Genetic architecture of vitamin B12 and folate levels uncovered applying deeply sequenced large datasets
Grarup, Niels; Sulem, Patrick; Sandholt, Camilla H; Thorleifsson, Gudmar; Ahluwalia, Tarun Veer Singh; Steinhorsdottir, Valgerdur; Bjarnason, Helgi; Gudbjartsson, Daniel F; Magnusson, Olafur T; Sparso, Thomas; Albrechtsen, Anders; Kong, Augustine; Masson, Gisli; Tian, Geng; Cao, Hongzhi; Nie, Chao; Kristiansen, Karsten; Husemoen, Lise Lotte; Thuesen, Betina; Li, Yingrui; Nielsen, Rasmus; Linneberg, Allan René; Olafsson, Isleifur; Eyjolfsson, Gudmundur I; Jorgensen, Torben; Wang, Jun; Hansen, Torben; Thorsteinsdottir, Unnur; Stefansson, Kari; Pedersen, Oluf Borbye

Published in:
P L o S Genetics

DOI:
10.1371/journal.pgen.1003530

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Abstract

Genome-wide association studies have mainly relied on common HapMap sequence variations. Recently, sequencing approaches have allowed analysis of low frequency and rare variants in conjunction with common variants, thereby improving the search for functional variants and thus the understanding of the underlying biology of human traits and diseases. Here, we used a large Icelandic whole genome sequence dataset combined with Danish exome sequence data to gain insight into the genetic architecture of serum levels of vitamin B12 (B12) and folate. Up to 22.9 million sequence variants were analyzed in combined samples of 45,576 and 37,341 individuals with serum B12 and folate measurements, respectively. We found six novel loci associating with serum B12 ([CD320, TCN2, ABCD4, MMAA, MMAPCH] or folate levels ([FOLR3]) and confirmed seven loci for these traits ([TCN1, FUT6, FUT2, CUBN, CLYBL, MUT, MTHFR]). Conditional analyses established that four loci contain additional independent signals. Interestingly, 13 of the 18 identified variants were coding and 11 of the 13 target genes have known functions related to B12 and folate pathways. Contrary to epidemiological studies we did not find consistent association of the variants with cardiovascular diseases, cancers or Alzheimer’s disease although some variants demonstrated pleiotropic effects. Although to some degree impeded by low statistical power for some of these conditions, these data suggest that sequence variants that contribute to the population diversity in serum B12 or folate levels do not modify the risk of developing these conditions. Yet, the study demonstrates the value of combining whole genome and exome sequencing approaches to ascertain the genetic and molecular architectures underlying quantitative trait associations.


Editor: Scott M. Williams, Dartmouth College, United States of America

Received December 20, 2012; Accepted April 11, 2013; Published June 6, 2013

Copyright: © 2013 Grarup et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This project was funded by the ENGAGE project (HEALTH-F4-2007-201413) and by the Lundbeck Foundation (The Lundbeck Foundation Centre for Applied Medical Genomics in Personalised Disease Prediction, Prevention and Care (LuCamp), www.lucamp.org). The Novo Nordisk Foundation Center for Basic Metabolic Research is an independent Research Center at the University of Copenhagen partially funded by an unrestricted donation from the Novo Nordisk Foundation (www.metabol.ku.dk). Further funding came from the Danish Council for Independent Research (Medical Sciences). The Inter99 study was financially supported by grants from the Velux Foundation and the Danish Medical Research Council. Danish Agency for Science, Technology and Innovation: The Aase and Ejner Danielsens Foundation; ALK-Abelló A/S, Harsholm, Denmark, and Research Centre for Prevention and Health, the Capital Region of Denmark. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: I have read the journal’s policy and declare that the authors from deCODE Genetics (P. Sulem, G. Thorleifsson, V. Steinthorsdottir, H. Bjarnason, D. F. Gudbjartsson, O. T. Magnusson, A. Kong, G. Masson, U. Thorsteinsdottir, and K. Stefánsson) are employees of deCODE Genetics or own stock options in deCODE Genetics. The remaining authors have declared that no competing interests exist.

* E-mail: kstefan@decode.is (K. Stefánsson); oluf@sund.ku.dk (O. Pedersen)

† These authors contributed equally to this work.
Author Summary

Genome-wide association studies have in recent years revealed a wealth of common variants associated with common diseases and phenotypes. We took advantage of the advances in sequencing technologies to study the association of low frequency and rare variants in conjunction with common variants with serum levels of vitamin B\textsubscript{12} (B\textsubscript{12}) and folate in Icelanders and Danes. We found 18 independent signals in 13 loci associated with serum B\textsubscript{12} or folate levels. Interestingly, 13 of the 18 identified variants are coding and 11 of the 13 target genes have known functions related to B\textsubscript{12} and folate pathways. These data indicate that the target genes at all of the loci have been identified. Epidemiological studies have shown a relationship between serum B\textsubscript{12} and folate levels and the risk of cardiovascular diseases, cancers, and Alzheimer’s disease. We investigated association between the identified variants and these diseases but did not find consistent association.

Introduction

One-carbon metabolism (OCM) is a process whereby folate transfers one-carbon groups in a range of biological processes including DNA synthesis, methylation and homocysteine metabolism [1,2]. The water-soluble B vitamins, vitamin B\textsubscript{12} (B\textsubscript{12}) and folate play key roles as enzyme cofactors or substrates in OCM. Individuals with deficiencies in these vitamins can develop anemia. The water-soluble B vitamins, vitamin B\textsubscript{12} (B\textsubscript{12}) and folate play key roles as enzyme cofactors or substrates in OCM. Individuals with deficiencies in these vitamins can develop anemia. Individuals with subclinical B\textsubscript{12} or folate deficiencies are associated with increased levels of B\textsubscript{12} and folate are in addition to nutrition influenced by several biological processes including absorption, transportation and cellular uptake, as well as processing of precursors into active molecules. Heritability, utilizing di- and monozygotic twins, is estimated to be 59% and 56% for B\textsubscript{12} and folate levels, respectively, indicating that there is a substantial genetic component to the population diversity in these physiological variables [8].

Identification of sequence variants that affect circulating levels of B\textsubscript{12} and folate can thus give insights into the interplay of diet, genetics and human health. Genome-wide association studies (GWAS) have yielded some sequence variants influencing B\textsubscript{12} levels [9–12], but have been less successful in identifying variants affecting folate levels [10,11]. Thus, genome-wide significant associations with serum B\textsubscript{12} levels have been convincingly reported for four loci, FUT2, MUT, CUBN and TCN1 in European populations [9–11] and additional four loci, MS4A3, CLYBL, FUT6 and 5q32 in a Chinese population [12]. No genome-wide significant GWAS associations have been reported for serum folate levels, however, significant association with the MTHFR A222V variant was demonstrated prior to the GWAS era [13,14] and suggestive associations have been reported in European populations for two loci (FIGV and PRICKLE2) [10,11].

The classic GWAS applied commercial chip-based genotyping and imputation of HapMap variants of which a majority were common single nucleotide variants (SNVs) with very few rare variants with minor allele frequency (MAF) <1% [15,16]. However, the search for the truly associated functional variants and the targeted gene at each locus has been hindered by the lack of coverage of the full spectrum of the sequence variation of the human genome. Recently, focus has turned to the use of next generation sequencing of whole genomes (WGS) [17], exomes (WES) [18] or specific targets [19], all contributing to a better understanding of the spectrum of allelic variations in the human genome. We expect that attempts to directly cover low frequency and rare sequence variants through next generation sequencing, in addition to the common variants, will improve the search for functional variants and thus the understanding of the underlying biology of human traits and diseases.

Here we aimed to identify and characterize associations of SNVs across the allele frequency spectrum with serum levels of B\textsubscript{12} and folate by compiling data in up to 45,576 individuals based on sequencing initiatives in Iceland and Denmark. For the first time, we apply next generation sequence data to identify sequence variants affecting serum levels of B\textsubscript{12} and folate and the present datasets are the largest utilized to date for the analysis of these traits.

Results

Heritability of serum B\textsubscript{12} and folate levels

We estimated the heritability of B\textsubscript{12} and folate serum levels based on 38,229 and 21,708 Icelandic sibling pairs, respectively. Our analysis revealed estimates of 27% for B\textsubscript{12} and 17% for folate which are lower than previously reported [8].

Experimental design

To search for sequence variants affecting serum B\textsubscript{12} and folate levels we compiled data from two sequencing initiatives in Iceland and Denmark. In Iceland, a large population-based resource has been generated applying WGS and highly accurate imputation of the sequence information into a large fraction of the population [20,21]. Utilizing this resource many low frequency and rare causative sequence variants have recently been discovered that affect the risk of common diseases [22–26]. In the Danish samples, WES was used to search for low frequency variation associated with complex traits [27,28]. The outline of the present study is depicted in Figure 1. In the Icelandic study sample, 1,176 individuals were whole genome sequenced to an average depth of >10 \times 22.9 million SNVs were identified. These variants were then imputed into 25,960 and 20,717 chip-genotyped Icelanders with serum B\textsubscript{12} and folate measurement, respectively, using highly accurate long-range phasing based imputation [20]. The Icelandic genealogical database allowed for further propagation of the sequence information, applying genealogy based imputation, into 11,323 and 8,196 relatives of the chip-genotyped individuals, for a total sample size of 37,283 and 29,913, respectively, for the two phenotypes [25] (Text S1 and Table S1). In the Danish part of the study whole exomes of 2,000 Danes were sequenced to an average sequencing depth of 8 \times 28. From that effort, 16,192 coding SNVs with allelic frequency above 1% were selected for Illumina iSelect genotyping in two Danish population-based cohorts of 8,293 individuals with measurements of serum B\textsubscript{12} and 8,428 individuals with measurement of serum folate (Table S2). Of the 16,192 SNVs, 15,994 overlapped with the Icelandic variants.

A generalized form of linear regression was used to test for association of serum levels of B\textsubscript{12} or folate with SNVs, taking into account relatedness and population stratification within each sample set, applying the method of genomic control (GC). Analyses were performed in three steps; sequence variants were analyzed in the Icelandic and Danish samples separately, then by combining in a meta-analysis the overlapping sequence variants identified in both study samples. Loci that associated significantly with B\textsubscript{12} or folate levels from these studies were fine mapped using the Icelandic WGS data imputed into chip genotyped individuals and the same data set was used to identify additional signals at...
each of these loci through conditional analysis. Finally, the full Icelandic data of 22.9 million SNVs were used in GWAS to identify additional loci represented by non-coding variants or rare coding signals not genotyped in the Danish design. Genome-wide significance (GWS) level in the study was set at $P = 2.26 \times 10^{-9}$, based on Bonferroni correction for the 22.9 million SNVs (Figure 1).


doi:10.1371/journal.pgen.1003530.g001

Figure 1. Schematic overview of the study.

Discovery analyses for serum B12 and folate

In the separate and combined analyses of SNVs with serum B12 and serum folate levels in the Icelandic and Danish data, a total of 13 genetic loci were found to associate at GWS, $P < 2.2 \times 10^{-9}$ (Table 1 and 2, Figure S1 and S2). Of the 11 loci associated with serum B12, five (CD320, TCN2, ABCD4, MMAA and MMACHC) were novel and six were previously reported either in populations of European or East-Asian ancestry [9–12] (Table 1). Association analyses with serum folate yielded one novel locus (FOLR3) and confirmed the reported MTHFR locus (Table 2).

Since only coding variants were in the combined analysis we used the Icelandic WGS-based data to screen for stronger non-coding signals at the loci identified in meta-analysis of coding variants. Interestingly, the strongest signal at 10 of the 11 B12-associated loci in the Icelandic data corresponded to missense ($n = 9$) or nonsense ($n = 1$) mutations with only the FUT6 locus having a stronger non-coding signal (rs708686) than the missense P124S mutation (Table S3). As only SNVs had been called from the WGS data and imputed into the Icelandic samples we reassessed each of the 13 B12 and folate loci with INDEL data called using the GATK algorithm (http://www.broadinstitute.org/gatk/). None of the INDELs detected at the 11 B12 loci associated more strongly than the lead SNVs. However, when reassessing each of the two folate-associated loci we detected a two nucleotide insertion (rs139130389, NM_000804: exon3:c.318_319insTA) encoding a common (MAF 10.0%) frameshift mutation in exon 3 of FOLR3, that associated more strongly with folate levels than the intronic SNV rs652197 identified in the initial scan (rs139130389: $P = 2.45 \times 10^{-12}$, effect = 0.087 SD, Table 2). The insertion and rs652197 are in linkage disequilibrium (LD) in the Icelandic sequencing data ($r^2 = 0.51$). Upon further inspection, we found that the ancestral sequence contained the insertion indicating the occurrence of a two base deletion in humans. The deletion with an allelic frequency of 90% in Iceland creates a premature stop codon at amino acid position 107 compared to the full-length protein consisting of 245 amino acids. Coding variants are thus lead signal of both folate loci (FOLR3 and MTHFR).

The lead SNVs included both rare, low frequency and common variants with MAFs ranging from 0.2% to 48% (Table 1 and 2). Of the six novel loci, four contained a lead variant with MAF below 6% with the rare missense rs12272669 variant (MAF 0.22%) in MMACHC that associates with B12 found in the Icelandic data being at the extreme (Table 1). This variant has been observed in other populations than the Icelandic, albeit at much lower frequency (MAF 0.02%) (Exome Variant Server, http://evs.gs.washington.edu/EVS/). For TCN1 and FUT6 previously reported to associate with serum B12 levels we confirmed the association, yet with different SNVs than reported. At the TCN1 locus the strongest associated SNV in the Icelandic data was rs34324219 (Table 1) encoding a D301Y missense mutation, whereas the reported [10,11] and correlated ($r^2 = 0.28$) non-coding rs526934 was more weakly associated (Table S4). At the FUT6 locus, the P124S missense mutation (rs778805) identified in the combined analysis of Icelandic and Danish data associated more strongly (Table 1) than the previously reported promoter rs3760776 variant (Table S4). For the remaining four reported B12-associated loci, MUT, FUT2, CUBN and CLYBL, we
Table 1. Novel and previously reported genomic loci that associate with serum B12 levels at $P < 2.2 \times 10^{-9}$.

<table>
<thead>
<tr>
<th>SNV name</th>
<th>Locus</th>
<th>Chr.</th>
<th>Position (build 36/hg18)</th>
<th>Annotation</th>
<th>Alleles $^2$</th>
<th>EAF Icelandic</th>
<th>Danish - Health2006</th>
<th>Combined $^4$</th>
<th>$P$ (HET)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2336573</td>
<td>CD320</td>
<td>19</td>
<td>8,273,709 G220R T/C</td>
<td></td>
<td>0.031</td>
<td>0.32</td>
<td>1.1x10^{-2}</td>
<td>1.7x10^{-8}</td>
<td>4.5x10^{-5}</td>
</tr>
<tr>
<td>rs1131603</td>
<td>TCN2</td>
<td>22</td>
<td>29,348,975 L376S C/T</td>
<td></td>
<td>0.055</td>
<td>0.19</td>
<td>4.3x10^{-2}</td>
<td>1.8x10^{-3}</td>
<td>5.3x10^{-2}</td>
</tr>
<tr>
<td>rs3742801</td>
<td>ABCD4</td>
<td>14</td>
<td>73,828,759 E368K T/C</td>
<td></td>
<td>0.294</td>
<td>0.046</td>
<td>4.3x10^{-2}</td>
<td>1.8x10^{-3}</td>
<td>5.3x10^{-2}</td>
</tr>
<tr>
<td>rs2270655</td>
<td>MMAA</td>
<td>4</td>
<td>46,747,242 R206Q A/G</td>
<td></td>
<td>0.082</td>
<td>0.51</td>
<td>3.6x10^{-2}</td>
<td>1.3x10^{-2}</td>
<td>4.5x10^{-2}</td>
</tr>
<tr>
<td>rs12272669</td>
<td>TCN1</td>
<td>11</td>
<td>59,379,954 D301Y C/A</td>
<td></td>
<td>0.082</td>
<td>0.21</td>
<td>8.8x10^{-2}</td>
<td>3.2x10^{-2}</td>
<td>4.5x10^{-2}</td>
</tr>
<tr>
<td>rs77880533</td>
<td>FUT6</td>
<td>19</td>
<td>5,783,209 P124S A/G</td>
<td></td>
<td>0.254</td>
<td>0.046</td>
<td>2.1x10^{-2}</td>
<td>0.8x10^{-2}</td>
<td>0.8x10^{-2}</td>
</tr>
<tr>
<td>rs11413214</td>
<td>MUT</td>
<td>6</td>
<td>49,320,382 R324H C/T</td>
<td></td>
<td>0.048</td>
<td>0.17</td>
<td>9.6x10^{-3}</td>
<td>3.2x10^{-3}</td>
<td>4.5x10^{-3}</td>
</tr>
</tbody>
</table>

**Association results for serum B12 in Icelandic and Danish study samples separately and combined. The effect allele is the allele associated with increased serum B12 levels. The effect is on a quantile normalized scale. Data were combined in fixed effect meta-analyses based on $P$-value and direction of effect adjusted for the number of individuals in each sample. Values of $I^2$ are percentages.**

**Chr., chromosome; EAF, effect allele frequency; HET, heterogeneity; SNV, single nucleotide variant.**

$^1$The annotation is based on the RefSeq hg18.

$^2$The reference alleles based on Build 36/hg18 are shown in bold.

$^3$In the Icelandic data the strongest signal at the FUT6 locus is for rs708686 located 5kb to the FUT6 gene (see Table S3).

$^4$Danish data are given for the perfect proxy rs4267943 (1000 Genomes data: $r^2 = 1.0$).
confirmed the association signal [9–11] (Table 1). At the MTHFR locus the strongest folate association was for the major allele of the common A222V (rs1801133) for which previous association with serum folate has been reported [10,13,14] (Table 2).

For the two loci reported to associate with B12 levels in individuals of East-Asian ancestry (MSRA and 5q32) the variant was either not present in the Icelandic data or at very low frequency (Table S4) whereas the reported non-coding folate signals at FIGN and PRICKLE2 loci did not replicate in the Icelandic folate data (Table S5).

At a less stringent significance level of P<1×10^-6 we found three additional loci, CPS1, SPG51 and ZBTB10 with suggestive associations with serum B12 levels (Table S6) while suggestive association with folate levels at P<1×10^-6 was found for eight additional loci (Table S7).

Analyses conditional on the identified associated sequence variants

For the 13 loci associated with serum B12 or folate levels we performed stepwise conditional analyses to search for secondary signals applying Icelandic WGS data imputed into the 25,960 and 20,717 chip-genotyped Icelanders with serum B12 and folate information. We detected additional signals at five loci, CUBN, TCN1, TCN2, FUT6 and MTHFR (Figure 2). For the serum B12-associated loci, secondary independent association signals at P<5×10^-8 were detected at three, CUBN, TCN1 and TCN2 (Figure 2, Table 3, Table S8), while the secondary independent signal at FUT6 (observed for the reported B12-associated rs3760776 upstream of FUT6 [12]) did not reach the threshold of significance (P=4.4×10^-5). The secondary signal at the CUBN locus was found in a group of correlated markers represented by rs56077122 (located in an intron of the neighboring TRDMT1) (Figure 2). In TCN1 two additional independent signals at P<5×10^-8 for serum B12 were found including a missense variant (R35H) and an intergenic variant whereas one secondary signal in the TCN2 locus, represented by rs5753231, was located immediately 5′ to TCN2 (Figure 2, Table 3). In the folate-associated loci, a secondary independent signal was found at the MTHFR locus represented by rs17421511 located in intron 4 of the MTHFR gene (Figure 2, Table 3). In contrast to the lead SNVs a large fraction of the secondary B12 or folate signals were non-coding. Of the identified variants (lead and secondary) the fraction of variance in serum B12 or folate levels explained is estimated to be 6.3% for B12 and 1.0% for folate (Text S1).

Mapping effects of associated sequence variants on gene expression

To determine whether any of the lead or secondary association signals at the B12 or folate loci affect the expression of the target gene we analyzed genome-wide expression QTL (eQTL) data from white blood cells (n = 1,001) and adipose tissue (n = 673) from Icelanders with information on 22.9 million SNVs [29]. Of the lead and secondary B12 or folate signals that are coding (Tables 1–3) two showed strong association with the expression of the target gene; the R352H missense variant in MUT (P = 9.1×10^-59) in white blood cells and P = 2.5×10^-46 in adipose tissue) and the frameshift INDEL in FOLR3 (P = 7.1×10^-110 in white blood cells and P = 2.5×10^-46 in adipose tissue; Table S9). Of all the cis variants at the MUT locus the R352H missense mutation had by far the strongest effect on MUT expression indicating that this effect is not mediated by a non-coding regulatory variant in LD with the R352H mutation. The large effect of the frameshift mutation on FOLR3 expression is likely caused by nonsense-

<table>
<thead>
<tr>
<th>SNV name</th>
<th>Locus</th>
<th>Position (build 36)</th>
<th>Allele2 (effect/other)</th>
<th>EAF</th>
<th>Icelandic</th>
<th>Danish – Health2006</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr:515079700</td>
<td>FOLR3</td>
<td>11,722,389</td>
<td>Intron</td>
<td>C/T</td>
<td>0.179</td>
<td>0.069</td>
<td>2.5×10^-10</td>
</tr>
<tr>
<td>rs1801133</td>
<td>FOLR3</td>
<td>11,778,965</td>
<td>A222V</td>
<td>G/A</td>
<td>0.668</td>
<td>0.096</td>
<td>10^-48</td>
</tr>
<tr>
<td>rs56077122</td>
<td>FUT6</td>
<td>37,337,939</td>
<td>9.5</td>
<td>37,337,939</td>
<td>9.5×10^-12</td>
<td>62.0×10^-12</td>
<td>62.0×10^-12</td>
</tr>
</tbody>
</table>

Table 2. Novel and previously reported genomic loci that associate with serum folate levels at P<2.2×10^-9.

The annotation is based on the RefSeq hg18.

2The reference allele based on Build 36 hg18 is shown in bold.

3In the Icelandic data a 2 bp INDEL in exon 3 of FOLR3 associated more strongly with serum folate levels. As only SNVs were analyzed in the Danish data this data was not available for the Danish samples.

4The rs652197 variant was initially discovered in the Icelandic samples but subsequently genotyped in Danish samples to confirm the association.

5The reference allele based on Build 36 hg18 is shown in bold.
mediated decay of transcripts containing the premature termination mutation [30]. A similar effect was not seen for the nonsense mutation in the \textit{CLYBL} gene which can likely be explained by the closeness of the mutation to the N-terminal of the CLYBL protein (amino acid 259 of 340) (Table S9). Of the non-coding lead or secondary B12 or folate signals a statistically significant effect on

<table>
<thead>
<tr>
<th>Table 3. Novel secondary association signals at the serum B\textsubscript{12} or serum folate loci that associate at $P&lt;5\times10^{-8}$.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SNV #</strong></td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>\textit{CUBN} region (B\textsubscript{12})</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>\textit{TCN1} region (B\textsubscript{12})</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>\textit{TCN2} region (B\textsubscript{12})</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>\textit{MTHFR} region (folate)</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Conditional analyses were performed using imputed sequence data from chip-genotyped Icelanders with information on serum B\textsubscript{12}, or folate levels. Results for SNV #1 (lead SNVs) at each loci are unconditional on other SNVs. Analysis of SNV #2 is conditional on SNV #1 and SNV #3 is conditional on SNV #1 and #2. The LD between the SNVs at each locus was estimated from the sequence information of the 1,179 whole genome sequenced Icelanders.

doi:10.1371/journal.pgen.1003530.t003
expression was only seen for the TCN2 promoter variant, however, other markers in the region, that had no effect on serum B12 levels associated more strongly with TCN2 expression. Although lack of appropriate tissue to evaluate the effect of the B12 and folate mutations on expression cannot be excluded, these data suggest that except for the MUT gene the effects of both the coding and non-coding mutations are unlikely to be through expression.

**Association of identified sequence variants with other traits linked to B12 and folate levels**

Rare mutations in some of the B12 genes described here i.e. MMAICHC, MMA4, MUT, CD320, TCN2 and CUBV have been described in connection with rare conditions of methylmalonic aciduria and megaloblastic anemia that all relate to defects in B12 metabolism (OMIM database, http://www.ncbi.nlm.nih.gov/omim/). In addition, epidemiological studies have suggested a link between reduced B12 and folate levels and the risk of common conditions such as cardiovascular diseases [3,4], cancers [5,6] and neurodegenerative disorders [7]. To evaluate the effect of the B12 or folate variants on these conditions we analyzed the association with coronary artery disease (CAD), stroke, colon cancer, prostate cancer and Alzheimer’s disease in data obtained from deCODE’s phenotype database. As outlined in Table S10, variants associated with serum B12 or folate levels did not consistently affect the risk of the diseases tested; the B12 or folate increasing allele for some variants was weakly protective and for others weakly at risk, and only two loci (CUBV associated with CAD and MTHFR with stroke) were statistically significant (P<0.0018) but with opposite effects on these diseases. B12 or folate deficiencies can lead to increased serum homocysteine [2], yet of all the B12 or folate loci associated with CAD and stroke, one B12-increasing allele was protective (Table S11). These associations could be mediated through the effect of these variants on B12 and folate levels as reduced levels of B12 and folate have been linked to the development of aortic aneurysm [42].

**Discussion**

Here we performed association analyses of up to 22.9 million SNVs, identified through WGS and WES, in up to 45,576 individuals to identify and characterize genetic variation influencing population diversity in serum levels of B12 and folate. We discovered five novel loci that associate with serum B12 levels and one novel locus for folate levels and replicated the six reported B12 loci and one folate locus. In addition, we identified five novel secondary independent signals at both the new and previously reported loci. The fraction of variance in serum B12 or folate levels explained by the identified variants is estimated to be 6.3% for B12 and 1.0% for folate (Text S1). Of the identified SNVs, both common and rare, we find that a large fraction (13 of 18) is represented by coding variants which is an unusually high fraction of coding variants compared to previous GWAS for other traits. Furthermore, of the 13 loci that associate with serum B12 and folate levels the genes at 11 of them can be directly linked to the current understanding of B12 and folate metabolism such as absorption, transport or enzymatic processes and one (FUT2) has potential links with these processes (Figure 3). Only CLBL has a function that cannot be directly related to these pathways. Specifically, eight loci are involved in transporting B12 and folate between different tissues, four of them TCN1, FUT2, FUT6 and TCN2 as co-factors or regulators of co-factors necessary for the transport and the other four, CUBN, CD320, ABCD4 [43] and FOLR3 as membrane transporters actively facilitating membrane crossing. MUT and MTHFR catalyze enzymatic reactions in the OCM where MMAICHC and MMA4 are involved in co-enzymatic processes (Figure 3). Moreover, we note that of the 13 genes, two (TCN2 and CD320) are known and two (MUT and MMA4) are suggested to interact in vivo [44] (Figure 3). Together with the high fraction of coding mutations these data indicate that the target genes at all of the loci have been identified.
By screening the deCODE database for pleiotropic effects of the B12 and folate variants we replicated some of the previous associations of the FUT2 gene and detected novel suggestive association with increased risk of osteoporosis (low BMD) potentially mediated through diminished bone remodeling as a consequence of reduced ALP activity. We also detected suggestive associations of the FUT6 and the MTHFR variants with AAA and TA, respectively. However, we did not demonstrate association of any of the variants with the cardiovascular diseases, CAD and stroke, colorectal cancer, prostate cancer or Alzheimer’s disease and only two of the variants associated with homocysteine levels. Although to some degree impeded by low statistical power for some of these conditions, these data suggest that sequence variants that contribute to the population diversity in serum B12 or folate levels do not modify the risk of developing these conditions.
Materials and Methods

Ethics statement
All participants gave written informed consent. The studies were conducted in accordance with the Declaration of Helsinki II and were approved by the local Ethical Committees (approval numbers Denmark: H-3-2012-155, KA 98155 and KA-20060011, DeCode 08-105-V3-S1 (issued 30.08.2011) ref. VSNb2008060006/03.1).

Study participants in Iceland
For the Icelandic samples, serum B12 and folate levels were assessed in blood samples from Icelanders from the Landspitali University Hospital Laboratory or at the Icelandic Medical Center (Laeknasetrid) Laboratory in Mjodd (RAM), between the years 1990 and 2011. B12 and folate levels were normalized to a standard normal distribution using quantile normalization and then adjusted for sex, year of birth and age at measurement. For individuals for which more than one measurement was available we used the average of the normalized value.

Study participants in Denmark
The Danish data were generated in two population-based study samples recruited in Copenhagen. The Inter99 cohort is a randomized, non-pharmacological intervention study for the prevention of ischaemic heart disease, conducted on 6,784 randomly ascertained participants aged 30 to 60 years at the Research Centre for Prevention and Health in Glostrup, Denmark [45] (ClinicalTrials.gov: NCT002979237). Detailed characteristics of Inter99 have been published previously [45–47]. The Inter99 cohort included 5,481 and 5,624 individuals with genotypes and measurement of serum B12 and folate, respectively. Health2006 is a population-based epidemiological study of general health, diabetes and cardiovascular disease of 3,471 individuals aged 18–74 years [48]. Health2006 was also conducted at the Research Centre for Prevention and Health in Glostrup, Denmark. The Health2006 cohort included 2,812 and 2,804 individuals with valid genotypes and measurement of serum B12 and folate, respectively. In Inter99 serum B12 and folate were measured by a competitive chemiluminescent enzyme immunoassay (Immulite 2000 System; Siemens Medical Solutions Diagnostics, Los Angeles, CA, USA) as previously reported [14]. In Health2006, serum B12 and folate were measured by chemiluminescent immunoassay (Dimension Vista platform, Siemens Healthcare Diagnostics GmbH, Eschborn, Germany).

Genotype data generation
In the Icelandic part, SNVs were identified through the Icelandic WGS project. A total of 1,176 Icelanders were selected for sequencing based on having various neoplastic, cardiovascular and psychiatric conditions. All of the individuals were sequenced to a depth of at least 10x. The generation of genotypic data in Iceland is detailed in earlier reports [23] and in Text S1, and consisted of the following steps: SNV calling and genotyping in WGS, long range phasing, genotype imputation and in silico genotyping.

In the Danish part of the study 16,192 SNVs for genotyping were selected from a WES study of 2,000 individuals [28]. In brief, exon capture and Illumina sequencing to a depth of at least 10x were performed in 2,000 Danes by methods previously described [27]. The exome was captured by a NimbleGen 2.1M HD array with a target region of 34.1 Mb including 18,954 genes defined by CCDS (Consensus Coding Sequence database). The average number of reads sequenced for each individual was 22.3 million with most reads being 30 to 80 bases long. After alignment to the human reference genome (assembly hg18, NCBI build 36.3) and stringent quality assurance, including uniqueness of genomic mapping and Q-score >20, the median coverage per individual was 91% of the target region and had an average depth of 8x (96% coverage and 11x depth before filtering). After applying quality criteria 70,182 SNVs with an estimated MAF above 1% based on the reads using maximum likelihood were identified [49]. The details of the WES have been described previously [28]; 20,065 SNVs were, as part of a published study, selected from the exome sequencing for genotyping in 16,888 samples by a custom-designed Illumina Select array. First, 18,358 SNVs annotated to the most likely deleterious categories (179 nonsense, 15,789 nonsynonymous, 219 located in splice sites and 2,171 in untranslated regions) were prioritized. Second, 1,048 SNVs nominally associated with type 2 diabetes (P<0.05) in a sequencing-based association study were selected. Finally, we selected 599 synonymous variants in 192 loci previously associated with common metabolic traits at GWAS. Genotype data was obtained for 18,744 SNVs. Quality control of samples included removing closely related individuals, individuals with an extreme inbreeding coefficient, individuals with a low call rate, individuals with a mislabeled sex and individuals with a high discordance rate to previously genotyped SNVs. 15,989 individuals passed all quality control criteria. The SNVs were filtered based on their MAF (>0.5%), genotype call rate (>95%), Hardy-Weinberg equilibrium (P>10^-5) or cross-hybridization with the X-chromosome. 16,192 SNVs passed all filters [28]. Genotyping of FOLR3 rs652197 in Danish samples was done by KASPar SNP Genotyping System (KBioscience, Hoddesdon, UK).

Statistical analyses

Icelandic analyses and quantitative trait association testing. A generalized form of linear regression was used to test for association of serum B12 and folate with SNVs. Let y be the vector of quantitative measurements, and let g be the vector of expected allele counts for the SNV being tested. We assume the quantitative measurements follow a normal distribution with a mean that depends linearly on the expected allele at the SNV and a variance covariance matrix proportional to the kinship matrix:

\[ y \sim N(\beta g, 2\sigma^2 \Phi), \]

where

\[ \Phi_{ij} = \begin{cases} 0.5 & i = j \\ 2 \rho \delta_{ij} & i \neq j \end{cases} \]

is based on the kinship between individuals as estimated from the Icelandic genealogical database (kij) and estimate of the heritability of the trait (\(\rho\)). It is not computationally feasible to use this full model and we therefore split the individuals with in silico genotypes and serum B12 and folate measurements into smaller clusters. Here we chose to restrict the cluster size to at most 300 individuals.

The maximum likelihood estimates for the parameters \(\beta\), \(\phi\), and \(\sigma^2\) involve inverting the kinship matrix. If there are \(n\) individuals in the cluster, then this inversion requires \(O(n^3)\) calculations, but since these calculations only need to be performed once the computational cost of doing a GWAS will only be \(O(n^2)\) calculations; the cost of calculating the maximum likelihood estimates if the kinship matrix has already been inverted.
Multivariate regression and conditional analyses. For the multivariate regression analysis we only used Icelandic individuals which have been genotyped using the Illumina chip-genotyping platform. The multivariate linear regression analysis was performed conditioning for a given marker by adjusting for the estimated allele count based on imputation of this marker. The GC correction factor was the same as used for the unadjusted association analysis. A forward selection multiple logistic regression model was used to further define the extent of the genetic association. Briefly, all imputed SNVs located within an interval around the lead SNVs were tested for possible incorporation into a multiple regression model. In a stepwise fashion, a SNV was added to the model if it had the smallest $P$-value among all SNVs not yet included in the model and if it had a $P<5 \times 10^{-6}$. In the last step none of the SNVs remained significant at this threshold.

Association analyses of serum $B_{12}$ and folate in Danish samples. Association analysis of each SNV in the Danish data was performed using linear regression assuming an additive model. Principal component analysis was performed using the covariance matrix and the first principal component and sex were included in the model as covariates. All quantitative traits were quantile-normalized to a normal distribution prior to analysis. Association analyses were done using PLINK software (version 1.07, http://pngu.mgh.harvard.edu/~purcell/plink/). All $P$-values were corrected by GC. Inflation factors ($\lambda$) were at acceptable levels: B$_{12}$: Inter99: 1.027, Health2006: 1.014 and folate: Inter99: 1.024, Health2006: 1.010.

Meta-analyses. For all SNVs with data from more than one study sample (Icelandic, Inter99 and/or Health2006) we performed meta-analyses of summary association data where we estimated the combined effect in a fixed-effects meta-analysis using the METAL software (http://www.sph.umich.edu/csg/abecasis/Metal/) [50]. An overall z-statistic relative to each reference allele was estimated based on $P$-value and direction of effect adjusted for the number of individuals in each sample.

Supporting Information

Figure S1 Regional plots of the 11 loci associated with serum $B_{12}$. Genotyped and imputed SNVs passing quality control measures are plotted with their meta-analysis $P$-values ($-\log_{10}$ values) as a function of genomic position (NCBI Build 36). Only SNVs with $P<0.01$ are plotted. The lead SNV with the lowest combined $P$-value is indicated by the rs-number. Estimated recombination rates (HapMap CEU) are plotted to reflect the local LD structure. Gene annotations were obtained from RefGene.

Table S2 Clinical characteristics of the Danish samples. Data are mean ± standard deviation or median (interquartile range). (PDF)

Table S3 Overview of the most significantly associated SNV for each of the identified $B_{12}$ or folate loci in the Icelandic data. For each of the identified $B_{12}$ or folate loci presented in Tables 1 and 2 the Icelandic association data for the lead SNV is shown. Moreover, the strongest associations at these loci in the Icelandic data are shown. The lead SNVs presented in Tables 1 and 2 are either the strongest signal at each of the loci or highly correlated with the strongest signal except at the $FUT6$ locus were rs708686 located 5’ of $FUT6$ gives the strongest signal. (PDF)


Table S6 Suggestive loci in the Icelandic or the Icelandic and Danish data associated with serum $B_{12}$ levels ($2.2 \times 10^{-5}>P<10^{-6}$). (PDF)

Table S7 Results in the Icelandic data for loci with suggestive association with serum folate levels ($2.2 \times 10^{-5}>P<10^{-6}$). (PDF)

Table S8 Results from stepwise conditional analyses using the Icelandic data at loci associated with serum $B_{12}$ or folate levels for signals with $P<5 \times 10^{-10}$. Conditional analyses were performed using imputed sequence data from chip-genotyped Icelanders with information on serum $B_{12}$ or folate levels. Results for SNV #1 (lead SNVs) at each loci are unconditional on other SNVs. Analysis of SNV #2 is conditional on SNV #1 and SNV #3 is conditional on SNV #1 and #2. The LD between the SNVs at each locus was estimated from the sequence information of the 1,179 whole genome sequenced Icelanders. (PDF)

Table S9 $\text{cis}$-effect of the $B_{12}$ and folate SNVs on the expression of the target gene in white blood cells and adipose tissue. Correlation between SNVs that associate with increased $B_{12}$ or folate and mRNA expression in blood and adipose tissue from 1,001 and 673 individuals, respectively. The correlations are tested...
by regression analysis adjusted, for age, sex and differential cell counts (blood only), and inverse normal transformed relative expression values on the estimated genotype dosage. No other SNV shows significantly higher correlation with the expression in adipose or blood of MUT than rs1141321. \( ^6 \)\cite{theINDEL} chr11:71527804 is the most significant \( c/o \) variant for FOLR3. \( ^7 \)For TCN2 there are \( c/o \) variants both in blood and adipose tissue that have stronger correlation than rs5753231 with its expression, while having little effect on B12 levels. (PDF)

**Table S10**

Association results for the B12 and folate associated markers with potential co-morbid conditions in Icelanders. \(^{\text{1}}\) Effect size and effect allele frequency from the Icelandic population. Associations at \( P<0.001 \) are shown in bold. EA, effect allele frequency; MCV, mean corpuscular volume; MCH, mean corpuscular hemoglobin.

<table>
<thead>
<tr>
<th>Marker</th>
<th>Disease</th>
<th>Size</th>
<th>Anno.</th>
<th>EAF</th>
<th>( c/o )</th>
<th>( c/o )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(PDF)

**Table S11**

Association results for the B12 and folate variants with diseases and traits in deCODE database. Shown are the strongest association results for the folate and B12 variants, genome-wide significant or suggestive, with diseases and traits in deCODE’s database. EAF, effect allele frequency; BMD, bone mineral density; AAA, abdominal aortic aneurysm; TA, thoracic aneurysm. The annotation is based on the RefSeq hg18; 2The reference alleles based on Build 36 hg18 are shown in bold; 3The low BMD phenotypes are defined as those BMD values that are below \( -1 \) standard deviation (SD) from the mean.

(PDF)

**Text S1**

Supplementary Text S1 contains additional description of Methods. (PDF)

**Acknowledgments**

The authors wish to thank A. Forman, T. Lorentzen, B. Andreassen, and G.J. Klavsen for technical assistance and A.L. Nielsen, G. Lademann, and M.M.H. Kristensen for management assistance. The Inter99 was initiated by Torben Jørgensen (PI), Knut Borch-Johnsen (co-PI), Hans Ibsen, and Troels F. Thomsen. The steering committee comprises the former two and Charlotte Fisinger. The Health2006 was initiated by Allan Linneberg (PI) and Torben Jørgensen (co-PI).

**Author Contributions**


**References**


