrix are constant w.r.t the mixing parameters. Set

$$\mathbf{y}' = \mathbf{\Gamma}^T \mathbf{y},$$

$$\mathbf{x}' = \mathbf{\Gamma}^T \mathbf{x}.$$

Since the likelihood is invariant to rotations, we have

$$f(\mathbf{y}') = f(\mathbf{y})$$

and

$$f(\mathbf{y}') = \frac{1}{(2\pi)^{n/2}(\sigma_e^2 \mathbf{\Lambda}_e + \sigma_r^2 \mathbf{I}_n)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{x}_i \beta^i)^T (\sigma_e^2 \mathbf{\Lambda}_e + \sigma_r^2 \mathbf{I}_n)^{-1} (\mathbf{y} - \mathbf{x}_i \beta^i)\right).$$

Reparametrizing with

$$\gamma = \sigma_r^2 / \sigma_e^2,$$

the log-likelihood becomes

$$l(\mathbf{y}') = \frac{n}{2} \log(2\pi) - \frac{n}{2} \sum \log(\sigma_e^2(\lambda_i + \gamma)) - \frac{n}{2} \sum \frac{(y'_k - x'_{ki} \beta_i)^2}{(2\sigma_e^2(\lambda_i + \gamma))}$$

Partial derivation shows that the maximum of the log likelihood is reached at

$$\hat{\beta}^i = \sum_k \frac{y'_k x'_{ki}}{(\lambda_k + \hat{\gamma})} / \sum_k \frac{(x'_{ki})^2}{(\lambda_k + \hat{\gamma})},$$

and

$$\hat{\sigma}_e^2 = \sum_k \frac{(y'_k - x'_{ki} \hat{\beta}^i)^2}{(\lambda_k + \hat{\gamma})} / \sum_k \frac{n}{(\lambda_k + \hat{\gamma})}.$$

Reducing the 3 parameter optimization problem to a one parameter optimization over  $\gamma$ .  $p$ -values can be obtained by the likelihood ratio test for null hypothesis that  $\beta = 0$ .

## F. Supplementary Information for Chapter 4

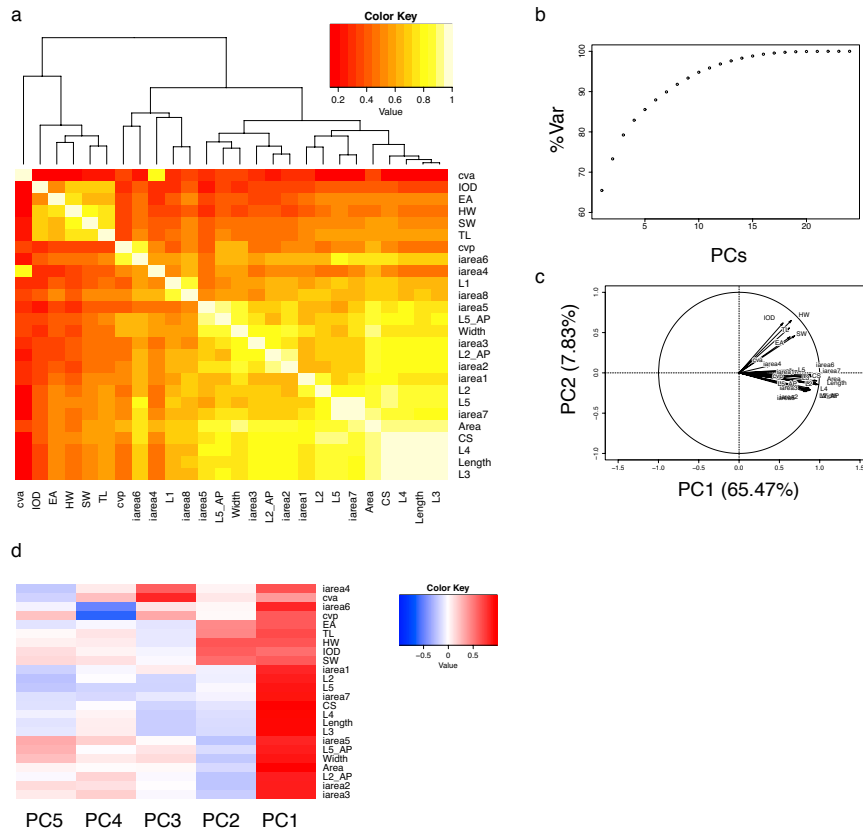


Fig S1: Analysis of the male dataset.

a) Genetic correlation between morphometric traits in males. The two modules of higher correlation observed in females are still visible (bright yellow in the upper left and lower right corners) but the overall clustering is more influenced by the more inaccurately measured smaller veins and areas. b) Cumulative variance explained in male data by increasing number of principal components. As in the female dataset, the first two PCs explain almost 75% of the variance in the data. c) Factor map for the variables. PCs 1 and 2 split the data into two groups. d) Correlation between PCs and traits. PC1 reflects a general size component and PC2 is highly correlated with head/thorax traits, effectively splitting the data in two groups.

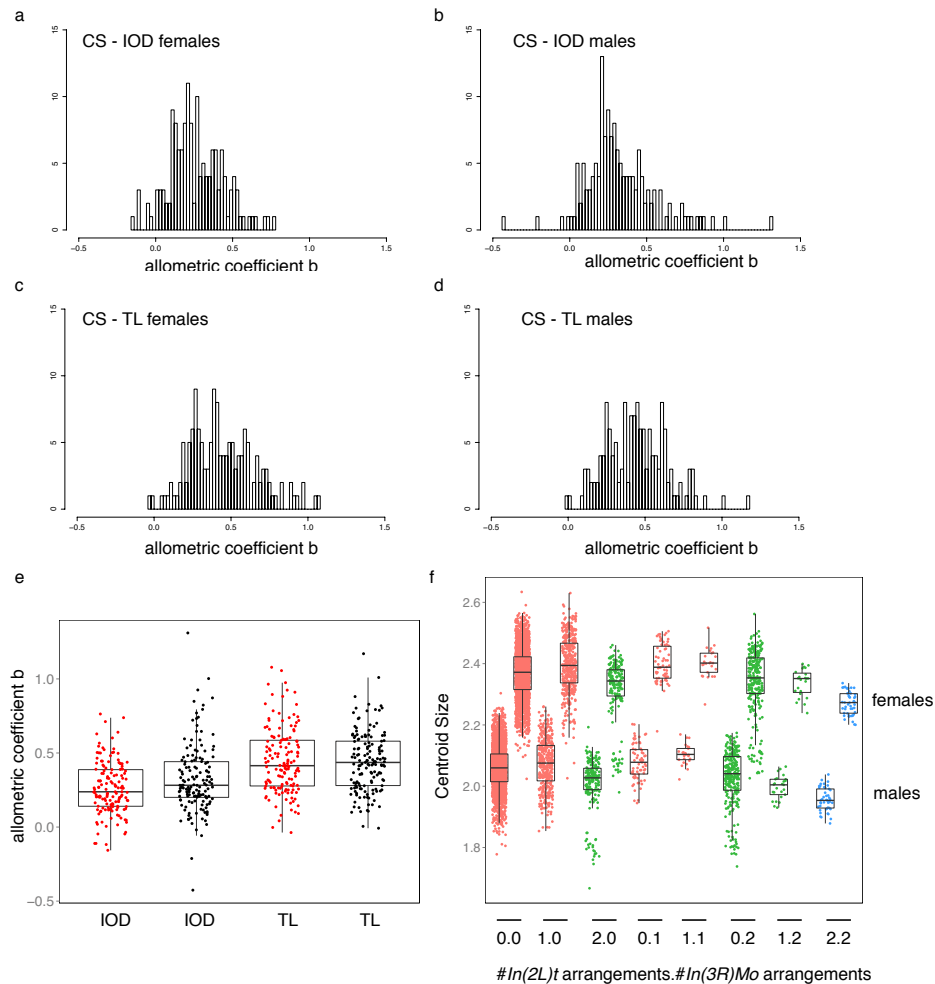


Fig S2: Allometry and inversions.

Histograms of the estimates for the allometric coefficient  $b$  for the relationship between CS and IOD in females (a), in males (b) and between CS and TL in females (c) and males (d). e) Boxplot and individual datapoints of the data in a-d. Red = females and black = males. 95% confidence intervals for  $b$  are very broad for some lines due to few datapoints used for fitting, so these are just very rough estimates for the allometric relationship. Nevertheless there is variation among lines for all evaluated relationships. f) The effect of cosmopolitan inversions on wing size. Lines are plotted according to the number of homozygous inversion arrangements they have: 0 (red) = neither  $In(2L)t$  nor  $In(3R)Mo$  present, 1 (green) = homozygous for either  $In(2L)t$  or  $In(3R)Mo$ , 2 (blue) = homozygous for both  $In(2L)t$  and  $In(3R)Mo$ . Datapoints are individual flies.

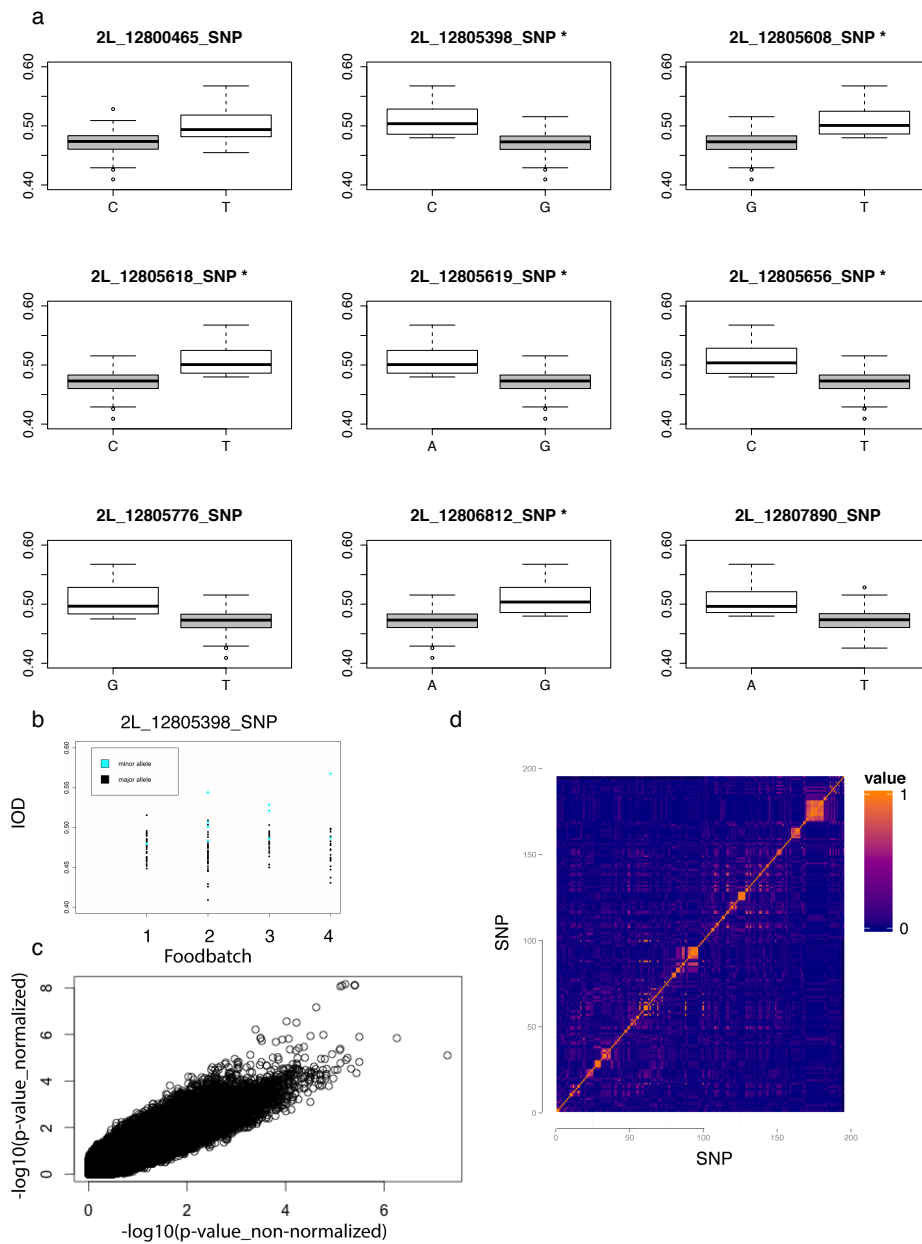


Fig S3: The minor and major haplotype of genome-wide significant SNPs show differential association with female IOD.

a) The minor allele haplotype of the genome-wide significant cluster is associated with an increased IOD in females. Boxplots of female IOD by genotype at the nine SNPs annotated to *kek1*. SNPs marked by a star pass Bonferroni correction. Grey=major allele, white = minor allele. b) Lines with the minor haplotype are distributed across all four foodbatches. Black dots = major allele, blue dots = minor allele. The IOD distribution for each foodbatch is plotted for females for the most significant SNP. The distribution is the same for all other SNPs of the cluster as all minor alleles form a haplotype. c) Correlation between  $p$ -values from GWAS with normalized IOD (y-axis) and non-normalized iod (x-axis) in females. Axes are on the  $-\log_{10}$  scale. d) Several blocks of higher LD are visible in the region 20kb upstream of *kek1*. Blue = no correlation, orange = complete correlation.

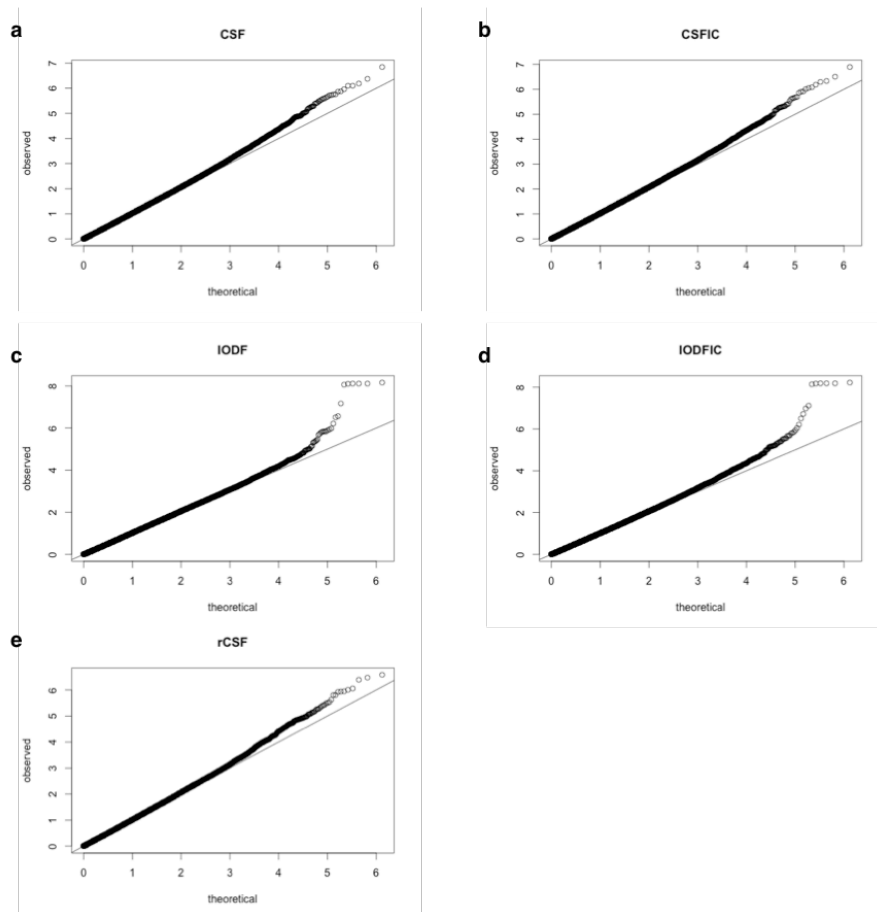


Fig S4: QQ-plots from GWA in females for all traits show a departure from uniformity of top associations.

Observed association  $p$ -values are  $-\log_{10}$  transformed (y-axis) and plotted against the  $-\log_{10}$  transformed theoretically expected  $p$ -values under the assumption of no association (uniform distribution, x-axis). Centroid size (a), inversion corrected centroid size (b), interocular distance (c), inversion corrected interocular distance (d) and relative centroid size (e).



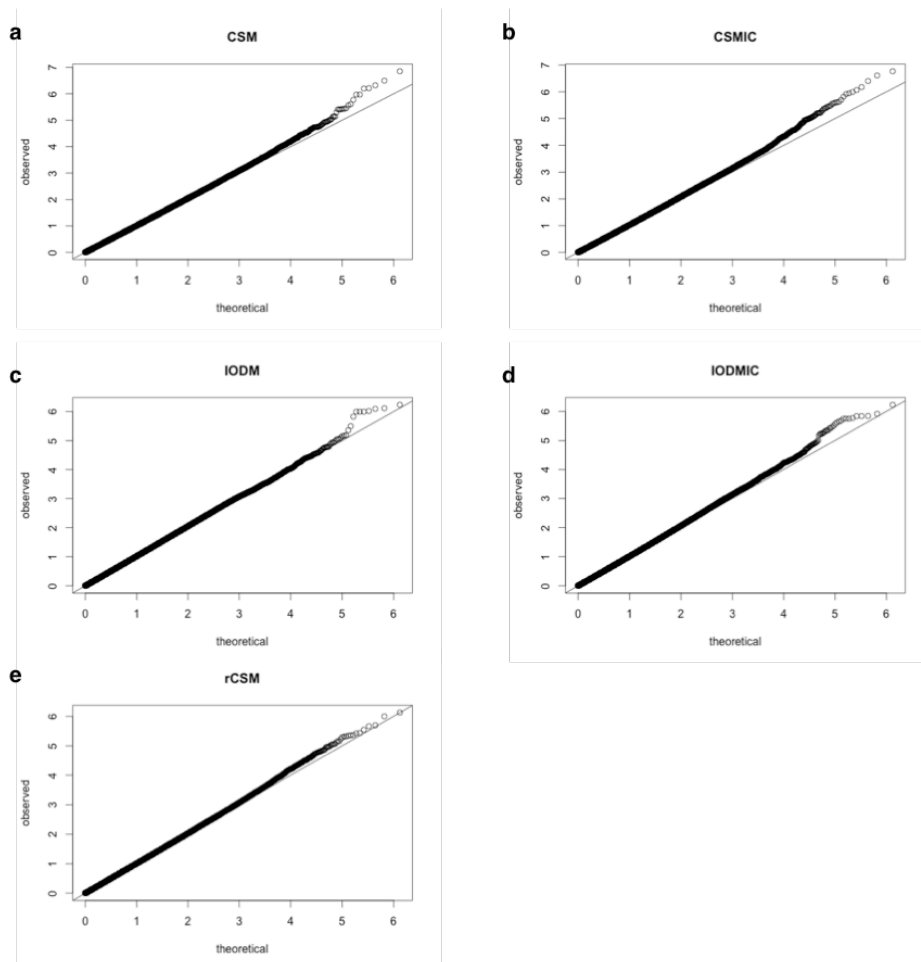


Fig S5: QQ-plots from GWAS in males for all traits show a departure from uniformity of top associations.

Observed association  $p$ -values are  $-\log_{10}$  transformed (y-axis) and plotted against the  $-\log_{10}$  transformed theoretically expected  $p$ -values under the assumption of no association (uniform distribution, x-axis). Centroid size (a), inversion corrected centroid size (b), interocular distance (c), inversion corrected interocular distance (d) and relative centroid size (e).

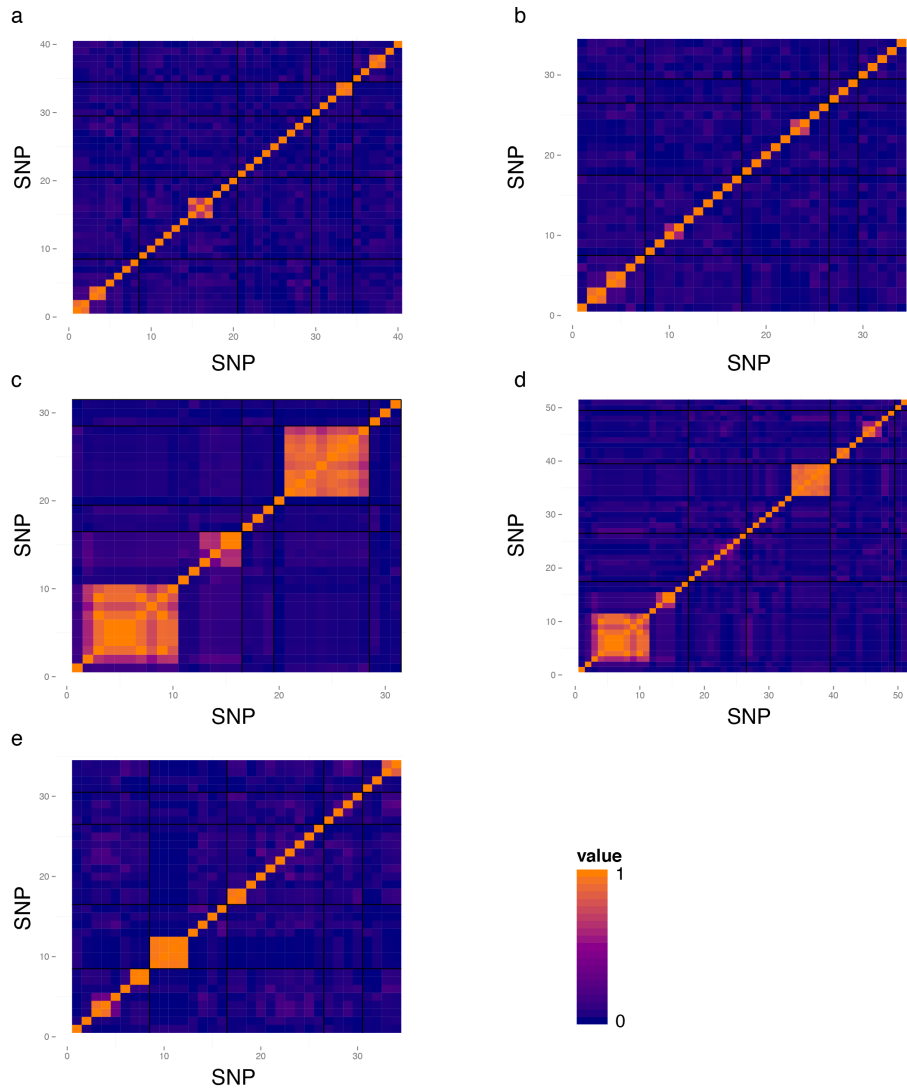


Fig S6: Correlation between associated ( $p < 10^{-05}$ ) SNPs in females. The SNPs are ordered according to chromosome arm (2L, 2R, 3L, 3R, X) and black dividers separate chromosomes. Within one chromosome arm SNPs are ordered according to their position on that chromosome with each tile representing one SNP. The color code is depicted on the right: orange = complete correlation (1) and blue = no correlation (0). Centroid size (a), inversion corrected centroid size (b), interocular distance (c), inversion corrected interocular distance (d) and relative centroid size (e).

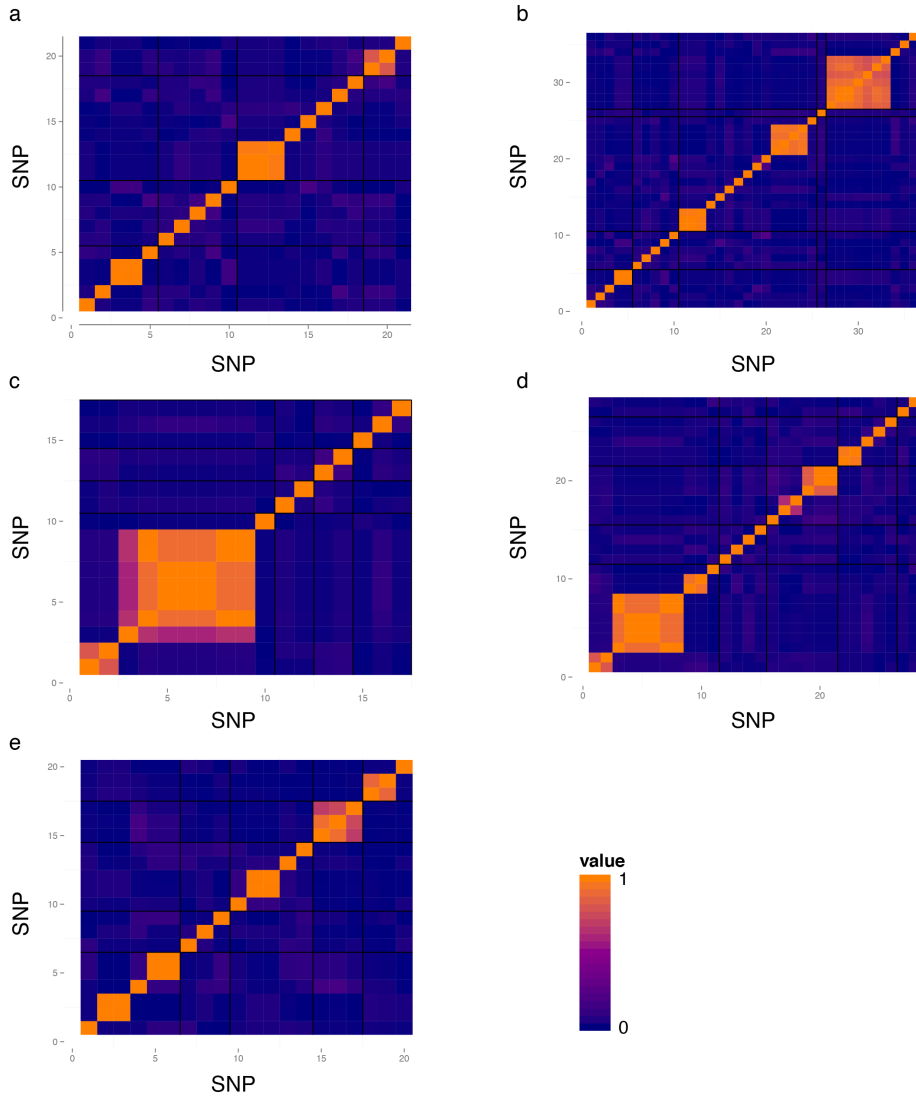


Fig S7: Correlation between associated ( $p < 10^{-05}$ ) SNPs in males. The SNPs are ordered according to chromosome arm (2L, 2R, 3L, 3R, X) and black dividers separate chromosomes. Within one chromosome arm SNPs are ordered according to their position on that chromosome with each tile representing one SNP. The color code is depicted on the right: orange = complete correlation (1) and blue = no correlation (0). Centroid size (a), inversion corrected centroid size (b), interocular distance (c), inversion corrected interocular distance (d) and relative centroid size (e).

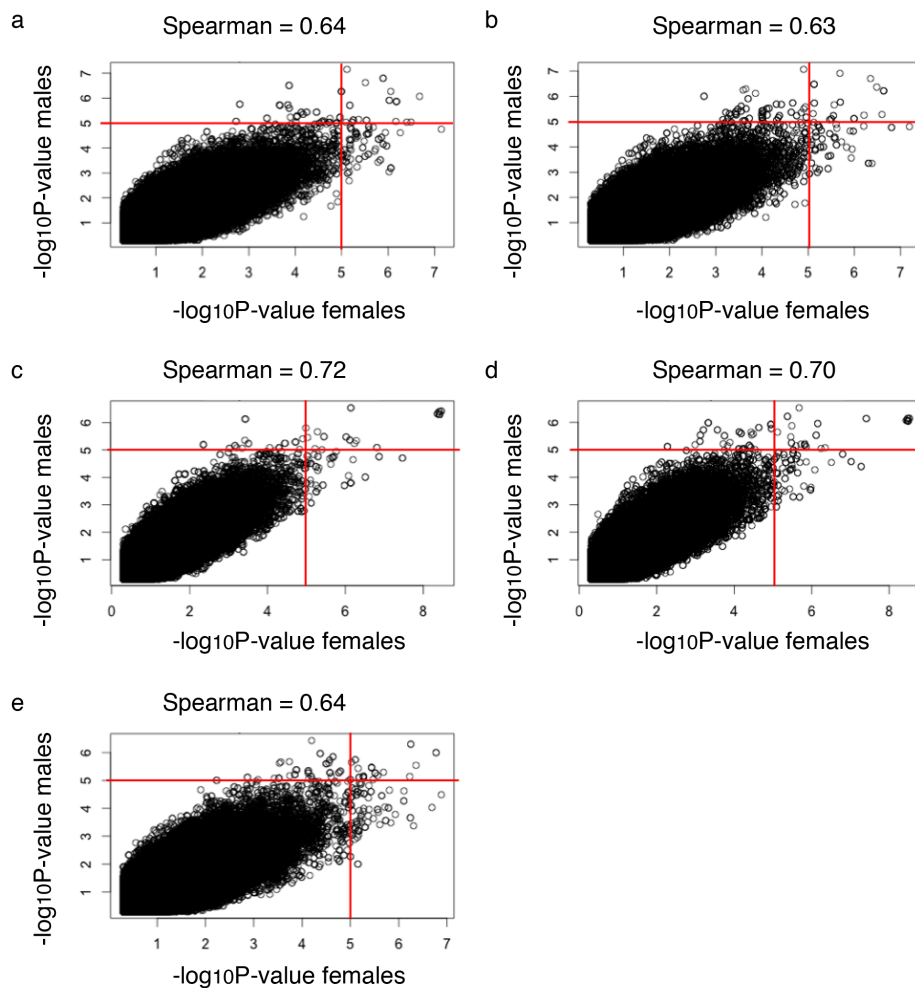


Fig S8: Correlation of SNP  $p$ -values between the sexes. SNP  $p$ -values in females (x-axis) are plotted against their respective  $p$ -values in males (y-axis). The Spearman rank correlation is given for each trait and the red lines denote the significance cutoff. a = CS, b = CSIC, c = IOD, d = IODIC, e = rCS.

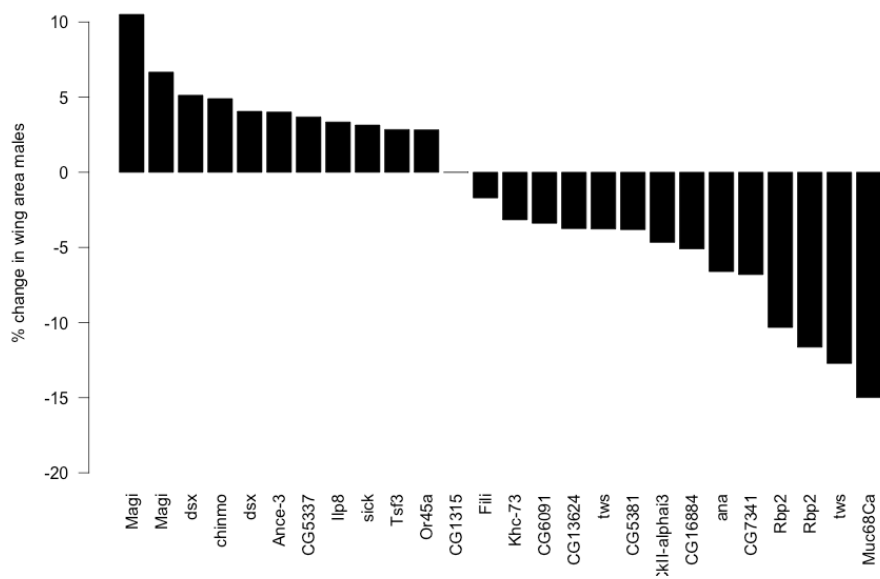


Fig S9: RNAi knockdown results males.

Percent change in median wing area compared to CG1315 RNAi upon wing-specific knockdown of the validated candidate genes in males. Only the lines yielding a significant wing size change ( $p < 0.001$ , Wilcoxon rank sum test) are depicted.

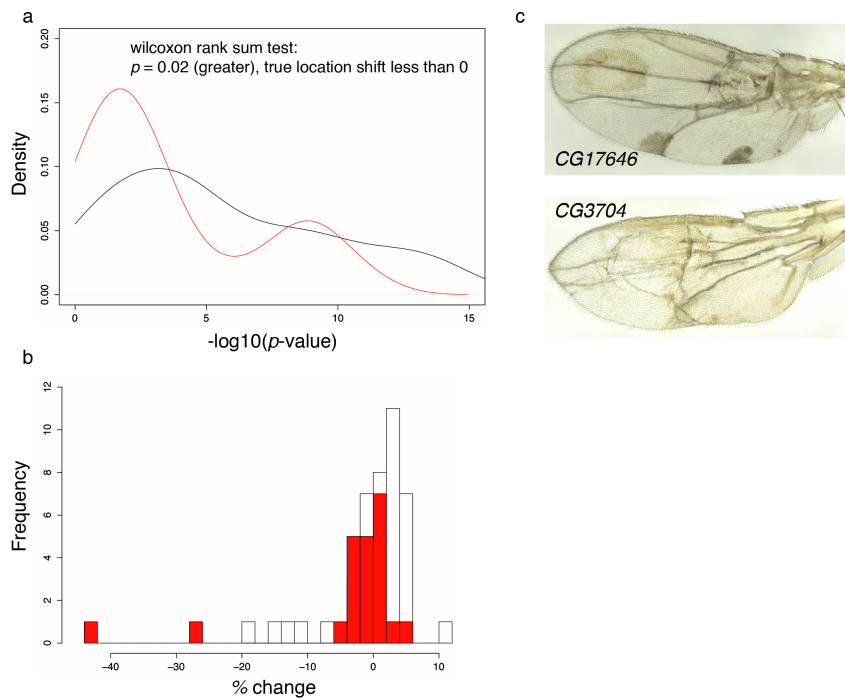


Fig S10: Comparison of  $p$ -values and effect sizes between candidate and control RNAi. a:  $-\log_{10}$  transformed  $p$ -value densities of the candidate (black) and combined control (red) data sets. The two  $p$ -value distributions differ by a location shift that is not zero (i.e. are not the same); specifically, the  $-\log_{10}$  transformed control  $p$ -value distribution (red) is shifted towards the left of the  $-\log_{10}$  transformed candidate  $p$ -value distribution (black) (one sided Wilcoxon rank sum test  $p = 0.02$ ). b: The distribution of candidate effect sizes (percent change in wing size upon knockdown) is shifted towards positive effect sizes (white boxes), whereas the control knockdown effect size distribution (red) is more centered on 0. The two exceptions at -28% (CG17646) and -42% (CG3704) are lines whose wings not only show a size reduction but also considerable morphological defects (c).  $N = 43$  candidates (white),  $N = 22$  control (red); only data from females was used for these analyses.

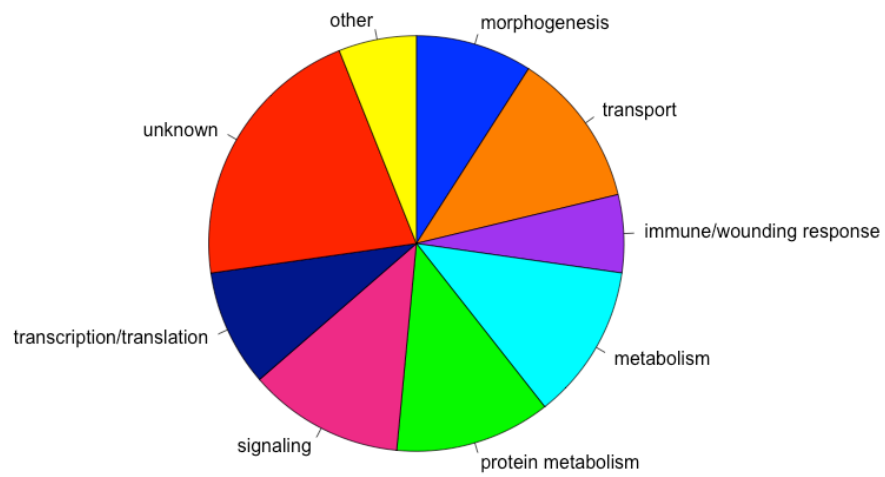


Fig S11: Functional annotation of the 33 validated candidate genes based on DAVID GO annotation.

## References

- [1] David Lamparter, Daniel Marbach, Rico Rueedi, Zoltán Kutalik, and Sven Bergmann. Fast and rigorous computation of gene and pathway scores from snp-based summary statistics. *PLoS Comput Biol*, 12(1):e1004714, Jan 2016.
- [2] David Lamparter, Daniel Marbach, Rico Rueedi, Sven Bergmann, and Zoltán Kutalik. Genome-wide association between transcription factor expression and chromatin accessibility reveals regulators of chromatin accessibility. *PLoS Comput Biol*, 13(1):e1005311, Jan 2017.
- [3] Sibylle Chantal Vonesch, David Lamparter, Trudy F C Mackay, Sven Bergmann, and Ernst Hafen. Genome-wide analysis reveals novel regulators of growth in drosophila melanogaster. *PLoS Genet*, 12(1):e1005616, Jan 2016.
- [4] Jorg D. Hoheisel. Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet*, 7(3):200–210, 03 2006.
- [5] Andrew R Wood et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet*, 46(11):1173–86, Nov 2014.
- [6] Fan Liu, Fedde van der Lijn, Claudia Schurmann, Gu Zhu, M Mallar Chakravarty, Pirro G Hysi, Andreas Wollstein, Oscar Lao, Marleen de Bruijne, M Arfan Ikram, Aad van der Lugt, Fernando Rivadeneira, André G Uitterlinden, Albert Hofman, Wiro J Niessen, Georg Homuth, Greig de Zubicaray, Katie L McMahon, Paul M Thompson, Amro Daboul, Ralf Puls, Katrin Hegenscheid, Liisa Bevan, Zdenka Pausova, Sarah E Medland, Grant W Montgomery, Margaret J Wright, Carol Wicking, Stefan Boehringer, Timothy D Spector, Tomáš Paus, Nicholas G Martin, Reiner Biffar, and Manfred Kayser. A genome-wide association study identifies five loci influencing facial morphology in europeans. *PLoS Genet*, 8(9):e1002932, Sep 2012.
- [7] Harm-Jan Westra, Marjolein J Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann,



- Johannes Kettunen, Mark W Christiansen, Benjamin P Fairfax, Katharina Schramm, Joseph E Powell, Alexandra Zhernakova, Daria V Zhernakova, Jan H Veldink, Leonard H Van den Berg, Juha Karjalainen, Sebo Withoff, André G Uitterlinden, Albert Hofman, Fernando Rivadeneira, Peter A C 't Hoen, Eva Reinmaa, Krista Fischer, Mari Nelis, Lili Milani, David Melzer, Luigi Ferrucci, Andrew B Singleton, Dena G Hernandez, Michael A Nalls, Georg Homuth, Matthias Nauck, Dörte Radke, Uwe Völker, Markus Perola, Veikko Salomaa, Jennifer Brody, Astrid Suchy-Dacey, Sina A Gharib, Daniel A Enquobahrie, Thomas Lumley, Grant W Montgomery, Seiko Makino, Holger Prokisch, Christian Herder, Michael Roden, Harald Grallert, Thomas Meitinger, Konstantin Strauch, Yang Li, Ritsert C Jansen, Peter M Visscher, Julian C Knight, Bruce M Psaty, Samuli Ripatti, Alexander Teumer, Timothy M Frayling, Andres Metspalu, Joyce B J van Meurs, and Lude Franke. Systematic identification of trans eqtls as putative drivers of known disease associations. *Nat Genet*, 45(10):1238–43, Oct 2013.
- [8] 1000 Genomes Project Consortium, Gonçalo R Abecasis, David Altshuler, Adam Auton, Lisa D Brooks, Richard M Durbin, Richard A Gibbs, Matt E Hurles, and Gil A McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, Oct 2010.
- [9] 1000 Genomes Project Consortium, Goncalo R Abecasis, Adam Auton, Lisa D Brooks, Mark A DePristo, Richard M Durbin, Robert E Handsaker, Hyun Min Kang, Gabor T Marth, and Gil A McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, Nov 2012.
- [10] Albert Tenesa, Pau Navarro, Ben J Hayes, David L Duffy, Geraldine M Clarke, Mike E Goddard, and Peter M Visscher. Recent human effective population size estimated from linkage disequilibrium. *Genome Res*, 17(4):520–6, Apr 2007.
- [11] Joseph Lachance. Disease-associated alleles in genome-wide association studies are enriched

- for derived low frequency alleles relative to hapmap and neutral expectations. *BMC Med Genomics*, 3:57, Dec 2010.
- [12] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy F C Mackay, Steven A McCarroll, and Peter M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–53, Oct 2009.
- [13] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael E Goddard, and Peter M Visscher. Common snps explain a large proportion of the heritability for human height. *Nat Genet*, 42(7):565–9, Jul 2010.
- [14] Christian Fuchsberger, Jason Flannick, Tanya M Teslovich, Anubha Mahajan, Vineeta Agarwala, Kyle J Gaulton, Clement Ma, Pierre Fontanillas, Loukas Moutsianas, Davis J McCarthy, Manuel A Rivas, John R B Perry, Xueling Sim, Thomas W Blackwell, Neil R Robertson, N William Rayner, Pablo Cingolani, Adam E Locke, Juan Fernandez Tajés, Heather M Highland, Josee Dupuis, Peter S Chines, Cecilia M Lindgren, Christopher Hartl, Anne U Jackson, Han Chen, Jeroen R Huyghe, Martijn van de Bunt, Richard D Pearson, Ashish Kumar, Martina Müller-Nurasyid, Niels Grarup, Heather M Stringham, Eric R Gamazon, Jaehoon Lee, Yuhui Chen, Robert A Scott, Jennifer E Below, Peng Chen, Jinyan Huang, Min Jin Go, Michael L Stitzel, Dorota Pasko, Stephen C J Parker, Tibor V Varga, Todd Green, Nicola L Beer, Aaron G Day-Williams, Teresa Ferreira, Tasha Fingerlin, Momoko Horikoshi, Cheng Hu, Iksoo Huh, Mohammad Kamran Ikram, Bong-Jo Kim, Yongkang Kim, Young Jin Kim, Min-Seok Kwon, Juyoung Lee, Selyeong Lee, Keng-Han Lin, Taylor J Maxwell, Yoshihiko Nagai, Xu Wang, Ryan P Welch, Joon Yoon, Weihua

Zhang, Nir Barzilai, Benjamin F Voight, Bok-Ghee Han, Christopher P Jenkinson, Teemu Kuulasmaa, Johanna Kuusisto, Alisa Manning, Maggie C Y Ng, Nicholette D Palmer, Beverley Balkau, Alena Stancáková, Hanna E Abboud, Heiner Boeing, Vilmantas Giedraitis, Dorairaj Prabhakaran, Omri Gottesman, James Scott, Jason Carey, Phoenix Kwan, George Grant, Joshua D Smith, Benjamin M Neale, Shaun Purcell, Adam S Butterworth, Joanna M M Howson, Heung Man Lee, Yingchang Lu, Soo-Heon Kwak, Wei Zhao, John Danesh, Vincent K L Lam, Kyong Soo Park, Danish Saleheen, Wing Yee So, Claudia H T Tam, Uzma Afzal, David Aguilar, Rector Arya, Tin Aung, Edmund Chan, Carmen Navarro, Ching-Yu Cheng, Domenico Palli, Adolfo Correa, Joanne E Curran, Denis Rybin, Vidya S Farook, Sharon P Fowler, Barry I Freedman, Michael Griswold, Daniel Esten Hale, Pamela J Hicks, Chiea-Chuen Khor, Satish Kumar, Benjamin Lehne, Dorothée Thuillier, Wei Yen Lim, Jianjun Liu, Yvonne T van der Schouw, Marie Loh, Solomon K Musani, Sobha Puppala, William R Scott, Loïc Yengo, Sian-Tsung Tan, Herman A Taylor, Jr, Farook Thameem, Gregory Wilson, Sr, Tien Yin Wong, Pål Rasmus Njølstad, Jonathan C Levy, Massimo Mangino, Lori L Bonnycastle, Thomas Schwarzmayer, João Fadista, Gabriela L Surdulescu, Christian Herder, Christopher J Groves, Thomas Wieland, Jette Bork-Jensen, Ivan Brandslund, Cramer Christensen, Heikki A Koistinen, Alex S F Doney, Leena Kinnunen, Tõnu Esko, Andrew J Farmer, Liisa Hakaste, Dylan Hodgkiss, Jasmina Kravic, Valeriya Lyssenko, Mette Hollensted, Marit E Jørgensen, Torben Jørgensen, Claes Ladenvall, Johanne Marie Justesen, Annemari Käräjämäki, Jennifer Kriebel, Wolfgang Rathmann, Lars Lannfelt, Torsten Lauritzen, Narisu Narisu, Allan Linneberg, Olle Melander, Lili Milani, Matt Neville, Marju Orho-Melander, Lu Qi, Qibin Qi, Michael Roden, Olov Rolandsson, Amy Swift, Anders H Rosengren, Kathleen Stirrups, Andrew R Wood, Evelin Mihailov, Christine Blancher, Mauricio O Carneiro, Jared Maguire, Ryan Poplin, Khalid Shakir, Timothy Fennell, Mark DePristo, Martin Hrabé de Angelis, Panos Deloukas, Anette P Gjesing, Goo Jun, Peter Nilsson, Jacquelyn Murphy, Robert Onofrio, Barbara Thorand, Torben Hansen, Christa Meisinger, Frank B Hu, Bo Isomaa, Fredrik Karpe, Liming Liang, Annette Peters, Cornelia

Huth, Stephen P O’Rahilly, Colin N A Palmer, Oluf Pedersen, Rainer Rauramaa, Jaakko Tuomilehto, Veikko Salomaa, Richard M Watanabe, Ann-Christine Syvänen, Richard N Bergman, Dwaipayan Bharadwaj, Erwin P Bottinger, Yoon Shin Cho, Giriraj R Chandak, Juliana C N Chan, Kee Seng Chia, Mark J Daly, Shah B Ebrahim, Claudia Langenberg, Paul Elliott, Kathleen A Jablonski, Donna M Lehman, Weiping Jia, Ronald C W Ma, Toni I Pollin, Manjinder Sandhu, Nikhil Tandon, Philippe Froguel, Inês Barroso, Yik Ying Teo, Eleftheria Zeggini, Ruth J F Loos, Kerrin S Small, Janina S Ried, Ralph A DeFronzo, Harald Grallert, Benjamin Glaser, Andres Metspalu, Nicholas J Wareham, Mark Walker, Eric Banks, Christian Gieger, Erik Ingelsson, Hae Kyung Im, Thomas Illig, Paul W Franks, Gemma Buck, Joseph Trakalo, David Buck, Inga Prokopenko, Reedik Mägi, Lars Lind, Yossi Farjoun, Katharine R Owen, Anna L Gloyn, Konstantin Strauch, Tiinamaija Tuomi, Jaspal Singh Kooner, Jong-Young Lee, Taesung Park, Peter Donnelly, Andrew D Morris, Andrew T Hattersley, Donald W Bowden, Francis S Collins, Gil Atzmon, John C Chambers, Timothy D Spector, Markku Laakso, Tim M Strom, Graeme I Bell, John Blangero, Ravindranath Duggirala, E Shyong Tai, Gilean McVean, Craig L Hanis, James G Wilson, Mark Seielstad, Timothy M Frayling, James B Meigs, Nancy J Cox, Rob Sladek, Eric S Lander, Stacey Gabriel, Noël P Burt, Karen L Mohlke, Thomas Meitinger, Leif Groop, Goncalo Abecasis, Jose C Florez, Laura J Scott, Andrew P Morris, Hyun Min Kang, Michael Boehnke, David Altshuler, and Mark I McCarthy. The genetic architecture of type 2 diabetes. *Nature*, 536(7614):41–7, Aug 2016.

- [15] Christoph Lippert, Jennifer Listgarten, Robert I Davidson, Scott Baxter, Hoifung Poon, Hoifung Poong, Carl M Kadie, and David Heckerman. An exhaustive epistatic snp association analysis on expanded wellcome trust data. *Sci Rep*, 3:1099, 2013.
- [16] Gibran Hemani, Konstantin Shakhbazov, Harm-Jan Westra, Tonu Esko, Anjali K Henders, Allan F McRae, Jian Yang, Greg Gibson, Nicholas G Martin, Andres Metspalu, Lude Franke, Grant W Montgomery, Peter M Visscher, and Joseph E Powell. Detection and replication

- of epistasis influencing transcription in humans. *Nature*, 508(7495):249–53, Apr 2014.
- [17] Andrew R Wood, Marcus A Tuke, Mike A Nalls, Dena G Hernandez, Stefania Bandinelli, Andrew B Singleton, David Melzer, Luigi Ferrucci, Timothy M Frayling, and Michael N Weedon. Another explanation for apparent epistasis. *Nature*, 514(7520):E3–5, Oct 2014.
- [18] Simon G Thompson and Peter Willeit. Uk biobank comes of age. *Lancet*, 386(9993):509–10, Aug 2015.
- [19] Margit Sutrop and Kadri Simm. The estonian healthcare system and the genetic database project: from limited resources to big hopes. *Camb Q Healthc Ethics*, 13(3):254–62, 2004.
- [20] Peter M Visscher, William G Hill, and Naomi R Wray. Heritability in the genomics era—concepts and misconceptions. *Nat Rev Genet*, 9(4):255–66, Apr 2008.
- [21] B Devlin, M Daniels, and K Roeder. The heritability of iq. *Nature*, 388(6641):468–71, Jul 1997.
- [22] Jenny van Dongen, Gonneke Willemsen, Wei-Min Chen, Eco J C de Geus, and Dorret I Boomsma. Heritability of metabolic syndrome traits in a large population-based sample. *J Lipid Res*, 54(10):2914–23, Oct 2013.
- [23] Christof Geisen, Matthias Watzka, Katja Sittinger, Michael Steffens, Laurynas Daugela, Erhard Seifried, Clemens R Müller, Thomas F Wienker, and Johannes Oldenburg. Vkorc1 haplotypes and their impact on the inter-individual and inter-ethnic variability of oral anticoagulation. *Thromb Haemost*, 94(4):773–9, Oct 2005.
- [24] Sandhya Pruthi, Bobbie S Gostout, and Noralane M Lindor. Identification and management of women with brca mutations or hereditary predisposition for breast and ovarian cancer. *Mayo Clin Proc*, 85(12):1111–20, Dec 2010.
- [25] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative

- analysis of large gene lists using david bioinformatics resources. *Nat Protoc*, 4(1):44–57, 2009.
- [26] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–50, Oct 2005.
- [27] Bradley Efron and Robert Tibshirani. On testing the significance of sets of genes. *The Annals of Applied Statistics*, 1(1):107–129, 2007.
- [28] Jelle J Goeman and Peter Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–7, Apr 2007.
- [29] Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1 of *Institute of mathematical statistics monographs*. Cambridge University Press, Cambridge, 2010.
- [30] B Devlin and K Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, Dec 1999.
- [31] Bradley Efron. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004.
- [32] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8):904–9, Aug 2006.
- [33] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, Matthew Stephens,

- and Carlos D Bustamante. Genes mirror geography within europe. *Nature*, 456(7218):98–101, Nov 2008.
- [34] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjálmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, Nick Patterson, and Alkes L Price. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet*, 47(3):284–90, Mar 2015.
- [35] L P O’Neill and B M Turner. Histone h4 acetylation distinguishes coding regions of the human genome from heterochromatin in a differentiation-dependent but transcription-independent manner. *EMBO J*, 14(16):3946–57, Aug 1995.
- [36] L P O’Neill and B M Turner. Immunoprecipitation of chromatin. *Methods Enzymol*, 274:189–97, 1996.
- [37] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–37, May 2007.
- [38] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–502, Jun 2007.
- [39] Tarjei S Mikkelsen, Manching Ku, David B Jaffe, Biju Issac, Erez Lieberman, Georgia Giannoukos, Pablo Alvarez, William Brockman, Tae-Kyung Kim, Richard P Koche, William Lee, Eric Mendenhall, Aisling O’Donovan, Aviva Presser, Carsten Russ, Xiaohui Xie, Alexander Meissner, Marius Wernig, Rudolf Jaenisch, Chad Nusbaum, Eric S Lander, and Bradley E Bernstein. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–60, Aug 2007.
- [40] S C Elgin. The formation and function of dnase i hypersensitive sites in the process of gene activation. *J Biol Chem*, 263(36):19259–62, Dec 1988.

- [41] W A Scott and D J Wigmore. Sites in simian virus 40 chromatin which are preferentially cleaved by endonucleases. *Cell*, 15(4):1511–8, Dec 1978.
- [42] C Wu. The 5' ends of drosophila heat shock genes in chromatin are hypersensitive to dnase i. *Nature*, 286(5776):854–60, Aug 1980.
- [43] D J Galas and A Schmitz. Dnase footprinting: a simple method for the detection of protein-dna binding specificity. *Nucleic Acids Res*, 5(9):3157–70, Sep 1978.
- [44] J D McGhee, W I Wood, M Dolan, J D Engel, and G Felsenfeld. A 200 base pair region at the 5' end of the chicken adult beta-globin gene is accessible to nuclease digestion. *Cell*, 27(1 Pt 2):45–55, Nov 1981.
- [45] T Grange, E Bertrand, M L Espinás, M Fromont-Racine, G Rigaud, J Roux, and R Pictet. In vivo footprinting of the interaction of proteins with dna and rna. *Methods*, 11(2):151–63, Feb 1997.
- [46] Alan P Boyle, Sean Davis, Hennady P Shulha, Paul Meltzer, Elliott H Margulies, Zhiping Weng, Terrence S Furey, and Gregory E Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–22, Jan 2008.
- [47] Jay R Hesselberth, Xiaoyu Chen, Zhihong Zhang, Peter J Sabo, Richard Sandstrom, Alex P Reynolds, Robert E Thurman, Shane Neph, Michael S Kuehn, William S Noble, Stanley Fields, and John A Stamatoyannopoulos. Global mapping of protein-dna interactions in vivo by digital genomic footprinting. *Nat Methods*, 6(4):283–9, Apr 2009.
- [48] Shane Neph, Jeff Vierstra, Andrew B Stergachis, Alex P Reynolds, Eric Haugen, Benjamin Vernot, Robert E Thurman, Sam John, Richard Sandstrom, Audra K Johnson, Matthew T Maurano, Richard Humbert, Eric Rynes, Hao Wang, Shiny Vong, Kristen Lee, Daniel Bates, Morgan Diegel, Vaughn Roach, Douglas Dunn, Jun Neri, Anthony Schafer, R Scott Hansen, Tanya Kutuyavin, Erika Giste, Molly Weaver, Theresa Canfield, Peter Sabo, Miaohua Zhang,



- Gayathri Balasundaram, Rachel Byron, Michael J MacCoss, Joshua M Akey, M A Bender, Mark Groudine, Rajinder Kaul, and John A Stamatoyannopoulos. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90, Sep 2012.
- [49] Robert E Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T Maurano, Eric Haugen, Nathan C Sheffield, Andrew B Stergachis, Hao Wang, Benjamin Vernot, Kavita Garg, Sam John, Richard Sandstrom, Daniel Bates, Lisa Boatman, Theresa K Canfield, Morgan Diegel, Douglas Dunn, Abigail K Ebersol, Tristan Frum, Erika Giste, Audra K Johnson, Ericka M Johnson, Tanya Kutuyavin, Bryan Lajoie, Bum-Kyu Lee, Kristen Lee, Darin London, Dimitra Lotakis, Shane Neph, Fidencio Neri, Eric D Nguyen, Hongzhu Qu, Alex P Reynolds, Vaughn Roach, Alexias Safi, Minerva E Sanchez, Amartya Sanyal, Anthony Shafer, Jeremy M Simon, Lingyun Song, Shinny Vong, Molly Weaver, Yongqi Yan, Zhancheng Zhang, Zhuzhu Zhang, Boris Lenhard, Muneesh Tewari, Michael O Dorschner, R Scott Hansen, Patrick A Navas, George Stamatoyannopoulos, Vishwanath R Iyer, Jason D Lieb, Shamil R Sunyaev, Joshua M Akey, Peter J Sabo, Rajinder Kaul, Terrence S Furey, Job Dekker, Gregory E Crawford, and John A Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, Sep 2012.
- [50] Roger Pique-Regi, Jacob F Degner, Athma A Pai, Daniel J Gaffney, Yoav Gilad, and Jonathan K Pritchard. Accurate inference of transcription factor binding from dna sequence and chromatin accessibility data. *Genome Res*, 21(3):447–55, Mar 2011.
- [51] Myong-Hee Sung, Michael J Guertin, Songjoon Baek, and Gordon L Hager. Dnase footprint signatures are dictated by factor dynamics and dna sequence. *Mol Cell*, 56(2):275–85, Oct 2014.
- [52] Eduardo G Gusmao, Manuel Allhoff, Martin Zenke, and Ivan G Costa. Analysis of computational footprinting methods for dnase sequencing experiments. *Nat Methods*, 13(4):303–9, Apr 2016.

- [53] ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.
- [54] Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, Viren Amin, John W Whitaker, Matthew D Schultz, Lucas D Ward, Abhishek Sarkar, Gerald Quon, Richard S Sandstrom, Matthew L Eaton, Yi-Chieh Wu, Andreas R Pfenning, Xinchun Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R Alan Harris, Noam Shores, Charles B Epstein, Elizabetha Gjoneska, Danny Leung, Wei Xie, R David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K Canfield, R Scott Hansen, Rajinder Kaul, Peter J Sabo, Mukul S Bansal, Annaick Carles, Jesse R Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R Mercer, Shane J Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C Sallari, Kyle T Siebenthal, Nicholas A Sinnott-Armstrong, Michael Stevens, Robert E Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E Beaudet, Laurie A Boyer, Philip L De Jager, Peggy J Farnham, Susan J Fisher, David Haussler, Steven J M Jones, Wei Li, Marco A Marra, Michael T McManus, Shamil Sunyaev, James A Thomson, Thea D Tlsty, Li-Huei Tsai, Wei Wang, Robert A Waterland, Michael Q Zhang, Lisa H Chadwick, Bradley E Bernstein, Joseph F Costello, Joseph R Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A Stamatoyannopoulos, Ting Wang, and Manolis Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–30, Feb 2015.
- [55] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyaev, Sandra Stehling-Sun, Audra K Johnson, Theresa K Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R Scott Hansen, Shane Neph,

- Peter J Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R Sunyaev, Rajinder Kaul, and John A Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–5, Sep 2012.
- [56] Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verner Anttila, Han Xu, Chongzhi Zang, Kyle Farh, Stephan Ripke, Felix R Day, ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, RACI Consortium, Shaun Purcell, Eli Stahl, Sara Lindstrom, John R B Perry, Yukinori Okada, Soumya Raychaudhuri, Mark J Daly, Nick Patterson, Benjamin M Neale, and Alkes L Price. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet*, 47(11):1228–35, Nov 2015.
- [57] C Nüsslein-Volhard and E Wieschaus. Mutations affecting segment number and polarity in drosophila. *Nature*, 287(5785):795–801, Oct 1980.
- [58] O. Loudet, S. Chaillou, C. Camilleri, D. Bouchez, and F. Daniel-Vedele. Bay-0 x shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in arabidopsis. *Theor Appl Genet*, 104(6-7):1173–1184, May 2002.
- [59] GTEx Consortium. Human genomics. the genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–60, May 2015.
- [60] Julien Bryois, Alfonso Buil, David M Evans, John P Kemp, Stephen B Montgomery, Donald F Conrad, Karen M Ho, Susan Ring, Matthew Hurles, Panos Deloukas, George Davey Smith, and Emmanouil T Dermitzakis. Cis and trans effects of human genomic variants on gene expression. *PLoS Genet*, 10(7):e1004461, Jul 2014.
- [61] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. Ld score regression distinguishes

- confounding from polygenicity in genome-wide association studies. *Nat Genet*, 47(3):291–5, Mar 2015.
- [62] Iain Mathieson and Gil McVean. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet*, 44(3):243–6, Feb 2012.
- [63] Jennifer Listgarten, Christoph Lippert, and David Heckerman. Fast-lmm-select for addressing confounding from spatial structure and rare variants. *Nat Genet*, 45(5):470–1, May 2013.
- [64] B. Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics Monographs. Cambridge University Press, 2010.
- [65] Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–93, Oct 2009.
- [66] Daniel Marbach, David Lamparter, Gerald Quon, Manolis Kellis, Zoltán Kutalik, and Sven Bergmann. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat Methods*, 13(4):366–70, Apr 2016.
- [67] Lisa Ann Cirillo, Frank Robert Lin, Isabel Cuesta, Dara Friedman, Michal Jarnik, and Kenneth S Zaret. Opening of compacted chromatin by early developmental transcription factors hnf3 (foxa) and gata-4. *Mol Cell*, 9(2):279–89, Feb 2002.
- [68] Abdenour Soufi, Meilin Fernandez Garcia, Artur Jaroszewicz, Nebiyu Osman, Matteo Pellegrini, and Kenneth S Zaret. Pioneer transcription factors target partial dna motifs on nucleosomes to initiate reprogramming. *Cell*, 161(3):555–68, Apr 2015.

- [69] Makiko Iwafuchi-Doi and Kenneth S Zaret. Pioneer transcription factors in cell reprogramming. *Genes Dev*, 28(24):2679–92, Dec 2014.
- [70] Richard I Sherwood, Tatsunori Hashimoto, Charles W O’Donnell, Sophia Lewis, Amira A Barkal, John Peter van Hoff, Vivek Karun, Tommi Jaakkola, and David K Gifford. Discovery of directional and nondirectional pioneer transcription factors by modeling dnase profile magnitude and shape. *Nat Biotechnol*, 32(2):171–8, Feb 2014.