Whole-exome sequencing identifies common and rare variant metabolic QTLs in a Middle Eastern population

Noha A. Yousri1,2, Khalid A. Fakhro1,3, Amal Robay1, Juan L. Rodriguez-Flores4, Robert P. Mohney5, Hassina Zeriri1, Tala Odeh1, Sara Abdul Kader6, Eman K. Aldous7, Gaurav Thareja6, Manish Kumar6, Alya Al-Shakaki1, Omar M. Chidiac1, Yasmin A. Mohamoud7, Jason G. Mezey4, Joel A. Malek1,7, Ronald G. Crystal4 & Karsten Suhre6

Metabolomics-genome-wide association studies (mGWAS) have uncovered many metabolic quantitative trait loci (mQTLs) influencing human metabolic individuality, though predominantly in European cohorts. By combining whole-exome sequencing with a high-resolution metabolomics profiling for a highly consanguineous Middle Eastern population, we discover 21 common variant and 12 functional rare variant mQTLs, of which 45% are novel altogether. We fine-map 10 common variant mQTLs to new metabolite ratio associations, and 11 common variant mQTLs to putative protein-altering variants. This is the first work to report common and rare variant mQTLs linked to diseases and/or pharmacological targets in a consanguineous Arab cohort, with wide implications for precision medicine in the Middle East.
Metabolites represent functional intermediates to the end phenotype, can be conserved over several years’ time frame, and can uniquely identify individuals. Several studies have shown that they are influenced by a combination of genetics and environment, the latter comprising both lifestyle exposures and microbial interactions. Recent technological improvements have enabled the accurate detection of thousands of metabolites (collectively, the metabolome), adding highly informative downstream read-outs supporting genetic and transcriptomic signatures in the study of personalized medicine. With the abundance of such omics data, it will become possible to infer causal relationships between constitutive genetic variants and metabolite levels to accurately predict the likelihood of developing pathophysiologic signatures, as a normal individual progresses into a disease state.

To date, there have been several large-scale genome-wide association studies for metabolic traits (mGWAS) the largest (with broad non-targeted metabolomics) of which interrogated 7000 individuals of European ancestry and discovered 145 significant metabolomics quantitative trait loci (mQTLs). While studies to date have uncovered hundreds of mQTLs, they have also faced certain limitations. First, they relied on imputed genotyping array data for the discovery of common variant mQTLs, where high-quality SNPs are in non-coding regions. More recently, next-generation sequencing (NGS) technologies have become more affordable and begun to identify protein-coding variants largely affecting metabolite levels, yet on a small scale of effect, which is more consistent with the small effect sizes seen in GWAS studies. In particular, recent Metabolon DiscoveryHD platform with deep whole-exome sequencing data of 614 Qataris for discovery, and imputed array data to discover loci affecting metabolites and metabolite ratios in this population, while integrating this data to discover loci affecting metabolites and metabolite ratios, underscoring the metabolic individuality of subjects from this ethnic population.

Pre-discovery exploration of replication in Caucasians. As an investigation step prior to discovery, we attempted to replicate 145 previously known loci in the discovery cohort. We found exact or proxy SNPs (SNPs in LD) for 101 of 145 loci in our dataset, of which sentinel metabolites in 80 loci were matched to metabolites in our cohort. For 13 loci, we replicated associations between the reported SNP and the reported metabolite at a Bonferroni p-value (p ≤ 0.05/80 = 6.25 × 10^-4), and for another 15 loci, we replicated the association at the same/proxy SNP but with another metabolite (p < 0.05/(101 × 826) = 5.9 × 10^-7) (Supplementary Data 2). In total, we replicated 28 loci (19.3% of 145 loci, 35% of 80 replicable loci)—including 11 of the 20 most significant loci.

21 common variant loci influence metabolites and ratios. To discover all metabolites associated with exome variants in 614 Qataris, 1.6 million imputed exome variants (272,061 SNPs after LD pruning) were tested for association with 826 metabolites (see Supplementary Fig. 1 for distribution of kinship-based heritability estimates for each super pathway). This step uncovered 3127 significant associations (Bonferroni p ≤ 2.2 × 10^-10) (Supplementary Data 7), with an average inflation factor of 0.98 (range: 0.83–1.05). Those associations collapsed into 19 independent loci (see Methods), which we attempted to replicate (based on the sentinel SNP and the sentinel metabolite) in the imputed array data set (n = 382) and found that 17 of them replicated (p ≤ 0.05/19 = 0.0026) and one nominally replicated at p ≤ 0.05 (Table 2, Supplementary Data 3–5).

In addition to single-metabolite-variant associations, we examined each locus to identify significantly associated metabolite ratios. To limit multiple testing, two approaches were used to examine associations of metabolic ratios. First, we computed the associations between SNPs (within 100 Kb of the sentinel SNP from single-metabolite association analysis) and the ratio of the sentinel metabolite to all remaining 825 metabolites. Second, for all SNPs where two metabolites had been nominally associated in the discovery phase (p ≤ 10^-4) but in opposite directions (opposite beta signs), we computed the association of the given SNP to the ratio of that pair of metabolites. Both p-value and p-gain thresholds were used to find significant ratios. A total of 11 significant SNP to metabolite ratio associations were discovered with p ≤ 0.05/(826 × 18 + 826 × 272,061) = 2.2 × 10^-10 and p-gain ≥ (1/(2 × 0.05)) × (826 × 18) = 1.48 × 10^5 (see Methods). Seven of these resulted in metabolic fine-mapping of seven of the loci discovered in the single-metabolite phase (i.e. where a ratio replaced the single metabolite as the sentinel association), whereas the remaining four were new associations at loci not associated with single metabolites. Only three of those four replicated in the cohort of 382 individuals (Table 2, and Supplementary Data 3–5), resulting in a total of 10 SNP to metabolite ratio associations.

### Results

**Subject selection and genotyping.** A total of 996 Qatari subjects were selected for this study (Table 1), of whom 614 were whole-exome sequenced (WES) for the discovery step, and 382 were array genotyped for replication. For the purpose of replication, both data sets were imputed using a reference set of 108 deeply covered, phased Qatari genomes. A total of 1,650,892 imputed exome variants were available for the analysis after Quality Control (see Methods). All samples were analyzed on Metabolon’s DiscoveryHD4 platform, where a total of 1303 metabolites were detected (Supplementary Data 1). After applying strict QC, 826 metabolites (including 249 unknowns) remained for the association analysis (Fig. 1 gives a schematic representation of the study design).
Combined, we discovered 21 unique metabolite and metabolite ratio quantitative trait loci (mQTLs) in Qataris (Fig. 2, and Supplementary Figs. 2 and 3). The variance explained by these genetic variants ranges from the highest value of 31% (5-acetylamino-6-amino-3-methyluracil/1-methylxanthine with rs4646257 in \(NAT2\) locus) to 7% (association of undecanedioate with c15p90683852 in \(SEMA4B\) locus) with an average of 14.8% (Table 2, Fig.3). Of the 21 loci, 7 (31%) were unknown to studies that used older metabolomics platforms (studies prior to 2017) and are defined here as newly identified loci (four loci were concurrently identified in a study published while finalizing this manuscript\(^{15}\), and three loci are novel (Table 2, Supplementary Data 3). Five of those seven loci might be new due to the use of the new metabolomics platform, another two (\(TMPRS511E\) and \(SEMA4B\)) were not discovered in ref.\(^{11}\), and other known loci were fine-mapped to new metabolites, and not reported elsewhere (as rs2147896 in \(PYROXD2\) with N-methylpipecolate, and \(UGT3A1\) with X-24348). We explored the frequencies of the sentinel SNPs in the Genome Aggregation Database (gnomAD)\(^{20}\) (Supplementary Data 3), and found that among the seven loci, the sentinel SNP in two novel loci, the sentinel SNP in two novel ones are not reported in the database (\(PHYHD1\)/\(NUP188\), \(SEMA4B\)), and those in another two loci (\(SLC22A24\), \(TTC38\)/\(PKDREJ\)) are rare in other populations.

All 10 reported metabolite ratio associations represent metabolic fine-mapping of loci discovered here or in previous studies. According to this cohort, those ratios associate more strongly than the single metabolites previously reported\(^{11,15}\). These include the association of rs4646257 in \(NAT2\) locus with the ratio 5-acetylamino-6-amino-3-methyluracil/1-methylxanthine, rs375811360 in \(NAT8\) locus with 2-aminooctanoate/X-12511 (X-12511 is possibly 2-acetamidoocanoic acid as identified by Metabolon), rs3756669 in \(UGT3A1\) locus with X-24348/pregn-
Table 2 21 Unique locus-metabolite pairs, indicating 7 newly identified and novel loci, 10 loci fine mapped with new metabolite ratio associations, and 11 protein coding variants

| Locus     | rsID        | Metabolite/Ratio                                      | p-value (pgain)<sup>a</sup> | Beta       | Annotation<sup>b</sup> | r2 (%)<sup>c</sup> | Ref/NL/ NAS<sup>d</sup> | Replication p-value<sup>e</sup> |
|-----------|-------------|------------------------------------------------------|-----------------------------|------------|-------------------------|-----------------|------------------------|-------------------------------|
| NAT2      | rs4664257   | 5-acetylamino-6-amino-3-methyluracil                 | 2.23 × 10⁻⁵⁸               | 0.970      | IG, NFS: rs1801280, p.Gly209Ser | 37              | 7.8                    | 9.6 × 10⁻⁹                   |
| ACADS     | rs4921913   | 5-acetylamino-6-amino-3-methyluracil                 | 2.14 × 10⁻¹²               | 0.499      | p.Gly209Ser             | 28.5             | 2.1 × 10⁻⁹             |
| NAT8      | rs13358     | 2-aminoantranilic                                   | 4.4 × 10⁻⁴⁷               | 0.885      | p.Phe143Ser             | 26.6             | 13.1 × 10⁻⁸            |
| TMPRSS11E | rs34109652  | N-acetylglutamine                                    | 5.5 × 10⁻¹⁶               | -0.781     | INT                     | 21.4             | 3.8 × 10⁻⁶             |
| SLC19A1   | rs4149056   | (dehydroxycorticosterone acid glucuronide or isomer) | 3.28 × 10⁻¹⁶               | -0.737     | INT                     | 21.4             | 3.8 × 10⁻⁶             |
| PYROXD2   | rs9139969   | N-acetylpeptidyl                                      | 3.06 × 10⁻³¹               | 0.833      | p.Val174Ala, p.Met461Thr | 13.8             | 7.4 × 10⁻¹⁹            |
| UGT3A1    | rs3756669   | X-24348                                              | 1.55 × 10⁻²⁵               | -0.915     | p.Cys121Gly              | 16.6             | 7.1 × 10⁻⁶             |
| FADS2     | rs28456     | [1-(1-enyl-palmitoyl+2-arachidonoyl-GPC(P16:0/20:4)] / X-24438 | 9.81 × 10⁻²⁵               | -0.641     | INT                     | 17.8             | 7.1 × 10⁻⁶             |
| AGXT2     | rs37370     | 3-aminoisobutyrate                                    | 9.09 × 10⁻⁴⁴               | 0.457      | INT                     | 8.9              | 1.6 × 10⁻⁸             |
| PHYHD7<sup>f</sup> | rs18185093 | X-22145 (20'-2-arachidonoyl-GPC) / X-24438 | 8.45 × 10⁻²⁵               | 0.810      | p.Ser102Asn, p.Asp141Thr | 13               | 4.3 × 10⁻⁸             |
| THEM4     | rs6690449   | X-18921                                              | 9.35 × 10⁻¹⁰               | -0.547     | INT                     | 14.1             | 7.1 × 10⁻⁶             |
| UGT1A1    | rs2099934   | 1-bilirubin (E, E)<sup>+</sup>                       | 1.15 × 10⁻¹¹               | 0.426      | IG                      | 7.2              | 2.05 × 10⁻⁴             |
| SULT2A1   | rs62199970  | X-11440 (tentatively steroid)                         | 7.59 × 10⁻¹⁷               | -0.951     | IG                      | 11               | 1.4 × 10⁻⁷             |
| SLC22A2d  | rs78176967  | X-22379 (androsterone glucuronide)                    | 1.43 × 10⁻⁵⁶               | 0.995      | IG                      | 10.8             | 7.1 × 10⁻⁶             |
| SPTLC1d<sup>f</sup> | rs2069258 | X-23293 (cis-4-decenoyl carnitine)                            | 1.71 × 10⁻¹⁶               | 0.425      | IG                      | 10.6             | 2.4 × 10⁻⁶             |
| TTC3b<sup>g</sup> | rs17135869 | X-22162                                              | 4.26 × 10⁻¹⁰               | 0.623      | p.Ala12Val              | 10.4             | 7.1 × 10⁻⁶             |
| SLC22A5   | rs274554    | Tryptophan betaine                                    | 8.8 × 10⁻¹⁰               | 0.616      | INT, INS                 | 6.7              | 3.3 × 10⁻⁵             |
| CCBL2<sup>f</sup> | rs7530513 | Imidazole lactate                                     | 5.6 × 10⁻¹³               | 0.425      | INT                     | 8.1              | 3.2 × 10⁻²             |
| SLC7A1    | rs1651961   | X-12824 (hexanoylglutamine)                           | 1.46 × 10⁻¹²               | -0.543     | p.Thr269Ile, p.Lys280Ser | 13               | 8.5 × 10⁻⁸             |
| CYP3A5    | rs776746    | X-12063                                              | 1.54 × 10⁻¹⁰               | -0.620     | SA, c.219-237A>G        | 7.1              | 1.2 × 10⁻¹²            |
| SEMA4A<sup>b</sup> | c1sp90683852 | Undecanidoate                                         | 1.18 × 10⁻¹⁰               | 0.421      | IG                      | 7                | 9.7 × 10⁻⁴             |

<sup>a</sup>pgain was introduced in<sup>19</sup> as “an ad-hoc measure to determine whether a ratio between two metabolite concentrations carries more information than the two corresponding metabolite concentrations alone,” calculated as pgain = min (pval(m1)/pval(m2), pval(m2)/pval(m1)), given two metabolites m1 and m2.

<sup>b</sup>SNP Annotation and Nearest Functional SNP (NFS), function, mutation, NFS refers to intergenic, INT (Intron), INS (Intron near splice), SA (Splice Acceptor)

<sup>c</sup>r2 is percent of variance explained

<sup>d</sup>Reference or Novel Locus (NL) or Novel Association (NAS)

<sup>e</sup>Exact replication SNP is indicated in Supplementary Dataset 4

<sup>f</sup> Newly identified/novel loci

<sup>g</sup> Novelly identified and novel loci

<sup>h</sup> IC50 (missense), bold font indicates functional variant

(continued)
Rare variants associated with metabolites in Qataris. In addition to common-variant mQTL discovery, we hypothesized that with a modest sample size, one may find rare functional variants influencing metabolite levels in homozygous state shared by more than one individual due to high levels of consanguinity. We selected high-quality protein-altering rare variants (MAF < 5%; n = 21,933 SNPs, see Methods). First, we performed gene-based burden testing for all genes harboring at least one of these rare variants (n = 9,823 genes; Bonferroni p ≤ 6.16 × 10^-9); and second, we tested single variant associations for SNPs with at least two homozygotes for the rare variant (n = 2,660 SNPs in 2,119 genes; Bonferroni p ≤ 2.27 × 10^-8). In this analysis, we focused on rare, homozygous variants shared by two or more individuals, whose metabolite values were at either tail of the distribution (highest or lowest). After stringent QC and filtering, we discovered two genes with rare variants influencing metabolite levels through gene-burden analysis. In contrast to burden testing, we identified 10 variants significantly influencing metabolite levels through single-variant testing (Table 3, Fig. 6, Supplementary Fig. 5).

Discussion

We describe the first large-scale (n = 996 individuals) metabolite association study in a Middle Eastern population, combining deep WES data with an updated metabolomics platform. Altogether, we discover and replicate 21 common variant mQTLs. Amongst those, seven are novel or newly identified loci (associations that...
have been concurrently discovered\textsuperscript{15} and completely new ones, as indicated in Table 2), and 10 represent fine-mapping of previously reported loci to new metabolite ratios that associate more significantly than previously discovered single metabolites. Importantly, by having deep exome data, we fine-map a total of 11 loci to candidate protein-altering variants, with biological implications described below. Further, in the rare variant spectrum, we present 12 novel mQTL associations, all not previously reported.

\textit{Fig. 4} Boxplots for the loci NAT2, FADS2, and UGT3A1. Boxplots showing metabolite/ratio levels and number of samples for each genotype group and comparing ratios to single metabolites for NAT2 (a, b), FADS2 (c, d), and UGT3A1 (e, f) loci, where the percent of variance explained by the ratio is 3.9-, 1.99-, and 1.94-fold greater, respectively, than that explained by the single metabolite.

Being able to replicate 35\% of the 80 replicable loci reported in the largest mGWAS study to date\textsuperscript{11} with one tenth the sample size may be attributable to the high levels of consanguinity in our population, which allows for sufficient numbers of alternate allele homozygotes to be present for metabolic associations. Loci replicated among different ethnicities are likely to belong to pathways common to human metabolism (7 out of the 10 most significant loci reported in\textsuperscript{11}—PYROXD2, ACADS, NAT8, 

---

\textbf{ARTICLE NATURE COMMUNICATIONS} | DOI: 10.1038/s41467-017-01972-9 | www.nature.com/naturecommunications
FADS2, SLCO1B1, SULT2A1, and UGT1A1—being among the top associations in our study).

Several of the newly discovered associations may reveal putative biological links between SNPs and metabolites. One such example is the association of rs34109652 in TMPRSS11E with X-11491 (tentatively identified by Metabolon as the bile acid deoxycholic acid glucuronide). This particular SNP had been reported as an eQTL for UGT2B15 (GTEx version 4.1, Build #201), a gene at which CpG methylation was also significantly associated with X-11491 levels27, consistent with the gene’s putative function of conjugating bile acids (GenAtlas(http://genatlas.medecine.univ-paris5.fr/fiche.php?n=6643)). Another biological association is that of the missense variant rs117135869 (p.Ala12Val) in TTC38 (TTC38/PKDREJ locus) associating significantly to X-22162, and more significantly, to the ratio of this metabolite to X-24513. TTC38 (Tetratricopeptide Repeat Domain 38) expression is significantly positively correlated with age28, and levels of both metabolites are also significantly associated with age by regression analysis in our cohort (see Methods). Additionally, X-24513 is significantly partially correlated with C-mannosyltryptophan (also known as C-glycosyltryptophan) (see Methods), which was reported to be associated with aging29. More interestingly, the fact that this locus encompasses PKDREJ (Polycystic Kidney Disease and Receptor for Egg Jelly related protein) gene (regional association plot—Supplementary Fig. 3), a homolog of PKD genes associating with kidney disease30, and that X-24513 might be similar in characteristics to C-mannosyltryptophan, reportedly elevated in chronic kidney disease31, highlights the importance of investigating the biological implication of this association in relation to both kidney diseases and ageing.

Finally, the association of rs7530513 in the aminotransferase CCBL2 (Cysteine-S-conjugate beta-lyase 2, also known as KYAT3

Fig. 5 Regional association plots for the loci UGT3A1 and ACADS. UGT3A1 (a) and ACADS (b) loci missense SNPs showing the strength of the association (−log10 (p-value)) for X-24348/pregn-sterol-monosulfate* and ethylmalonate, respectively, on the Y-axis and the genes on the X-axis. The colors correspond to different LD thresholds, where LD is computed between the sentinel SNP (lowest p-value, colored in blue) and all SNPs. Shapes of markers correspond to their functionality as described in the legend
(Kynurenine–oxoglutarate transaminase 3) with imidazole lactate is another biologically relevant association; CCBL2 enzyme [EC2.6.1.7] has a transaminase activity towards histidine (Uniprot) and it has been reported that histidine transaminase [EC2.6.1.38], takes L-histidine and 2-oxoglutarate as substrates and produces imidazole pyruvate that is later converted to imidazole lactate. However, we are unable to confirm that the same histidine transaminase is the one involved in both processes.

A main advantage of using WES data in an association analysis is the enhanced resolution at associated loci that could reveal protein-altering variants affecting metabolite levels. We were able to identify such sentinel SNPs for over half of all loci (11 of 21), possibly revealing functional relationships between genes harboring these variants and their associated metabolites. One such example is the fine-mapping of a previously identified signal in the 3′UTR of ACADS to a missense mutation (rs1799958, p.Gly209Ser) in exon 6 (Fig. 5b). This mutation (MIM 608885.0007) has been reported independently as causing mild SCAD deficiency, with the mutant allele predicted to have 85% of wild-type activity. Further, the population-specificity analysis is the enhanced resolution at associated loci that could reveal protein-altering variants affecting metabolite levels. Notably, the mutation appears to have population-specific allele distribution, having been observed in up to 40% of Asians, 15% of Europeans, but only 3% of Africans—a distribution which is maintained in our study cohort (overall Qatari MAF 12.8%), being present in Q1 and Q2 Qataris (Bedouin and Persian/South Asian), but not in Q3, who are of African descent.

In addition to identifying functional SNPs, the use of the updated metabolomics platform (DiscoveryHD4) enabled the refinement of previously reported loci to new metabolites or metabolite ratios in the same or in different biological pathways from those originally reported. For example, NAT2 encodes an arylamine transferase that controls the conversion of paraxanthine to 5-acetylamino-6-aminomethyluracil, and its locus has previously been reported to be associated with 1-methylxanthine. In this study, we discovered a stronger association with the ratios 5-acetylamino-6-aminomethyluracil/1-methylxanthine and 5-acetylamino-6-aminomethyluracil-paraxanthine, which are consistent with the known activity of the NAT2 enzyme in caffeine metabolism (see network of metabolites linked to NAT2, and KEGG pathway in Supplementary Fig. 4b, c). Being the top most significant association in this population in comparison to the top most ones previously reported, it might have several biological implications for the studied population. It is thus worthy to note that the two missense SNPs at this locus are identified in OMIM as associated with the rate of acetylation, presenting a possible mechanism for affecting drug metabolism.

### Table 3 12 novel functional rare variant mQTLs

| Gene-based burden test | SNP name (c#chr p#position) | rsID | Gene | Metabolite | EA/OA | C-MAF | p-value |
|------------------------|-----------------------------|------|------|------------|-------|-------|---------|
| c4p57221348            | rs3179958, p.Gly209Ser       | 11   | 21   | 2.6        | 0.036 | 4.1 x 10^-09 |
| c4p57248716            | rs34228795                  | 12   | 15   | 1.88       | 0.102 | 2.1 x 10^-10 |
| c4p57250285            | rs34543011                  | 6    | 6    | 1.88       | 0.102 | 2.1 x 10^-10 |
| c12p56075599           | rs21995109                  | 11   | 11   | 1.88       | 0.102 | 2.1 x 10^-10 |
| c12p56075915           | rs75289684                  | 11   | 11   | 1.88       | 0.102 | 2.1 x 10^-10 |
| c12p56077768           | rs11568788                  | 11   | 11   | 1.88       | 0.102 | 2.1 x 10^-10 |

| Single-variant analysis | SNP name (c#chr p#position) | rsID | Gene | Metabolite | EA/OA | EAF | N | Beta (s.e.) | p-value |
|-------------------------|-----------------------------|------|------|------------|-------|-----|---|-----------|---------|
| c12p11506141             | rs510988100                  | 12   | 12   | 1.88       | 0.102 | 2.1 x 10^-10 |
| c15p89939112             | rs56332208                  | 12   | 12   | 1.88       | 0.102 | 2.1 x 10^-10 |
| c16p81876624             | rs150988100                  | 12   | 12   | 1.88       | 0.102 | 2.1 x 10^-10 |
| c18p48256030             | rs37730168                   | 12   | 12   | 1.88       | 0.102 | 2.1 x 10^-10 |
| c20p18295959             | rs73730168                   | 12   | 12   | 1.88       | 0.102 | 2.1 x 10^-10 |
| c11p18158958             | rs61733595                   | 12   | 12   | 1.88       | 0.102 | 2.1 x 10^-10 |

EA effective allele, OA observed allele, EAF effective allele frequency, C-MAF cumulative minor allele frequency
All SNPs have a call rate of 100% except for the SNP marked (*), which has a call rate of 97%
metabolism; specifically, slow acetylation was reported for individuals harboring either of the two missense SNPs rs1799930 and rs1801280 (MIM: 612182.0001 and 612182.0002). Additionally, we discovered associations between SLC17A1 and the ratio hexanoylglutamine/X-16087. Previously, rs1185567 in SLC17A3 was reported to be associated with steroid levels. However, since the SLC17 family are vesicular glutamate transporters, our association signal, linking the missense variant p.Thr269Ile (rs1165196) to a ratio containing hexanoylglutamine may reflect a direct biological relationship. We also discovered mQTLs associating known loci to metabolites in pathways different from those previously reported. Once such example is the association between the missense variant rs2147896 (p Met461Thr) in PYROXD2 with levels of N-methylpyrocollate. This

Fig. 6 Boxplots for the rare variant loci AASDH, PRB1, ACAN, and OTOF, indicating the metabolite level and the number of samples for each genotype group. Boxplots of rare variants associations of AASDH with thyroxine (3 SNPs) (a-c), PRB1 with mannose (d), ACAN with X-12844 (e), and OTOF with retinol (vitamin A) (f)
gene had previously been linked to levels of urinary trimethylamine and dimethylamine, yet interestingly not found significantly associated with any of the mentioned metabolites in previous large scale studies. Finally, our study expands the associations of several loci with newly identified metabolites or ratios as in the NAT2 locus and FADS2 locus among others (Supplementary Data 5). One example that supports the functional importance of the discovered loci is the SLC01B1 locus. Its sentinel SNP, rs4149056, was previously associated with the ratio isoleucine/X-11529 and other unknowns, and found in our study to be associated with the bile acids glycochenodeoxycholate-glucuronide and glycochenolate. That supports the functionality of SLC01B1 as a bile acid transporter, suggesting that the mutation p.Val174Ala alters the gene’s native function.

In addition to common mQTLs, we discovered a total of 12 rare mQTLs—10 by single-variant analysis and 2 by gene-burden analysis—with interesting biological implications (Table 3 and Supplementary Fig. 5). First, the ACAN gene encodes a major component of the extracellular matrix, lending important biochemical properties of cartilage, which explains its role in diseases such as osteoarthritis (MIM: 155760). We discovered an association of a missense variant p.Ala766Thr (rs150988100) in ACAN with a steroid (X-12844, tentatively identified by Metabolon as glucurononlated steroid); it is well-established that circulating steroid levels are linked to inflammation in joints and diseases such as arthritiss. Additional investigations would be needed to uncover the relation between the variant and the metabolite since one of the two subjects having this mutation have musculoskeletal problems. Similarly, the association of rs377301648 in ZNF133 with a 3-methyl-2-oxovalerate, a metabolite in the branched-chain amino acid pathway might be due to the role of the gene in osteoblast differentiation and the previously reported involvement of 3-methyl-2-oxovalerate in osteoarthritis. Additionally, association of the c12p11506114 mutation in PRB1 with mannose (Fig. 6) might be due to the involvement of this glycosylated, proline-rich protein in the salivary secretion pathway (Michael W King, PhD 1996–2016 themedicalbiochemistrypage.org).

Another interesting rare variant association is that between rs56332208 in OTOF with retinol (vitamin A). This might be a disease-relevant link because retinoic acid mediates the regeneration of specialized mechano-sensory hair cells in the inner ear that capture auditory and balance sensory inputs, and which die after acoustic trauma, ototoxic drugs or aging diseases, leading to progressive hearing loss. Previous studies have also described mutations in OTOF causing recessive neurosensory non-syndromic deafness in patient cohorts from many different ethnicities via the gene’s role in exocytosis of inner and outer hair cells. Thus, our study may provide a mechanistic link between this gene and hair cell development via modulation of retinol levels. Finally, burden-testing revealed mQTLs in AASDH and METTL7B with decreased levels of thyroxine and andosterone sulfate respectively. The AASDH gene plays an active role in the pipecolate pathway in which thyroxine is a by-product. Thus, these mutations may impact protein function leading to significant reduction of thyroid hormone and metabolism. Regarding the METTL7B mQTL, one of the three mutations in this gene (rs115687886) is a nonsense mutation, suggesting the other two are also loss-of-function mutations as they influence the metabolite in the same direction. Increases in METTL7B expression had been previously observed in patients with acute respiratory distress syndrome, involving tissue injury and inflammation, whereas adrenal androgens as andosterone sulfate have been observed to decrease in stress and inflammation. The fact that one of the two subjects with the rs199581976 mutation had recurrent chest infections and pneumonia and the other subject had bronchial asthma, recurrent pneumonia, and right lung lower lobe collapse, supports the functionality of the gene, and thus provides another possible link between a rare variant and disease via metabolite-level modulation.

To summarize, our study revealed 21 mQTLs in Qataris, among which 7 are unknown to studies that used older metabolomics platforms, 10 are metabolically fine-mapped with new metabolite ratios, and 11 in which the sentinel SNP was at or in complete LD with protein-altering variants. We also discovered 12 novel functional rare variant mQTLs that are likely specific to this ethnic population. We believe that it is important to replicate rare variant associations, yet the low frequency of these variants and absence of similar cohorts makes it challenging. This study demonstrates the efficiency of using WES for mQTL discovery, which could be more convenient compared to the more expensive WGS, while providing deep coverage of protein coding variants, for fine mapping of common mQTLs and being suitable for rare variant mQTL discovery. The use of WES is strongly supported by the ability to replicate a fair fraction of the mQTLs in Caucasians (19% of all loci and 35% of replicable loci). Finally, this study is the largest and the first of its kind in a Middle Eastern population, where we show that studies in consanguineous populations offer a large insight with modest sample sizes and have the potential to reveal loci linked to disease and pharmacological/drug targets important to precision medicine in this region of the world.

Methods

Study cohort. Human subjects were recruited and written informed consent was obtained at Hamad Medical Corporation (HMC) and HMC Primary Health Care Centers in Doha, Qatar and approved by the Institutional Review Boards of Hamad Medical Corporation and Weill Cornell Medicine in Qatar. Briefly, a total of 614 subjects were recruited for exome sequencing, and another 382 subjects for array genotyping. Subjects were included on the basis of being 3-generation Qataris (four grandparents born in Qatar) and being healthy or diabetics, above the age of 30 years. Sample demographic characteristics are given in Table 1.

Exome sequencing. DNA of 614 subjects was extracted from blood using the QIAamp DNA Blood Maxi Kit (Qiagen Sciences Inc, Germantown, MD) and subjected to exome sequencing on the Illumina HiSeq 2000 platform using standard methods. Each subject was sequenced to a minimum mean depth of 70X. Genotyped were generated using the GATK Best Practices workflow (Whole exome sequencing data for a set of recently received 64 samples were merged with those of the 550 samples after calling variants using GATK on each and finding that the majority of SNPs detected in both sets overlap. We limited the analysis to variants in both sets by setting a genotype call rate of 98% for common variant analysis and 90% for rare variant analysis). Detailed preparation methods and genotypes are available in. The DNA of 382 subjects was extracted and subjected to array genotyping using Illumina Omni 2.5M array kits.

Imputation and filtering. Exome and array data were imputed after being filtered (MAF ≥ 0.05, R2 > 0.6, genotype call rate >98%) based on phased 108 Qatari whole genomes as a reference panel, using shapeit2 and Impute2 software packages. Throughout all the manuscript and supplementary data sets/information, two unique notations of “c<sup>y</sup>p<sup>y</sup>” and “<sup>fl</sup>” for indicating a chromosome, position pair (for example c1p100 and 1:100) are used to distinguish original finding that the majority of SNPs detected in both sets overlap. We limited the analysis to variants in both sets by setting a genotype call rate of 98% for common variant analysis and 90% for rare variant analysis). Detailed preparation methods and genotypes are available in. The DNA of 382 subjects was extracted and subjected to array genotyping using Illumina Omni 2.5M array kits.

Metabolomics data. Serum samples were prepared for metabolomics analysis as follows: 200 μL of serum were aliquoted, barcoded and transported on dry ice to...
Metabolon Inc. for analysis. Some samples were found hemolyzed, and the degree of hemolysis was recorded for each sample (according to a hemolysis chart as given by Metabolon Inc.), for correcting its effect on metabolomics measurements. The Metabolon DiscoveryHD4 platform was used (details are given in Supplementary Note 1). This platform utilized a Waters ACQUITY ultra-performance liquid chromatography (UPLC) and a Thermo Scientific Q-Exactive high resolution/accurate mass spectrometer interfaced with a heated electrospray ionization (HESI II) source and Orbitrap mass analyzer (for correcting metabolite levels for covariates and kinship matrix, and regression function in GenABEL). A total of 1303 metabolites were measured on that platform. Outlier metabolite measurements (3 standard deviations larger than the mean) were replaced by a missing value to avoid biasing the results. Metabolites with more than 20% missing values were removed from the data. A total of 249 unknown metabolites survived as high quality and observed in >80% of all individuals. Metabolite measurements were log-scaled and z-score normalized. Since samples were collected over different periods of time, and there was no adopted fasting criterion, we addressed that issue by investigating whether there are batch effects in the PCA of metabolic data, and did not find any grouping for any of the 614 exome samples or the 382 array samples. Several unknown metabolites were investigated by the aid of Metabolon, and which revealed their identities (see below).

Identification of unknown metabolites by Metabolon. Identification of tentative structural features for unknown biochemicals incorporates a detailed analysis of mass spec data, i.e., gathering information such as the accurate monoisotopic mass, the elution time, and fragmentation pattern of the primary ion, and correlation to other molecules. The accurate monoisotopic mass is used to identify a likely structural formula for the unknown biochemical, which is then used to search against chemical structure databases (e.g. ChemSpider, SciFinder). When a candidate structure fits the accurate monoisotopic mass and fragmentation data, an authentic standard is commercially purchased or synthesized (when possible). Conformation of a proposed structure is based on a match to three primary criteria, including co-elution with the unknown molecule of interest, and a high degree match to both the accurate monoisotopic mass and fragmentation pattern. When a standard is not available to confirm the identity of the unknown biochemical, we seek high degree of confidence in its identity, the unknown biochemical may be retired for a named structure that is differentiated from other named metabolites confirmed with authentic standards by the addition of an asterisk after the biochemical name. When a high degree of confidence in the identity of the unknown molecule is not obtained, the molecule retains its unknown status in the molecule by unknowns list. Refer to Supplementary Note 2 for the particular identification information on each of the identified unknown metabolites, and Supplementary Fig. 6 for the comparison of MS² fragmentation spectrum and Extracted Ion Chromatogram of the candidate and unknown metabolite that were structurally confirmed by Metabolon.

mGWAS analysis. All associations were performed on imputed exome or imputed array genotype data. P-values and effect sizes were calculated using functions from both GenABEL and ProbABEL packages in R (version 3.1.2) that were used for computing associations between SNPs and metabolite/ratio levels (after being pre-processed as described in the previous sections) while correcting for age, gender, computing associations between SNPs and metabolite/ratio levels (after being pre-processed as described in the previous sections) while correcting for age, gender, environmental and population stratification (for inclusion of ethnicity) for all 996 individuals using lm function in R. Bonferroni p-value of a significant association with a phenotype (p ≤ 0.05/826). Partial correlations between two metabolites are calculated using the GeneNet package in R, and significant partial correlations are those which pass a Bonferroni p-value of (p ≤ 0.05/826 × 25/2). Rare variants analysis. High-quality rare variants were selected from the original non-imputed exome data, with MAF ≤ 5%, p<0.01, and genotype call rate ≥ 90%. Single-variant analysis was done using GenABEL package in R (as mentioned above for common variants), and burden test was done using seqMeta package in R (as in ref. 69,64). For burden tests, we accounted for the number of genes (n = 9823 genes; Bonferroni p ≤ 5.6 × 10⁻⁷), and for single-variant analysis, we accounted for the number of SNPs where we tested single-variant associations for SNPs with at least two homozygotes for the rare variant (n = 2660 SNPs in 2119 genes; Bonferroni p ≤ 2.27 × 10⁻⁶). All SNPs in the reported 12 rare variant mQTLs have a call rate of 100% except for c12p11506114 with a call rate of 97%. A filtering step was done by visual inspection of genotype-metabolite boxplots of significant associations to consider only associations where metabolite values of the minor homozygotes are at the extreme tail of the metabolite distribution (i.e., they should be the lowest or highest metabolite values for that metabolite).

URLs of databases used for annotation. GTEx portal (version 2.1, Build #201) [www.gtexportal.org], OMIM diseases database [www.omim.org], Orphanet disease database [http://www.orpha.net/CHEM/ targets_disease.php?chembl=PharmaGKB [www.pharmgkb.org], SNIPA [http://snipa.helmholtz-muenchen.de/snipa/], GenAtlas [http://genatlas.medecine.univ-paris5.fr/fiche.php?n=6643], and GenOMAD: [http://genomad.broadinstitute.org/].

Data availability. All information on metabolites are in Supplementary Data 1, expanded association results for 21 common variant mQTLs are in Supplementary Data 5, and the detailed biological annotations/interpretations of associations are in Supplementary Data 6, list of all associations (p < 1.4 × 10⁻¹⁰) are in Supplementary Data 7. All plots are also available in Supplementary Information and detailed information on Metabolon’s method for identification of unknowns is available in the Supplementary Note. Exome data used in this project were selected from a pool of four samples of which more than 1000 samples are deposited in SRA accession SRP060765, SRP061943, and SRP061463, accessible online at http://www.ncbi.nlm.nih.gov/Taxonomy/study?acc=SRP060765%2C%SRP061943%2C%SRP061463&go=). (SRA accession SRP061943).

Received: 9 May 2017 Accepted: 30 October 2017 Published online: 23 January 2018
References
1. Youssi, N. A. et al. Long term conservation of human metabolic phenotypes and link to heritability. *Metabolomics* 10, 1005–1017 (2014).
2. Beger, R. D. et al. Metabolomics enables precision medicine: “A white paper, community perspective”. *Metabolomics* 12, 149 (2016).
3. Gieger, C. et al. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet.* 4, e1000282 (2008).
4. Illig, T. et al. A genome-wide perspective of genetic variation in human metabolism. *Nat. Genet.* 42, 137–141 (2010).
5. Nicholson, G. et al. A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. *PLoS Genet.* 7, e1002270 (2011).
6. Suhre, K. et al. A genome-wide association study of metabolic traits in human urine. *Nat. Genet.* 43, 565–569 (2011).
7. Draisma, H. H. et al. Genome-wide association study identifies novel genetic variants contributing to variation in blood lipid levels. *Nat. Commun.* 6, 7208 (2015).
8. Kastenmüller, G., Rafter, J., Gieger, C. & Suhre, K. Genetics of human metabolism: an update. *Hum. Mol. Genet.* 24, R93–R101 (2015).
9. Yet, I. et al. Genetic influences on metabolite levels: a comparison across metabolic phenotypes. *PLoS ONE* 11, e0153672 (2016).
10. Kettunen, J. et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat. Commun.* 7, 11122 (2016).
11. Shin, Y. S. et al. An atlas of genetic influences on human blood metabolites. *Nat. Genet.* 46, 543–550 (2014).
12. Demirkan, A. et al. Insight in genome-wide association of metabolic quantitative traits by exome sequence analyses. *PLoS Genet.* 11, e1004835 (2015).
13. Guo, L. et al. Plasma metabolomic profiles enhance precision medicine for volunteers of normal health. *Proc. Natl Acad. Sci. USA* 112, E4901–E4910 (2015).
14. Yazdani, A., Yazdani, A., Liu, X. & Boerwinkle, E. Identification of rare variants in metabolites of the carnitine pathway by whole genome sequencing analysis. *Genet. Epidemiol.* 40, 486–491 (2016).
15. Long, T. et al. Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nat. Genet.* 49, 568–578 (2017).
16. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* 538, 161–164 (2016).
17. Rafter, J. et al. Genome-wide association study with targeted and non-targeted nmr metabolomics identifies 15 novel loci of urinary human metabolic individuality. *PLoS Genet.* 11, e1005487 (2015).
18. Ruedi, R. et al. Genome-wide association study of metabolic traits reveals novel gene-metabolite disease links. *PLoS Genet.* 10, e1004312 (2014).
19. Petersen, A. K. et al. On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies. *BMC Bioinformatics.* 13, 120 (2012).
20. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291 (2016).
21. McKusick, V. A. Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.* 80, 588–604 (2007).
22. Whirl-Carrillo, M. et al. Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* 92, 414–417 (2012).
23. Fakhro, K. A. et al. Rare copy number variations in congenital heart disease patients identify unique genes in left-right patterning. *Proc. Natl Acad. Sci. USA* 108, 2915–2920 (2011).
24. Jian, H., Liu, B. & Zhang, J. Hypoxia and hypoxia-inducible factor 1 repress SEMA4B expression to promote non-small cell lung cancer invasion. *Tumour Biol.* 35, 4949–4955 (2014).
25. Lang, I. C. & Schuller, D. E. Differential expression of a novel serine protease homologue in squamous cell carcinoma of the head and neck. *Br. J. Cancer* 84, 237–243 (2001).
26. Mackenzie, P. I. et al. Identification of UDP glycosyltransferase 3A1 as a UDP-N-acetylgalactosaminyltransferase. *J. Biol. Chem.* 283, 36205–36210 (2008).
27. Petersen, A. K. et al. Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits. *Hum. Mol. Genet.* 23, 534–545 (2014).
28. van den Akker, E. B. et al. Meta-analysis on blood transcriptomic studies identifies consistently coexpressed protein-protein interaction modules as robust markers of human aging. *Aging Cell* 13, 216–225 (2014).
29. Menii, C. et al. Metabolomic markers reveal novel pathways of ageing and early development in human populations. *Int. J. Epidemiol.* 42, 1111–1119 (2013).
30. Veldhuisen, B., Spruit, L., Daauwser, H. G., Breming, M. H. & Peters, D. J. Genes homologous to the autosomal dominant polycystic kidney disease genes (PKD1 and PKD2). *Eur. J. Hum. Genet.* 7, 860–872 (1999).
61. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* 83, 311–321 (2008).

62. Yousri, N. A. et al. Large scale metabolic profiling identifies novel steroids linked to rheumatoid arthritis. *Sci. Rep.* 7, 9137 (2017).

Acknowledgements
This work was supported by Biomedical Research Program (BMRP) in Weill Cornell Medicine-Qatar, funded by Qatar Foundation. It was also supported by Qatar National Research Funds (QNRF) grants numbers: 09-740-3-192, 09-741-3-793 and in part by grant number 7-272-1-041. We acknowledge the efforts of Greg Schaaf (Metabolon) for synthesizing standards used in the identification of structurally unknown molecules described in this manuscript.

Author contributions
N.A.Y., K.A.F., K.S. and R.G.C. conceived the study. N.A.Y. and K.A.F. led the study, interpreted genetic associations and wrote the paper. N.A.Y. performed all statistical and computational analysis, imputation, compilation and annotation of results, and interpretation of metabolic associations. K.A.F., J.L.R.-F., A.R., J.G.M. and R.G.C. provided all genotype and phenotype information for the Qatari subjects. R.P.M. provided Metabolon support for interpretation of unknown metabolites. A.A.-S. and O.M.C. performed sample collection and handling. S.A.K., T.O. and H.Z. aliquoted samples for Metabolon, and performed DNA extraction and processing in the genomic core. J.A.M., Y.A.M. and E.K.A. were responsible for the Genomic core for Illumina genotyping and NGS. G.T. provided pipeline tools for imputation and provided phased reference genome. M.K. provided plotting tools and aided with plotting results.

Additional information
Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-017-01972-9.

Competing interests: R.P.M. is an employee of Metabolon Inc. The remaining authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2017