# geno$^{5}$mC: A Database to Explore the Association between Genetic Variation (SNPs) and CpG Methylation in the Human Genome

**C. Gómez-Martín [1,2], E. Aparicio-Puerta [1,2,3,4], J. M. Medina [1,2], Guillermo Barturen [5], J. L. Oliver [1,2] and M. Hackenberg [1,2,3,4]***

*1 - **Dpto. de Genética,** Facultad de Ciencias, Universidad de Granada, Campus de Fuentenueva s/n, 18071 Granada, Spain*
*2 - **Lab. de Bioinformática,** Instituto de Biotecnología, Centro de Investigación Biomédica, PTS, Avda. del Conocimiento s/n, 18100 Granada, Spain*
*3 - **Instituto de Investigación Biosanitaria (IBS) Granada,** University Hospitals of Granada-University, Granada, Spain, Conocimiento s/n, 18100 Granada, Spain*
*4 - **Excellence Research Unit "Modeling Nature" (MNat),** University of Granada, 18071 Granada, Spain*
*5 - **Centro Pfizer-Universidad de Granada-Junta de Andalucía de Genómica e Investigación Oncológica,** Genetics of Complex Diseases, 18016 Granada, Spain*

***Correspondence to M. Hackenberg:*** Dpto. de Genética, Facultad de Ciencias, Universidad de Granada, Campus de Fuentenueva s/n, 18071 Granada, Spain. *hackenberg@ugr.es (M. Hackenberg)*

https://doi.org/10.1016/j.jmb.2020.11.008
***Edited by Rita Casadio***

## Abstract

Genetic variation, gene expression and DNA methylation influence each other in a complex way. To study the impact of sequence variation and DNA methylation on gene expression, we generated *geno$^{5}$mC*, a database that contains statistically significant SNP-CpG associations that are biologically classified either through co-localization with known regulatory regions (promoters and enhancers), or through known correlations with the expression levels of nearby genes. The SNP rs727563 can be used to illustrate the usefulness of this approach. This SNP has been associated with inflammatory bowel disease through GWAS, but it is not located near any gene related to this phenotype. However, *geno$^{5}$mC* reveals that rs727563 is associated with the methylation state of several CpGs located in promoter regions of genes reported to be involved in inflammatory processes. This case exemplifies how *geno$^{5}$mC* can be used to infer relevant and previously unknown interactions between described disease-associated SNPs and their functional targets.

## Introduction

Although several human diseases clearly show Mendelian inheritance, virtually all quantitative traits are complex, i.e. multi-genic and influenced by the environment. Those include complex diseases such as Alzheimer, autoimmune diseases[1] and most cancer types[2]. Genome Wide Association Studies (GWAS), which statistically relate allele frequencies at certain loci to phenotypes[3], have been widely used to identify the genetic component of complex traits. These studies contributed to the knowledge of the genetic predisposition to complex diseases and therefore the discovery of genetic variants (mostly SNPs, Single Nucleotide Polymorphisms) with a potential diagnostic and prognostic value.

However, statistically associated SNPs frequently locate outside coding or known regulatory regions[4]. Therefore, in many cases the mechanistic relationships between SNPs and the phenotype cannot be easily established. Expression quantitative trait loci (eQTL) provide a statistical link between genetic variation and gene expression[5]. For example, a recent study found 202,489 variants associated with the expression levels of 1959 genes in liver, which suggests that a substantial proportion of human genes might have at least one associated eQTL[6].

DNA methylation is well known for its implication in gene regulation. Early studies showed that DNA methylation generally relates to repression of gene expression, but recent experiments suggest that methylation associates with both reduced and increased levels of gene expression[7]. Furthermore, there are several lines of evidence that suggest that sequence variation can trigger changes in DNA methylation. Further studies found that changes in both transcription factor (TF) abundance and binding are associated with changes in DNA methylation[8–10]. And more recently, genome wide quantitative trait studies associated sequence variants with changes in methylation levels (mQTLs, Methylation Quantitative Trait Loci)[11], and many of them were found to co-localize with known TFBSs (Transcription Factor Binding Sites) or other regulatory sites[12].

Altogether, TFs might play a key role in linking sequence variants, DNA methylation and gene expression levels in a complex way. Indeed, recent observations show that eQTLs often co-localize with mQTLs, thus suggesting that these QTLs are connected at a functional level[13]. Pierce *et al.* identify more than 400 co-localized eQTL-mQTL pairs likely to share a common causal variant. Two possibilities exist: (1) changes in DNA methylation drive changes in gene expression or (2) DNA methylation is the result of gene regulation. By means of partial correlation and mediation analysis, these authors found that both possibilities co-exist in the genome, although many SNPs affect multiple CpGs in opposite directions.

Although eQTLs and mQTLs can notably extend our understanding of the molecular mechanisms of complex traits, these studies generally involve a high number of samples which makes them costly and labor intensive. Furthermore, it is often not possible to study eQTLs and mQTLs in the relevant tissue or cell type. Therefore, we present here a novel approach to explore the possible impact of sequence variants on methylation and thus on gene expression levels. It's known that SNPs can both, affect TF binding locally and provoke changes on DNA methylation levels[5]. DNA methylation in turn can influence gene expression as recently shown by[14,15] for TL-CpGs (Traffic Light CpG, i.e. the methylation level correlates with gene expression)[14]. Therefore, we postulate that a subset of eQTLs are mQTLs + TL-CpGs. This mechanistic interplay would also allow for causal relations between SNPs and gene expression at larger distances.

Using 58 publically available whole genome bisulfite sequencing datasets, we determined not only the methylation state of all CpG dinucleotides but also all sequence variants by means of *MethylExtract*[16]. We detected a total of 506,041,598 significant SNPs-CpG associations (Fisher exact test FDR $\leq$ 0.05) which are further classified by their biological relevance through co-localization with promoters[17], enhancers[18] and TL-CpGs [14]. To our knowledge, this is the first time that whole-genome bisulfite sequencing datasets (not limited to microarray probes[19,20]) are associated with GWAS results, unraveling new genetic regulatory associations previously unknown for complex traits.

The results of this study were compiled into *geno<sup>5</sup>mC*, the database presented here. The users can query the database using SNP identifiers, gene IDs or phenotype/syndrome traits to obtain statistically associated DNA methylation of CpGs located in functionally relevant genomic regions. We demonstrate the usefulness of the database by showing how a SNP that was found to be associated with Inflammatory Bowel Disease[21] but lacking a functional relationship, can be connected to relevant genes. This approach may also be useful to guide future population-based studies onto how epigenetic variation modulates risk of disease.

## Results

### geno$^5$mC database

After applying a minor allele frequency filter of 0.1 and removing all variant positions not known in dbSNP version 151 we obtain 4,086,616 SNPs for the analysis. Out of those 51,585 (1.3%) are associated with at least one CpG. On the other hand, we found that the methylation levels of 5,417,468 (19.3%) CpG dinucleotides are associated with at least one SNP. Please note that the number of associated CpGs is five times higher than the total number of CpGs interrogated by Infinium MethylationEPIC array.

The distance distribution is not monotonically decreasing and shows clear differences compared to the expected distribution. All chromosomes show an overrepresentation of short distances (<2 Mb in most cases) as depicted in Figure 1(a) for chr22. Up to distances of 2 Mb the number of observed CpG-SNP pairs are significantly higher than expected by chance alone ($z$-score > 3.3). At larger distances we can observe both, over and underrepresented distance ranges. Surprisingly, a very pronounced peak was found for distances between 30 Mb and 33 Mb in chr22. Blocks of SNPs associated to several CpGs are responsible for this peak. Figure 1(b) shows a block of 36 SNPs

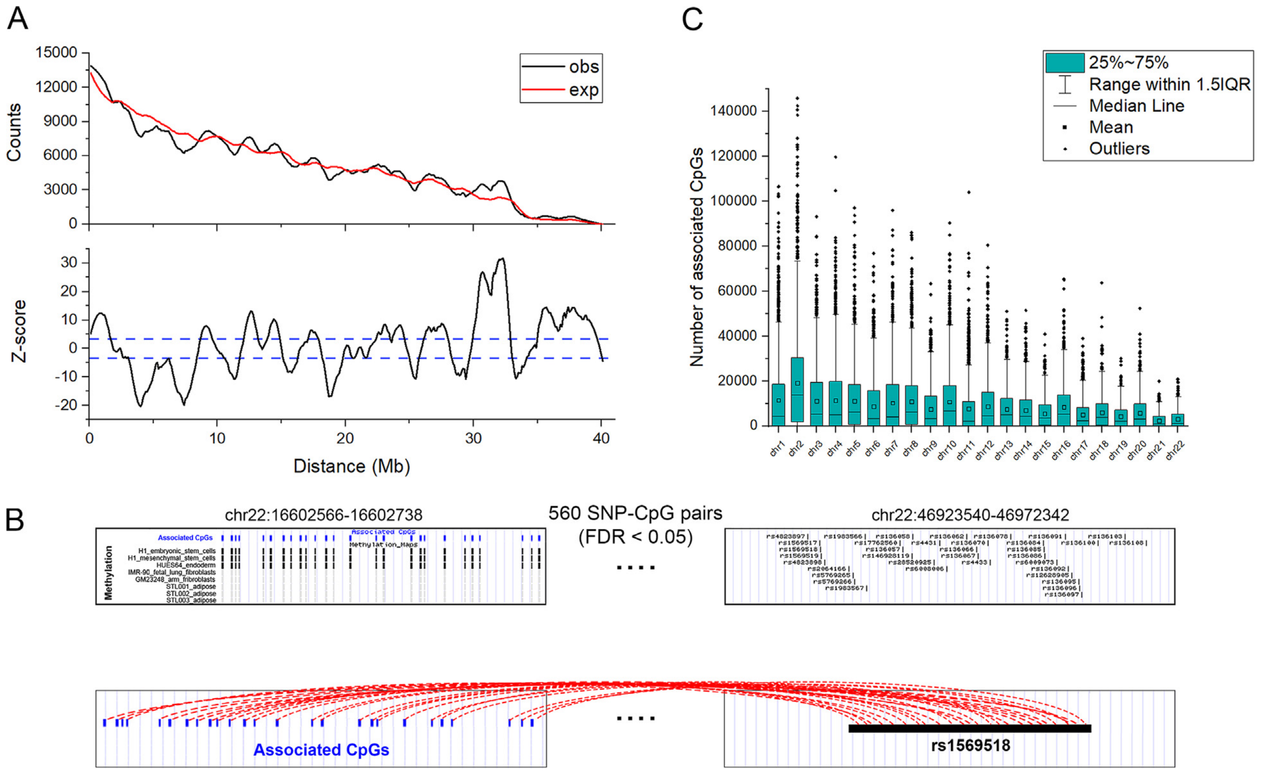


**Figure 1.** (a) Observed and expected distance distribution (SNP-CpG) and the statistical significance expressed in *z*-scores (bottom). We mark with dashed lines those differences between observed and expected counts that are highly significant (|*z*| > 3.3). (b) Two genome regions, one very dense in CpGs (26) and one in SNPs (36). A total of 560 SNP-CpG pairs do exist between these two regions, explaining the high number of distances observed between 30 Mb and 33 Mb. (c) The distribution of the number of associated CpGs per SNP as a function of chromosome.

(chr22:46923540-46972342) associated with 26 CpGs (chr22:16602566-16602738). The SNP rs1569518 is associated to the methylation values of 25 out of these 26 CpGs (bottom of Figure 1 (b)). The association between these two regions contributes with a total of 560 SNP-CpG pairs explaining the observed overrepresentation of these distances. Although these long range associations may appear as mere artefacts, it was recently shown that methylation domains can form long loops connecting loci that are several dozens of megabases apart[22].

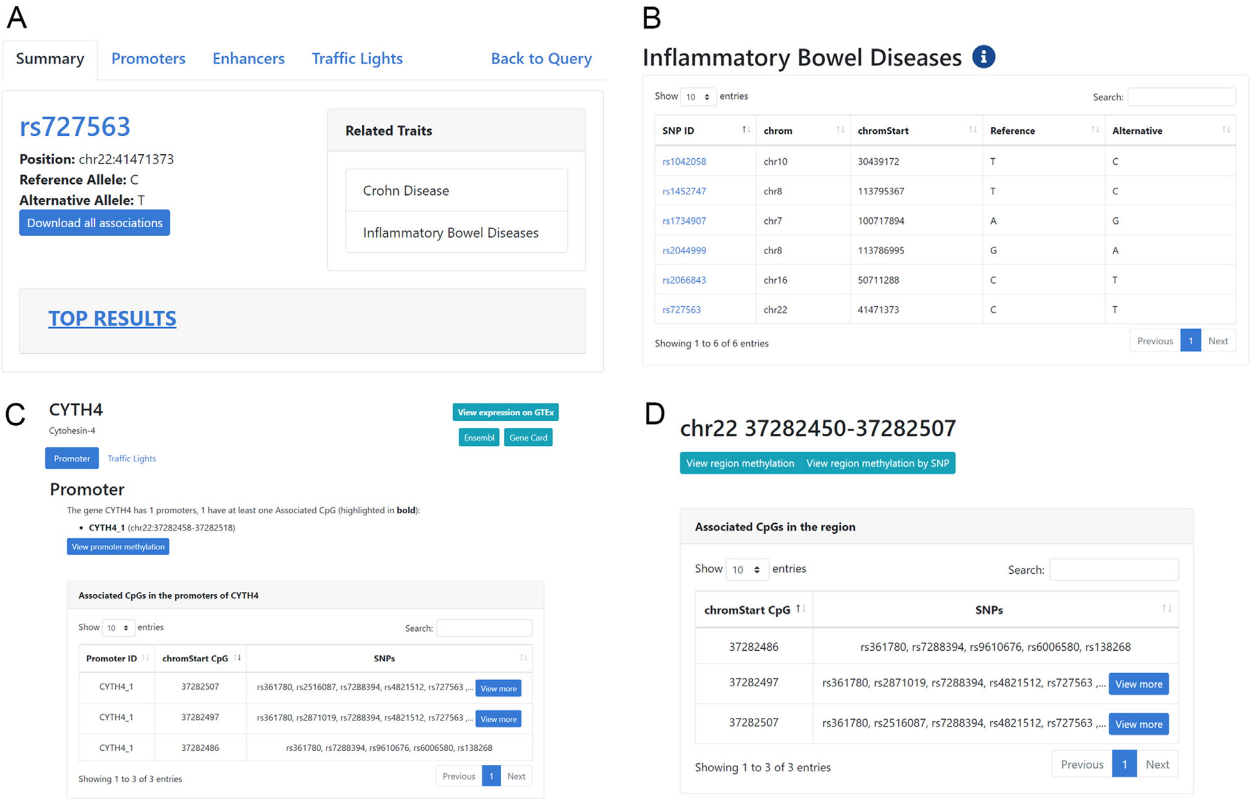
Finally, Figure 1(c) shows the distribution of the number of associated CpGs per SNP in the 22 autosomes. The median values range between 979 CpGs (chr21) and 13,684 (chr2). This shows that while only a minor fraction of SNPs (1.3%) are associated at all, most of them correlate with thousands of different CpGs. This distribution is clearly different than random expectation (noise) and therefore reinforces that the found statistical associations are not artefacts but of biological nature.

## Front and backend implementation

A MySQL database was used to store all data displayed at the website. The interactive web application was implemented using a Django framework, together with Bootstrap and Javascript. SQLAlchemy[23] was used as Object Relational Mapper (ORM) between MySQL and Python. The plotly package[24] was used for data visualization in order to improve the interactivity of the web application.

## Information extraction and workflow

The database can be queried in four different ways using: (i) a single SNP ID from dbSNP, (ii) a trait (iii) a gene symbol or (iv) a genomic region. Output page examples for these four different query types are shown in Figure 2. On the output page for a given SNP, the associated CpGs are grouped by their putative biological relevance, i.e. those that are located in known regulatory regions (promoter and enhancers) or those that are known to correlate with gene expression (TL-CpG). We highlight 'Top results' if an associated CpG dinucleotide correlates with the transcription levels of at least one gene and also lies within the promoter region of a gene or an enhancer. Given the hierarchical classification, different output levels are generated: (i) a summary tab that also contains the 'Top results', (ii) CpGs located in known promoters, (iii) CpGs located in enhancers



rs727563

**Position:** chr22:41471373
**Reference Allele:** C
**Alternative Allele:** T

Related Traits

- Crohn Disease
- Inflammatory Bowel Diseases

Inflammatory Bowel Diseases

Show 10 entries Search:

| SNP ID | chrom | chromStart | Reference | Alternative |
|---|---|---|---|---|
| rs1042058 | chr10 | 30439172 | T | C |
| rs1452747 | chr8 | 113795367 | T | C |
| rs1734907 | chr7 | 100717894 | A | G |
| rs2044999 | chr8 | 113786995 | G | A |
| rs2066843 | chr16 | 50711288 | C | T |
| rs727563 | chr22 | 41471373 | C | T |

Showing 1 to 6 of 6 entries

CYTH4
Cytohesin-4

Promoter

The gene CYTH4 has 1 promoters, 1 have at least one Associated CpG (highlighted in **bold**):

- **CYTH4_1** (chr22:37282458-37282518)

Associated CpGs in the promoters of CYTH4

Show 10 entries Search:

| Promoter ID | chromStart CpG | SNPs |
|---|---|---|
| CYTH4_1 | 37282507 | rs361780, rs2516087, rs7288394, rs4821512, rs727563 ,... View more |
| CYTH4_1 | 37282497 | rs361780, rs2871019, rs7288394, rs4821512, rs727563 ,... View more |
| CYTH4_1 | 37282486 | rs361780, rs7288394, rs9610676, rs6006580, rs138268 |

Showing 1 to 3 of 3 entries

chr22 37282450-37282507

Associated CpGs in the region

Show 10 entries Search:

| chromStart CpG | SNPs |
|---|---|
| 37282486 | rs361780, rs7288394, rs9610676, rs6006580, rs138268 |
| 37282497 | rs361780, rs2871019, rs7288394, rs4821512, rs727563 ,... View more |
| 37282507 | rs361780, rs2516087, rs7288394, rs4821512, rs727563 ,... View more |

Showing 1 to 3 of 3 entries

**Figure 2.** Example of the output for different ways to query *geno$^5$mC: (a) Query SNP, (b) Query trait, (c) Query Gene, and (d) Query Region.*

and (iv) CpGs reported to correlate with gene expression (TL-CpGs).

For a trait based search (Figure 2(b)), associated SNPs are reported if they have at least one associated CpG. The corresponding SNPs can then be explored separately.

The output for a gene search is slightly different reporting (i) all associated CpGs located in the promoter region together with the associated SNPs (see Figure 2(c)) and (ii) associated CpGs that correlate with gene expression (TF-CpGs). Furthermore, the methylation values can be visualized for each promoter.

Finally we provide the possibility to query a user-defined genome region by its coordinates. The output is virtually identical to the gene-centered search with the exception that TL-CpGs cannot be reported as those are limited to regions centered around genes.

## A working example

To illustrate how *geno$^5$mC* can be used to extend the knowledge about putatively functional implications of SNPs we analyzed *rs727563*, reported to be statistically associated with inflammatory bowel disease.[21] This SNP has two possible alleles C/T, being C the risk allele. Although the GWAS showed a highly significant association, no mechanistic link is known between this SNP and the mentioned phenotype.

Located in chromosome 22 (chr22:41471373, GRCh38.p12), this SNP is an intron variant of the gene ACO2 which encodes for the aconitase two protein. The ACO2 belongs to the aconitase/IPM isomerase family, and catalyzes the inter-conversion of citrate to isocitrate via cis-aconitate in the second step of the citric acid cycle. Diseases associated with ACO2 include Infantile Cerebellar-Retinal Degeneration and Optic Atrophy 9[25,26]; but it has no apparent relation to inflammatory bowel disease.

*geno$^5$mC* reports a total of 6280 associated CpGs for this SNP which can be downloaded from the output page. In order to provide more concise results, the output is centered on the most relevant CpGs. We find 2299 CpGs located in enhancers, 16 in promoters and 53 that are TL-CpGs. Figure 3(a) shows that three genes are found with at least one associated TL-CpG. The gene CYTH4 (Figure 3(b)) with two TL-CpGs encodes for the Cytohesin-4 protein, which has been related to inflammatory bowel diseases[27]. On the other hand, promoter co-localization analysis revealed that the promoter region of gene SLC5A1 presents three associated CpGs (not reported as TL-CpGs) (Figure 3(c)). SLC5A1 encodes a member of the sodium-dependent glucose transporter (SGLT) family. The encoded integral membrane protein is the primary mediator of dietary glucose and galactose uptake from the intestinal lumen, being mainly expressed in the intestine. Interestingly, its function has also been related to inflammatory bowel disease[28,29]. In the enhancer section this gene appears again showing that six enhancers contain associated CpGs, reinforcing therefore a possible functional link between this SNP and SLC5A1 through DNA methylation changes.

Taken together, this example demonstrates that although no molecular link has previously been described between *rs727563* and inflammatory bowel disease, some of its associated CpGs are located in the promoter regions of several genes reported to have a role in inflammatory processes. Therefore, this analysis suggests an implication of this SNP in the regulation of gene expression of a handful of genes involved with inflammation.

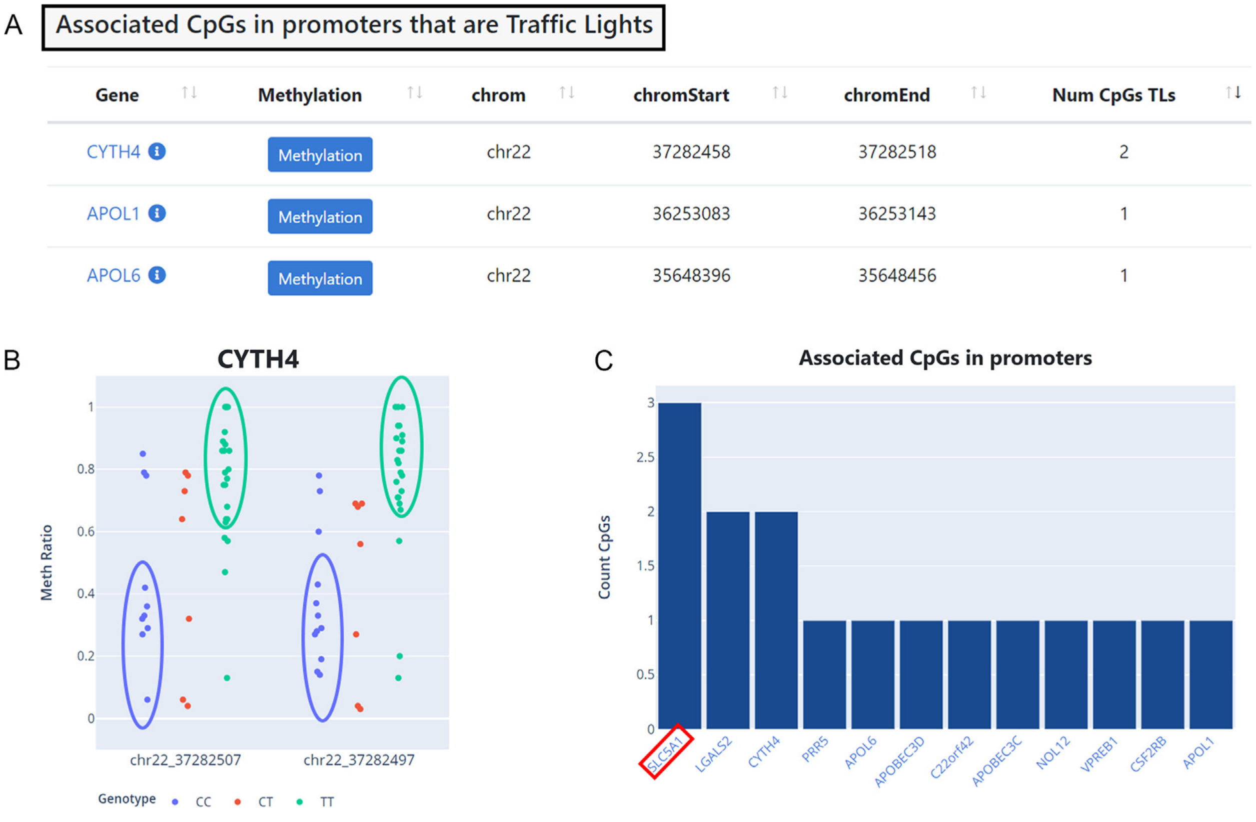
To show that this SNP is not an exception we analyzed some more phenotype associated SNPs that are not located in any known regulatory regions. For example the SNP rs4780401 is related to Rheumatoid Arthritis by GWAS[30] but is not located within or near any related gene. By means of our database we have found that it is associated with a CpG-TL in the promoter of the gene MLKL which has been related to the same disease[31]. The SNP rs10746333 is related to Diabetes Mellitus by GWAS. *geno$^5$mC* reports that it is associated with several CpGs-TL in the promoters of the genes LRMP and MGP that have been described as associated with the same disease[32,33].

## Discussion

Our database *geno$^5$mC* allows to connect sequence variation (SNPs) to genes through the statistical association with CpG methylation and their correlation with gene expression values. In this way, it can hint towards putative functional implications of SNPs known to be associated to specific phenotypes but without any known molecular links.

Note that *geno$^5$mC* is based on statistical association and therefore the putative mechanistic link is not confirmed. Just like in eQTL and mQTL studies, the associations reported by *geno$^5$mC* still need to be confirmed by independent functional assays.

One important last note is that if a causal mechanistic link exists, then some kind of effector



Associated CpGs in promoters that are Traffic Lights

| Gene | Methylation | chrom | chromStart | chromEnd | Num CpGs TLs |
|---|---|---|---|---|---|
| CYTH4 | Methylation | chr22 | 37282458 | 37282518 | 2 |
| APOL1 | Methylation | chr22 | 36253083 | 36253143 | 1 |
| APOL6 | Methylation | chr22 | 35648396 | 35648456 | 1 |

**Figure 3.** (a) Three genes with associated CpGs in their promoter regions. These CpGs also correlate with gene expression (TL-CpGs), (b) the distribution of methylation values of the two associated TL-CpGs in the CYTH4 promoter region. It can be seen clearly that the *CC* genotype is associated to unmethylation (chr22:3728507 p-value = $6 \cdot 10^{-4}$, chr22:37282497 p-value = $1.3 \cdot 10^{-4}$). (c) Associated CpGs in promoter regions that were not previously reported as correlating to gene expression.

molecule must be present. Strong candidates are transcription factors with different affinity to different binding sites defined by SNPs. It was shown that this scenario can impact on methylation (mQTL) and transcription levels (eQTL). Consequently, this effector needs to be present in the analyzed cell type, otherwise the SNP genotype will not have any impact on the methylation level. Since our approach is based on many different tissue types the effector might be absent in a subset of samples and, therefore, the associations are likely to be missed. However, it is important to highlight that this will not increase the number of false positives but the number of false negatives.

Up to our knowledge this is the first mQTL database for whole genome sequencing data, what makes it an important resource to leverage GWAS results and identify new genes and CpGs associated with complex traits.

## Materials and Methods

### Whole genome bisulfite sequencing data

Raw whole genome bisulfite sequencing data were downloaded from the NCBI Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra). To avoid technical bias and in order to screen all possible SNP-CpG combinations, only Whole Genome Bisulfite Sequencing (WGBS) data were selected. A total of 58 samples from different tissues and projects, summarized in Supp. Table 1, were selected. We made sure that each sample comes from a different individual or cell line in order to maximize the number of different haplotypes.

### DNA methylation and genotype profiling

To obtain DNA methylation levels and sequence variation from the same samples, we developed a pipeline based on *Trimmomatic*[34], *Bowtie2*[35], *Bismark*[36], *BSeQC*[37] and *MethylExtract*[16]. *MethylExtract* provides methylation levels as the ratio between 5mC reads and total number of reads mapped to each cytosine position. Therefore, values close to one imply methylation while values close to 0 indicate unmethylation. Additionally, sequence variation is determined using a Fisher Exact test like performed in *VarScan*[38,39]. Both, methylation values for all CpG dinucleotides and sequence variants are stored in a *MySQL* database.

### Statistical model

At a single-cell level, only three different methylation values are biologically possible: 0 (unmethylated), 0.5 (allele specific methylation) and 1 (methylated). However in a cell population different values can be biologically meaningful, like partial methylation values at enhancers[40,41]. We first classify the methylation values (Mv) for each CpG dinucleotide covered by at least five reads into three groups: M (methylated): $Mv > 0.65$, I (Intermediate): $0.35 \le Mv \le 0.65$ and U (unmethylated): $Mv < 0.35$.

After filtering out SNPs with a minor allele frequency below 0.1 in our set of 58 samples, we obtain a 3 × 3 contingency table for each SNP-CpG pair. This table is then reduced to a 2 × 2 table by using only homozygotes for the reference or alternative allele and methylated or unmethylated for the CpG methylation state. In this way, heterozygotes and intermediately methylated samples are not used to determine statistical significance. The exact p-value was then calculated using a Fisher Exact test. The false discovery rate (FDR) is finally obtained by correcting the exact p-value with the number of tests performed for each SNP. Only SNP-CpG pairs significant at $FDR \le 0.05$ are included into the database.

### Annotation and classification data

Sequence variants detected within the 58 samples were filtered using dbSNP[42] version 151 (all), i.e. only known sequence variants are further considered. Associated CpGs are classified according to their genome location using:

- EPD promoters version 006 downloaded from (https://epd.epfl.ch/human/human_database.php?db=human)[17].
- Enhancers from *GeneHancer* version 4.4 downloaded from (https://www.genecards.org/GeneHancer_version_4-4)[18].
- CpGs previously found to correlate with gene expression, i.e. TL-CpGs[14] (Supp. Table II).

Trait-SNP information was downloaded from the Phenotype-Genotype integrator (PheGenI)[43]. Only those SNPs reported as associated with both, a phenotype and DNA methylation were considered.

### Distance distribution

We carry out randomization experiments in order to obtain an expected distance distribution under the assumption that the found SNP-CpG associations are by chance alone. This allows us to determine statistically significant over and underrepresentation of certain distances. We first determine the number of SNP-CpG pairs for distance bins of 100 kb (observed counts). The expected values are then calculated by randomly shuffling 100 times the labels associated/not-associated among the CpGs associated to a given SNP. This is performed for each SNP separately preserving the number of associated CpGs. For each distance bin we obtain the observed and expected number of SNP-CpG distances and the standard deviation. Finally we calculate a *z*-score to determine at which distances statistically significant associations are over and under-represented.

**Database URL:** https://arn.ugr.es/geno5mc/

## Funding

This work was supported by Spanish Government [AGL2017-88702-C2-2-R] to M.H. and JLO; Instituto de Salud Carlos III [IFI16/00041] to E.A.

## CRediT authorship contribution statement

**C. Gómez-Martín:** Software, Validation, Formal analysis, Investigation, Visualization, Writing - review & editing. **E. Aparicio-Puerta:** Software, Validation, Writing - review & editing. **J.M. Medina:** Validation. **Guillermo Barturen:** Conceptualization. **J.L. Oliver:** Conceptualization, Writing - Original Draft, Writing - review & editing, Supervision, Investigation. **M. Hackenberg:** Conceptualization, Writing - Original Draft, Writing - review & editing, Software, Supervision, Investigation.

### DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2020.11.008.

*Received 7 August 2020;*
*Accepted 6 November 2020;*
*Available online 12 November 2020*

***Keywords*:**
SNPs;
DNA methylation;
gene regulation

## References

1. Goddard, M.E., Kemper, K.E., MacLeod, I.M., Chamberlain, A.J., Hayes, B.J., (2016). Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. *Proc. R. Soc. B Biol. Sci.*, **283**, 20160569. https://doi.org/10.1098/rspb.2016.0569.
2. Wu, S., Powers, S., Zhu, W., Hannun, Y.A., (2016). Substantial contribution of extrinsic risk factors to cancer development. *Nature.*, **529**, 43–47. https://doi.org/10.1038/nature16166.
3. Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., Meyre, D., (2019). Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.*,, 1. https://doi.org/10.1038/s41576-019-0127-1.
4. Tak, Y.G., Farnham, P.J., (2015). Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics Chromatin*, **8** https://doi.org/10.1186/s13072-015-0050-4.
5. Nica, A.C., Dermitzakis, E.T., (2013). Expression quantitative trait loci: present and future. *Philos. Trans. R. Soc. B Biol. Sci.*, **368** https://doi.org/10.1098/rstb.2012.0362.
6. Strunz, T., Grassmann, F., Gayán, J., Nahkuri, S., Souza-Costa, D., Maugeais, C., Fauser, S., Nogoceke, E., et al., (2018). A mega-analysis of expression quantitative trait loci (eQTL) provides insight into the regulatory architecture of gene expression variation in liver. *Sci. Rep.*, **8**, 1–11. https://doi.org/10.1038/s41598-018-24219-z.
7. Fisher, V.A., Wang, L., Deng, X., Sarnowski, C., Cupples, L.A., Liu, C.T., (2018). Do changes in DNA methylation mediate or interact with SNP variation? A pharmacoepigenetic analysis 06 Biological Sciences 0604 Genetics. *BMC Genet.*, **19**, 15–19. https://doi.org/10.1186/s12863-018-0635-6.
8. Bird, A., (2011). Putting the DNA back into DNA methylation. *Nat. Genet.*, **43**, 1050–1051. https://doi.org/10.1038/ng.987.
9. Banovich, N.E., Lan, X., McVicker, G., van de Geijn, B., Degner, J.F., Blischak, J.D., Roux, J., Pritchard, J.K., et al., (2014). Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet.*, **10**, https://doi.org/10.1371/journal.pgen.1004663 e1004663.
10. Zhu, H., Wang, G., Qian, J., (2016). Transcription factors as readers and effectors of DNA methylation. *Nat. Rev. Genet.*, **17**, 551–565. https://doi.org/10.1038/nrg.2016.83.
11. McRae, A.F., Marioni, R.E., Shah, S., Yang, J., Powell, J. E., Harris, S.E., Gibson, J., Henders, A.K., et al., (2018). Identification of 55,000 replicated DNA methylation QTL. *Sci. Rep.*, **8**, 1–9. https://doi.org/10.1038/s41598-018-35871-w.
12. Corradin, O., Scacheri, P.C., (2014). Enhancer variants: evaluating functions in common disease. *Genome Med.*, **6**, 85. https://doi.org/10.1186/s13073-014-0085-3.
13. Pierce, B.L., Tong, L., Argos, M., Demanelis, K., Jasmine, F., Rakibuz-Zaman, M., Sarwar, G., Islam, M.T., et al., (2018). Co-occurring expression and methylation QTLs allow detection of common causal variants and shared biological mechanisms. *Nat. Commun.*, **9**, 804. https://doi.org/10.1038/s41467-018-03209-9.
14. Lioznova, A.V., Khamis, A.M., Artemov, A.V., Besedina, E., Ramensky, V., Bajic, V.B., Kulakovskiy, I.V., Medvedeva, Y.A., (2019). CpG traffic lights are markers of regulatory regions in human genome. *BMC Genom.*, **20**, 102. https://doi.org/10.1186/s12864-018-5387-1.
15. Yang, L., Chen, Z., Stout, E.S., Delerue, F., Ittner, L.M., Wilkins, M.R., Quinlan, K.G.R., Crossley, M., (2020). Methylation of a CGATA element inhibits binding and

regulation by GATA-1. *Nat. Commun.*, **11**, 2560. https://doi.org/10.1038/s41467-020-16388-1.

16. Barturen, G., Rueda, A., Oliver, J.L., Hackenberg, M., (2013). MethylExtract: high-quality methylation maps and SNV calling from whole genome bisulfite sequencing data. *F1000 Res.*, **2**, 1–23.
17. Dreos, R., Ambrosini, G., Groux, R., Cavin Périer, R., Bucher, P., (2017). The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic Acids Res.*, **45**, D51–D55. https://doi.org/10.1093/nar/gkw1069.
18. Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., et al., (2017). GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database*, **2017** https://doi.org/10.1093/database/bax028.
19. Gong, J., Wan, H., Mei, S., Ruan, H., Zhang, Z., Liu, C., Guo, A.-Y., Diao, L., et al., (2019). Pancan-meQTL: a database to systematically evaluate the effects of genetic variants on methylation in human cancer. *Nucleic Acids Res.*, **47**, D1066–D1072. https://doi.org/10.1093/nar/gky814.
20. Gaunt, T.R., Shihab, H.A., Hemani, G., Min, J.L., Woodward, G., Lyttleton, O., Zheng, J., Duggirala, A., et al., (2016). Systematic identification of genetic influences on methylation across the human life course. *Genome Biol.*, **17**, 61. https://doi.org/10.1186/s13059-016-0926-z.
21. Liu, J.Z., van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., et al., (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.*, **47**, 979–986. https://doi.org/10.1038/ng.3359.
22. Zhang, X., Jeong, M., Huang, X., Wang, X.Q., Wang, X., Zhou, W., Shamim, M.S., Gore, H., et al., (2020). Large DNA methylation nadirs anchor chromatin loops maintaining hematopoietic stem cell identity. *Mol. Cell.*, **78**, 506–521.e6. https://doi.org/10.1016/j.molcel.2020.04.018.
23. Michael Bayer, SQLAlchemy, in: A.B. and G. Wilson (Ed.), Archit. Open Source Appl. Vol. II Struct. Scale, a Few More Fearless Hacks, 2012, p. 20.
24. Plotly Technologies, Collaborative data Science, 2015. https://plot.ly.es.
25. Metodiev, M.D., Gerber, S., Hubert, L., Delahodde, A., Chretien, D., Gérard, X., Amati-Bonneau, P., Giacomotto, M.-C., et al., (2014). Mutations in the tricarboxylic acid cycle enzyme, aconitase 2, cause either isolated or syndromic optic neuropathy with encephalopathy and cerebellar atrophy. *J. Med. Genet.*, **51**, 834–838. https://doi.org/10.1136/jmedgenet-2014-102532.
26. Spiegel, R., Pines, O., Ta-Shma, A., Burak, E., Shaag, A., Halvardson, J., Edvardson, S., Mahajna, M., et al., (2012). Infantile cerebellar-retinal degeneration associated with a mutation in mitochondrial aconitase, ACO2. *Am. J. Hum. Genet.*, **90**, 518–523. https://doi.org/10.1016/j.ajhg.2012.01.009.
27. Peters, L.A., Perrigoue, J., Mortha, A., Iuga, A., Song, W. M., Neiman, E.M., Llewellyn, S.R., Di Narzo, A., et al., (2017). A functional genomics predictive network model identifies regulators of inflammatory bowel disease. *Nat. Genet.*, **49**, 1437–1449. https://doi.org/10.1038/ng.3947.
28. Lee, C.Y., (2013). Chronic restraint stress induces intestinal inflammation and alters the expression of hexose and lipid transporters. *Clin. Exp. Pharmacol. Physiol.*, **40**, 385–391. https://doi.org/10.1111/1440-1681.12096.
29. Brzozowski, B., Mazur-Bialy, A., Pajdo, R., Kwiecien, S., Bilski, J., Zwolinska-Wcislo, M., Mach, T., Brzozowski, T., (2016). Mechanisms by which stress affects the experimental and clinical inflammatory bowel disease (IBD): role of brain-gut axis. *Curr. Neuropharmacol.*, **14**, 892–900. https://doi.org/10.2174/1570159x14666160404124127.
30. Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., et al., (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, **506**, 376–381. https://doi.org/10.1038/nature12873.
31. Wang, X., Gessier, F., Perozzo, R., Stojkov, D., Hosseini, A., Amirshahrokhi, K., Kuchen, S., Yousefi, S., et al., (2020). RIPK3–MLKL–mediated neutrophil death requires concurrent activation of fibroblast activation protein-$\alpha$. *J. Immunol.*, **205**, 1653–1663. https://doi.org/10.4049/jimmunol.2000113.
32. Sardana, M., Vasim, I., Varakantam, S., Kewan, U., Tariq, A., Koppula, M.R., Syed, A.A., Beraun, M., et al., (2017). Inactive matrix gla-protein and arterial stiffness in type 2 diabetes mellitus. *Am. J. Hypertens.*, **30**, 196–201. https://doi.org/10.1093/ajh/hpw146.
33. Grimm, C.H., Rogner, U.C., Avner, P., (2003). Lrmp and Bcat1 are candidates for the type I diabetes susceptibility locus Idd6. *Autoimmunity*, **36**, 241–246. https://doi.org/10.1080/0891693031000141068.
34. Bolger, A.M., Lohse, M., Usadel, B., (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
35. Langmead, B., Salzberg, S.L., (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359. https://doi.org/10.1038/nmeth.1923.
36. Krueger, F., Andrews, S.R., (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571–1572. https://doi.org/10.1093/bioinformatics/btr167.
37. Lin, X., Sun, D., Rodriguez, B., Zhao, Q., Sun, H., Zhang, Y., Li, W., (2013). BSeQC: quality control of bisulfite sequencing experiments. *Bioinformatics*, **29**, 3227–3229. https://doi.org/10.1093/bioinformatics/btt548.
38. Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., Weinstock, G.M., Wilson, R.K., et al., (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, **25**, 2283–2285. https://doi.org/10.1093/bioinformatics/btp373.
39. Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., et al., (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576. https://doi.org/10.1101/gr.129684.111.
40. Aran, D., Sabato, S., Hellman, A., (2013). DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biol.*, **14**, R21. https://doi.org/10.1186/gb-2013-14-3-r21.

41. Bell, R.E., Golan, T., Sheinboim, D., Malcov, H., Amar, D., Salamon, A., Liron, T., Gelfman, S., et al., (2016). Enhancer methylation dynamics contribute to cancer plasticity and patient mortality. *Genome Res.*, **26**, 601–611. https://doi.org/10.1101/gr.197194.115.

42. S.T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, K. Sirotkin, dbSNP: the NCBI database of genetic variation, 2001.

43. Ramos, E.M., Hoffman, D., Junkins, H.A., Maglott, D., Phan, L., Sherry, S.T., Feolo, M., Hindorff, L.A., (2014). Phenotype-genotype integrator (PheGenI): Synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur. J. Hum. Genet.*, **22**, 144–147. https://doi.org/10.1038/ejhg.2013.96.