A large-scale multivariate pQTL study sheds light on the genetic architecture of obesity

Hélène Ruffieux[1,2], Jérôme Carayol[2], Mary Ellen Harper[3], Robert Dent[4], Wim H. M. Saris[5], Arne Astrup[6], Anthony C. Davison[1], Jörg Hager[2], Armand Valsesia[2,*]

January 18, 2019

Abstract

The genetic contribution to obesity has been widely studied, yet the functional mechanisms underlying metabolic states remain elusive. This has prompted analysis of endophenotypes via quantitative trait locus studies, which assess how genetic variants affect intermediate gene (eQTL) or protein (pQTL) expression phenotypes. To leverage shared regulatory patterns, we present the first integrative multivariate pQTL analysis, performed with our scalable Bayesian framework LOCUS on plasma mass-spectrometry and aptamer-based proteomic datasets. We identify 136 pQTL associations in the Ottawa obesity clinical practice, of which $> 80\%$ replicate in the DiOGenes obesity cohort and show significant functional enrichments; 16% of the validated hits would be missed by standard univariate methods. By also exploiting extensive clinical data, our methods and results reveal the implication of proteins under genetic control in low-grade inflammation, insulin resistance, and dyslipidemia, thereby opening new perspectives for diagnosing and treating metabolic disorders. All results are freely accessible online from our searchable database.

Keywords: Metabolic Syndrome; Multivariate Bayesian modelling; Proteomic quantitative trait locus analysis; Scalable variational algorithm; Stratified obesity cohorts; Two-stage integrative study.

Genome-wide association studies (GWAS) have identified hundreds of loci associated with obesity susceptibility[1], yet their functional impact on metabolism remains poorly understood. The analysis of endophenotypes via molecular quantitative trait locus (QTL) studies may provide deeper insight into the biology underlying clinical traits. While gene expression QTL (eQTL) studies are now routinely performed, protein expression QTL (pQTL) studies have emerged only recently[2–4]. These studies allow the exploration of the functional bases of obesity, as certain proteins act as proxies for metabolic endpoints[3,5,6]. However two major hurdles hamper pQTL analyses. First, owing to the number of tests that they entail, conventional univariate approaches lack statistical power for uncovering weak

* Corresponding author: armand.valsesia@rd.nestle.com.

[1]Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

[2] Nestlé Research, EPFL Innovation Park, Lausanne, Switzerland

[3]Bioenergetics Laboratory, University of Ottawa, Ottawa, Ontario, Canada

[4]Weight Management Clinic, The Ottawa Hospital, Ottawa, Ontario, Canada

[5]Department of Human Biology, NUTRIM, School of Nutrition and Translational Research in Metabolism, Maastricht University Medical Centre, Maastricht, Netherlands

[6]Department of Nutrition, Exercise and Sports, Faculty of Science, University of Copenhagen, Copenhagen, Denmark

associations, such as *trans* and pleiotropic effects[7,8], while better-suited multivariate methods fail to scale to the dimensions of QTL studies. Second, the clinical data complementing QTL data are often very limited, restricting subsequent investigation to external information from unrelated populations, health status or study designs, and rendering some degree of speculation unavoidable.

Here we aim to address both concerns in an integrative study of two obesity clinical cohorts, the Ottawa clinical practice cohort[9] ($n = 1,644$) and the DiOGenes cohort[10] ($n = 789$), each with protein plasma levels quantified by both mass-spectrometry and aptamer-based assays.

We present a multivariate genome-wide pQTL analysis, using our variational Bayesian method LOCUS[11], which simultaneously accounts for all the genetic variants and proteomic outcomes, thereby leveraging the similarity across proteins controlled by shared regulatory mechanisms (Figure 1). We analyze the data from each proteomic technology in a two-stage design, using LOCUS for discovery with the Ottawa cohort and replicating our findings with the independent DiOGenes cohort. Our rich mass-spectrometry and SomaLogic proteomic data permit both cross- and intra-platform validation.

Pertinent interpretation of pQTL effects for complex diseases hinges on a careful examination of metabolic and clinical parameters from the same subjects or, at a minimum, from a population presenting similar clinical characteristics. We demonstrate the biomedical potential of several replicated pQTL hits, using comprehensive clinical data from the two pQTL obesity cohorts. Our results reveal novel protein biomarkers under genetic control, in the context of obesity co-morbidities; they are available from our online browser `https://locus-pqtl.epfl.ch`.

Results

Two-stage pQTL analyses. We used LOCUS for multivariate analyses of two proteomic datasets from the Ottawa cohort, comprising 133 and 1,096 proteins measured by mass spectrometry (MS) and the multiplexed aptamer-based technology SomaLogic[12], respectively. We analyzed about 275,000 single nucleotide polymorphisms (SNPs) capturing information from about 5M common variants for nearly 400 subjects, adjusting for age, gender and body mass index (BMI); see Methods and Figure 1d. At false discovery rate (FDR) 5%, LOCUS identified 18 pQTL associations from the MS analysis, corresponding to 14 unique proteins and 18 SNPs, and 118 pQTLs from the SomaLogic analysis, corresponding to 99 proteins and 111 SNPs; see Online Table S1.

We undertook to replicate all uncovered pQTLs in the independent DiOGenes cohort, using MS and SomaLogic data for $n = 400$ and $n = 548$ subjects, respectively (Figure 1d). The DiOGenes cohort recruited overweight/obese, non-diabetic subjects, while the Ottawa study was led in a specialized obesity practice where subjects had severe obesity, dyslipidemia and insulin resistance disorders (Online Table S4). We validated 15 of the 18 discovered MS pQTLs, and 98 of the 118 discovered SomaLogic pQTLs at FDR 5% (Online Table S5), yielding a replication rate of 83% in both cases. While the two platforms had inherent differences, 72 proteins were quantified by both, enabling cross-platform comparison. Eight of the MS pQTLs could be assessed with SomaLogic (i.e., had protein levels available), and 7 of them replicated at FDR 5%. Likewise, of the 20 SomaLogic associations having MS measurements, 14 were confirmed, demonstrating appreciable cross-technology replication.

We evaluated replication rates separately for *cis* and *trans* effects. With the MS data, all 15 *cis* Ottawa pQTLs replicated in DiOGenes, while the 3 *trans* pQTLs did not. With the SomaLogic data, 78 of 81 *cis* and 20 of 37 *trans* pQTLs could be validated. We reached overall replication rates of 97% for *cis* pQTLs and 50% for *trans* pQTLs; the *trans*-pQTL rate is in line with other pQTL studies[2,4,13].

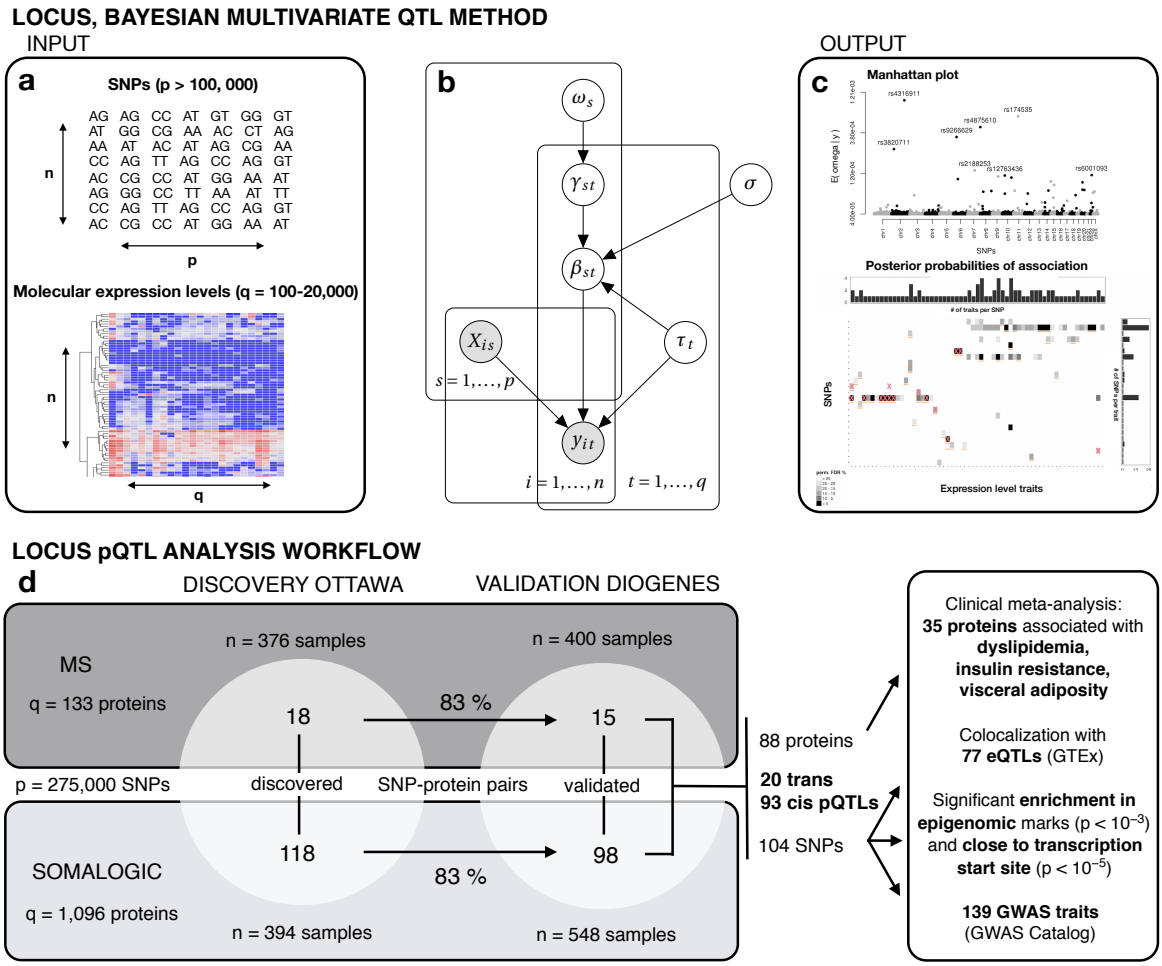

Figure 1: LOCUS model overview and study workflow. (a) Inputs to LOCUS are an $n \times p$ design matrix $\boldsymbol{X}$ of p SNPs, and an $n \times q$ outcome matrix $\boldsymbol{y}$ of q molecular traits, e.g., gene, protein, lipid or methylation levels, for n individuals. The model is multivariate; it accounts for all the SNPs and molecular traits jointly. (b) Graphical model representation of LOCUS. The effect size between a SNP s and a trait t is modelled by β_{st}, and γ_{st} is a latent variable taking value unity if they are associated, and zero otherwise. The parameter ω_s controls the pleiotropic level of each SNP, i.e., the number of traits with which it is associated. The parameter σ represents the typical size of effects, and the parameter τ_t is a precision parameter that relates to the residual variability of each trait t. LOCUS enforces sparsity on the QTL effects, so it identifies just one or few markers per relevant locus, even in regions of high linkage disequilibrium (LD). Univariate screening approaches do not exploit association patterns common to multiple outcomes or markers; they analyze the outcomes one by one, and do not account for LD structures, thereby highlighting redundant signals at loci with strong LD structures (see, e.g., Figure 2). (c) Outputs of LOCUS are posterior probabilities of associations, $\mathrm{pr}(\gamma_{st} = 1 \mid \boldsymbol{y})$, for each SNP and each trait ($p \times q$ panel), and posterior means for the pleiotropy propensity of each SNP, $\mathrm{E}(\omega_s \mid \boldsymbol{y})$ (Manhattan plot). (d) Workflow of the pQTL study. The mass-spectrometry and SomaLogic pQTL data are analyzed in parallel. LOCUS is applied on the Ottawa data for discovery, and 83% of the 18 and 118 pQTL associations discovered with the mass-spectrometry (MS) and SomaLogic data replicate in the independent study DiOGenes. The relevance of the validated pQTLs in the obese population is assessed via analyses of clinical parameters from the Ottawa and DiOGenes cohorts. Further support is obtained by evaluating colocalization with eQTLs, epigenomic marks and GWAS risk loci.

Finally, we found that 73 of our validated pQTLs overlap with pQTLs previously identified in the general population (using proxy search $r^2 > 0.8$, and reporting associations at $p < 1 \times 10^{-5}$; Online Table S6). The remaining 40 pQTLs are, to our knowledge, new.

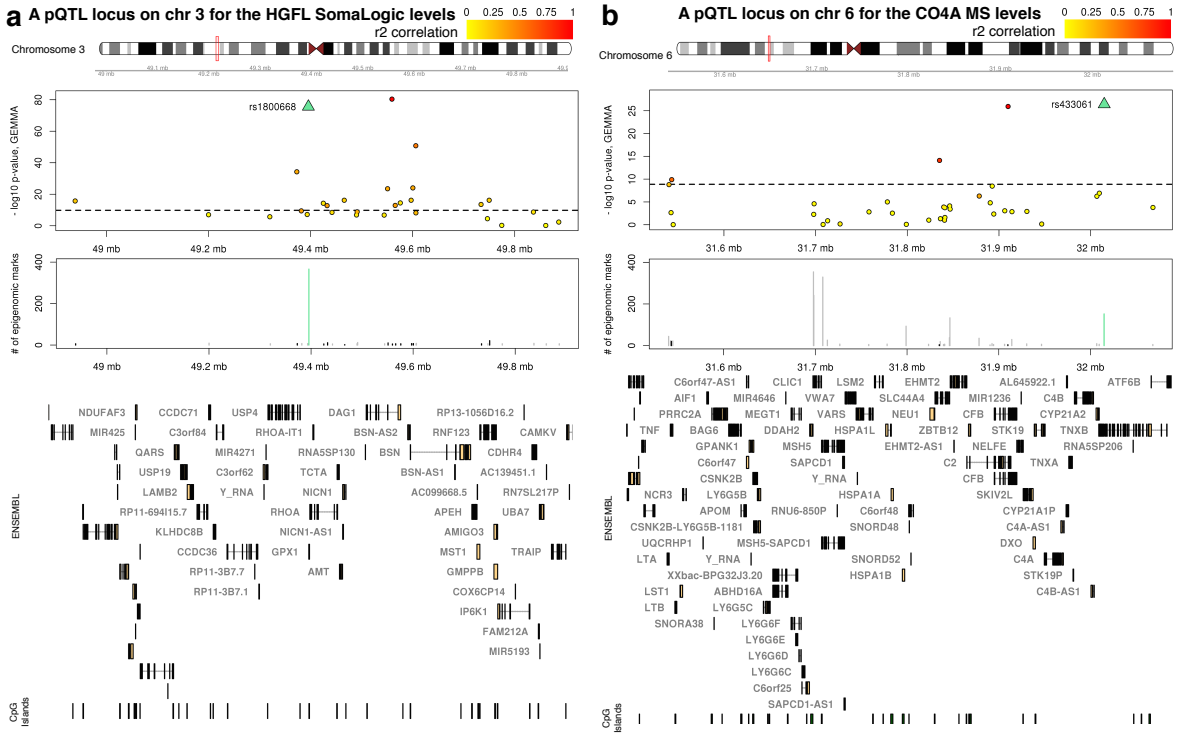

Figure 2: Regional association plots for two loci, identified by the (a) SomaLogic and (b) MS pQTL analyses. In each case, the top panel displays the nominal $-\log_{10} p$-values obtained when re-analyzing the region with GEMMA [14]; the dashed horizontal line corresponds to the Bonferroni level with $\alpha = 0.05$. The top SNP identified by LOCUS is marked with a green triangle, and its correlation in r^2 with the surrounding SNPs is indicated by the yellow to red colors. The second-row panel shows the cumulated numbers of annotation marks for each SNP. The green bars correspond to LOCUS top SNPs, the black bars, to the SNPs found significant with GEMMA (Bonferroni adjustment $\alpha = 0.05$). The bottom panel shows the transcript and CpG island positions.

Performance of LOCUS multivariate analyses over standard approaches. The high replication rates achieved using LOCUS are largely attributable to its flexible hierarchical sparse regression model which exploits shared association patterns across all genetic variants and proteomic levels (Figures 1a–c). Indeed, the increased statistical power of LOCUS over standard univariate approaches has been extensively assessed [11], and additional evidence using genetic variants from the Ottawa cohort and synthetic outcomes emulating the proteomic data is available in Appendix A.

The univariate approach GEMMA [14] would have missed 18 of the 113 validated hits (16%) using a conservative yet standard genome-wide Bonferroni correction of $\alpha = 0.05$ ($p < \alpha/275,297/133$ for MS and $p < \alpha/275,297/1,096$ for SomaLogic), and 14 hits (12%) with a permissive Bonferroni correction $\alpha = 0.25$; see Online Table S5. Moreover, the sparse selection of LOCUS highlights candidate variants with promising functional evidence, even in regions with strong linkage disequilibrium (LD) structures; we next provide two illustrations.

The first example concerns a locus associated with the SomaLogic levels of the HGFL (hepatocyte growth factor-like) protein, encoded by the macrophage-stimulating *MST1* gene (Figure 2a). At FDR 5%, two variants (rs1800668 and rs56116382) were associated with the HGFL levels. GEMMA highlighted a large LD block, with 15 SNPs significant at Bonferroni level $\alpha = 0.05$. The pQTLs selected by LOCUS corresponded to the second and third most significant hits of GEMMA. One of these two, rs1800668, is located 326 Kb upstream of the *MST1* gene, within a gene-dense region (> 40 genes). It had the highest overlap in epigenomic annotation marks (336 out of 450 marks,

enrichment $p = 2.06 \times 10^{-3}$) and is a known eQTL for many genes (25 including *MST1*) in several tissues (Online Table S2). Interestingly, public pQTL studies reported associations of this SNP with 23 distinct proteins (Online Table S3), but not with HGFL. The top hit identified by the univariate analysis, rs13062429, had no significant epigenomic enrichment (5 out of 450 marks); it was not picked by LOCUS.

The second example concerns the MHC region, with evidence of *cis* regulation of the CO4A MS protein levels (Figure 2b). Here, the LD structure is slightly simpler, and GEMMA identified four SNPs after Bonferroni adjustment with $\alpha = 0.05$. At FDR 5%, LOCUS selected two variants, one of which, rs433061, was the top hit from univariate analyses ($p = 3.94 \times 10^{-27}$). This variant colocalized with 156 epigenomic marks (enrichment $p = 0.0184$) and was 442 base pairs away from a transcription start site (compared to random SNPs $p = 0.0253$). This SNP is a known *cis* eQTL for the *C4A* gene in many tissues, including liver, arteries and adipose tissue, as well as for > 70 other transcripts (Online Table S2). It has already been described as a pQTL for CO4A and 27 other proteins (Online Table S3), suggesting a pleiotropic role.

These two examples indicate that the parsimonious selection of LOCUS can uncover SNPs that colocalize with many epigenomic marks and eQTLs, which supports possible regulatory roles. The next two sections generalize these observations for all validated pQTLs identified by LOCUS.

Colocalization with eQTLs and evidence for regulatory impact. We assessed the overlap of the 113 validated pQTLs with known eQTLs. Seventy-seven of the 104 SNPs involved in our pQTL associations had one or more eQTL associations in at least one tissue. These SNPs have been implicated in 83 eQTL associations, representing a significant enrichment ($p < 2.2 \times 10^{-16}$). Forty-nine of these 77 SNPs were eQTL variants for the gene coding for the protein with which they were associated in our datasets. Our pQTLs were also enriched in epigenome annotation marks ($p = 9.20 \times 10^{-4}$) and significantly closer to transcription start sites compared to randomly chosen SNP sets ($p = 9.99 \times 10^{-6}$). These observations suggest potential functional consequences of our pQTL hits.

Colocalization with GWAS risk loci. A total of 217 previously reported genome-wide associations overlapped our validated pQTL loci, corresponding to 139 unique traits mapping to 68 distinct regions (based on LD $r^2 > 0.8$). Nineteen *sentinel* SNPs, i.e., SNPs specifically identified by LOCUS pQTL analyses, were directly involved in these associations (Online Table S8) representing a significant enrichment ($p < 2.2 \times 10^{-16}$). Some of these results generate useful hypotheses to be explored in future research.

For instance, our aforementioned HGFL pQTL, rs1800668, is in strong LD ($r^2 > 0.95$) with rs9858542 and rs3197999, which are known to associate with Crohn's disease [15,16]. While gene causality remains to be demonstrated, our pQTL finding may be of clinical relevance given the prevalence of Crohn's disease in overweight and obese subjects [17]; the region would merit follow-up in inflammatory bowel disease cohorts.

Another example concerns an association between rs3865444 and the Siglec-3 protein, whose coding gene, *CD33*, has been reported as a risk factor for Alzheimer's disease [18]. As subjects obese in midlife are more at risk of developing late-life Alzheimer's [19], this pQTL may help to better understand the genetic bases of Alzheimer's disease and dementia; its potential as a prognosis biomarker should be studied in Alzheimer's cohorts, ideally using weight records.

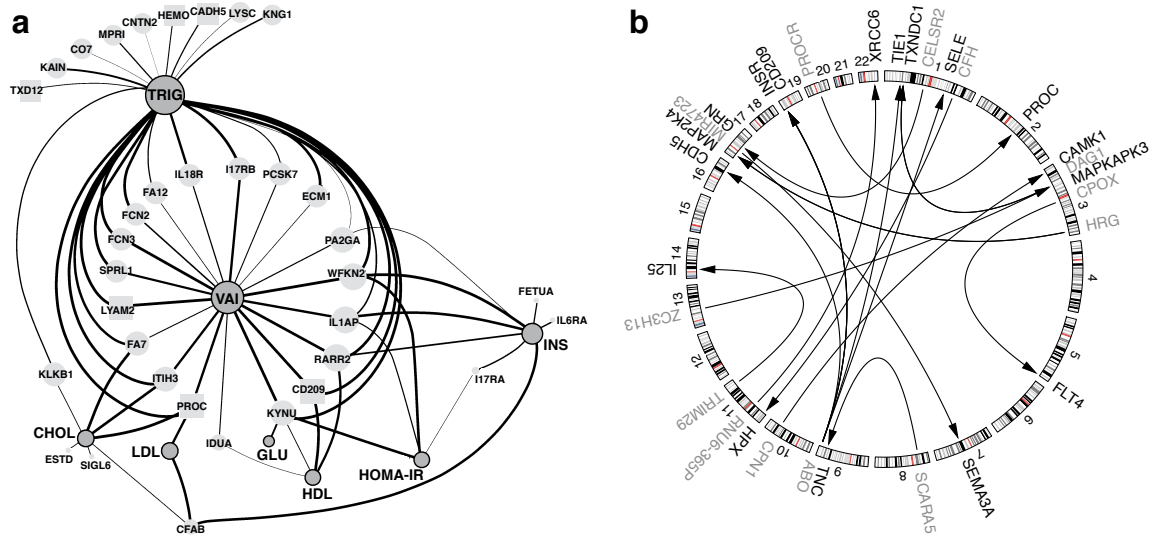

Figure 3: Associations of proteins under genetic control with clinical parameters, and *trans*-pQTLs associations. (a) Network displaying the associations (FDR < 5%) between protein levels and clinical variables obtained by meta-analysis, adjusting for age, gender and BMI. Nodes for clinical parameters are in dark grey with black borders (fasting glucose, HDL, HOMA-IR, insulin resistance, LDL, total cholesterol, triglycerides, visceral adiposity index); proteins are in light grey, and type of genetic control, *cis* or *trans*, is depicted with circular or square nodes, respectively. The edge thickness is proportional to the significance of association, and the node size is proportional to its connectivity. (b) Circular plot showing the *trans*-pQTL associations uncovered by LOCUS (FDR < 5% for discovery and validation). Each arrow starts from the pQTL SNP with label indicating its closest gene (grey) and points to the gene (black) coding for the controlled protein.

Proteins as endophenotypes to study the genetics of obesity. Annotation queries suggested that most pQTLs had implications in inflammation, insulin resistance, lipid metabolism or cardiovascular diseases (CVD). We performed a more systematic evaluation of their clinical relevance in a meta-analysis of the DiOGenes and Ottawa clinical and proteomic data, and found that 35 of the 88 proteins under genetic control had associations with dyslipidemia, insulin resistance or visceral fat-related measurements at FDR 5%, with consistent directions of effects in the two cohorts (Online Table S9). These associations should be attributable metabolic factors independently of overall adiposity, as we controlled for BMI as a potential confounder. Remarkably, we found that the 88 genetically-driven proteins are significantly more associated with the clinical variables than randomly chosen protein sets ($p = 0.014$, see Methods); this enrichment suggests that the primary pQTL analyses can help uncover potential proteomic biomarkers for the metabolic syndrome.

Figure 3a displays the associations as a network. The triglyceride measurements and visceral adiposity index (VAI) had the highest degree of connectivity and were connected with measures of insulin resistance and other lipid traits via proteins such as FA7, IL1AP, KYNU, PROC, RARR2 and WFKN2. CFAB, FETUA, PA2GA had lower connectivity, yet are relevant in the context of obesity[20–22]. Finally, several *trans*-regulated proteins were implicated in clinical associations: CADH5, CD209 and LYAM2, all controlled by the pleiotropic *ABO* locus; HEMO (Hemopexin), a liver glycoprotein controlled by the *CFH* locus, itself coding for another liver glycoprotein; PROC controlled by its own receptor *PROCR*; and TXD12 (thioredoxin domain containing 12), controlled by the *DAG1/BSN* locus (see also Figure 3b).

The pQTL associations involving proteins with clinical associations at FDR 5% are listed in Table 1. Subsequent sections discuss the possible functional and biomedical relevance of a selection of pQTL

Protein	Protein name	Clin.	SNP	Chr	Position	LOCUS PPI	pQTL valid. p-value
CADH5	**Cadherin-5**	**L**	**rs8176741**	**9**	**136131461**	**1.00**	$\mathbf{4.68 \times 10^{-30}}$
CD209	**DC-SIGN**	**L/V**	**rs8176741**	**9**	**136131461**	**1.00**	$\mathbf{7.74 \times 10^{-10}}$
			rs2519093	**9**	**136141870**	**1.00**	$\mathbf{6.24 \times 10^{-26}}$
CFAB	Factor B	G/L	rs150132450	6	31906334	0.85	9.61×10^{-4}
			rs641153	6	31914180	1.00	3.92×10^{-12}
CNTN2	CNTN2	L	rs11240396	1	205205081	1.00	6.82×10^{-14}
CO7	C7	L	rs71623870	5	40966676	0.83	4.03×10^{-4}
ECM1	ECM1	L/V	rs34964511	1	150298015	1.00	3.77×10^{-6}
			rs71578487	1	150340059	1.00	1.07×10^{-11}
			rs72696900	1	150425256	0.82	1.5×10^{-6}
			rs11802612	1	150427279	1.00	3.7×10^{-6}
			rs35094010	1	150449557	1.00	3.76×10^{-6}
ESTD	Esterase D	L	rs73193065	13	47383681	0.90	2.31×10^{-15}
FA12	Coagulation factor XII	L/V	rs55785724	5	176817583	1.00	1.34×10^{-5}
FA7	Coagulation Factor VII	L/V	rs3093233	13	113758130	1.00	3.11×10^{-88}
FCN2	FCN2	L/V	rs3811140	9	137772111	1.00	9.66×10^{-14}
FCN3	Ficolin-3	L/V	rs10902652	1	27558522	1.00	1.62×10^{-3}
FETUA	a2-HS-Glycoprotein	G	rs2593813	3	186332571	1.00	2.47×10^{-10}
			rs2593813	3	186332571	1.00	4.51×10^{-8}
HEMO	**Hemopexin**	**L**	**rs10801560**	**1**	**196714600**	**1.00**	$\mathbf{2.36 \times 10^{-26}}$
I17RA	IL-17 sR	G	rs738035	22	17594886	1.00	1.48×10^{-20}
I17RB	IL-17B R	L/V	rs35518479	3	53873814	0.76	9.98×10^{-6}
IDUA	IDUA	L/V	rs10017289	4	943534	1.00	1.22×10^{-11}
IL18R	IL-18 Ra	L/V	rs3836108	2	103037742	1.00	5.22×10^{-26}
IL1AP	IL-1 R AcP	G/L/V	rs724608	3	190348810	1.00	8.7×10^{-114}
IL6RA	IL-6 sRa	G	rs4845372	1	154415396	1.00	1.72×10^{-81}
ITIH3	Inter-alpha-trypsin inhibitor heavy chain H3	L/V	rs736408	3	52835354	0.97	1.46×10^{-6}
KAIN	Kallistatin	L	rs5511	14	95033595	1.00	9.9×10^{-24}
KLKB1	Prekallikrein	L	rs80177406	4	187166024	0.99	3.54×10^{-6}
KNG1	Kininogen HMW	L	rs1621816	3	186439173	1.00	1.44×10^{-13}
KYNU	KYNU	G/L/V	rs6741488	2	143793701	1.00	3.22×10^{-20}
LYAM2	**sE-Selectin**	**L/V**	**rs2519093**	**9**	**136141870**	**1.00**	$\mathbf{6.81 \times 10^{-62}}$
LYSC	Lysozyme	L	rs71094714	12	69790495	1.00	8.41×10^{-19}
MPRI	IGF-II receptor	L	rs3777411	6	160476945	1.00	4.95×10^{-11}
PA2GA	NPS-PLA2	G/L/V	rs6672057	1	20293791	1.00	3.86×10^{-15}
PCSK7	PCSK7	L/V	rs11216284	11	117003060	1.00	8.17×10^{-31}
PROC	**Protein C**	**L/V**	**rs141091409**	**20**	**33739915**	**0.43**	$\mathbf{1.66 \times 10^{-18}}$
RARR2	TIG2	G/L/V	rs1047586	7	150035459	0.96	2.39×10^{-11}
SIGL6	Siglec-6	L	rs77561179	19	52029477	1.00	3.39×10^{-14}
SPRL1	SPARCL1	L/V	rs7681694	4	88462729	0.99	5.70×10^{-14}
TXD12	**TXD12**	**L**	**rs13062429**	**3**	**49559485**	**1.00**	$\mathbf{2.26 \times 10^{-5}}$
			rs34519883	**3**	**49575831**	**1.00**	$\mathbf{5.39 \times 10^{-33}}$
WFKN2	WFKN2	G/L/V	rs9303566	17	48922281	1.00	3.38×10^{-11}

Table 1: Proteins associated with clinical parameters (Figure 3a) and controlled by pQTL variants. All associations were detected at FDR $< 5\%$. Associations with glycemic traits (fasting glucose, insulin, HOMA-IR) are indicated by G, with total lipid traits (HDL, LDL, triglycerides, total cholesterol), by L, and with visceral fat (visceral adiposity index), by V. *Trans*-pQTL associations are in bold.

associations based on their connection with clinical variables, as summarized by Figure 3a. Forest plots for this selection are given in Figure 4 to help visualize the effect directions. Unless otherwise specified, all associations described have meta-analysis FDR corrected p-value below 5%, and we provide their nominal p-values in parentheses.

CFAB and RARR2, mediators of adipogenesis are under genetic control. CFAB (complement factor B) and RARR2 (Retinoic acid receptor responder protein 2) levels associate with distinct clinical parameters (Figures 3a and 4), yet both play a role in adipogenesis and hence are particularly interesting in the context of obesity and related co-morbidities.

The CFAB protein controls the maturation of adipocytes[20]. Both the MS and SomaLogic measurements were positively associated with BMI (MS: $p = 2.08 \times 10^{-8}$, SomaLogic: $p = 2.23 \times 10^{-13}$) and with fasting insulin (adjusting for BMI; MS: $p = 4.45 \times 10^{-5}$, SomaLogic: $p = 3.44 \times 10^{-4}$).

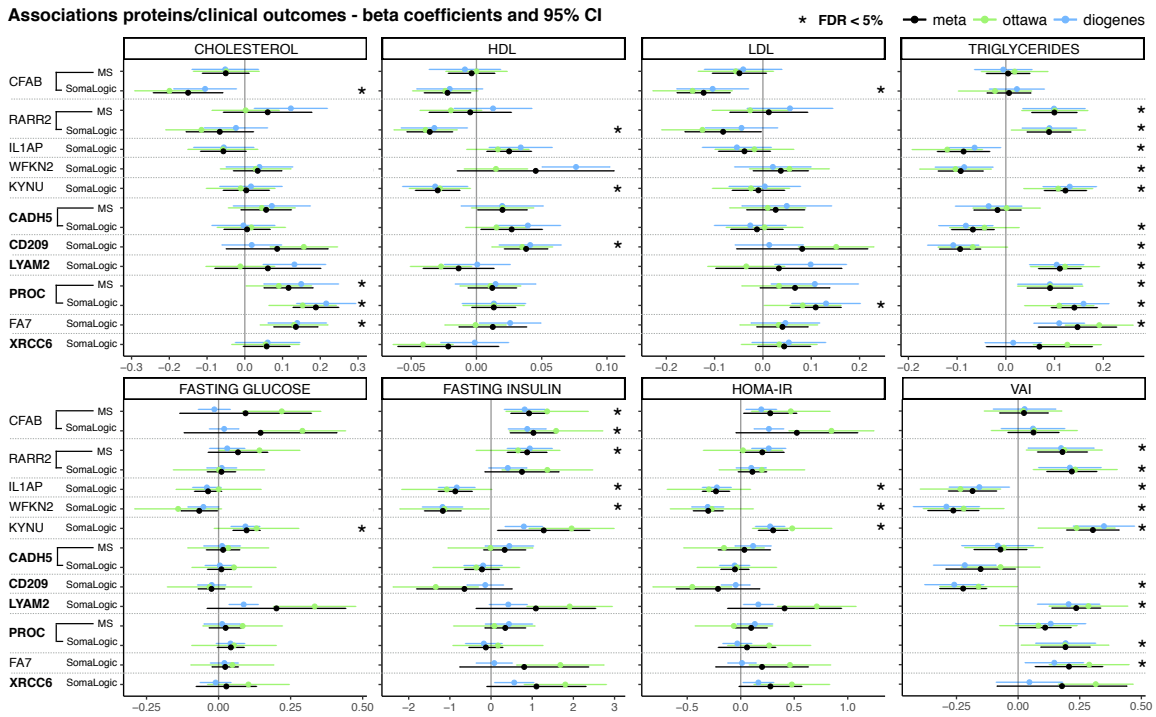

Figure 4: Forest plots for associations between a selection of proteins under genetic control and clinical parameters, adjusting for age, gender and BMI (Methods). All endpoints are measured in both the Ottawa and DiOGenes cohorts; they correspond to total lipid levels (first row: total cholesterol, HDL, LDL, triglycerides), glucose/insulin resistance (second row: fasting glucose, fasting insulin, HOMA-IR) and VAI. In each case, regression coefficients with 95% confidence intervals are shown for the Ottawa and DiOGenes analyses, and for the meta-analysis. The stars indicate associations with meta-analysis FDR $< 5\%$ (correction applied across all proteins under genetic control, not only those displayed; see Figure 3a). The order of appearance of the proteins follows that in the text. For proteins with measurements in the MS and SomaLogic platforms, association results are displayed for both; *trans*-regulated proteins are in bold.

The CFAB SomaLogic levels were negatively associated with cholesterol ($p = 1.43 \times 10^{-3}$), LDL ($p = 1.30 \times 10^{-5}$), and with HDL at higher FDR (nominal $p = 1.47 \times 10^{-2}$, corrected $p = 0.11$). This is consistent with previous gene expression findings[23].

The MS and SomaLogic analyses independently highlighted the same *cis*-acting locus as putative regulator of the CFAB protein. In particular, the sentinel pQTL SNP detected in the SomaLogic analysis, rs641153, is a missense variant located in the MHC region, 180 base pairs away from a transcription binding site (significantly closer than other SNPs, $p = 1.16 \times 10^{-2}$). Further investigation using JASPAR and SNP2TFBS indicated that rs641153 may affect the binding sites of four transcription factors (EBF1, TFAP2A, TFAP2C and HNFA). In GTEx[24], rs641153 is described as an eQTL for the *NELFE* and *SKIV2L* genes in multiple tissues, but not for the *CFB* gene.

RARR2 (Chemerin protein) is encoded by an essential adipogenesis gene, *RARRES2*, and regulates glucose and lipid metabolism by altering the expression of adipocyte genes[25]. We found significant associations with triglycerides, fasting insulin and HDL (Figure 4, Online Table S9), which is consistent with previously described pleiotropic associations of *RARRES2* variants with circulating RARR2, triglyceride levels and diverse measurements related to inflammation[26]. The MS and SomaLogic RARR2 levels were strongly associated with BMI and visceral fat, even when controlling for BMI (Figure 4; Online Table S9); this clarifies the as-yet unclear relation between *RARRES2* and visceral fat mass in obese subjects[25].

Our pQTL analyses indicated a *cis* association between a missense variant, rs1047586, and RARR2.

This variant was described as an eQTL for multiple genes and as associated with epigenomic marks (DNA methylation and histone modifications, including H3K27ac and H3K4me1 enhancers)[27].

Our analyses illustrate the relevance of CFAB and RARR2 for better understanding metabolic complications in obese subjects, and provide evidence in favour of their genetic control; both pQTLs colocalize with several epigenomic marks.

The importance of IL1AP for metabolic syndrome. The IL-1 pathway plays a critical role in the immune-response associated with obesity and type 2 diabetes[28]; other IL-1 related cytokines, such as IL-1ra, are also well documented in the context of type 1 and type 2 diabetes[29]. The IL1AP (IL-1 receptor accessory) protein is a co-receptor of the IL-1 receptor, and its soluble levels were found reduced in obese subjects[30]. Our analyses found an association between rs724608 and IL1AP, corroborating previously identified associations with SNPs in LD ($r^2 = 0.93$)[30].

We found associations between IL1AP expression and measures of fasting insulin levels ($p = 3.88 \times 10^{-5}$), HOMA-IR ($p = 3.89 \times 10^{-4}$), triglycerides ($p = 1.61 \times 10^{-3}$) and visceral fat ($p = 2.1 \times 10^{-4}$) (Figures 3a and 4). Moreover, worsened metabolic syndrome scores[31] were associated with lower protein levels ($p = 1.20 \times 10^{-3}$ in Ottawa and $p = 2.50 \times 10^{-4}$ in DiOGenes).

WFKN2, a TGFβ-activity protein with protective effect against metabolic disorders. The role of the WFKN2 protein and of its coding gene, *WFIKKN2*, in regulating TGFβ activity has been extensively studied in muscle and skeletal muscle[32], but, to our knowledge, not in other tissues. We describe it for the first time in the context of obesity and metabolic disorders. We found that higher protein levels were associated with lower levels of fasting insulin, triglycerides, HOMA-IR and visceral fat (Figure 4), suggesting a protective role against metabolic dysregulation.

Our analyses suggested that the WFKN2 levels are controlled by rs9303566, which is consistent with other p- and eQTL studies (Online Tables S6–S7). This SNP was found to be associated with DNA methylation and histone marks[27,33], and is located within 100 base pairs of a transcription factor binding site, with numerous factors such as MYBL2, NFIC, EP300 and MXI1. It is in strong LD with other SNPs with potential regulatory impact; for instance, it is located 9Kb upstream to rs8072476 ($r^2 = 0.97$), which overlaps another cluster of transcription factor binding sites (FOXA1, ESR1, USF1 & 2, TFAP2A & 2C).

Inflammation mediated proteins and their role in insulin resistance. We found a *cis* effect of rs6741488 on KYNU (Kynureninase) plasmatic levels. KYNU is an enzyme involved in the biosynthesis of nicotinamide adenine dinucleotide (NAD) cofactors from tryptophan. This protein and its pathway have been found to be particularly relevant for obesity and associated metabolic disorders. KYNU was up-regulated by pro-inflammatory cytokines in human primary adipocytes, and more so in the omental adipose tissue of obese compared to lean control subjects[34]. Other studies indicated that the kynurenine pathway (KP) may act as an inflammatory sensor, and that increased levels of its catabolites may be linked with several cardiometabolic defects, including CVD, diabetes and obesity[35]. In our cohorts, higher KYNU levels were associated with decreased HDL levels ($p = 6.66 \times 10^{-4}$), and increased triglycerides levels ($p = 3.43 \times 10^{-8}$), visceral fat ($p = 2.51 \times 10^{-8}$) and insulin resistance (marginally, nominal $p = 2.53 \times 10^{-2}$, corrected $p = 0.17$), see Figure 4; as expected, higher protein levels were associated with a worsened metabolic syndrome score (Ottawa $p = 8.23 \times 10^{-5}$; DiOGenes $p = 3.62 \times 10^{-6}$).

Recent work suggested a causal link between obesity and cancer, mediated by KP activation through inflammatory mechanisms[36]. Interestingly, our analyses highlighted two soluble interleukin receptor antagonist proteins, namely IL6RA and I17RA, that were both under genetic control and associated with insulin resistance (Figure 3a). We did not find significant correlation between the I17RA and KYNU protein levels, but we did observe a significant negative correlation between IL6RA and KYNU (Ottawa $p = 0.01$ and DiOGenes $p = 4 \times 10^{-3}$). We found a link between the plasma levels of KYNU and pro-inflammatory molecules, namely, IL6, IFNG and TNFα. In the Ottawa cohort, where subjects displayed high low-grade inflammation status, KYNU was positively associated with IL6 and IFNG at FDR 5%, while in DiOGenes, we found a positive association with IFNG only (Appendix C). Finally, metabolic dysfunctions mediated via KP may relate to another inflammatory pathology, namely, psoriasis[37], a skin disease aggravated by obesity and improved by weight loss[38,39].

Our results thus highlighted pQTLs with probable roles in inflammation and subsequent metabolic dysfunctions, reinforcing previous discussion[35,40] of the potential of KP therapeutic inhibitors against CVD and metabolic disorders.

Below, we focus on *trans*-regulatory mechanisms that may be relevant to metabolic disorders in the obese population. Indeed, owing to its potential for the detection of weak effects, LOCUS identified several *trans* and pleiotropic effects that suggest novel metabolic pathways (Figure 3b).

Pleiotropic effects from the ABO locus onto CADH5, CD209, INSR, LYAM2 and TIE1. ABO is a well-known pleiotropic locus associated with coronary artery diseases, type 2 diabetes, liver enzyme levels (alkaline phosphatase) and lipid levels[2–4]. Our analyses highlighted two independent sentinel SNPs in the ABO region: rs2519093 and rs8176741 ($r^2 = 0.03$). The former SNP is *trans*-acting on E-selectin (protein LYAM2 encoded by SELE), the Insulin Receptor and the CD209 antigen. The latter SNP is *trans*-acting on the Tyrosine-protein kinase receptor (Tie-1), Cadherin-5 and CD209. Both SNPs were reported as *cis*-acting eQTL variants for ABO, OBP2B and SURF1, and further queries in public databases indicated that rs8176741 may affect the binding sites for three transcription factors (Myc, MYC-MAX and Arnt), suggesting a complex gene regulation circuitry.

Our clinical analyses indicated associations of CD209 and LYAM2 with triglycerides and visceral fat, and CADH5 with triglycerides only (Figure 4). All were associated with triglyceride levels, and LYAM2 and CD209 were associated with visceral fat (Figures 3a and 4). Moreover, CD209 may have an important role in controlling lipid levels as it was associated with HDL ($p = 7.58 \times 10^{-6}$): higher CD209 levels had higher HDL, lower triglyceride levels, and, consistently with these effects, lower visceral fat index. Dyslipidemia is a risk factor for Non-Alcoholic Fatty Liver Disease (NAFLD)[41], and the CD209 gene levels have been reported as differentially expressed in patients with Non-Alcoholic Steatohepatitis (NASH) compared to healthy subjects[42]. The role of circulating protein levels of CD209 could be further studied in NASH/NAFLD cohorts.

Finally, the LYAM2 levels were associated with all the glycemic variables in the Ottawa cohort (fasting glucose: $p = 6.43 \times 10^{-6}$, fasting insulin: $p = 3.54 \times 10^{-4}$, HOMA-IR: $p = 1.8 \times 10^{-4}$), but only with fasting glucose in the DiOGenes cohort ($p = 8.91 \times 10^{-4}$), although we observed a suggestive association with HOMA-IR (nominal $p = 0.02$, corrected $p = 0.15$). Since the Ottawa subjects are more insulin-resistant than the DiOGenes subjects (average HOMA-IR with standard deviation: 4.97(3.88) versus 3.00(1.71), $p = 2.52 \times 10^{-18}$; Online Table S4), LYAM2 might represent a marker of insulin-resistance severity. Consistent with this hypothesis, the plasma levels of LYAM2 are employed as a biomarkers of endothelial dysfunction and risk of type 2 diabetes[43].

Complement/coagulation: a *trans*-acting insertion linking PROC and its receptor. PROC (Protein C, coding gene PROC on chromosome 2) and its paralog protein FA7 (Coagulation Factor 7, coding gene F7 on chromosome 13) regulate the complement and the coagulation systems. Both systems promote inflammation[44] and contribute to metabolic dysfunction in the adipose tissue and liver[45]. Our analyses suggested novel pQTLs for these proteins (Online Table S5): FA7 was associated with rs3093233, which is a known eQTL of *F7* and *F10* in several tissues (Online Table S7). PROC may be controlled by *trans*-regulatory mechanisms, initiated in its receptor gene, *PROCR*, on chromosome 20; it was indeed associated with an insertion, rs141091409, located 20Kb upstream of *PROCR*, an association observed with both our proteomic platforms. Previous studies found associations between CVD and variants located in the *PROC* or *PROCR* genes[46,47]. Interestingly, our hit, rs141091409, was in strong LD ($r^2 > 0.95$) with the missense variant rs867186, previously identified as associated with coronary heart disease[47].

Our clinical analyses support the relation of PROC and FA7 levels with lipid traits: both were positively associated with cholesterol, triglycerides and visceral fat (Figures 3a and 4). PROC levels were quantified by both platforms, and displayed consistent results. The SomaLogic measurements of PROC were positively associated with LDL ($p = 5.39 \times 10^{-5}$). The role of these proteins for CVD and NAFLD diseases in the overweight/obese population would merit further investigation.

XRCC6, a DNA repair protein as putative biomarker for metabolic disorders. We identified a novel *trans* pQTL for XRCC6 (X-Ray Repair Complementing Defective Repair In Chinese Hamster Cells; also known as Ku70). The *XRCC6* gene activates DNA-dependent protein kinases (DNA-PK) to repair double-stranded DNA breaks by nonhomologous end joining. DNA-PKs have been linked to lipogenesis in response to feeding and insulin signaling[48]. DNA-PK inhibitors may reduce the risk of obesity and type 2 diabetes by activating multiple AMPK targets[49]. A recent review discussed the role of DNA-PK in energy metabolism, and in particular, the conversion of carbohydrates into fatty acids in the liver, in response to insulin[50]. It described increased DNA-PK activity with age, and links with mitochondrial loss in skeletal muscle and weight gain. Finally, *XRCC6* functions have been reported as associated with regulation of beta-cell proliferation, islet expansion, increased insulin levels and decreased glucose levels[49,51].

We observed significant associations between the XRCC6 protein levels and several clinical variables in the Ottawa cohort (FDR $< 5\%$). Higher expression was associated with decreased HDL ($p = 5.83 \times 10^{-4}$), as well as with higher triglycerides ($p = 4.39 \times 10^{-4}$), insulin levels ($p = 4.50 \times 10^{-4}$) and visceral adiposity ($p = 5.94 \times 10^{-5}$; Figure 4). We only found marginal associations using the DiOGenes data for insulin levels (nominal $p = 0.02$, corrected $p = 0.14$) and HOMA-IR (nominal $p = 0.02$, corrected $p = 0.16$). The directionality of these effects was consistent in both cohorts. As the Ottawa subjects were more severely obese, the effects might be larger for subjects with pronounced metabolic syndrome, but this would require confirmation.

Our pQTL sentinel SNP, rs4756623, is intronic and located within the *LRRC4C* gene, a binding partner for Netrin G1 and member of the axon guidance[52]. To our knowledge, *LRRC4C* has not been previously described in the context of obesity, insulin resistance or type 2 diabetes. However, its partner Netrin G1 is known to promote adipose tissue macrophage retention, inflammation and insulin resistance in obese mice[53]. The underlying regulatory mechanisms between rs4756623 and the *XRCC6* locus should be clarified, and functional studies will be required to understand their physiological impact.

Discussion

Despite important technological advances, large-scale pQTL studies remain infrequent, owing to their high costs[2–4,13,54]. To date, all but our recent study[3] have focused on the general population and have assessed links with diseases by relying on information from different studies.

Here, we described the first integrative pQTL study that relates the associations discovered to metabolic disorders, such as insulin resistance and dyslipidemia, in the obese population considered. Our Bayesian method LOCUS confirmed many pQTLs highlighted in previous studies, despite our sample sizes 2.5 to 18 times smaller, and revealed a number of novel pQTLs, with sound evidence for functional relevance and implications for the development of the metabolic syndrome. Our two-stage approach achieved very high replication rates ($> 80\%$), and validated new findings, which standard univariate approaches would have missed (e.g., the aforementioned *cis* and *trans* associations with CO7, INSR and XRCC6). This corroborates numerical experiments demonstrating the increased statistical power of LOCUS over existing approaches[11]. Owing to its joint modelling of all proteins and genetic variants, LOCUS both accounts for linkage disequilibrium and exploits the shared regulatory architecture across molecular entities; this drastically reduces the multiplicity burden and enhances the detection of weak effects. Finally, our analyses indicated that proteins under genetic control are enriched in associations with clinical parameters pertaining to obesity co-morbidities, which further supports a genetic basis of these parameters and emphasizes the advantages of pQTL studies for elucidating the underlying functional mechanisms. Our complete pQTL and clinical association results offer opportunities to generate further hypotheses about therapeutic options; they are accessible from the searchable online database `https://locus-pqtl.epfl.ch`.

The applicability of LOCUS goes beyond pQTL studies, as it is tailored to any genomic, proteomic, lipidomic or methylation QTL analyses and can be used for genome-wide association with several clinical endpoints. Its multivariate framework is made efficient at a genome-wide scale thanks to a scalable batch-wise variational algorithm and an effective C++/R implementation. Our MS and SomaLogic analyses completed in a few hours for 275K tag SNPs representing information from about 5M common markers. Moreover, our method scales linearly in terms of memory and CPU usage; for instance, analyses of 2M SNPs and 1000 proteins run in less than 40 hours and with a memory footprint smaller than 256Gb (see profiling in Appendix B). To our knowledge, no other fully multivariate method is applicable to large molecular QTL studies without drastic preliminary dimension reduction; LOCUS therefore opens new perspectives for uncovering weak and complex effects.

Methods

Ethics. The study was approved by the local human research ethic committees. Participants provided informed written consent, and all procedures were conducted in accordance with the Declaration of Helsinki.

Study samples. The *Ottawa* study was a medically supervised program set up by the Weight Management Clinic of Ottawa[9]. Subjects under medication known to affect weight, glucose homeostasis or thyroid indices were excluded from all analyses, and subjects who were not under fasting conditions at plasma sample collection were excluded from the proteomic analyses.

The *DiOGenes* study was a multi-center pan-European program[10]. Eight partner states participated to the study: Bulgaria, the Czech Republic, Denmark, Germany, Greece, the Netherlands, Spain and the United Kingdom. Participants were overweight/obese (BMI between 27 and 45 kg/m^2), non-diabetic and otherwise healthy.

The main clinical characteristics of both cohorts are given in Online Table S4.

Proteomic data. Plasma protein expression data were obtained using two types of technologies: mass-spectrometry (MS) and a multiplexed aptamer-based assay developed by SomaLogic[12]. Samples were randomized, ensuring that the plate numbers were not associated with age, gender, ethnicity, weight-related measures, glycemic indices, measures of chemical biochemistry, and, for the DiOGenes samples, collection centers.

The MS proteomic quantification used plasma samples spiked with protein standard lactoglobulin (LACB). Samples were immuno-depleted, reduced, digested, isobarically 6-plex labeled and purified. They were analyzed in duplicates on two separate but identical systems using linear ion trap with Orbitrap Elite analyzer and Ultimate 3000 RSLCnano System (Thermo Scientific). Protein identification was done with the UniProtKB/Swiss-Prot database, using Mascot 2.4.0 (Matrix Sciences) and Scaffold 4.2.1 (Proteome Software). Both peptide and protein false discovery rates (FDR) were set to 1%, with a criterion of two unique peptides. The relative quantitative protein values corresponded to the $\log_2$-transformation of the protein ratio fold changes with respect to their measurements in the biological plasma reference sample. The sample preparation and all other manipulations relative to the MS measurements are detailed further in previous work[55–57].

The SomaLogic protein measurements were characterized using the SOMAscan assay[12], which relies on fluorescent labelling of poly-nucleotide aptamers targeting specific protein epitopes. Protein measurements were obtained in relative fluorescence unit and were then $\log_2$-transformed.

We discarded MS-based proteins if their measurements were missing for more than 5% of the samples, leaving 210 proteins in the Ottawa cohort and 136 in the DiOGenes cohort; we restricted all downstream analyses to the 133 proteins available for both cohorts. The SomaLogic measurements had no missing values. Totals of $1,100$ and $1,129$ proteins were assayed in the Ottawa and DiOGenes cohorts. All our analyses focused on the $1,096$ proteins quantified for both cohorts. The overlap between the MS and SomaLogic panels was of 72 proteins only.

We excluded samples with extreme expression values in more than 5% of the proteins, i.e., values beyond the outer fences of the empirical distribution ($q_{0.25} - 3 \times \text{IQR}$, $q_{0.75} + 3 \times \text{IQR}$, where $q_{0.25}$, $q_{0.75}$ are the lower and upper quartiles, and IQR is the interquartile range). After this quality control procedure, approximately 10 to 20 samples were removed from each of the four datasets; 577 and 428 Ottawa samples remained in the MS and SomaLogic datasets, respectively, and 481 and 563 DiOGenes samples remained in the MS and SomaLogic datasets, respectively.

Genotyping. Genotypes were generated using HumanCoreExome-12 v1.1 Illumina SNP arrays (Illumina, Inc., San Diego, CA), according to their manufacturer's instructions and were called with the GenomeStudio Software provided by Illumina. Preprocessing steps, including imputation and quality control, have been previously documented[58]. We discarded SNPs with call rate $< 95\%$, violating Hardy–Weinberg equilibrium (FDR $< 20\%$), and we discarded subjects with low call rate ($< 95\%$), abnormally high autosomal heterozygosity (FDR $< 1\%$), an XXY karyotype, or gender inconsistencies between genotype data and clinical records. For subjects with identity-by-state IBS$> 95\%$, we kept

only the one with the highest call rate. The subjects from both cohorts were of European ancestry and the two cohorts had similar genetic structure. We used principal component analyses separately on each cohort to exclude subjects that were extremely heterogeneous genetically. We performed genotype imputation using SHAPEIT and IMPUTE2, based on the European reference panel from the 1,000 Genome project (March 2012 release, phase 1 version 3). We then discarded SNPs with INFO score < 0.8, which left 4.9M imputed SNPs in both datasets. In order to avoid near-collinearity, which may render multivariate analyses unstable, we applied a light linkage disequilibrium (LD) pruning with PLINK using pairwise r^2 threshold 0.95. We applied a minor allele frequency threshold of 5%, after having restricted the genotype data to the subjects with available proteomic data.

The above steps were performed separately for the Ottawa and the DiOGenes cohorts, so in order to define a common set of SNPs for discovery and replication, we restricted each dataset to the SNPs available for both cohorts. After all genetic quality controls, and in both cohorts, $p = 275,485$ SNPs remained for the SomaLogic analysis and $p = 275,297$ remained for the MS analysis. In the Ottawa cohort $n = 376$ subjects had both genotype and MS proteomic data, and $n = 394$ subjects had both genotype and MS proteomic data. In the DiOGenes cohort, these numbers were $n = 400$ and 548.

Clinical data. Both cohorts had records on age, gender, anthropometric traits (weight and BMI), glycemic variables (fasting glucose, fasting insulin, HOMA-IR), and total lipid levels obtained from blood biochemistry (total cholesterol, triglycerides, HDL). We derived LDL values using the Friedewald formula[59], and obtained gender-specific *visceral adiposity index* (VAI) values using the formula of Amato et al.[60]. In each cohort and for each clinical variable, we removed a few samples with extreme measurements, similarly as for the proteomic data quality control.

Overview of LOCUS. LOCUS is an efficient Bayesian approach for estimating QTL associations jointly from $p = 10^5 - 10^6$ genetic variants, typically SNPs, and $q = 10^2 - 10^4$ expression outcomes, for $n = 10^2 - 10^4$ individuals; see Figure 1a. It is based on a hierarchical sparse regression model that involves a collection of high-dimensional regressions, one per outcome $\boldsymbol{y}_t$ (centered),

$$\boldsymbol{y}_t = \boldsymbol{X}\boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{N}_n\left(\boldsymbol{0}, \tau_t^{-1}\boldsymbol{I}_n\right), \qquad t = 1, \ldots, q. \tag{1}$$

Each outcome, $\boldsymbol{y}_t$, is related linearly to all p candidate predictor SNPs (centered), $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$, and has a specific residual precision, τ_t, to which we assign a Gamma prior, $\tau_t \sim \text{Gamma}(\eta_t, \kappa_t)$. As $p, q \gg n$, we enforce sparsity of the $p \times 1$ regression parameters $\boldsymbol{\beta}_t$ by placing a spike-and-slab prior on each of their components, namely, for $s = 1, \ldots, p$,

$$\beta_{st} \mid \gamma_{st}, \sigma^2, \tau_t \sim \gamma_{st}\,\mathcal{N}\left(0, \sigma^2\,\tau_t^{-1}\right) + (1 - \gamma_{st})\,\delta_0, \qquad \gamma_{st} \mid \omega_s \sim \text{Bernoulli}\left(\omega_s\right), \tag{2}$$

where δ_0 is the Dirac distribution. Hence, to each regression parameter β_{st} corresponds a binary latent parameter γ_{st}, which acts as a "predictor-response association indicator": the predictor $\boldsymbol{X}_s$ is associated with the response $\boldsymbol{y}_t$ if and only if $\gamma_{st} = 1$. The model thus enforces sparsity on the regression coefficients, and LOCUS identifies just one or a few markers per relevant locus, even in regions of high LD. The parameter σ represents the typical size of nonzero effects and is modulated by the residual scale, $\tau_t^{-1/2}$, of the response concerned by the effect; we infer σ from the data using a Gamma prior specification, $\sigma^{-2} \sim \text{Gamma}\left(\lambda, \nu\right)$. Finally, we let the probability parameter ω_s have a Beta distribution,

$$\omega_s \sim \text{Beta}\left(a_s, b_s\right), \qquad a_s, b_s > 0. \tag{3}$$

As it is involved in the Bernoulli prior specification of all $\gamma_{s1}, \ldots, \gamma_{sq}$, the parameter ω_s controls the proportion of responses associated with the predictor $\boldsymbol{X}_s$, and hence directly represents the propensity of predictors to be pleiotropic "hotspots". Both ω_s and σ^2 allow the leveraging of shared association patterns across all molecular variables, and enhances the estimation of weak *trans* and pleiotropic QTL effects. A graphical representation of the model is provided in Figure 1b; see Ruffieux et al.[11] for details. LOCUS estimates interpretable posterior probabilities of association for all SNP-outcome pairs (Figure 1c), from which Bayesian false discovery rates are easily calculated.

Inference on high-dimensional Bayesian models is both computationally and statistically difficult. Previous joint QTL approaches[61,62] are based on sampling procedures, such as Markov Chain Monte Carlo (MCMC) algorithms, and require prohibitive computational times on data with more than few hundreds of SNPs or outcomes. LOCUS uses a fast deterministic variational inference algorithm, which scales to the typical sizes of QTL problems. Previous work[11] compared LOCUS with existing QTL methods, whether sampling-based or deterministic, univariate or multivariate. We recently augmented our algorithm with a simulated annealing procedure[63] to enhance exploration of multimodal parameter spaces, as induced by strong LD structures. LOCUS is tailored to genomic, proteomic, lipidomic and methylation QTL analyses; it can also be used for genome-wide association with several clinical endpoints. Details and extensive performance studies are available[11]; see also Appendices A and B for simulations based on the Ottawa pQTL data and for a runtime profiling.

The applicability of a fully multivariate method to large molecular QTL data also hinges on the effective computational implementation of its algorithmic procedure. The annealed variational updates of LOCUS are analytical and performed by batches of variables. The software is written in R with C++ subroutines; it is publicly available at `https://github.com/hruffieux/locus`.

Proteomic quantitative trait locus analyses. We performed pQTL analyses separately for each platform, i.e., one analysis for the MS proteomic dataset, and another for the SomaLogic proteomic dataset. Each analysis comprised two stages: a discovery stage using the Ottawa cohort and a replication stage based on the DiOGenes cohort.

For discovery, we used the multivariate Bayesian method LOCUS on both the MS and the SomaLogic datasets, with an annealing schedule of 50 geometrically-spaced temperatures, and set the initial temperature to 20; pilot experiments indicated that estimation was not sensitive to these choices. We used a convergence tolerance of 10^{-3} on the absolute changes in the objective function as the stopping criterion. The algorithm can handle missing data in the response matrix, so no imputation was necessary for the MS proteomic data.

We adjusted all analyses for age, gender, and BMI at baseline. No important stratification was observed in the genotype data; the first ten principal components together explained little of the total variance ($< 4\%$), so we did not include them as covariates. We derived FDR values from the posterior probabilities of association obtained between each SNP and each protein, and reported pQTL associations using an FDR threshold of 5%. Both LOCUS runs completed within hours; convergence was reached after 2 hours (79 iterations) for the MS dataset, and after 10 hours and 20 minutes (72 iterations) for the SomaLogic dataset, on an Intel Xeon CPU, 2.60 GHz. The method can handle larger datasets, e.g., it takes less than a day to run the MS data with 3 million SNPs and on the SomaLogic data with 1 million SNPs.

We performed a validation study of the pQTLs discovered using the DiOGenes cohort with GEMMA[64], with centered relatedness matrix (default) and p-values from (two-sided) Wald tests.

We then obtained adjusted p-values using Benjamini–Hochberg false discovery rates, and validated our hits using again an FDR threshold of 5%.

pQTL annotation. We used the Ensembl database (GRCh37, release 94) to retrieve the list of genes within 2Mb of each sentinel SNP (i.e., involved in the pQTL associations identified by LOCUS), and retrieved the SNPs in LD ($r^2 > 0.8$), limiting the search to 500Kb upstream and downstream of the sentinel SNP position. We called *cis pQTLs*, all sentinel SNPs located within ±1Mb of the gene encoding for the controlled protein, and *trans pQTLs*, all other pQTLs.

We evaluated the overlap between our pQTL associations and previously reported pQTL signals with the PhenoScanner database[65], using the default p-value threshold $p < 1 \times 10^{-5}$ and an LD proxy search ($r^2 > 0.8$).

Epigenomic annotation. We retrieved epigenomic annotations of 1,000 Genomes Project (release 20110521) from Pickrell[66]. The data covered 450 annotation features, each binary-coded according to the presence or absence of overlap with the SNPs. The features corresponded to DNase-I hypersensitivity, chromatin state, SNP consequences (coding, non-coding, 5'UTR, 3'UTR, etc), synonymous and nonsynonymous status and histone modification marks. We obtained distances to the closest transcription start site from the UCSC genome browser. Ninety-seven of our 104 validated sentinel SNPs had annotation data; to evaluate their functional enrichment, we resampled SNP sets of size 97 from our initial SNP panel, and, for each set, we computed the cumulated number of annotations. We did the same for the distances to transcription start sites. We repeated this 10^5 times to derive empirical p-values.

Colocalization with known eQTLs and with GWAS risk loci. We evaluated the overlap of our pQTLs with the eQTL variants reported by GTEx Consortium[24] (release 7) at q-value < 0.05. We considered all 49 tissues listed by GTEx but eQTL SNPs for several tissues were counted only once. We made both general queries and queries asking whether a pQTL uncovered by LOCUS was an eQTL for the gene coding for the controlled protein.

We retrieved known associations between the validated sentinel pQTLs and diseases or clinical traits, based on the GWAS catalog[67] (v1.0 release e92), and also using an LD proxy search ($r^2 > 0.8$).

We evaluated enrichment for eQTL and risk loci using one-sided Fisher exact tests based on the 104 validated sentinel pQTLs.

Associations with clinical variables. We tested associations between the proteins under genetic control and clinical parameters separately in each cohort. For the DiOGenes data, we used linear mixed-effect models, adjusting for age, gender as fixed effects, and center as a random effect. For the Ottawa data, we used linear models, adjusting for age and gender. Except when testing associations with anthropomorphic traits, all analyses were also adjusted for BMI. For the clinical variables available in the two cohorts (total cholesterol, HDL, LDL, fasting glucose, fasting insulin, HOMA-IR, triglycerides and VAI), we performed meta-analyses using the R package `metafor`. We used random-effects models to account for inter-study variability, which may in part result from geographical differences, and employed two-sided Wald tests for fixed effects, and Cochran Q-tests for measuring residual heterogeneity; we did not interpret the results if between-study heterogeneity estimates were high ($I^2 > 80\%$), and evaluated the directional consistency of the effects between Ottawa

and DiOGenes. We adjusted for multiplicity using Benjamini–Hochberg correction across all tests, i.e., involving the 88 tested proteins and the two proteomic technologies, and reported associations using a 5% FDR threshold.

We assessed whether the proteins under genetic control were enriched in associations with the clinical variables. We randomly selected 10^5 sets of 88 proteins from the panel used for the pQTL analyses and derived an empirical p-value by counting, for each set, the number of proteins with at least one clinical association at FDR 5%.

Data availability. The MS proteomic data have been deposited on the ProteomeXchange Consortium via the PRIDE partner repository `http://www.proteomexchange.org` with the dataset identifiers PXD005216 for DiOGenes and PXD009350 for Ottawa. The SomaLogic proteomic data are available from the Open Science Framework, at `https://osf.io/v8mes/?view_only=13e4ccd127024ee7b4c819385325925c` and `https://osf.io/s4v8t/?view_only=90637f2941e14ec986e5888491fbdbbb`, respectively for Ottawa and DiOGenes. All pQTL and clinical association results can be browsed from our online database: `https://locus-pqtl.epfl.ch`. Other data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability. All statistical analyses were performed using the R environment (version 3.3.2). LOCUS and ECHOSEQ are freely available from Github.

URLs. ECHOSEQ: `https://github.com/hruffieux/echoseq`
Ensembl: `http://grch37.ensembl.org/index.html`
GEMMA: `http://www.xzlab.org/software.html`
GTEx: `https://gtexportal.org/home`
GWAS Catalog: `https://www.ebi.ac.uk/gwas`
IMPUTE2: `http://mathgen.stats.ox.ac.uk/impute/impute_v2.html`
JASPAR: `http://jaspar.genereg.net`
LOCUS: `https://github.com/hruffieux/locus`
Metafor: `https://cran.r-project.org/web/packages/metafor/index.html`
PhenoScanner: `http://www.phenoscanner.medschl.cam.ac.uk`
PLINK: `http://zzz.bwh.harvard.edu/plink`
ProteomeXchange: `http://www.proteomexchange.org`
R: `https://www.r-project.org`
SHAPEIT: `https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html`
SNP2TFBS: `https://ccg.vital-it.ch/cgi-bin/snp2tfbs/snpviewer_form_parser.cgi`
UCSC: `https://genome.ucsc.edu`
UniProt: `https://www.uniprot.org`

Acknowledgments. We thank Antonio Núñez Galindo, John Corthésy, Martin Kussmann, Ornella Cominetti and Loïc Dayon for designing and performing proteomic experiments, and the generation and cleaning of the MS-based proteomic datasets. We thank Radu Popescu for setting up the web server. We thank Zoltan Kutalik, Nele Gheldof and Ruth McPherson for their constructive comments, and thank Leonardo Bottolo and Sylvia Richardson for their contributions on the method's

simulated annealing procedure.

Author contributions. HR designed and developed the LOCUS method with input from AD and JH. AV supervised the omics data generation and preprocessing. HR and AV designed the pQTL study. HR implemented statistical analyses with input from AV and AD. WS and AA designed the DiOGenes clinical study; BD and MEH designed the Canadian program. HR and AV interpreted the results, wrote the manuscript with input from all authors, and had primary responsibility for final content.

Conflicts of interest. HR, JC, JH and AV are full-time employees at Nestlé Research. WS reports research support from several food companies (Nestlé, DSM, Unilever, Nutrition et Santé and Danone), and pharmaceutical companies (GSK, Novartis and Novo Nordisk). He is an unpaid scientific advisor for the International Life Science Institute, ILSI Europe.

AA reports personal fees from Acino, Switzerland, BioCare Copenhagen, DK , Dutch Beer Institute, NL, Gelesis, USA, Groupe Éthique et Santé, France, McCain Foods Limited, USA, Pfizer, USA, Weight Watchers, USA, Zaluvida, Switzerland, Navamedic, DK, Novo Nordisk, DK, and Saniona, DK; personal fees, grants and other from Gelesis, USA; grants from Arla Foods, DK, Danish Dairy Research Council, and Nordea Foundation. He is co-inventor/-owner of patents pending to University of Copenhagen; Co-owner of University of Copenhagen spin-outs Flaxslim and Gluco-diet.dk, recipient of stock options in Gelesis, USA, and co-author of books on diet and personalized nutrition for weight loss.

The DiOGenes project was supported by the European Commission (Food Quality and Safety Priority of the Sixth Framework Program: FP6-2005- 513946). Local sponsors made financial contributions to the shop centers, which received some foods free of charge from manufacturers. A full list of these sponsors can be seen at `www.DiOGenes-eu.org/sponsors`. The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the manuscript.

References

[1] Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of European ancestry. *Human Molecular Genetics*, 27:3641–3649, 2018.

[2] Suhre, K. et al. Connecting genetic risk to disease end points through the human blood plasma proteome. *Nature Communications*, 8:14357, 2017.

[3] Carayol, J. et al. Protein quantitative trait locus study in obesity during weight-loss identifies a leptin regulator. *Nature Communications*, 8:2084, 2017.

[4] Sun, B. B. et al. Genomic atlas of the human plasma proteome. *Nature*, 558:73–79, 2018.

[5] Hess, A. L. et al. Analysis of circulating angiopoietin-like protein 3 and genetic variants in lipid metabolism and liver health: the DiOGenes study. *Genes & Nutrition*, 13:7, 2018.

[6] Thrush, A. B. et al. Diet-resistant obesity is characterized by a distinct plasma proteomic signature and impaired muscle fiber metabolism. *International Journal of Obesity*, 42:353, 2018.

[7] Mackay, T. F. C., Stone, E. A., and Ayroles, J. F. The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, 10:565, 2009.

[8] Nica, A. C. and Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Philosophical Transaction of the Royal Society B*, 368:20120362, 2013.

[9] Dent, R. M., Penwarden, R. M., Harris, N., and Hotz, S. B. Development and evaluation of patient-centered software for a weight-management clinic. *Obesity Research*, 10:651–656, 2002.

[10] Larsen, T. M. et al. The Diet, Obesity and Genes (Diogenes) Dietary Study in eight European countries—a comprehensive design for long-term intervention. *Obesity Reviews*, 11:76–91, 2010.

[11] Ruffieux, H., Davison, A. C., Hager, J., and Irincheeva, I. Efficient inference for genetic association studies with multiple outcomes. *Biostatistics*, 18:618–636, 2017.

[12] Kraemer, S. et al. From SOMAmer-based biomarker discovery to diagnostic and clinical applications: a SOMAmer-based, streamlined multiplex proteomic assay. *PloS one*, 6:e26332, 2011.

[13] Yao, C. et al. Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nature Communications*, 9:3268, 2018.

[14] Zhou, X. and Stephens, M. Efficient algorithms for multivariate linear mixed models in genome-wide association studies. *Nature Methods*, 11:407, 2014.

[15] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.

[16] Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics*, 47:979–986, 2015.

[17] Singh, S., Dulai, P. S., Zarrinpar, A., Ramamoorthy, S., and Sandborn, W. J. Obesity in IBD: epidemiology, pathogenesis, disease course and treatment outcomes. *Nature Reviews Gastroenterology & Hepatology*, 14:110–121, 2017.

[18] Sims, R. et al. Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. *Nature Genetics*, 49:1373–1384, 2017.

[19] Lambert, J. C. et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*, 45:1452–1458, 2013.

[20] Matsunaga, H. et al. Adipose tissue complement factor B promotes adipocyte maturation. *Biochemical and Biophysical Research Communications*, 495:740–748, 2018.

[21] Goustin, A. and Abou-Samra, A. B. The "thrifty" gene encoding Ahsg/Fetuin-A meets the insulin receptor: Insights into the mechanism of insulin resistance. *Cellular Signalling*, 23:980–990, 2011.

[22] Monroy-Muñoz, I. E. et al. PLA2g2a polymorphisms are associated with metabolic syndrome and type 2 diabetes mellitus. Results from the genetics of atherosclerotic disease Mexican study. *Immunobiology*, 222:967–972, 2017.

[23] Moreno-Navarrete, J. M. et al. Complement factor H is expressed in adipose tissue in association with insulin resistance. *Diabetes*, 59:200–209, 2010.

[24] GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, 348:648–660, 2015.

[25] Müssig, K. et al. RARRES2, encoding the novel adipokine chemerin, is a genetic determinant of disproportionate regional body fat distribution: a comparative magnetic resonance imaging study. *Metabolism*, 58:519–524, 2009.

[26] Er, L. et al. Pleiotropic Associations of RARRES2 Gene Variants and Circulating Chemerin Levels: Potential Roles of Chemerin Involved in the Metabolic and Inflammation-Related Diseases. *Mediators of Inflammation*, 2018, 2018.

[27] Bonder, M. J. et al. Disease variants alter transcription factor levels and methylation of their binding sites. *Nature Genetics*, 49:131–138, 2017.

[28] Banerjee, M. and Saxena, M. Interleukin-1 (IL-1) family of cytokines: role in type 2 diabetes. *International Journal of Clinical Chemistry*, 413:1163–1170, 2012.

[29] Böni-Schnetzler et al. Beta Cell-Specific Deletion of the IL-1 Receptor Antagonist Impairs beta Cell Proliferation and Insulin Secretion. *Cell Reports*, 22:1774–1786, 2018.

[30] Bozaoglu, K. et al. Plasma Levels of Soluble Interleukin 1 Receptor Accessory Protein Are Reduced in Obesity. *The Journal of Clinical Endocrinology and Metabolism*, 99:3435–3443, 2014.

[31] Alberti, K. G. M. M. et al. Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. *Circulation*, 120:1640–1645, 2009.

[32] Monestier, O. and Blanquet, V. WFIKKN1 and WFIKKN2:"Companion" proteins regulating TGFB activity. *Cytokine & Growth Factor Reviews*, 32:75–84, 2016.

[33] Chen, L. et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell*, 167:1398–1414.e24, 2016.

[34] Favennec, M. et al. The kynurenine pathway is activated in human obesity and shifted toward kynurenine monooxygenase activation. *Obesity*, 23:2066–2074, 2015.

[35] Song, P., Ramprasath, T., Wang, H., and Zou, M. Abnormal kynurenine pathway of tryptophan catabolism in cardiovascular diseases. *Cellular and molecular life sciences: CMLS*, 74:2899–2916, 2017.

[36] Stone, T. W., McPherson, M., and Gail Darlington, L. Obesity and Cancer: Existing and New Hypotheses for a Causal Connection. *EBioMedicine*, 30:14–28, 2018.

[37] Harden, J. L. et al. The tryptophan metabolism enzyme L-kynureninase is a novel inflammatory factor in psoriasis and other inflammatory diseases. *The Journal of Allergy and Clinical Immunology*, 137:1830–1840, 2016.

[38] Armstrong, A. W., Harskamp, C. T., and Armstrong, E. J. The association between psoriasis and obesity: a systematic review and meta-analysis of observational studies. *Nutrition & Diabetes*, 2:e54, 2012.

[39] Jensen, P. et al. Long-term effects of weight reduction on the severity of psoriasis in a cohort derived from a randomized trial: a prospective observational follow-up study. *The American Journal of Clinical Nutrition*, 104:259–265, 2016.

[40] Jacobs, K. R., Castellano-González, G., Guillemin, G. J., and Lovejoy, D. B. Major Developments in the Design of Inhibitors along the Kynurenine Pathway. *Current Medicinal Chemistry*, 24:2471–2495, 2017.

[41] Bass, N. M. et al. Clinical, laboratory and histological associations in adults with nonalcoholic fatty liver disease. *Hepatology (Baltimore, Md.)*, 52:913–924, 2010.

[42] Sheldon, R. D. et al. Transcriptomic differences in intra-abdominal adipose tissue in extremely obese adolescents with different stages of NAFLD. *Physiological Genomics*, 48:897–911, 2016.

[43] Song, Y. et al. Circulating levels of endothelial adhesion molecules and risk of diabetes in an ethnically diverse cohort of women. *Diabetes*, 56:1898–1904, 2007.

[44] Ricklin, D., Hajishengallis, G., Yang, K., and Lambris, J. D. Complement: a key system for immune surveillance and homeostasis. *Nature Immunology*, 11:785–797, 2010.

[45] Phieler, J., Garcia-Martin, R., Lambris, J. D., and Chavakis, T. The role of the complement system in metabolic organs and metabolic diseases. In *Seminars in Immunology*, volume 25, pages 47–53. Elsevier, 2013.

[46] Reiner, A. P. et al. PROC, PROCR, and PROS1 polymorphisms, plasma anticoagulant phenotypes, and risk of cardiovascular disease and mortality in older adults: the Cardiovascular Health Study. *Journal of Thrombosis and Haemostasis*, 6:1625–1632, 2008.

[47] van der Harst, P. and Verweij, N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circulation Research*, 122:433–443, 2018.

[48] Wong, R. H. F. et al. A role of DNA-PK for the metabolic gene regulation in response to insulin. *Cell*, 136:1056–1072, 2009.

[49] Park, S. et al. DNA-PK promotes the mitochondrial, metabolic and physical decline that occurs during aging. *Cell Metabolism*, 25:1135–1146.e7, 2017.

[50] Chung, J. H. The role of DNA-PK in aging and energy metabolism. *The FEBS Journal*, 285:1959–1972, 2018.

[51] Tavana, O. et al. Ku70 functions in addition to nonhomologous end joining in pancreatic beta-cells: a connection to beta-catenin regulation. *Diabetes*, 62:2429–2438, 2013.

[52] Lin, J. C., Ho, W., Gurney, A., and Rosenthal, A. The netrin-G1 ligand NGL-1 promotes the outgrowth of thalamocortical axons. *Nature Neuroscience*, 6:1270–1276, 2003.

[53] Ramkhelawon, B. et al. Netrin-1 promotes adipose tissue macrophage retention and insulin resistance in obesity. *Nature Medicine*, 20:377–384, 2014.

[54] Folkersen, L. et al. Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLoS Genetics*, 13:e1006706, 2017.

[55] Cominetti, O. et al. Proteomic biomarker discovery in 1000 human plasma samples with mass spectrometry. *Journal of Proteome Research*, 15:389–399, 2015.

[56] Oller Moreno, S. et al. The differential plasma proteome of obese and overweight individuals undergoing a nutritional weight loss and maintenance intervention. *PROTEOMICS*, 12:1600150, 2018.

[57] Cominetti, O. et al. Obesity shows preserved plasma proteome in large independent clinical cohorts. *Scientific Reports*, 8:16981, 2018. just accepted.

[58] Valsesia, A. et al. Genome-wide gene-based analyses of weight loss interventions identify a potential role for NKX6.3 in metabolism. *Nature Communications*, in press.

[59] Friedewald, W. T., Levy, R. I., and Fredrickson, D. S. Estimation of the concentration of low-density lipoprotein cholesterol in plasma without use of the preparative ultracentrifuge. *Clinical Chemistry*, 18:499–502, 1972.

[60] Amato, M. C. et al. Visceral adiposity index (VAI): a reliable indicator of visceral fat function associated with cardiometabolic risk. *Diabetes Care*, 2010.

[61] Jia, Z. and Xu, S. Mapping quantitative trait loci for expression abundance. *Genetics*, 176:611–623, 2007.

[62] Bottolo, L. et al. Bayesian detection of expression quantitative trait loci hot spots. *Genetics*, 189:1449–1459, 2011.

[63] Ueda, N. and Nakano, R. Deterministic annealing EM algorithm. *Neural Networks*, 11:271–282, 1998.

[64] Zhou, X. and Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44:821, 2012.

[65] Staley, J. R. et al. PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics*, 32:3207–3209, 2016.

[66] Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics*, 94:559–573, 2014.

[67] Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42:D1001–D1006, 2014.

A Statistical performance of LOCUS

We evaluated the expected performance of LOCUS on our data by conducting two simulation studies. We compared its statistical power to detect pQTL associations with that of GEMMA [14], a univariate linear mixed model approach. We used the R package `echoseq` (`https://github.com/hruffieux/echoseq`) to generate synthetic data that emulate real data.

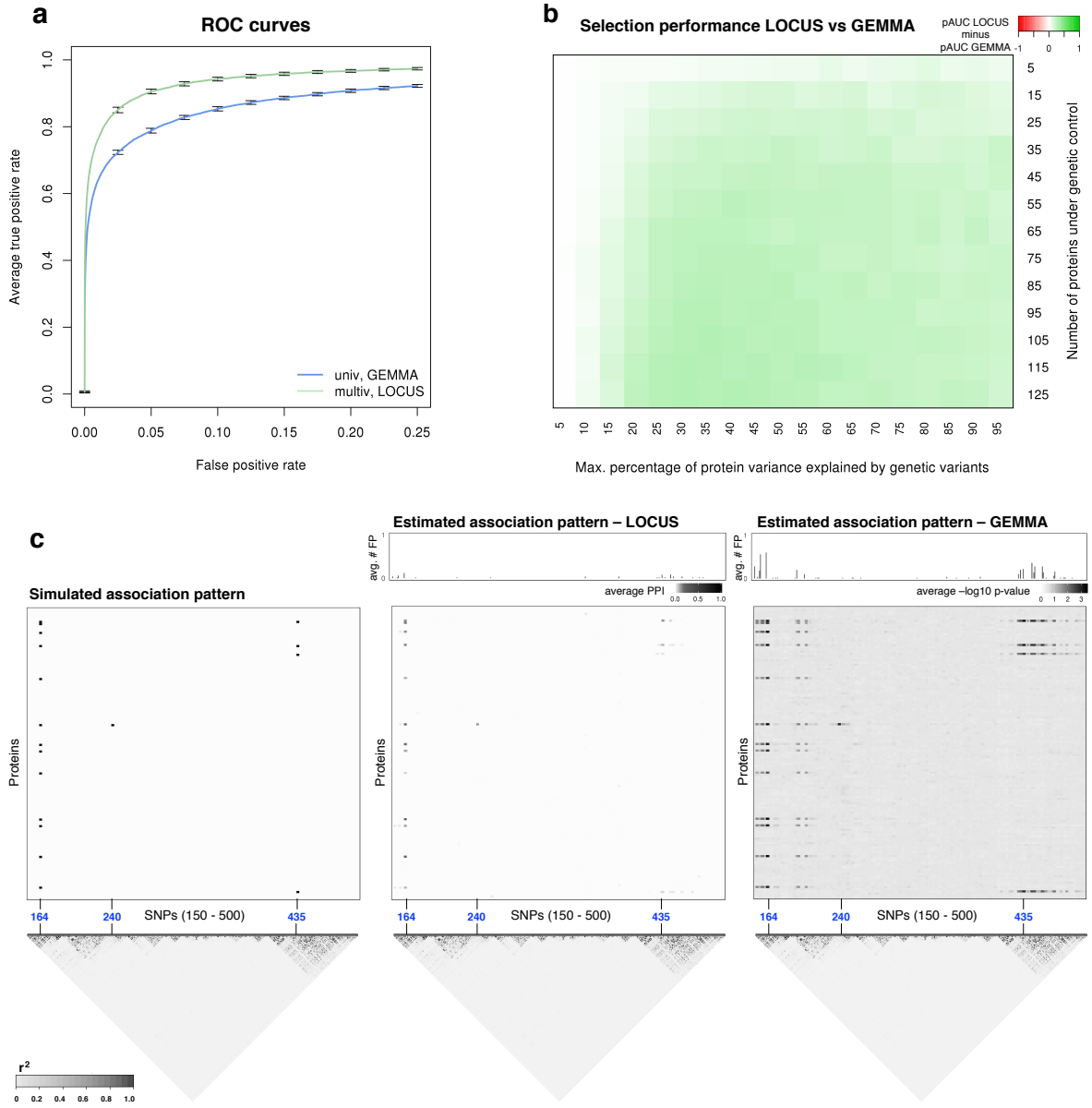

Figure A.1: Selection performance of LOCUS and GEMMA. (a) Truncated average ROC curves with 95% confidence intervals, obtained from 50 replications, for identification of SNP-trait associations. (b) Difference of average standardized pAUC of LOCUS and GEMMA for a grid of effect sizes (x-axis) and signal sparsity (y-axis), using 20 replications for each scenario. (c) Simulated pattern, and patterns recovered by LOCUS and GEMMA, averaged over the 50 replications. The plots display a window of 350 SNPs (x-axis) containing the first three SNPs having simulated associations (blue labels), along with their linkage disequilibrium (LD) pattern. The top panels of the middle and right plots display the average number of false positives when selecting SNPs at FDR 25%. GEMMA indicates many false positive associations in regions of high LD, while LOCUS better pinpoints the relevant SNPs.

We ran LOCUS and GEMMA on the SNPs of all $n = 376$ Ottawa subjects, and on simulated expression outcomes with residual dependence replicating that of the $q = 133$ mass-spectrometry pro-

teomic expression levels. We first used the SNPs from chromosome one ($p = 20,900$), and generated associations between 20 SNPs and 25 proteins chosen randomly, leaving the remaining variables unassociated. Some proteins were under pleiotropic control; we drew the degree of pleiotropy of the 20 SNPs from a positively-skewed Beta distribution, so only a few SNPs were hotspots, i.e., were associated with many proteins. We generated associations under an additive dose-effect scheme and drew the proportions of outcome variance explained by a given SNP from a Beta$(2, 5)$ distribution to give more weight to smaller effect sizes. We then rescaled these proportions so that the variance of each protein attributable to genetic variation was below 35%. These choices led to an inverse relationship between minor allele frequencies and effect sizes, which is to be expected under natural selection. We generated 50 replicates, re-drawing the protein expression levels and effect sizes for each.

The ROC curves of Figure A.1a show a net gain in power for selections with LOCUS compared to GEMMA. The average standardized partial areas under the curve (pAUC) with 95% confidence intervals are 0.926 ± 0.005 for LOCUS and 0.840 ± 0.005 for GEMMA, using a false positive threshold of 25%.

In the second simulation, we re-assessed the performance of LOCUS for a grid of data generation scenarios. We considered a wide range of sparsity levels (numbers of proteins under genetic control) and effect sizes (proportions of outcome variance explained by the genetic variants). Given the large number of configurations (247), and in order to limit the computational burden, we used the first $p = 2,000$ SNPs, and ran LOCUS and GEMMA on 20 replicates for each configuration. Figure A.1b indicates that the superiority of LOCUS over GEMMA generalizes to all data generation scenarios, as the average standardized pAUC is everywhere greater for LOCUS than for GEMMA.

The performance of LOCUS is largely attributable to its multivariate modelling of all the SNPs and proteomic outcomes, which allows sharing of information across and within loci, as well as across different proteins under common genetic regulation. By design, univariate screening approaches do not exploit association patterns common to multiple outcomes or markers; they analyze the outcomes individually, and do not account for LD, which increases false discoveries at loci with strong LD (Fig. A.1c). At a given FDR, such spurious associations hamper the detection of weak but genuine signals. Owing to its simulated annealing procedure that improves exploration at loci with strong LD, LOCUS better discriminates truly associated SNPs from their correlated neighbours (Fig. A.1c).

B Computational performance of LOCUS

The runtime of LOCUS for the simulations of Section A was similar to that of GEMMA. On average, for one replicate, LOCUS took 5 minutes and 26 seconds to complete, while GEMMA took 7 minutes and 4 seconds, running in parallel on four cores of an Intel Xeon CPU, 2.60 GHz.

Figure B.1 presents runtime profiling for LOCUS run on different numbers of SNPs and molecular traits on Intel Xeon CPU, 2.40 GHz machines, with 256 Gb RAM. All runs completed within hours.

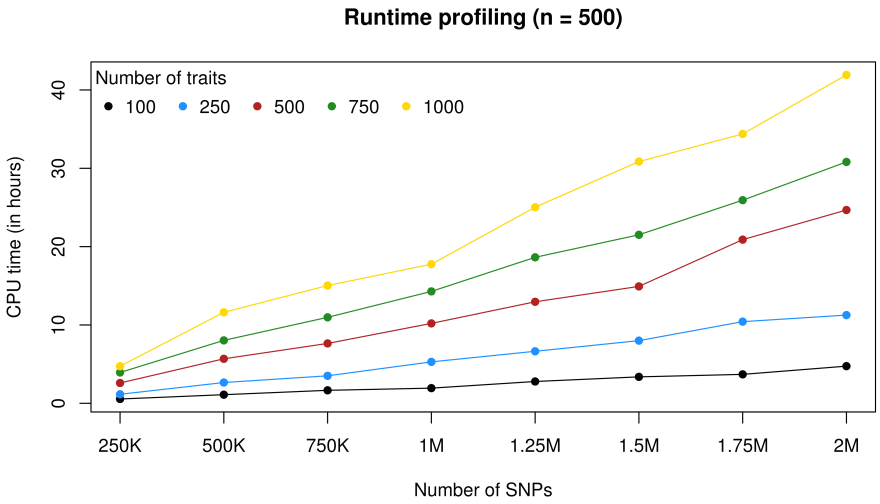

Figure B.1: Runtime profiling in CPU hours, for 2.5×10^5 to 2×10^6 SNPs and 100 to 1000 traits, on an Intel Xeon CPU at 2.40 GHz with 256 Gb RAM. In each case, the average runtime of five replications is displayed. The same annealing settings as in the MS and SomaLogic analyses were used (50 geometrically-spaced temperatures and initial temperature $T = 20$).

C Correlation between expression levels of KYNU and other inflammation mediated proteins

Figure C.1 shows the correlation between the expression levels of inflammation mediated proteins, in Ottawa and DiOGenes.

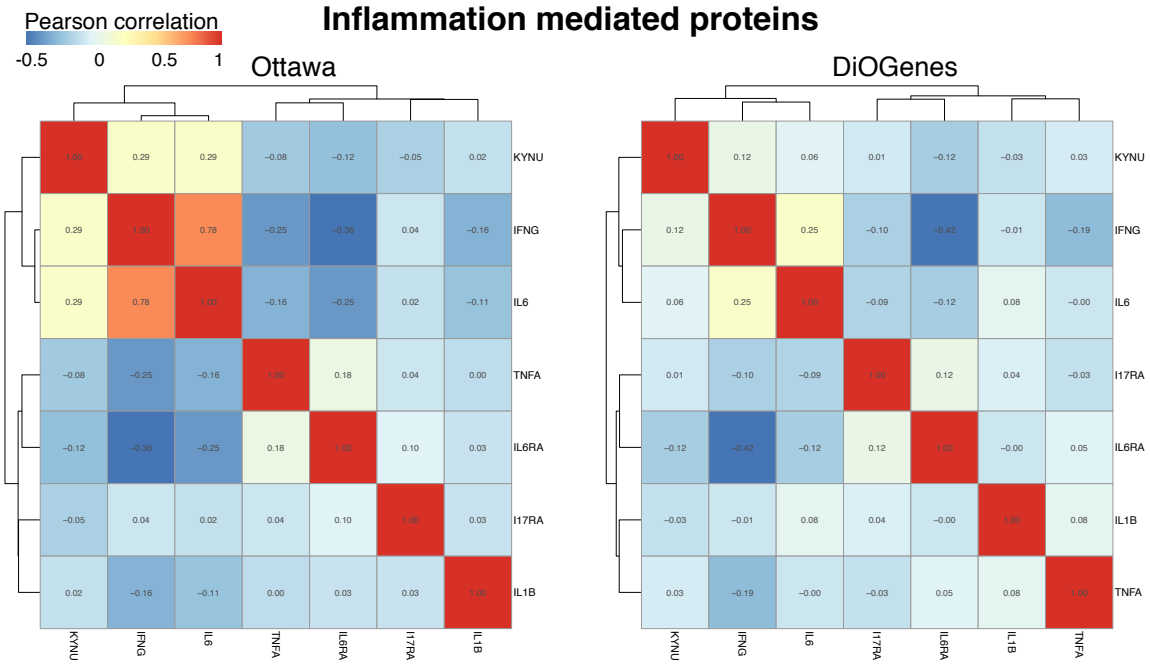

Figure C.1: Correlation of KYNU, IFNG, IL6, THFA, IL6RA, i17RA and IL1B in Ottawa (left) and in DiOGenes (right).