# The interplay between genetics, the microbiome, DNA-methylation & gene-expression

Bonder, Marc Jan

# Discussion

# 10

## 1. Understanding health and disease by studying genetics and the environment

In recent years there has been a focus on explaining phenotypes and diseases using genetic approaches. So far, hundreds of genome-wide association studies (GWAS) in many different disease phenotypes and traits have been performed and led to the identification of thousands of genetic variants that influence the predisposition to common diseases[1]. For instance, large case-control GWAS have been performed on inflammatory bowel disease[2] (IBD), type 2 diabetes[3] (T2D), and on common traits such as height[4] and body mass index[5] (BMI). Using variants identified by GWAS we can explain a proportion of the variation in traits between individuals. Expanding the GWAS studies with more samples leads to more loci being identified and more accurate effect-size estimates. For example, in 2008, three independent GWAS studies on height were published at the same time[6–8] and reported between 12 and 27 genome-wide significantly associated loci ($P < 5 \times 10\text{-}8$) in studies of between 15,000 and 30,000 samples. These height-associated loci explained up to 3.7% of individual variation in height. In the following years, the sample sizes increased almost 10 times, to 253,288 individuals, and the number of genome-wide significantly associated height loci (or SNPs) rose sharply to 697[4]. The explained variance in height has now risen to 16% when considering these 697 SNPs, or even to 29% if we also include SNPs that are likely relevant but that did not reach genome-wide significance4. It is to be expected that even larger GWAS studies on height and other phenotypes will yield many more variants and more accurate effect-sizes for the variants identified, such that the proportion of explained variance will rise even further.

However, recent studies have also shown there is a limit to the extent that common genetic variants can be used for predicting disease or other phenotypes. For instance, it has been estimated that approximately 30 – 40% of the variation in height is unlikely to be explained by genetic variants[9]. This is because, for nearly all complex diseases and traits (such as height), a substantial part of the phenotypic variation is determined by environmental factors and potentially by the interplay between genetic and environmental variation[10] (such as exposure to smoking or medication, degree of physical activity, and dietary habits).

Given the recent success in identifying genetic risk factors for disease, the clear importance of environmental risk factors, and, most importantly, the increasing availability of large cohort studies, it is becoming possible to identify which 'omics' layers are influenced by genetic and environmental factors, and how these interact. These large cohort studies have not only looked at genetic variation in the participants, but also generated data on gene expression, methylation, the microbiome, metabolites, and many other phenotypes for them. In this thesis we have identified many of these relations. I will discuss below how these studies have aided the interpretation of genetic risk factors and the usefulness of multi-omics in predicting disease and other phenotypes.

## 2. Integrating multi-omics data to gain more insight into the functional consequences of genetic variants

Despite the tremendous success in identifying genetic variants associated to traits and diseases, there are several challenges in interpreting the associated signals. It is important to realize that approximately 88% of variants associated to common diseases are located in non-coding regions[11], which hampers the interpretation of these variants because it is often very difficult to infer the likely causal gene for a genomic locus.

One way to overcome this difficulty, and to gain more knowledge on the likely causal genes and pathways, is to ascertain whether these genetic variants are also affecting gene expression or methylation levels, for instance[11]. Given the availability of cohort studies with one or multiple-omics datasets, it is now possible to ascertain such quantitative traits in a large number of samples and at high resolution. In this thesis we have conducted three quantitative trait studies: we studied the effects of *cis* (i.e. local) expression quantitative trait locus (*cis*-eQTL) and methylation quantitative trait locus (*cis*-meQTL) in four different tissues (chapter seven);

we studied *cis* and *trans* (distal) effects on DNA-methylation in blood (chapter nine); and we studied how genetic variants affect the human gut microbiome (chapter six).

We observed that many genetic risk factors affect gene expression and methylation in *cis*, but also identified *trans*-eQTLs and *trans*-meQTLs (chapter nine), indicating that it is also possible to identify the downstream pathways disrupted by these disease-associated variants. This was particularly clear for *trans*-meQTLs, where 31% of the tested genetic risk factors were affecting methylation in CpG sites in *trans*. However, since the biology of methylation is less well understood than gene expression, we overlapped the *trans*-meQTLs with identified eQTLs to aid in their interpretation. We found over 240 *trans*-eQTLs overlapping with *trans*-meQTL signals, i.e. roughly 9% of the total *trans*-meQTLs are also identified in expression, based on the links found between SNPs and methylation, and between methylation and gene expression. Over 90% of the identified *trans*-eQTLs had a concordant allelic direction based on the relations seen between methylation and gene-expression. Furthermore, we observed several instances where *trans*-meQTLs were likely caused by differences in transcription factor abundances, which were due to a *cis*-eQTL gene expression effect on the transcription factor gene (chapter nine, figure 4). Consequently, the binding sites for this DNA transcription factor are more or less occupied, and this leads to either methylation or demethylation through an unknown molecular mechanism. For example, we observed that a genetic variant, SNP rs3774959, that affects gene expression levels of the nearby transcription factor *NFKB1*[12] also influenced methylation at 413 downstream sites; nearly all these sites overlap with binding sites of the NFKB transcription factor. This example also shows the value of multi-omics studies, because the identified eQTLs were instrumental in helping us understand the nature of a massive number of *trans*-meQTLs from this SNP. Yet, since most of the *trans*-meQTLs that we detected remain unexplained, it is possible that some of the biology underlying these epigenetic signals will become clear from incorporating additional omics data, for instance, on protein or metabolite levels or on the microbiome.

## 3 Integration of multi-omics data to explain variation in phenotypes

Apart from using multi-omics datasets to better understand the downstream molecular consequences of genetic risk factors, these data can also be used to study non-genetic effects on diseases and traits in detail. Examples of such analyses are presented in chapter four, where we aimed to explain variation in lipid levels and BMI by using the gut microbiome as a predictor, and in chapter eight, where we aimed to explain variation in height and BMI by using methylation as a predictor. Both these studies were performed using data from the LifeLines-DEEP[13] project. In this cohort study we collected extensive phenotype information on 1,500 participants and on their genotypes, microbiome composition, gene expression, DNA-methylation, metabolite and cytokine levels. Since the data was all generated at the same time-point, this study is uniquely suited to conducting multi-omics integration studies, which link more than two data-layers to each other.

In chapter eight we describe a study on three groups of samples from the LifeLines-DEEP[13] population cohort, the Lothian Birth Cohorts[14], and the Brisbane systems genetic study[15]. All three datasets provided information on genetics and DNA-methylation. We first built a predictor for the height of an individual, based solely either on genetics or on DNA-methylation levels. By across dataset training and testing the models, we found that a DNA-methylation based predictor could explain 0.76% of the variation in height and that genetics could explain 19.8%. Since the genetic predictor explains almost 25 times more variation in height as compared to the methylation predictor, there was little to be gained by making a combined predictor for height. This was expected since height is known to be mostly heritable, with its heritability estimated at 60% – 80% by twin and family-based studies[9,16,17].

However, when applying the same method to BMI, we found that DNA-methylation levels could explain up to 7.3% of the variation in BMI, whereas genetics could explain up to 9.4%. A combined predictor could explain 13.6% of the variation in BMI. This shows clearly that prediction performance for BMI increases substantially when using multi-omics datasets.

Similarly, in chapter four we describe a study using combined genetic and microbiome-based predictors to explain variation in lipid levels and BMI based on the microbiome composition. In this study we used only data from the LifeLines-DEEP cohort and, in order to get accurate predictions, we used cross-validation to estimate the explained variation. We found that for three of the five traits tested (i.e. high density cholesterol, triglycerides, and BMI), the combined genetic and microbiome predictor performed significantly better than a predictor using solely genetic data. With the microbiome composition data we were able to explain up to 4.5% in variation in BMI, while with genetics we could explain 2.1%. (The difference in explained variance by genetic data in chapter eight and chapter four is due to a different significance threshold for selecting SNPs in the genetic predictor for BMI). In total, when we also took age and gender into account, we were able to explain 11.3% of the variation in BMI with genetics and microbiome composition.

When we combined the predictors for BMI described in chapter four and chapter eight into one predictor incorporating three levels of data (genetic variation, methylation and microbiome data; see figure 1), we observed that the combined predictor could explain even more of the variation in BMI than the predictors that used only two omics levels. The explained variation shown in figure 1 was re-estimated based on the original data reported for the three cohorts and our estimates were based on a subset of samples for which all three data layers were available. This means the explained variation is different than that reported in chapter four and chapter eight. In our combined analysis age and gender explain 4.2%, the microbiome composition explains 3.0%, DNA-methylation levels explain 6.9%, and genetic variation explains 8.6% of the variation in BMI. The predictor that uses age, gender, genetic variation, DNA-methylation and microbiome composition can explain 20.3% of the variation in BMI. It is evident that the predictors are not independent, as the summed explained variation is higher (22.8%) than the explained variation by the combined predictor (20.3%). This is because the different omics datasets do not provide fully independent information (for example, genetic variation can influence DNA methylation, as described in chapter nine, and can also affect microbiome composition, as described in chapter six). However, the difference between the five-level predictor and the best four-level predictor remains statistically significant, Anova P-value = 0.0003, indicating that a multi-omics predictor can be helpful in predicting complex disease and trait phenotypes.

In the BMI example above, we see that the use of non-genetic information in predicting risk can improve the phenotype prediction substantially. Although the prediction of BMI levels is not directly relevant in a clinical manner (because BMI can usually be measured), the same method can also be applied to predict disease phenotypes. Given that most large-scale GWAS analyses have so far found only a limited number of risk factors, which usually explain only a limited proportion of phenotype variation, the development of models that incorporate gene expression, methylation, metabolite or microbiome data or a subset of these is promising and can be informative in predicting disease and trait phenotypes. For instance, it would be clinically relevant to predict risk for cardiovascular diseases (CVD) or T2D; although enormous progress has been made in prevention and treatment since the 1960s, CVD remains the leading cause of death worldwide. It has been estimated that nearly 40% of all deaths in the US will be due to CVD by 2030[18]. Since CVD and T2D are chronic diseases and expensive to treat, early diagnosis would permit better monitoring of disease progression and might allow less expensive treatments if the diagnosis is made while symptoms are still mild.

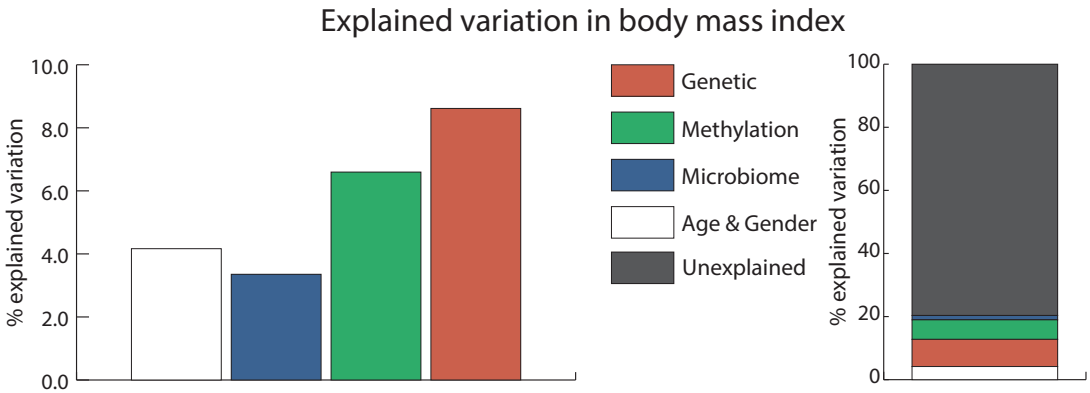


*Figure 1. Multi-omics data can explain some of the variation in body mass index. This figure shows the variation explained by the different biological data layers and phenotypes. In the left part of the figure, the variation explained by the separate predictors on each layer is shown. In the right part of the figure the total variation explained by the combined model is shown. Here we started with age and gender, then included genetics, methylation and finally the microbiome predictor, so that the explained variation is attributed to the features based on the individual explained prediction.*

Another use of integrative multi-omics predictors is to help identify environmental factors leading to disease in large cohorts, like smoking. Smoking has an influence on both the microbiome (see chapter five), and methylation levels[19,20]. Using the methylation and microbiome data, it might be possible to accurately predict smoking status without needing to ask participants in biobanks about their smoking behavior. By using such integrated approaches, we may also be able to identify mismatches between the different data-layers, which could then help correct potential sample mix-up problems, or identify which answers in the lifestyle questionnaires have not been filled out accurately. Furthermore, if we can infer clinically relevant diseases or phenotypes accurately, it might be possible to save biobanking costs since phenotypes are typically expensive to measure.

These multi-omics datasets also hold great promise for personalized approaches: to better tailor nutrition[21], treatment or medication to individual patients. Currently a lot of research is being conducted into such precision medicine[22]. For instance, genetic data can be used to partly determine the appropriate type of medication or dosage for treating patients. Such patient-specific prescriptions may also benefit from the use of other omics data, thereby leading to a more accurate type of medication prescription and dosage better tailored to individual patients, possible cost reductions and fewer side-effects. Likewise, we can use the omics data to offer a more personalized nutritional advice, as reported by Zeevi et al[23].

However, before we can start using these predictors in a diagnostic or clinical setting, it is crucial to test how they work thoroughly[24]. Firstly, these models have to reach a sufficiently high, prediction accuracy before they can be adopted for clinical use. With the exception of the genetic data, where it is clear that genetic variation is causal to the phenotype, the other omics data may not be so easy to use, for instance, expression data in blood is not identical at two different time-points, even if the blood samples have been drawn from an individual on the same day[25]. So relations identified at the first time-point might not be well reflected at the second time-point. Although we observed that these omics levels were informative for BMI and lipid levels, an important step is to perform longitudinal analyses and, ideally, also to analyze omics data from multiple time-points. By using more time-points we can learn about the stability of the various omics layers and phenotypes over time. And by using a predictor built on the expression and methylation levels at time-point X to predict the level of a trait at time-point Y, we can learn more about the usability and stability of the non-genetic predictors in general. This research will eventually lead to better predictions on the phenotypes of

interest and will also provide insight into whether each layer of omics data are informative for predicting disease development at a later stage.

For this purpose, it would be interesting to determine the power of the current predictors in the follow-up data from the LifeLines and LifeLines-DEEP cohorts. At this second time-point, five years after the initial sample collection, we could assess the stability of the non-genetic predictors, which would tell us more about the relevance of the information captured in the microbiome composition, expression data, and methylation data.

Furthermore, it is important to gain a better understanding of how the different biological data layers interact which each other and how they relate to confounding factors. In chapters two, three and five, we have studied potential confounding factors for studies on the microbiome. We identified relations between different dietary factors, commonly used drugs, intrinsic factors, diseases, and smoking on the microbiome composition in the LifeLines-DEEP cohort. More specifically, we modeled all of these factors simultaneously in chapter five, while taking into account relations between the individual factors, rather than studying all factors separately. Studies linking the microbiome or other biological omics data layers to phenotypes, or preferably to multiple phenotypes at the same time, are proving valuable in identifying unwanted confounders, which can vary per trait. We have provided insight into some of these confounders, which we hope will enable the creation of better models yielding better reproducible results.

While the generation of such multi-omics datasets is rather costly, this should be considered against their potential to aid the earlier diagnosis and treatment of patients. In the United States alone, healthcare costs related to T2D were estimated to be US$176 billion in 2012[26]. If we can achieve earlier diagnoses using good predictors, this may limit future treatment costs. Given that the risk of development of either T2D or CVD can be limited by lifestyle changes, earlier diagnosis could have a major impact on quality of life and reduce treatment costs. The LifeLines-DEEP dataset does not include a sufficient number of participants suffering from CVD or T2D for us to be able to build and test specific predictors. However, since we can better predict lipid levels and BMI with the multi-omics predictors, I believe this approach can also be applied to case-control analyses in both CVD and T2D cohorts.

## 4. Future perspectives

Several avenues are important for further research using multi-omics integration approaches in the interpretation of genetic risk variants and prediction of diseases and traits. Firstly, both the GWAS interpretation and predictor development will benefit from having **larger datasets** available. The larger datasets will aid identification of smaller effects and offer better effect size estimations. Secondly, by using **higher resolution**[27]**, multi-omics datasets**[28], such as single cell data, and the integration of data on multiple tissues, we can improve interpretation of GWAS data and explain variation in traits. Thirdly, by using **more different omics datasets simultaneously**, and fourthly, by studying the **most relevant tissues**[29]**,** we can improve both the interpretation of the GWAS results and the accuracy and power of multi-omics predictors. In the next two sections I will describe the prospects for two multi-omics integration strategies.

### 4.1 Multi-omics in interpretation of genetic variants

Specifically for the interpretation of GWAS results using multi-omics, there is much to gain by studying different omics layers simultaneously. In chapter nine, we describe an overlap we identified between *cis*-eQTLs on transcription factors and *trans*-methylation QTLs on downstream binding sites for these transcription factors.

The next step in these analyses would be scaling to a genome-wide *trans*-mapping to identify more relations between genetic variation and DNA-methylation changes. This would help identify genetic regulation of genome-wide binding proteins or DNA binding transcripts. By direct integration of genetic, expression and methylation data, we may be able to identify more

relations between genetic, (distal) methylation changes and expression. This could enable us to explain more relations between methylation and transcription, and to learn more about the relations where there is no clear transcription factor binding motif or footprint, which proved crucial in interpreting the trans-methylation QTLs (chapter nine). By directly integrating the data levels, it might even be possible to do a "DNA-binder wide" analysis, mapping all the DNA-binding proteins and transcripts on the whole genome in one analysis. This could make such an approach an alternative to CHromatin immunoPrecipitation sequencing (CHiP-seq) analysis, which is performed per protein. The multi-omics-based alternative would have an advantage over CHiP-seq methods, because it would not only work for proteins that bind to DNA but also for RNAs that bind to the DNA. However, it is important to note that we used the Illumina 450K array, which assays only part of the DNA-methylome[30]. Switching to genome-wide bisulfite sequencing, or a genome-wide sequencing approach that can identify methylation without bisulfite treatment, would provide a full picture, but this is likely to remain expensive for the next few years[30]. In addition to the drawback of genomic resolution, using combined expression and methylation QTL mapping to identify downstream binding of proteins or RNAs on the DNA may not have the same precision as CHiP-seq. Firstly, the precision is affected because there is a non-perfect relation between expression levels and protein levels, i.e. the functional level of transcription factors. Secondly, protein levels might not be directly representative for the binding of the same protein to the DNA. To get to a better "DNA-binder wide" analysis, it might be worthwhile to integrate protein levels of transcription factors in the analysis, thereby giving a better understanding of how expression levels of transcription factors relate to protein-levels. This would close the gap between the difference in expression and protein levels[31].

By moving to more precise data sets, like single cell data[27], where information on gene expression and/or DNA-methylation is generated at a single cell level, we can learn much more about the relation between different omics levels, because many of these relations are highly cell-type- and context-specific. In this thesis we only investigated information at a bulk level, thus missing much context-specific information, i.e. differences from cell to cell, or even information on differences per cell type. For our research we mainly used data derived from whole blood, which is a mixture of cell types, and since there are differences between the cells and the cell types that make up blood, it is possible that we missed genetic effects that are only present in the rarer cell types. This also has implications when trying to interpret GWAS signals. By using the appropriate (i.e. affected) cell type or cells for a disease or trait, more specific information on the effect of genetic variation will be revealed. We know that genetic regulation can be tissue-specific (chapter seven) and that using the correct tissue type is relevant, for example, when integrating GWAS data and multi-omics data derived from colonic biopsies with the gut microbiome composition. If we could perform such studies we could learn more about the relation between the host and microbiome, whereas using expression data derived from blood might not be truly reflective of the gut situation. Especially for gut diseases, like IBD and celiac disease, this could yield new insights. We envision that, using these data, we would also learn more about the microbiome QTLs identified in our microbiome quantitative trait study (chapter six). Currently, there are no such experimental results available in which there is data on intestinal DNA-methylation levels, gene-expression levels, and the gut microbiome. The generation of such datasets is challenging, since it is not possible to collect the stool microbiome at the same time-point as a biopsy; this is important to relate human gene expression, microbial gene activity, and methylation levels to each other properly.

## 4.2 Multi-omics prediction models

Many biobanks are now enriching their datasets with additional molecular levels. One issue is the actual usefulness of generating another molecular level on top of the multi-omics datasets that may already be available. To shed light on this, we have expanded our BMI predictor with the expression data available for the LifeLines-DEEP cohort. By using the same strategy as for microbiome composition (see methods, chapter four), I built a classifier based on expression

data. Then I combined the classifier into our multi-omics BMI predictor, which now integrates phenotypes, genetic variation, DNA-methylation, gene expression, and microbiome data to predict BMI. I observed that expression alone can explain 9.0% of the variation in BMI, and after including gene-expression data in the combined predictor, we could explain 26.5% of the total variation in BMI, which represents a 6.2% increase over the predictor shown in figure 1. This clearly shows that every additional data layer added in our model makes it possible to predict trait levels more accurately.

I therefore expect that by adding additional datasets, such as more detailed information on phenotype, on physical activity (measured by a wearable device for example), on food intake or quality of life, it is likely the prediction performance will be improved. Another avenue would be to generate molecular data on cell types and tissues that are more relevant for the disease or phenotype under investigation. For instance, for inflammatory bowel disease and celiac disease, the most relevant tissue to study would be the intestine, and ideally this should involve single-cell technologies that can provide information on every individual cell that is present in the intestine.

Another strategy would be to develop better predictive models for phenotypes by integrating the omics layers in a different way. In the analyses described in this thesis, we extracted relevant features from each of the omics layers independently: we identified methylation CpG sites that were informative, genes whose expression levels were informative, and individual bacteria that were informative for the trait of interest. We then built models per data-layer, which we subsequently integrated at model level to develop a final model based on all the individual predictors. While this approach works well as can be seen in the research chapters, it ignores more complex, but potentially informative features that may serve as features for phenotype prediction. For instance, we might well find that a certain combination of two variables, or a certain ratio of two variables that have been assayed using different techniques, could be even more informative for a phenotype, but that these features will have been missed in our current approaches. Methods that can resolve this issue and that identify such relations, while avoiding overfitting the data, could therefore prove highly valuable to better model and predict complex phenotypes. Another in the future.

Yet another strategy to improve these phenotype predictions would be to use larger sample sizes. In the risk prediction studies (chapters four and eight), we used information from large-scale GWAS meta-analyses and the reported effect-sizes on individual SNPs to build our genetic risk predictors. It is expected that the same strategy can be applied to other molecular levels: information obtained from large-scale epigenome-wide association studies (EWAS) or transcriptome-wide association studies (TWAS) could be used to select informative CpG sites and individual genes, along with reported effect-sizes to improve phenotype predictions in smaller multi-omics datasets, as in the LifeLines-DEEP. However, with ever-increasing model complexities, cross-validation or preferably replication of the models built or the effects identified, will be absolutely crucial. This is especially true if we try to generate more sophisticated integrative models, using several multi-omics datasets in one predictor.

Thus, multi-omics prediction of disease phenotypes is likely to be an interesting research avenue for the next few years. First of all, we have shown that these multi-omics predictors work substantially better in predicting BMI than methods that use more limited levels. An important question is to understand whether such predictors can also help to predict the onset of complex traits like obesity or diseases. If this turns out to be the case, it is likely that some of the features that have been selected for the predictor are indeed reflecting genes that could play a causal role in causing obesity, and which might therefore serve as potential drug targets. Given that we identified specific genes whose expression levels are associated to BMI, and that we could employ Mendelian randomization approaches to infer a causal role[32], we might well be able to correct the expression of these genes by medication, or we might be able to alter the gut microbiome in such a way that we promote or decrease the levels of BMI-relevant microbes.

# Conclusions

From the work presented in this thesis, we have gained better insight into the downstream effects of genetic risk factors for disease, through integration of different biological omics datasets. We further studied and identified factors which influence or shape the gut microbiome. Finally, we have shown that multi-omics datasets can be used to predict disease phenotypes and that each of these separate molecular levels can be informative. We are now beginning to use a wide-scale multi-omics approach for predicting traits and we expect that larger and more deeply characterized biobank cohorts will make it possible to build prediction algorithms to help diagnose patients earlier and to improve their treatment.

# References

1. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 42, D1001-6 (2014).
2. Jostins, L. et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491, 119–124 (2012).
3. Reynisdottir, I. et al. Localization of a susceptibility gene for type 2 diabetes to chromosome 5q34-q35.2. Am. J. Hum. Genet. 73, 323–35 (2003).
4. Wood, A. R. et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. 46, 1173–1186 (2014).
5. Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. Nature 518, 197–206 (2015).
6. Weedon, M. N. et al. Genome-wide association analysis identifies 20 loci that influence adult height. Nat. Genet. 40, 575–583 (2008).
7. Lettre, G. et al. Identification of ten loci associated with height highlights new biological pathways in human growth. Nat. Genet. 40, 584–591 (2008).
8. Gudbjartsson, D. F. et al. Many sequence variants affecting diversity of adult human height. Nat. Genet. 40, 609–615 (2008).
9. Yang, J. et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat. Genet. 47, 1114–1120 (2015).
10. Purcell, S. Variance Components Models for Gene–Environment Interaction in Twin Analysis. Twin Res. 5, 554–571 (2002).
11. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: illuminating the dark road from association to function. Am. J. Hum. Genet. 93, 779–797 (2013).
12. Zhernakova, D. et al. Hypothesis-free identification of modulators of genetic risk factors. bioRxiv 33217 (2015). doi:10.1101/033217
13. Scholtens, S. et al. Cohort Profile: LifeLines, a three-generation cohort study and biobank. Int. J. Epidemiol. 44, 1172–1180 (2015).
14. Deary, I. J., Gow, A. J., Pattie, A. & Starr, J. M. Cohort profile: the Lothian Birth Cohorts of 1921 and 1936. Int. J. Epidemiol. 41, 1576–1584 (2012).
15. McRae, A. F. et al. Contribution of genetic variation to transgenerational inheritance of DNA methylation. Genome Biol. 15, R73 (2014).
16. Silventoinen, K. et al. Heritability of Adult Body Height: A Comparative Study of Twin Cohorts in Eight Countries. Twin Res. 6, 399–408 (2003).
17. Hemani, G. et al. Inference of the genetic architecture underlying BMI and height with the use of 20,240 sibling pairs. Am. J. Hum. Genet. 93, 865–875 (2013).
18. Heidenreich, P. A. et al. Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association. Circulation 123, 933–44 (2011).
19. Dogan, M. V et al. The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women. BMC Genomics 15, 151 (2014).

20. Monick, M. M. et al. Coordinated changes in AHRR methylation in lymphoblasts and pulmonary macrophages from smokers. Am. J. Med. Genet. B. Neuropsychiatr. Genet. 159B, 141–151 (2012).
21. Zeevi, D. et al. Personalized Nutrition by Prediction of Glycemic Responses. Cell 163, 1079–1095 (2015).
22. Offit, K. Personalized medicine: new genomics, old lessons. Hum. Genet. 130, 3–14 (2011).
23. Zeevi, D. et al. Personalized Nutrition by Prediction of Glycemic Responses. Cell 163, 1079–1094 (2015).
24. Harper, R. & Reeves, B. Reporting of precision of estimates for diagnostic accuracy: a review. BMJ 318, 1322–3 (1999).
25. Whitney, A. R. et al. Individuality and variation in gene expression patterns in human blood. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 1896–901 (2003).
26. American Diabetes Assosiation. Economic costs of diabetes in the U.S. in 2012. Association, American Diabetes. Diabetes Care 36, 1033–46 (2013).
27. Eberwine, J., Sul, J.-Y., Bartfai, T. & Kim, J. The promise of single-cell sequencing. Nat. Methods 11, 25–27 (2013).
28. Vazquez, A. I. et al. Increased Proportion of Variance Explained and Prediction Accuracy of Survival of Breast Cancer Patients with Use of Whole-Genome Multiomic Profiles. Genetics 203, (2016).
29. Nica, A. C. & Dermitzakis, E. T. Using gene expression to investigate the genetic basis of complex disorders. Hum. Mol. Genet. 17, R129-34 (2008).
30. Teh, A. L. et al. Comparison of Methyl-capture Sequencing vs. Infinium 450K methylation array for methylome analysis in clinical samples. Epigenetics 11, 36–48 (2016).
31. Edfors, F. et al. Gene-specific correlation of RNA and protein levels in human cells and tissues. 1–10 (2016). doi:10.15252/msb.20167144
32. Smith, G. D. Mendelian randomization for strengthening causal inference in observational studies: application to gene x environment interactions. Perspectives on Psychological Science 5, 527–545 (2010).