# Genetic association of molecular traits: A help to identify causative variants in complex diseases

Claire Vandiedonck

Received: 31 August 2017 | Revised: 24 November 2017 | Accepted: 27 November 2017

DOI: 10.1111/cge.13187

WILEY CLINICAL GENETICS

INVITED REVIEW

# Genetic association of molecular traits: A help to identify causative variants in complex diseases

C. Vandiedonck

Univ Paris Diderot, Sorbonne Paris Cité, Paris, France

**Correspondence**
Claire Vandiedonck, Univ Paris Diderot, Sorbonne Paris Cité, UMRS 958 INSERM, F-75010 Paris, France.
Email: claire.vandiedonck@inserm.fr

**Funding information**
Université Paris Diderot

In the past 15 years, major progresses have been made in the understanding of the genetic basis of regulation of gene expression. These new insights have revolutionized our approach to resolve the genetic variation underlying complex diseases. Gene transcript levels were the first expression phenotypes that were studied. They are heritable and therefore amenable to genome-wide association studies. The genetic variants that modulate them are called expression quantitative trait loci. Their study has been extended to other molecular quantitative trait loci (molQTLs) that regulate gene expression at the various levels, from chromatin state to cellular responses. Altogether, these studies have generated a wealth of basic information on the genome-wide patterns of gene expression and their inter-individual variation. Most importantly, molQTLs have become an invaluable asset in the genetic study of complex diseases. Although the identification of the disease-causing variants on the basis of their overlap with molQTLs requires caution, molQTLs can help to prioritize the relevant candidate gene(s) in the disease-associated regions and bring a functional interpretation of the associated variants, therefore, bridging the gap between genotypes and clinical phenotypes.

**KEYWORDS**

causative variants, complex disease, eQTL, fine-mapping, GWAS, linkage disequilibrium, molecular trait, molQTL, quantitative trait

## 1 | INTRODUCTION

Common diseases are a major public health issue. Their etiology is complex, with a broad spectrum influenced by environmental and genetic factors. The arrival of high-throughput technologies together with the development of the Human Genome Project has promoted intensive research of the genetic factors by the means of genome-wide associations studies (GWAS) that can uncover common variants in large case-control cohorts. The last decade has thus witnessed the identification of thousands of genetic variants associated with common traits and diseases,[1] revealing an unprecedented polygenicity.[2] Nonetheless, these associations have shown small size effects on the phenotypic variance of these traits leaving a large amount of unexplained heritability. In addition, only few causative variants have been actually identified. So far, for Mendelian diseases, causative variants have been found to be mostly coding with deleterious impact on proteins,[3,4] whereas in complex traits, the associated variants predominantly fall in intronic or intergenic regions.[1,2,5,6] Unfortunately, the identification of the causative variants is the final step often missing as it is highly challenging due to polygenicity, linkage disequilibrium (LD), cost, time and availability of relevant tissues.

Fortunately, while GWAS expanded, progresses in omics technologies have allowed the study of the human genome at the level of both its sequence, revealing the genetic diversity across populations, and of its expression at the various functional levels including the transcriptome, the proteome or the regulome. Characterizing these molecular traits has been essential. Studying them in light of genetic diversity has been a further step. Indeed, applying genetic approaches to these traits has opened a new era of "genetical genomics," a concept first coined in 2001.[7] In this endeavor, blood has been most extensively studied for obvious reasons (ease of collection and culture, wealth of markers), resulting in a bias of current data towards the hematopoietic and immune system. However, recent collaborative projects have started to investigate large panels of tissues and

are closing the gap. This review outlines recent developments in the field along with its history, the methods behind it, and how it is applied to human genetics with emphasis on complex pathologies.

## 2 | LINKAGE DISEQUILIBRIUM, THE CORRELATION OF ALLELES IN A POPULATION, IS THE BASIS FOR GENETIC ASSOCIATION STUDIES

Genetic association is a powerful method to relate a genotype to a phenotype, the primary task of a geneticist. Its principle consists in correlating alleles or genotypes with a phenotype such as a disease or a quantitative trait (QT), for example, height or cholesterol level, in a population. It is based on linkage disequilibrium (LD), a fundamental relationship between alleles at pairs of genetically linked loci (Table 1). Association studies use available genetic variants as markers and rely on the hypothesis that at least one of them is in LD with the unknown causative allele. Unfortunately, an allele reported as associated with a trait is very rarely causal. Most likely, it is in LD with the causative allele and, therefore, the association is indirect. Another possibility not to forget is the existence of a false positive result.

On the basis outlined above, the principle of association tests is simply to assess the impact of marker alleles or genotypes on the phenotype. For diseases and binary traits, allele frequencies are compared between cases carrying the trait and matched controls using proportion comparison tests or logistic regression. For a continuous QT, a linear regression of the QT on the genotype is performed to identify the locus controlling it, termed QTL (quantitative trait locus) (Figure 1). Overall, association studies are powered to detect frequent variants having a small effect size on the trait by using large samples. Once replicated, the next step is to fine-map association signals to identify causative variants. While a strong LD facilitates the discovery of an association, it becomes a major obstacle to identify the causative variant that cannot be statistically distinguished from the tested variants, especially when LD extends over several genes. Functional approaches are then necessary (Figure 2).[8,9]

## 3 | GENOME-WIDE GENETIC STUDIES OF EXPRESSION PHENOTYPES

### 3.1 | eQTL (expression QTL): the first genomic endophenotypes

In multifactorial diseases, it is expected that most causative variants are regulatory, mainly acting through subtle modifications in the expression of messenger RNAs.[2] Two early examples were provided by the identification of the causative alleles at the *INS* and *CHRNA1* loci, encoding the autoantigens in type 1 diabetes and autoimmune myasthenia gravis, respectively. They were shown to finely modulate the mRNA levels of these genes in the thymus, thereby affecting the central tolerance.[10,11] The levels of gene expression can therefore be

**TABLE 1** Key concepts of linkage disequilibrium (LD)

| *Definition* |
|---|
| If we consider 2 genetically linked variants, each having 2 alleles, we expect to observe 4 pairwise combinations of these alleles, or haplotypes. If the alleles were independent or at linkage equilibrium, their frequencies should be the product of each allele frequency in the population. Likewise, for $n$ biallelic variants, we expect to see $2^n$ haplotypes in the population. In fact, this random association of alleles is rarely observed, especially if the variants are physically close. We rather observe preferential allelic associations (ie, haplotype frequencies deviate from those expected at linkage equilibrium). The difference $D$ between the observed and expected frequencies of a haplotype defines the LD of the 2 alleles on this haplotype. If 2 alleles are found together more often than expected, they are positively associated and their LD is positive. Conversely, the alternative combinations of alleles are less frequent than expected with the same absolute LD value but negative. |
| There are different measures of LD. The most intuitive one for biallelic variants is their correlation estimated by their coefficient of determination $r^2$ ranging between 0 and 1. The LD is "complete" when at least 1 haplotype is missing and the most extreme case of LD said "perfect" is when only 2 of the 4 haplotypes are observed, with an $r^2$ of 1. In the latter case, both variants are redundant such that the genotype at the second variant can be imputed from the genotype at the first variant for any individual in the population. This has major consequences for genetic association studies (see main text). |
| *Forces shaping the LD are intimately connected to the history of the population*: LD depends on when the variants appeared in the population and how they evolved across the subsequent generations (see below): |
| 1. *Mutations/variations* |
| Initially, when mutation creates a new allele at a locus near an established variant, there are only 3 haplotypes and therefore the LD is complete. |
| 2. *Meiotic recombination* |
| Following mutation, the major force shaping the LD is meiotic recombination that creates the fourth haplotypes and therefore dissipates the LD. The higher the number of generations since the mutation has appeared and the higher the recombination rate between 2 loci, the lower the LD between their alleles. Thus distant loci are generally in lower LD. However, the recombination activity is not homogeneous along the genome: at a small scale, hot-spots of recombination shape islands of LD or haplotype blocks where a smaller number of haplotypes than expected are observed in the population. Their average size is ~20 kb and varies between populations. However, it is important to remind that 2 variants in the same haplotype block are not always in LD and that, conversely, 2 variants may be in LD without being in the same haplotype block. |
| 3. *Demographic and evolutionary forces* |
| By influencing variant allele frequencies, these forces also strongly impact the LD. For example, in a growing population, the LD decreases by increasing the number of recombinations. On the contrary, the genetic drift can fix some alleles, and thus the corresponding haplotypes, leading to new LD patterns. Migrations and population admixture also modify allele frequencies and LD. In population genetics, the time of a migration can be deduced from the LD decay. Recent positive selection can be detected with a higher LD around the selected variants. Epistatic interactions may also be selected and keep alleles in LD even if they are distant. This explains why LD depends on the considered haplotypes with different patterns for multi-allelic variants. For example, in the highly polymorphic human Major Histocompatibility Complex (MHC), LD may extend over long distances for some specific ancestral haplotypes. Finally, as allele frequencies vary substantially worldwide, both the extent and strength of LD also differ among populations. |

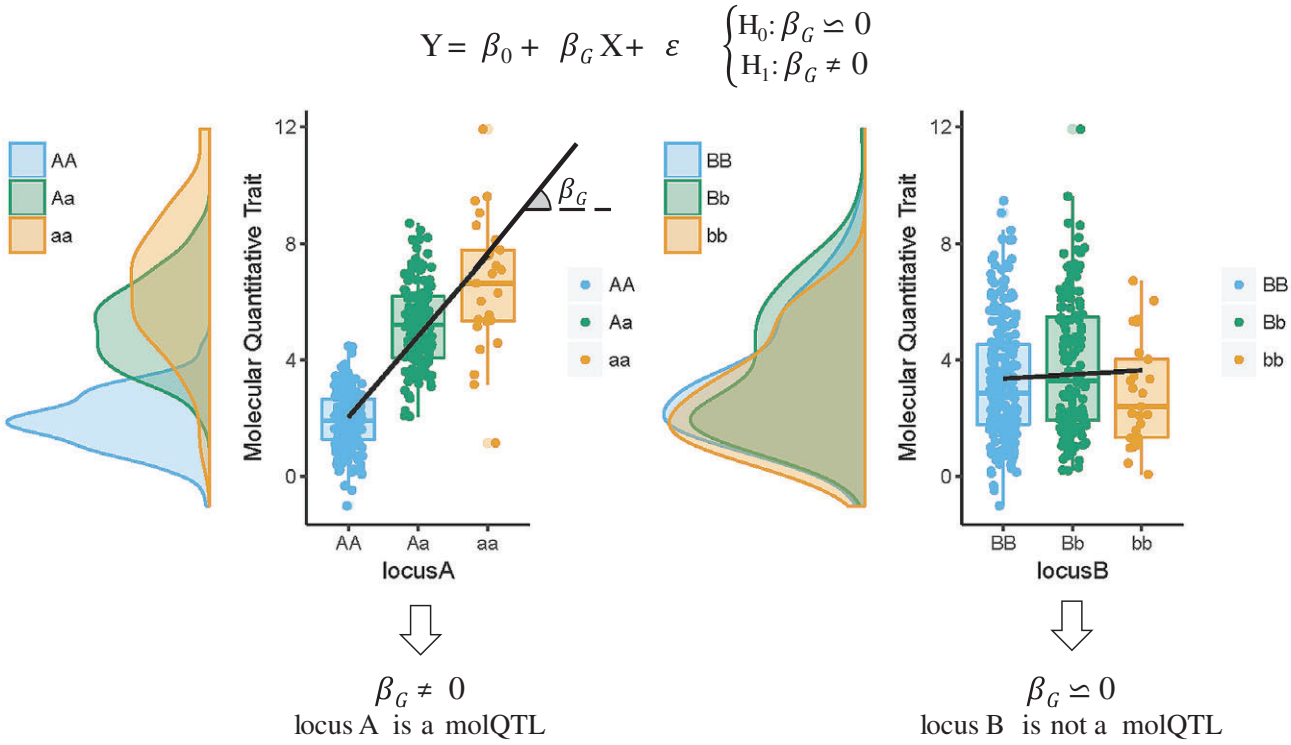


**FIGURE 1** Genetic association of a quantitative trait (QT). The association between a molecular QT (Y) and 2 loci A and B is being tested. Each panel shows the distribution of the QT depending on the genotype (X) as either density curves or boxplots. For locus A the regression slope $\beta_G$ significantly differs from 0, indicating that A is a QTL for the trait

regarded as intermediate phenotypes contributing to the disease outcome.

A major achievement was the demonstration that the variability of gene expression is heritable in humans.[12] These expression phenotypes are thus QTs amenable to genetic linkage and association analyses[13,14] leading to coin the term of expression QTL (eQTL)[14] (Table 2). When eQTLs are single nucleotide polymorphisms (SNPs), they are often called eSNPs. To conduct an eQTL mapping, the same samples are both genotyped genome-wide and characterized for gene expression. The first human eQTL studies benefited from the well-genotyped HapMap samples with available lymphoblastoid cell lines (LCLs).[32–34] They revealed the contribution of genetic diversity to the inter-individual variation of gene expression levels. In this context, searching for eQTLs has become a considerable endeavor requiring important resources and large consortia[35–43](Table 3).

However, because LCLs are Epstein-Barr-virus (EBV)-transformed and often oligoclonal, these initial findings raised questions regarding their relevance to *in vivo* cells and tissues. Nevertheless, further studies conducted on primary blood cells largely confirmed the results obtained with the LCLs[44,45] that eventually appeared to be good surrogates for B cells.[86] Subsequently, similar studies conducted on numerous primary tissues including adipose tissue, liver or human cortex, largely confirmed LCL findings and revealed that 30%-80% of eQTLs were cell- or tissue-specific notably in the brain.[35,45–48] A recent study even investigated induced pluripotent stem (iPS) cells. Comparisons with 44 human tissues revealed that 32% of the eQTLs were specific to iPS cells.[87] It is thus necessary to directly investigate gene expression in the tissue of interest, which raises complex issues for the central nervous system. Moreover, tissues are often heterogeneous and, without enrichment, eQTLs specific for rare cell types can be masked. A seminal demonstration of cell specificity was made in a study conducted in monocytes and B cells of 288 healthy volunteers[88] showing that the majority of eQTLs were cell-specific and that 6% of the eSNPs regulated the expression of a different gene in each cell type. Conversely, different eSNPs could affect the expression of the same gene in each cell type. Opposing effects of the same eSNP could also be observed, for example, rs2223286*C allele decreased expression of *SELL* in monocytes but increased it in B cells.

An important kind of eQTLs, called response eQTLs (reQTLs), occurs following exposure to an external signal (Table 2), in particular in the immune system in response to infectious agents and to various stimuli.[49,54,55] Other environmental factors, including hormones, drugs, diet, metals or pollutants, can also be the source of reQTLs.[56] Sex and age effects have also been identified.[57] One can anticipate that development-stage-specific eQTLs also exist in humans, as it has been reported in *Drosophila*.[58]

Altogether, these studies revealed important recurrent features of the thousands of eQTLs found in different tissues. Thus, almost 60%-80% of genes have at least one eQTL,[64] mostly *cis*-acting usually within 1 Mb of the transcription start site (TSS).[89] Distant eQTLs or even eQTLs lying on a different chromosome are less frequent, although a bias due to a lack of statistical power to detect them cannot be excluded.[90] A general rule is that the more distant the eQTLs, the more cell- and context-specific and the lower their effect sizes. These distant eQTLs may be *cis*-eQTLs of a transcription factor (TF) that in turn controls distant targets.[91] Besides, a given *trans*-eQTL may control several genes and thus operate as a master regulator like in the Major Histocompatibility Complex (MHC).[88]

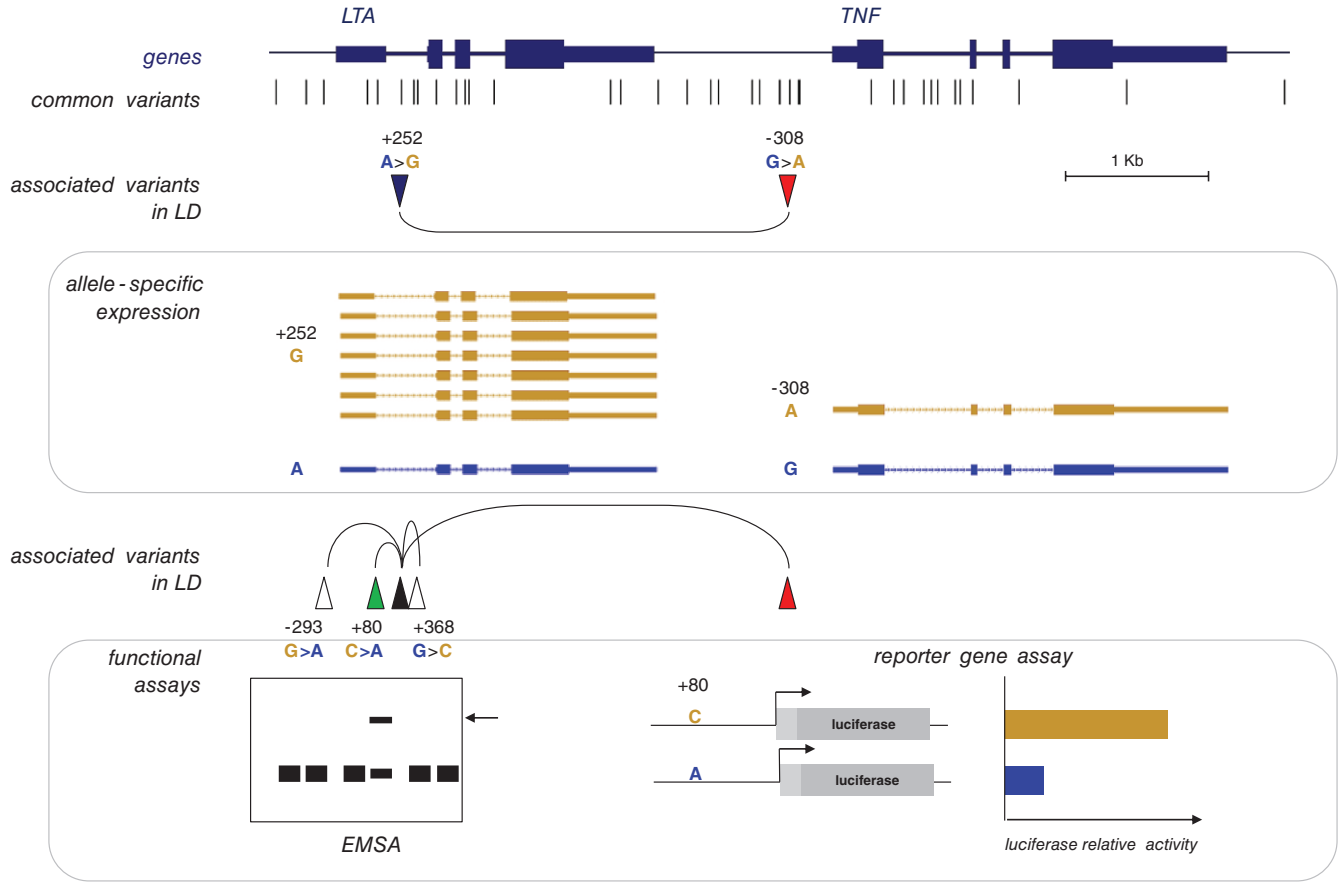


**FIGURE 2** The two main challenges in fine-mapping GWAS hits: identifying the modulated genes and the causative variant. The minor allele rs1800629*A, alias TNF-308*A (red triangle), in the *TNF* gene promoter had been associated with numerous autoimmune and infectious diseases but consensus lacked on its regulatory role. Using the variant rs909253 (LTA+252, black) in LD with it, allele-specific expression assay revealed that the risk haplotype (alleles in brown) rather affects the expression of the neighbor *LTA* gene.[8] Fine-mapping shortlisted 3 other variants in LD. Using functional assays, the causative variant rs2239074 was identified in 5′UTR of *LTA*+80 (green), with its A allele binding the ABF1 transcriptional repressor.[9] For each SNP, the major allele is named first

So far, eQTL variants were essentially eSNPs genotyped with microarrays while the contribution of copy number variants was evaluated at 17.7%.[92] The advent of next-generation sequencing might change the rules. The first study using a low-coverage whole-exome sequencing (WES) and RNASeq on the same 462 samples from the 1000 genomes project showed a significant enrichment for insertion/deletions (indels) among the top eQTLs.[36] More recently, using WES with higher coverage and RNASeq data from 13 tissues of the GTEx Consortium (Table 3), up to 6.8% of *cis*-eQTLs were found to be driven by structural variants, including large deletions, duplications, inversions, mobile-element insertions and copy-number variants, with effects several-fold larger than those of SNPs and short indels.[93] Most notably, rare structural variants contributed a large fraction of gene expression variation.

In the above studies, the considered parameter of the gene expression distribution was the mean. However, the variance of QTs may also depend on the genotype, as shown with an *FTO* polymorphism on body mass index (BMI).[85] When impacting on the variance of gene expression, these variance QTLs (vQTLs) are called evQTLs or veQTLs. The few studies conducted so far in LCLs, skin and fat cells, have revealed a non-negligible number of veQTLs, with a large proportion being *trans*-acting[51,52] and a parent-of-origin effect for 12.6% of genes.[53]

## 3.2 | Extension to other "omics" expression phenotypes

Gene expression is complex and diverse. Most of the above studies considered protein-coding genes. However, the idea of expression phenotypes can be extended to non-coding genes, to post-transcriptional RNA modifications and to the post-translational level, introducing the general concept of molecular QTLs (molQTLs, Table 2).

According to the last release of GENCODE (version 27, https://www.gencodegenes.org/releases/current.html),[94] there are currently twice as many non-coding RNAs genes as there are protein-coding genes, with a plethora of sizes, abundances and functions, amenable to fine characterization with next-generation sequencing. The most diverse are the long intergenic non-coding RNA (lincRNA) with about 7500 genes. They are implicated in both transcriptional and post-transcriptional regulation of gene expression and are generally less expressed than protein-coding transcripts. A first study mapped and replicated ~50 local lincRNA eQTLs.[68] Interestingly, 25% altered the expression of the neighboring protein-coding gene. Moreover, for half of them the expression levels of the lincRNAs and the protein-coding genes were negatively correlated, suggesting a negative regulation of the protein-coding gene by its lincRNA neighbor (Figure 3).

**TABLE 2** Glossary of expression QTLs (eQTLs) and other molecular quantitative trait loci (molQTLs)

| Cascade | Omic layer | Molecular quantitative trait | molQTLs | | Publications |
|---|---|---|---|---|---|
| | | | Recommended short name | Alternate names | |
| Regulation | Any | Regulatory quantitative trait | regQTL | regulatory QTL | 15 |
| | Epigenome | Epigenomic mark | epiQTL | | This review |
| | | DNA methylation | meQT | metQTL, mQTL[a] | 15–22 |
| | | DNAse I hypersensitivity/ chromatin accessibility | dsQTL | caQTL | 15,23–27 |
| | | Altered chromatin state | chromQTL | cQTL[b], chrQTL, chromatine QTL | 15,27–29 |
| | | Histone modification | hQTL | hmQTL | 15,20,21,24–27,29 |
| | | Histone acetylation | haQTL | | 15,30 |
| | | Transcription factor binding | bQTL | bindingQTL, tfQTL | 27,29,31 |
| | | Variable chromatin module | vcmQTL | cQTL[b] | 29 |
| Gene expression | Epigenome + transcriptome | Expression + DNA methylation | eQTM | expression QT methylation[c] | 19 |
| | Transcriptome | Expression | eQTL | eSNP if the variant is a SNP | 15,23,32–50 |
| | | Variance expression | evQTL | veQTL | 51–53 |
| | | Response expression | reQTL | response eQTL | 49,54–59 |
| | | Alternative splicing | sQTL | asQTL | 15,21,60–63 |
| | | Transcript ratio | trQTL | | 21,36,62–64 |
| | | Alternative polyadenylation | poly(A) ratio QTL | | 42 |
| | | MicroRNA binding | mirQTL | miRNA-binding QTLs | 65,66 |
| | | MicroRNA expression | miR-eQTL | mirQTL[d] | 36,66,67 |
| | | LincRNA expression | lincRNA eQTL | | 36,68,69 |
| | | Repeat element | repeat eQTL | Retrotransposon-derived | 36 |
| | | RNA decay | rdQTL | | 15,70 |
| | | Ribosome occupancy | rQTL | roQTL | 15,50,71 |
| | | Module expression QTL | emodQTL | mQTL[a] | 72 |
| | Proteome | Protein | pQTL | | 15,50,71,73,74 |
| | | Cytokine | cQTLs | | 75 |
| | | Immunoglobulin | IgQTL | | 76,77 |
| Higher-level intermediate phenotypes | Metabolome | Metabolite quantitative trait | mQTL | GIM | 72,78,79 |
| | Metagenome | Microbiota | mbQTL | | 80–82 |
| | Blood | Blood cell parameters | hemaQTL | Hematological QTL | 41,83,84 |
| Any | | Variance of a trait | vQTL | | 51–53,85 |
| | | Module of traits | modQTL | | 29,72 |

[a] Not recommended as it refers also to metabolite QTL.

[b] Also refers to cytokine QTL.

[c] With correlated eQTL and meQTL.

[d] Also refers to miRNA-binding QTLs.

The lincRNA eQTLs were mostly tissue-specific as the lincRNAs themselves. A more powered study highlighted some properties of lincRNAs eQTLs compared to those of protein-coding genes: *cis*-eQTLs are more prevalent, 2.5-4.5 times closer to the TSS and have larger effect sizes, suggesting a relaxed purifying selection.[69] It also confirmed that lincRNA *cis*-eQTL often modulate expression of the nearby protein-coding genes. It is worth noting that the Functional ANnoTation Of the Mammalian (FANTOM) genome collaborative project characterized an atlas of human long non-coding RNAs (lncRNAs) and found that expression of the lncRNAs overlapping an eQTL for protein-coding genes positively correlates with expression of these protein-coding genes, reflecting a general mode of co-regulation of neighbor loci.[95] A smaller group of non-coding RNAs are microRNAs (miRNAs) with 1881 genes and short transcripts involved in post-transcriptional regulation of presumably a third of mRNAs genes. A first category of eQTLs associated to miRNAs are

**TABLE 3** Some large-scale molecular quantitative trait loci (molQTLs) consortia

| Acronym | Full name | Samples | Webportal | Publications |
|---|---|---|---|---|
| MuTHER | Multiple Tissue Human Expression Resource | LCLs, skin, fat, ~60 subjects belonging to twin-pairs | http://www.muther.ac.uk/ | Nica et al[35] |
| GEUVADIS | Genetic European Variation in Health and Disease | LCLs, 462 subjects from 5 populations of the 1000 genome project | http://www.geuvadis.org/ | Lappalainen et al[36] |
| GTEx | Genotype-Tissue Expression | 11 688 samples, 714 individuals for 53 tissues, all major organs of the body were included plus detailed subregions of the brain obtained post-mortem | https://www.gtexportal.org/ | The GTEx Consortium[37]; Stranger et al[43] |
| STARNET | Stockholm-Tartu Atherosclerosis Reverse Networks Engineering Task | Blood, cardiovascular and metabolic tissues from 600 coronary artery disease patients | https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001203.v1.p1 | Franzén et al [38] |
| BLUEPRINT | Blueprint epigenome | Primary cells of the hematopoietic system (blood, tonsils, bone marrow), 50 healthy and 50 neoplasms | http://www.blueprint-epigenome.eu | Martens and Stunnenberg [40] |
| IHEC | International Human Epigenome Consortium | Key cellular stages relevant to health, diseases and senescence, 1000 epigenomes | http://ihec-epigenomes.org/ | Astle et al [41] |
| CMC | CommonMind | 600 prefrontal cortex from 291 healthy subjects, 275 schizophrenia and 47 bipolar disorder | http://commonmind.org | Fromer et al[39] |
| BIOS | Dutch Biobank-Based Integrative Omics Study | Peripheral-blood cells in 2116 healthy individuals | http://genenetwork.nl/biosqtlbrowser | Zhernakova et al[42] |

the miRNA-binding QTLs (mirQTLs) located on their target transcript explaining up to 25% of eQTLs in 3′ untranslated regions (UTRs) in LCLs and blood cells.[65,66] A second category is called miR-eQTLs. They directly regulate miRNA levels and are mostly located outside of the genomic region defined by the mature miRNA. Remarkably, 20%-33% of these were also associated to a variation of expression of their targets in immune cells.[36,66,67] Whether miR-eQTLs also affect protein levels remains to be explored.

Alternative splicing is a major mechanism driving transcriptome diversity, affecting 50%-74% of pre-mRNAs[96] and having a heritable pattern.[97,98] It is more prevalent in the MHC than in other genomic regions, thus adding to the huge polymorphism of this region.[99] The genetic variants modulating exon usage have first been mapped as spliceQTLs (sQTL or asQTL) in different tissues.[15,60,61] With more accurate isoform quantification using RNASeq, it has become possible to map eQTLs for transcript ratios (trQTLs)[36,64] and even for "composite" transcript ratios reflecting the equilibrium of all isoforms.[62,63] Overall, the sQTLs are 4 to 6 times less frequent than eQTLs and act independently of them, although they are uncommon in the absence of global gene-level variation. sQTLs usually are intronic, falling at binding sites for the splicing machinery. They are also enriched in alternative exons. Alternative poly-adenylation (poly(A)) is another mechanism of transcript diversification generating 5% of isoforms. By modifying the 3′ end of mRNAs, it alters their sensitivity to miRNAs and modulates their stability. About 2300 genes have a SNP that influences their poly(A) usage (poly(A) ratio QTLs) in whole-blood.[42] Finally, the aforementioned studies considered the steady-state transcript levels, without distinguishing mRNA production and decay rates. It is estimated that 19% of eQTLs could result from RNA decay. Thus, 195 loci affecting the mRNA decay rates (rdQTLs) have been found in 70 LCLs.[70]

Expression of proteins, the ultimate gene products in many cases, is also genetically regulated as showed by a study of 42 serum proteins in 1200 individuals that detected 8 *cis* protein QTLs (pQTLs) with large effect sizes.[73] Subsequently, the proteome of 95 HapMap LCLs was characterized by mass spectrometry.[74] The most variable proteins were immune-related. This study confirmed a marked inter-individual variability with 5.7% of proteins having at least a 1.5-fold variation in half of the sample pairs. However, only 77 *cis*-pQTLs were identified. Altogether, these pQTLs mapped preferentially in the

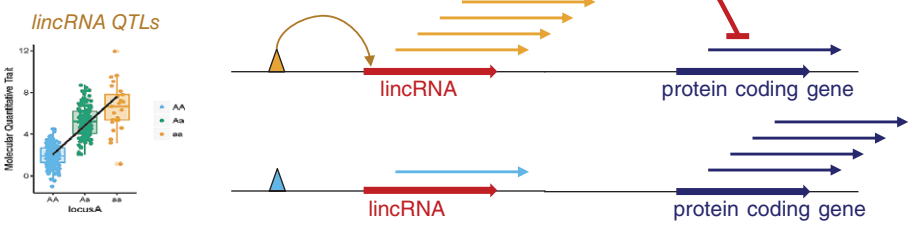


**FIGURE 3** Example of complex gene regulation by a molQTL. A molQTL controls the expression level of a *lincRNA* gene that, in turn, negatively modulates the expression of a neighbor target protein-coding gene

exonic parts and were enriched for missense SNPs.[15,50,74] Strikingly, only half of pQTLs were also eQTLs suggesting different genetic mechanisms regulating transcript and protein expression. When they were the same, the effect sizes of pQTLs were lower than those of eQTLs, indicating buffering at the protein level. However, small sample sizes, the different technologies used and their lack of sensitivity could also explain these findings and the difficulty to replicate pQTLs.

To better understand the genetic mechanisms at the interplay between transcripts and proteins, two studies combined RNASeq and mass spectrometry data and measured the translational activity by RiboSeq in LCLs.[50,71] The ribosome occupancy QTLs (rQTLs or roQTLs) overlapped with eQTLs with similar effect sizes, suggesting that specific pQTLs likely affect the post-translational stages. Finally, a landmark paper assessed the contribution of all molQTLs in LCLs, from the transcriptional rate, measured by 4sU-labeled RNAseq, to the protein abundance.[15] Effect sizes correlated across molQTLs, only being lower for the final pQTLs. In total, 85% of molQTLs were shared from one stage to the next into the regulatory cascade, with 73% of molQTLs that affect transcription rates also affecting protein expression.

### 3.3 | Extension to intermediate molecular phenotypes that are relevant to diseases

Besides proteins, cellular metabolites such as carbohydrates, amino acids or fatty acids provide a readout of the physiological status and cellular activity. Metabolite QTLs (mQTLs) were mapped in different tissues and body fluids.[78,79] Groups of mQTLs define genetically influenced metabotypes (GIMs). They usually have larger effect size than the clinical phenotype and are located close to the genes encoding, for instance, the transporters or the enzymes important for the homeostasis of the metabolites.

Cytokines and immunoglobulins (Ig) are proteins highly characteristic of the immune response. Seventeen cytokine QTLs (cQTLs) were identified in response to viral, bacterial or fungal stimuli.[75] They mapped to pathways including pattern recognition receptors, cytokines, complement inhibitors and kallikrein. They were enriched in SNPs previously associated with immune-mediated diseases. Regarding immunoglobulins, IgQTLs were characterized first for IgE plasma levels in relation to asthma.[76] A recent study explored IgA, IgM, IgG in a cohort of more than 20 000 healthy subjects.[77] A total of 43 IgQTLs were mapped explaining 4.3%-8.7% of the variance. They overlapped with variants influencing lymphoid malignancies, autoimmune diseases or immunodeficiencies. All associations were isotype-specific. Of interest, 10 IgQTLs had an impact on blood cell frequencies, including the B cell lineage.

Hematological blood traits, including counts of erythrocytes, leukocytes, platelets and related laboratory parameters, are tightly regulated with narrow normal physiological ranges and are important markers of medical conditions. Large-scale studies of thousands of samples have been performed worldwide.[41,83,84] The recent record-breaking study that was conducted on 36 hematological traits measured in 173 480 British individuals with 29.5 million imputed variants, brought the number of blood QTLs to 2706.[41] All but 5 were cell-specific. The associated variants were mostly frequent, with small effect sizes, and predominantly mapped to non-coding regions enriched for enhancers.

Another kind of molecular phenotype is the microbiome composition, that is, the collective genome of the microbiota that colonizes human gut, skin and mucosal surfaces. It substantially varies between individuals and can be altered in pathological conditions. It is influenced by genetic factors as showed by statistically powered twin studies.[100,101] QTs include microbiota diversity, abundance of taxa and of taxonomy groups and functional pathways. Three large GWAS identified up to 58 microbiote QTLs (mbQTLs).[80–82] Although there was little overlap between studies, the Vitamin D receptor gene was recurrently associated and variants lied in genes involved in host defense and immunity, cell adhesion and tissue barrier integrity.

## 4 | GENOME-WIDE GENETIC STUDIES OF EPIGENOMIC PHENOTYPES

Gene expression is primarily governed by chromatin structure, which determines DNA accessibility to regulatory transcriptional complexes. Major epigenomic marks include DNA methylation, histone modifications, chromatin accessibility or nucleosome occupancy and chromosomal conformation.[102] Binding of TFs to DNA also belongs to the epigenetic framework.[28] Altogether, they define the epigenome. The Encyclopedia of DNA Elements (ENCODE) has been the first major project to characterize it in-depth in a large panel of cell lines, providing the first genome-wide catalogue of functional regulatory elements.[103] Subsequently, international projects aiming at profiling epigenetic patterns in primary cells and tissues have relied on large consortia such as the NIH Roadmap Epigenomics Program, the International Human Epigenome Consortium or the BLUEPRINT project.[104]

Epigenomic marks can be themselves considered as QTs. In this review, we call the associated variants epigenomic QTLs (epiQTL, Table 2). Because DNA is usually available and its methylation can be easily quantified, 5-methylcytosine methylation at CpG islands has been the most studied epiQTL. The corresponding methylation QTLs (meQTLs)[16] remain consistent across the lifespan in blood.[17] They mostly act in *cis* within 10 kb of the relevant CpG. Remarkably, about 10% of *cis*-meQTLs are also *cis*-eQTLs, thereby called expression quantitative trait methylation (eQTM), 30% of them surprisingly being correlated with an increased gene expression. They are also enriched for CCCTC-binding factor (CTCF)- and TF-binding sites as shown in brain, purified neurons, glia, T cells and placenta.[18] *trans*-meQTLs are less frequent (3%-7%) but may target hundreds of CpGs and several of them appear to be eQTLs in *cis* for CTCF or TFs.[19] Similarly, chromatin accessibility can be studied by mapping DNAse I hypersensitivity sites. The corresponding dsQTL are often (39%) associated with gene expression variation in LCLs.[23] Conversely, 55% of eQTLs are also dsQTLs. These functional dsQTLs are enriched in allele-specific TF-binding sites. Further studies performed jointly on several chromatin marks in LCLs identified extensive allele-dependent variation defining histone QTLs (hQTL).[20,24–26] Histone marks at enhancers were the most variable ones. Enrichment of hQTLs in dsQTLs, eQTLs and meQTLs was observed. Of note, epiQTLs also correlated with

sQTLs, albeit to a lesser extent than with eQTLs.[15] These studies pointed to the disruption of TF-binding sites as the key mechanism driving these coordinated changes of epigenomic marks. Overall, 15% of the hQTLs can act both locally and distally. They define regions with coordinated chromatin state changes that are embedded within topologically associated domains (TADs).[27,29] In this context, TF-binding QTLs (bQTL) were shown to also influence long-range chromosomal contacts.[31] These findings could be extended to primary blood cell subsets and highlighted the cell specificity of this chromatin coordination.[21]

Altogether, current studies of eQTLs and other molQTLs, mostly bearing on immune cells, cast unprecedented light on the regulatory elements underlying gene expression, their spatial distribution across the genome and how they are correlated in the regulatory cascades, with half of the transcription variation explained by chromatin marks jointly with genetic variation.[15,21]

## 5 | TOWARDS THE IDENTIFICATION OF DISEASE-CAUSING VARIANTS

### 5.1 | Clear advantages of using molQTLs

In an effort to pinpoint the causative variants among the GWAS variants, the overlap with ENCODE elements provided interesting hints. Thus, 13.1% of GWAS hits were found to map directly to an ENCODE element, while 59% were in LD with a SNP mapping to an ENCODE element.[103,105] Similarly, in primary cells of the Epigenome Roadmap, 60% of GWAS variants for immune-mediated diseases mapped to "immune" enhancers and 10%-20% to TF-binding motifs.[6] However, connecting the associated alleles to phenotypes has been less straightforward because ENCODE and Epigenome Roadmap samples have a limited genetic diversity.

Catalogues of molQTLs thus provide invaluable tools to bridge the gap between genotypes and clinical phenotypes.[14,43] The rationale is to prioritize the most credible candidate variants by intersecting the list of GWAS SNPs with that of molQTLs (Figure 4A), using for example, webportals like RegulomeDB[106] or HaploReg.[107] Following a first success in asthma,[108] this approach has become "classical." Importantly, mapping eQTLs in healthy individuals is assumed to be equally informative, the only difference being the frequency of disease-risk alleles. Besides, tissues not available from patients can be studied, even post-mortem for the brain and other vital organs.[43] Finally, variation of gene expression is not biased by confounding factors due to the disease or its treatment. On this basis, wealth of studies have highlighted eQTLs at GWAS hits.[90,109] Overall, there is a significant enrichment for all types of molQTLs in GWAS hits and reciprocally.[110] A classic example is the dissection of the 1p13 locus encompassing 7 genes and associated with myocardial infarction and high LDL cholesterolemia. The causative variant was fine-mapped to the 3′UTR of *CELSR2* but it affected expression of *SORT1* at 40 kb by creating a binding site for CCAAT-enhancer-binding proteins (C/EBPs).[111] A more complete story is provided by the *FTO* gene that was consistently associated to type 2 diabetes (T2D) and obesity but lacked functional support. Two studies identified long-distance chromatin interactions with 7 genes.[112,113] An eQTL mapping analysis in adipose tissue shortlisted two of them, *IRX3* and *IRX5* whose expression was affected by the risk haplotype. The lead SNP in *FTO* disrupted a binding site for ARID5B repressor factor. Its impact on *IRX3* and *IRX5* transcripts was confirmed by CRISPR-Cas9 genome editing in adipocyte progenitors and for *IRX3* by germline disruption in mice. Other molQTLs can be useful for fine-mapping of disease variants. For example, open chromatin regions, histone 3 (H3) lysine modification and CTCF-binding in pancreatic islets were confronted to top T2D-associated SNPs and this led to identify new regulatory variants at *TCF7L2* and *WFS1* with allele-dependent enhancer activity.[114,115]

A "reverse" approach (Figure 4B) starts from the molecular trait showing difference between patients and controls, then identifies genetic variants controlling this difference and checks that the molQTLs are indeed associated with the clinical trait. Thus, the HDL-cholesterol plasma level could be correlated with 67 transcripts regulated by a *cis*-eSNP. *VNN1* was the best correlated transcript and showed 4 eSNPs in its promoter that were also associated with HDL-cholesterol concentrations.[44] In systemic lupus erythematosus (SLE), both classical and reverse approaches were used. According to the classical one, the causative variant on chromosome 7 was found to act in *cis* on *IKZF1*. According to the reverse approach, this eQTL also explained in *trans* the previously known differential expression in SLE of the complement *C1QB* gene and 5 type I interferon response genes.[116] In this case, no new genetic variant was associated with the disease, but we propose that this reverse approach starting from gene expression may be used to rescue variants missed by GWAS. Epigenomic traits, including DNA methylation and histone acetylation, have been also searched for differential effects by Epigenome-Wide Association Studies (EWAS) or by Histone Acetylome-Wide Association Study (HAWAS).[30,117] In complement of these studies, mapping of epiQTLs can help to discriminate the differential epigenomic marks due to the influence of the environment from those having a genetic basis, provided analyses are conducted in the same cell-context with careful experimental design.[28,118]

A third usage of molQTL in clinical genetics can be envisioned in the context of Mendelian traits (Figure 4C). Non-coding variants affecting gene expression can act as modifiers of the main disease-causing locus by affecting its penetrance or expressivity.[119–121] MolQTLs in relevant tissues are a mine to search for these modifiers.

### 5.2 | Caution and limitations in using molQTLs

The fact that the molQTLs and the GWAS hits colocalize or are in strong LD does not imply that the disease association and the molecular trait are explained by the same variant. MolQTLs are many and widespread across the genome. Moreover, databases of molQTLs report all associated SNPs regardless of LD, and not only the lead molSNP, thus considerably increasing the probability of wrongly assigning a GWAS hit to a molQTL. The difficulty is amplified by the fact that the samples used to generate the molQTL databases and the GWAS are rarely the same, although there are exceptions.[108,122,123] They may come from populations with different LD patterns. Even if the population is the same, sampling fluctuation may

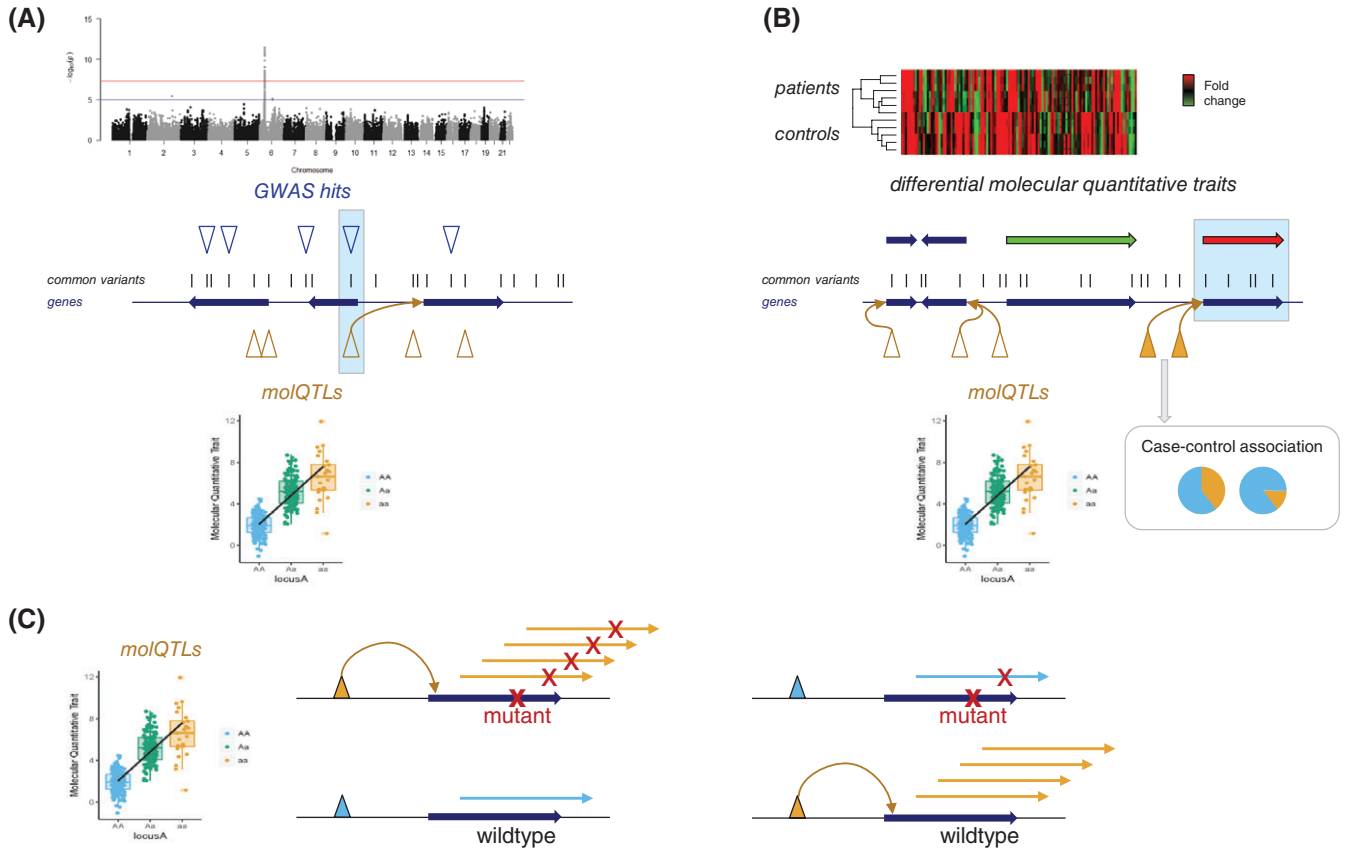


**FIGURE 4** Strategies for identification of the disease causative variants using molQTLs. (A) The most credible GWAS variants are prioritized by statistical methods of colocalization with molQTLs. Here, the candidate causative variant (blue box) modulates a distant gene. (B) Differential molecular traits are filtered by the presence of molQTLs regulating them. Here, the boxed gene is upregulated in patients and has 2 molQTLs (closed triangles) further tested for disease association. (C) Modulation of a Mendelian trait by a molQTL. Two scenarios are depicted. On the left, the molQTL is in *cis* with the disease-causing mutation and upregulates the mutant transcript, potentially exacerbating the disease. Note that if the mutation results in reduced function, it may instead rescue the phenotype. On the right, the molQTL upregulates the wild-type transcript, potentially attenuating the disease severity

affect LD estimates. To address this issue of causality inference, sophisticated statistical methods are being developed.[22,124–132] Such a method applied to 7 autoimmune diseases in 180 000 patients revealed that only 25% of the lead GWAS SNPs were eQTLs, 75% of them being cell-specific.[132] A similar proportion was obtained in inflammatory bowel disease (IBD) after fine-mapping the disease variants with high confidence.[133] These surprising results, which remain to be confirmed in other diseases, are at odds with the previously established idea that the majority of GWAS hits are molQTLs. That being said, even before these two studies, it was not possible to assign a molQTL to every GWAS hit. One possibility is that a fraction of the non-coding GWAS hits is actually not explained by a molQTL. Alternatively, the molQTLs explaining the GWAS hits have not been found yet. There are several reasons for that which are discussed below.

Strikingly, most of these studies have a blind spot, the MHC, despite the fact that this region is the most prominent one in terms of GWAS associations.[134] Indeed, the MHC holds a special status.[135] It is characterized by a strong LD over megabases that complicates its fine-scale analysis. In addition, its extreme polymorphism impairs the interpretation of molQTLs for molecular traits assessed by solely using the reference sequence, arguing for dedicated approaches.[99]

A second set of explanations pertains to the current incompleteness of molQTLs databases. Indeed, molQTLs studies are often underpowered, especially for molQTLs other than eQTLs and for *trans*-molQTLs whose effects are more often tissue-specific. In addition, quantification methodologies for a given molecular trait are heterogeneous and genome-wide significance thresholds vary between molQTL studies. This may explain, for instance, the limited overlap between 4 eQTL resources for blood pressure associations.[136] A final point is the frequent lack of the appropriate biological context in the molQTL databases. We have previously emphasized the importance of the tissue or cell type for molQTLs. The importance of the relevant context in fine-mapping of disease-causing variants has been underscored by the robust colocalization of hits from 57 GWAS with eQTLs from 24 different cells and tissues.[137] Likewise, GWAS hits are more enriched for reQTLs than for constitutive QTLs.[59] Ultimately, a first eQTL mapping recently performed on single cells from whole blood has strengthened this argument of cell specificity by identifying a significant proportion of cell-specific eQTLs undetected in bulk RNA.[138] Along this line, a method has been proposed to infer cell types to relevant complex diseases using single-cell eQTLs.[139]

The complex inheritance of molQTLs resulting from allelic heterogeneity,[140,141] pleiotropy[22,88] and epistasis[142,143] further

impacts on their interpretation and the causality inference to complex diseases. It is sometimes difficult to tease apart these effects. For instance, epistasis between two eSNPs without pairwise LD could be explained by a third eSNP in low LD with them.[144] Related to epistasis is the notion of context-dependency whereby an eQTL is influenced by the expression of another gene detected by interaction modeling. For example, expression of *NOD2* in whole blood is explained by an eQTL whose effect is substantially increased when *STX3A*, a neutrophils marker, is highly expressed.[42] Locus pleiotropy is illustrated by the *SLC16A11* haplotype, a major risk for T2D in Mexico that contains regulatory variants decreasing its transcript level in liver. But, in addition, coding variants act at the post-translational level by impairing binding to its chaperone, resulting in a decreased surface expression.[145] And similarly to complex traits, there is missing heritability with, for instance, current estimates of 69% for expression phenotypes in whole blood.[146]

A final difficulty stems from the interplay between the different molQTLs. There is no one-to-one relationship in their "percolation" cascade. This lack of linearity is exemplified by the non-independence of molecular traits. In fact, molecular traits show covariance both within and between each omic layer. To tackle this, *trans*-molQTLs are a first step forward to infer module networks. New methods are developed to integrate multiple molecular phenotypes.[29,72,91,147] Human quantitative genetics is thus leaping into the area of systems biology.[148]

## 6 | CONCLUSION/PERSPECTIVES

Studying the genetic basis of genomic and epigenomic traits has first yielded a wealth of basic information on the regulation of gene expression, on the dynamic structure of chromatin and the projection of regulatory variants onto cellular phenotypes. It also provides a wonderful tool to elucidate the genetic bases of complex diseases by reducing the list of disease-associated variants. Nonetheless, colocalization of GWAS hits and molQTLs is a complex process requiring sophisticated statistical methods. Strikingly, this approach could pinpoint the relevant gene, often at distance of the disease-associated variant. The next challenge is to better integrate the different molecular traits in gene and cellular networks to uncover the functional pathways modulated by the genetic component of common diseases.[2] This complexity emphasizes the absolute necessity to conduct functional studies to validate genetic findings. In this endeavor, the emergence of high-throughput genomics methods to edit the genome and to investigate the phenotypic impact of candidate variants will be essential to dissect the huge number of GWAS hits.[149,150]

## ACKNOWLEDGEMENTS

C.V. is supported by Université Paris Diderot. C.V. thanks Cécile Julier for reading the manuscript and the Editor for helpful comments. We apologize to those of our colleagues that we could not cite due to word count constraint.

## Conflict of interest

Nothing to declare.

## ORCID

*C. Vandiedonck* http://orcid.org/0000-0002-6669-6923

## REFERENCES

1. Manolio TA. In retrospect: a decade of shared genomic associations. *Nature*. 2017;546(7658):360-361.
2. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. *Cell*. 2017;169(7):1177-1186.
3. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nat Genet*. 2003;33(3s):228-237.
4. Stenson PD, Mort M, Ball EV, et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet*. 2017;136(6):665-677.
5. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci*. 2009;106(23):9362-9367.
6. Farh KK-H, Marson A, Zhu J, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*. 2014;518(7539):337-343.
7. Jansen RC, Nap J-P. Genetical genomics: the added value from segregation. *Trends Genet*. 2001;17(7):388-391.
8. Knight JC, Keating BJ, Rockett KA, Kwiatkowski DP. In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. *Nat Genet*. 2003;33(4):469-475.
9. Knight JC, Keating BJ, Kwiatkowski DP. Allele-specific repression of lymphotoxin-alpha by activated B cell factor-1. *Nat Genet*. 2004;36(4):394-399.
10. Giraud M, Taubert R, Vandiedonck C, et al. An IRF8-binding promoter variant and AIRE control CHRNA1 promiscuous expression in thymus. *Nature*. 2007;448(7156):934-937.
11. Vafiadis P, Bennett ST, Todd JA, et al. Insulin expression in human thymus is modulated by INS VNTR alleles at the IDDM2 locus. *Nat Genet*. 1997;15(3):289-292.
12. Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW. Allelic variation in human gene expression. *Science*. 2002;297(5584):1143.
13. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. *Science*. 2002;296(5568):752-755.
14. Schadt EE, Monks SA, Drake TA, et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature*. 2003;422(6929):297-302.
15. Li YI, van de Geijn B, Raj A, et al. RNA splicing is a primary link between genetic variation and disease. *Science*. 2016;352(6285):600-604.
16. Bell JT, Pai AA, Pickrell JK, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol*. 2011;12(1):R10.
17. Gaunt TR, Shihab HA, Hemani G, et al. Systematic identification of genetic influences on methylation across the human life course. *Genome Biol*. 2016;17:61.
18. Do C, Lang CF, Lin J, et al. Mechanisms and disease associations of haplotype-dependent allele-specific DNA methylation. *Am J Hum Genet*. 2016;98(5):934-955.
19. Bonder MJ, Luijk R, Zhernakova DV, et al. Disease variants alter transcription factor levels and methylation of their binding sites. *Nat Genet*. 2017;49(1):131-138.
20. Banovich NE, Lan X, McVicker G, et al. Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet*. 2014;10(9):e1004663.

21. Chen L, Ge B, Casale FP, et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell*. 2016;167(5):1398-1414.e24.
22. Hannon E, Weedon M, Bray N, O'Donovan M, Mill J. Pleiotropic effects of trait-associated genetic variation on DNA methylation: utility for refining GWAS loci. *Am J Hum Genet*. 2017;100(6):954-959.
23. Degner JF, Pai AA, Pique-Regi R, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*. 2012;482(7385):390-394.
24. Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, et al. Extensive variation in chromatin states across humans. *Science*. 2013;342(6159):750-752.
25. Kilpinen H, Waszak SM, Gschwind AR, et al. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*. 2013;342(6159):744-747.
26. McVicker G, van de Geijn B, Degner JF, et al. Identification of genetic variants that affect histone modifications in human cells. *Science*. 2013;342(6159):747-749.
27. Grubert F, Zaugg JB, Kasowski M, et al. Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell*. 2015;162(5):1051-1065.
28. Lappalainen T, Greally JM. Associating cellular epigenetic models with human phenotypes. *Nat Rev Genet*. 2017;18(7):441-451.
29. Waszak SM, Delaneau O, Gschwind AR, et al. Population variation and genetic control of modular chromatin architecture in humans. *Cell*. 2015;162(5):1039-1050.
30. Sun W, Poschmann J, Cruz-Herrera del Rosario R, et al. Histone acetylome-wide association study of autism spectrum disorder. *Cell*. 2016;167(5):1385-1397.e11.
31. Tehranchi AK, Myrthil M, Martin T, Hie BL, Golan D, Fraser HB. Pooled ChIP-Seq links variation in transcription factor binding to complex disease risk. *Cell*. 2016;165(3):730-741.
32. Cheung VG, Spielman RS. The genetics of variation in gene expression. *Nat Genet*. 2002;32 (suppl):522-525.
33. Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, Cheung VG. Common genetic variants account for differences in gene expression among ethnic groups. *Nat Genet*. 2007;39(2):226-231.
34. Stranger BE, Nica AC, Forrest MS, et al. Population genomics of human gene expression. *Nat Genet*. 2007;39(10):1217-1224.
35. Nica AC, Parts L, Glass D, et al. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet*. 2011;7(2):e1002003.
36. Lappalainen T, Sammeth M, Friedländer MR, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*. 2013;501(7468):506-511.
37. The GTEx Consortium. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. 2015;348(6235):648-660.
38. Franzén O, Ermel R, Cohain A, et al. Cardiometabolic risk loci share downstream *cis*- and *trans*-gene regulation across tissues and diseases. *Science*. 2016;353(6301):827-830.
39. Fromer M, Roussos P, Sieberts SK, et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat Neurosci*. 2016;19(11):1442-1453.
40. Martens JHA, Stunnenberg HG. BLUEPRINT: mapping human blood cell epigenomes. *Haematologica*. 2013;98(10):1487-1489.
41. Astle WJ, Elding H, Jiang T, et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell*. 2016;167(5):1415-1429.e19.
42. Zhernakova DV, Deelen P, Vermaat M, et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat Genet*. 2017;49(1):139-145.
43. Stranger BE, Brigham LE, Hasz R, et al. Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. *Nat Genet*. 2017;49:1664-1670.
44. Göring HHH, Curran JE, Johnson MP, et al. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet*. 2007;39(10):1208-1216.
45. Emilsson V, Thorleifsson G, Zhang B, et al. Genetics of gene expression and its effect on disease. *Nature*. 2008;452(7186):423-428.
46. Myers AJ, Gibbs JR, Webster JA, et al. A survey of genetic human cortical gene expression. *Nat Genet*. 2007;39(12):1494-1499.
47. Schadt EE, Molony C, Chudin E, et al. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol*. 2008;6(5):e107.
48. Dimas AS, Deutsch S, Stranger BE, et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*. 2009;325(5945):1246-1250.
49. Fairfax BP, Knight JC. Genetics of gene expression in immunity to infection. *Curr Opin Immunol*. 2014;30:63-71.
50. Battle A, Khan Z, Wang SH, et al. Impact of regulatory variation from RNA to protein. *Science*. 2015;347(6222):664-667.
51. Hulse AM, Cai JJ. Genetic variants contribute to gene expression variability in humans. *Genetics*. 2013;193(1):95-108.
52. Wang G, Yang E, Brinkmeyer-Langford CL, Cai JJ. Additive, epistatic, and environmental effects through the lens of expression variability QTL in a twin cohort. *Genetics*. 2014;196(2):413-425.
53. Brown AA. veqtl-mapper: variance association mapping for molecular phenotypes. *Bioinformatics*. 2017;33(17):2772-2773.
54. Quach H, Rotival M, Pothlichet J, et al. Genetic adaptation and Neanderthal admixture shaped the immune system of human populations. *Cell*. 2016;167(3):643-656.e17.
55. Nédélec Y, Sanz J, Baharian G, et al. Genetic ancestry and natural selection drive population differences in immune responses to pathogens. *Cell*. 2016;167(3):657-669.e21.
56. Moyerbrailean GA, Richards AL, Kurtz D, et al. High-throughput allele-specific expression across 250 environmental conditions. *Genome Res*. 2016;26(12):1627-1638.
57. Yao C, Joehanes R, Johnson AD, et al. Sex- and age-interacting eQTLs in human complex diseases. *Hum Mol Genet*. 2014;23(7):1947-1956.
58. Cannavò E, Koelling N, Harnett D, et al. Genetic variants regulating expression levels and isoform diversity during embryogenesis. *Nature*. 2016;541(7637):402-406.
59. Kim-Hellmuth S, Bechheim M, Pütz B, et al. Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. *Nat Commun*. 2017;8(1):266.
60. Zhang X, Joehanes R, Chen BH, et al. Identification of common genetic variants controlling transcript isoform variation in human whole blood. *Nat Genet*. 2015;47(4):345-352.
61. Takata A, Matsumoto N, Kato T. Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nat Commun*. 2017;8:14519.
62. Monlong J, Calvo M, Ferreira PG, Guigó R. Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nat Commun*. 2014;5:4698.
63. Yang Q, Hu Y, Li J, Zhang X. ulfasQTL: an ultra-fast method of composite splicing QTL analysis. *BMC Genomics*. 2017;18(S1):963.
64. Battle A, Mostafavi S, Zhu X, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res*. 2014;24(1):14-24.
65. Lu J, Clark AG. Impact of microRNA regulation on variation in human gene expression. *Genome Res*. 2012;22(7):1243-1254.
66. Gamazon ER, Ziliak D, Im HK, et al. Genetic architecture of microRNA expression: implications for the transcriptome and complex traits. *Am J Hum Genet*. 2012;90(6):1046-1063.
67. Huan T, Rong J, Liu C, et al. Genome-wide identification of microRNA expression quantitative trait loci. *Nat Commun*. 2015;6:6601.
68. Kumar V, Westra H-J, Karjalainen J, et al. Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genet*. 2013;9(1):e1003201.
69. Popadin K, Gutierrez-Arcelus M, Dermitzakis ET, Antonarakis SE. Genetic and epigenetic regulation of human lincRNA gene expression. *Am J Hum Genet*. 2013;93(6):1015-1026.
70. Pai AA, Cain CE, Mizrahi-Man O, et al. The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. *PLoS Genet*. 2012;8(10):e1003000.

71. Cenik C, Cenik ES, Byeon GW, et al. Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. *Genome Res*. 2015;25(11):1610-1621.
72. Nath AP, Ritchie SC, Byars SG, et al. An interaction map of circulating metabolites, immune gene networks, and their genetic regulation. *Genome Biol*. 2017;18(1):146.
73. Melzer D, Perry JRB, Hernandez D, et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet*. 2008;4(5):e1000072.
74. Wu L, Candille SI, Choi Y, et al. Variation and genetic control of protein abundance in humans. *Nature*. 2013;499(7456):79-82.
75. Li Y, Oosting M, Smeekens SP, et al. A functional genomics approach to understand variation in cytokine production in humans. *Cell*. 2016;167(4):1099-1110.e14.
76. Moffatt MF, Gut IG, Demenais F, et al. A large-scale, consortium-based genomewide association study of asthma. *N Engl J Med*. 2010;363(13):1211-1221.
77. Jonsson S, Sveinbjornsson G, de Lapuente Portilla AL, et al. Identification of sequence variants influencing immunoglobulin levels. *Nat Genet*. 2017;49(8):1182-1191.
78. Illig T, Gieger C, Zhai G, et al. A genomewide perspective of genetic variation in human metabolism. *Nat Genet*. 2010;42(2):137-141.
79. Kastenmüller G, Raffler J, Gieger C, Suhre K. Genetics of human metabolism: an update. *Hum Mol Genet*. 2015;24(R1):R93-R101.
80. Bonder MJ, Kurilshikov A, Tigchelaar EF, et al. The effect of host genetics on the gut microbiome. *Nat Genet*. 2016;48(11):1407-1412.
81. Turpin W, Espin-Garcia O, Xu W, et al. Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat Genet*. 2016;48(11):1413-1417.
82. Wang J, Thingholm LB, Skiecevičienė J, et al. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. *Nat Genet*. 2016;48(11):1396-1406.
83. Andrews NC. Genes determining blood cell traits. *Nat Genet*. 2009;41(11):1161-1162.
84. Reiner AP, Lettre G, Nalls MA, et al. Genome-wide association study of white blood cell count in 16,388 African Americans: the continental origins and genetic epidemiology network (COGENT). *PLoS Genet*. 2011;7(6):e1002108.
85. Yang J, Loos RJF, Powell JE, et al. FTO genotype is associated with phenotypic variability of body mass index. *Nature*. 2012;490(7419):267-272.
86. Caliskan M, Cusanovich DA, Ober C, Gilad Y. The effects of EBV transformation on gene expression levels and methylation profiles. *Hum Mol Genet*. 2011;20(8):1643-1652.
87. Kilpinen H, Goncalves A, Leha A, et al. Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature*. 2017;546(7658):370-375.
88. Fairfax BP, Makino S, Radhakrishnan J, et al. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat Genet*. 2012;44(5):502-510.
89. Liu X, Finucane HK, Gusev A, et al. Functional architectures of local and distal regulation of gene expression in multiple human tissues. *Am J Hum Genet*. 2017;100(4):605-616.
90. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. *Nat Rev Genet*. 2015;16(4):197-212.
91. Yao C, Joehanes R, Johnson AD, et al. Dynamic role of trans regulation of gene expression in relation to complex traits. *Am J Hum Genet*. 2017;100(4):571-580.
92. Stranger BE, Forrest MS, Dunning M, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*. 2007;315(5813):848-853.
93. Chiang C, Scott AJ, Davis JR, et al. The impact of structural variation on human gene expression. *Nat Genet*. 2017;49(5):692-699.
94. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*. 2012;22(9):1760-1774.
95. Hon C-C, Ramilowski JA, Harshbarger J, et al. An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature*. 2017;543(7644):199-204.
96. Johnson JM, Castle J, Garrett-Engele P, et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*. 2003;302(5653):2141-2144.
97. Hull J, Campino S, Rowlands K, et al. Identification of common genetic variation that modulates alternative splicing. *PLoS Genet*. 2007;3(6):e99.
98. Kwan T, Benovoy D, Dias C, et al. Heritability of alternative splicing in the human genome. *Genome Res*. 2007;17(8):1210-1218.
99. Vandiedonck C, Taylor MS, Lockstone HE, et al. Pervasive haplotypic variation in the spliceo-transcriptome of the human major histocompatibility complex. *Genome Res*. 2011;21(7):1042-1054.
100. Goodrich JK, Waters JL, Poole AC, et al. Human genetics shape the gut microbiome. *Cell*. 2014;159(4):789-799.
101. Xie H, Guo R, Zhong H, et al. Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell Syst*. 2016;3(6):572-584.e3.
102. Stricker SH, Köferle A, Beck S. From profiles to function in epigenomics. *Nat Rev Genet*. 2016;18(1):51-66.
103. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. *Genome Res*. 2012;22(9):1748-1759.
104. Satterlee JS, Schubeler D, Ng HH. Tackling the epigenome: challenges and opportunities for collaboration. *Nat Biotechnol*. 2011;28(10):1039-1044.
105. Maurano MT, Humbert R, Rynes E, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012;337(6099):1190-1195.
106. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*. 2012;22(9):1790-1797.
107. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res*. 2012;40(D1):D930-D934.
108. Moffatt MF, Kabesch M, Liang L, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature*. 2007;448(7152):470-473.
109. Montgomery SB, Dermitzakis ET. From expression QTLs to personalized transcriptomics. *Nat Rev Genet*. 2011;12(4):277-282.
110. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet*. 2010;6(4):e1000888.
111. Musunuru K, Strong A, Frank-Kamenetsky M, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*. 2010;466(7307):714-719.
112. Claussnitzer M, Dankel SN, Kim K-H, et al. FTO obesity variant circuitry and adipocyte browning in humans. *N Engl J Med*. 2015;373(10):895-907.
113. Smemo S, Tena JJ, Kim K-H, et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*. 2014;507(7492):371-375.
114. Gaulton KJ, Nammo T, Pasquali L, et al. A map of open chromatin in human pancreatic islets. *Nat Genet*. 2010;42(3):255-259.
115. Stitzel ML, Sethupathy P, Pearson DS, et al. Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci. *Cell Metab*. 2011;12(5):443-455.
116. Westra H-J, Peters MJ, Esko T, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet*. 2013;45(10):1238-1243.
117. Do C, Shearer A, Suzuki M, et al. Genetic-epigenetic interactions in cis: a major focus in the post-GWAS era. *Genome Biol*. 2017;18(1):120.
118. Birney E, Smith GD, Greally JM. Epigenome-wide association studies and the interpretation of disease—Omics. *PLoS Genet*. 2016;12(6):e1006105.
119. Dimas AS, Stranger BE, Beazley C, et al. Modifier effects between regulatory and protein-coding variation. *PLoS Genet*. 2008;4(10):e1000244.
120. Li X, Montgomery SB. Detection and impact of rare regulatory variants in human disease. *Front Genet*. 2013;4:67.

121. Lappalainen T, Montgomery SB, Nica AC, Dermitzakis ET. Epistatic selection between coding and regulatory variation in human evolution and disease. *Am J Hum Genet*. 2011;89(3):459-463.
122. Davenport EE, Burnham KL, Radhakrishnan J, et al. Genomic landscape of the individual host response and outcomes in sepsis: a prospective cohort study. *Lancet Respir Med*. 2016;4(4):259-271.
123. Ram R, Mehta M, Nguyen QT, et al. Systematic evaluation of genes and genetic variants associated with type 1 diabetes susceptibility. *J Immunol*. 2016;196(7):3043-3053.
124. Nica AC, Montgomery SB, Dimas AS, et al. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet*. 2010;6(4):e1000895.
125. He X, Fuller CK, Song Y, et al. Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *Am J Hum Genet*. 2013;92(5):667-680.
126. Giambartolomei C, Vukcevic D, Schadt EE, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet*. 2014;10(5):e1004383.
127. Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet*. 2015;47(9):1091-1098.
128. Pickrell JK, Berisa T, Liu JZ, Ségurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet*. 2016;48(7):709-717.
129. Hormozdiari F, van de Bunt M, Segrè AV, et al. Colocalization of GWAS and eQTL signals detects target genes. *Am J Hum Genet*. 2016;99(6):1245-1260.
130. Zhu Z, Zhang F, Hu H, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet*. 2016;48(5):481-487.
131. Wen X, Pique-Regi R, Luca F. Integrating molecular QTL data into genome-wide genetic association analysis: probabilistic assessment of enrichment and colocalization. *PLoS Genet*. 2017;13(3):e1006646.
132. Chun S, Casparino A, Patsopoulos NA, et al. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat Genet*. 2017;49(4):600-605.
133. Huang H, Fang M, Jostins L, et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature*. 2017;547(7662):173-178.
134. de Bakker PIW, Raychaudhuri S. Interrogating the major histocompatibility complex with high-throughput genomics. *Hum Mol Genet*. 2012;21(R1):R29-R36.
135. Vandiedonck C, Knight JC. The human major histocompatibility complex as a paradigm in genomics research. *Brief Funct Genomic Proteomic*. 2009;8(5):379-394.
136. Wain LV, Vaez A, Jansen R, et al. Novel blood pressure locus and gene discovery using genome-wide association study and expression data sets from blood and the kidney novelty and significance. *Hypertension*. 2017;70(3):e4-e19.
137. Hauberg ME, Zhang W, Giambartolomei C, et al. Large-scale identification of common trait and disease variants affecting gene expression. *Am J Hum Genet*. 2017;100(6):885-894.
138. van der Wijst MGP, Brugge H, de Vries DH, Franke LH. Single-cell RNA sequencing reveals cell-type specific *cis*-eQTLs in peripheral blood mononuclear cells. bioRxiv. August 2017.
139. Calderon D, Bhaskar A, Knowles D, et al. Inferring relevant cell types for complex traits using single-cell gene expression. bioRxiv. 2017:136283.
140. Hormozdiari F, Zhu A, Kichaev G, et al. Widespread allelic heterogeneity in complex traits. *Am J Hum Genet*. 2017;100(5):789-802.
141. Wood AR, Hernandez DG, Nalls MA, et al. Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association. *Hum Mol Genet*. 2011;20(20):4082-4092.
142. Buil A, Brown AA, Lappalainen T, et al. Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat Genet*. 2015;47(1):88-91.
143. Hemani G, Shakhbazov K, Westra H-J, et al. Detection and replication of epistasis influencing transcription in humans. *Nature*. 2014;508(7495):249-253.
144. Wood AR, Tuke MA, Nalls MA, et al. Another explanation for apparent epistasis. *Nature*. 2014;514(7520):E3-E5.
145. Rusu V, Hoch E, Mercader JM, et al. Type 2 diabetes variants disrupt function of SLC16A11 through two distinct mechanisms. *Cell*. 2017;170(1):199-212.e20.
146. Lloyd-Jones LR, Holloway A, McRae A, et al. The genetic architecture of gene expression in peripheral blood. *Am J Hum Genet*. 2017;100(2):228-237.
147. Brynedal B, Choi J, Raj T, et al. Large-scale trans-eQTLs affect hundreds of transcripts and mediate patterns of transcriptional co-regulation. *Am J Hum Genet*. 2017;100(4):581-591.
148. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet*. 2015;16(2):85-97.
149. Dao LTM, Galindo-Albarrán AO, Castro-Mondragon JA, et al. Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat Genet*. 2017;49(7):1073-1081.
150. Xie S, Duan J, Li B, Zhou P, Hon GC. Multiplexed engineering and analysis of combinatorial enhancer activity in single cells. *Mol Cell*. 2017;66(2):285-299.e5.


**How to cite this article:** Vandiedonck C. Genetic association of molecular traits: A help to identify causative variants in complex diseases. *Clin Genet*. 2018;93:520–532. https://doi.org/10.1111/cge.13187

## Graphical abstract