Bioinformatics, 32(14), 2016, 2136–2142
doi: 10.1093/bioinformatics/btw125
Advance Access Publication Date: 21 March 2016
Original Paper

Genetics and population analysis

Haplotype synthesis analysis reveals functional variants underlying known genome-wide associated susceptibility loci

André Lacour[1], David Ellinghaus[2], Stefan Schreiber[2], Andre Franke[2] and Tim Becker[3,*]

[1]German Center for Neurodegenerative Diseases (DZNE), Bonn 53127, Germany, [2]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel 24105, Germany and [3]Institute for Community Medicine, Ernst Moritz Arndt University Greifswald, Greifswald 17475, Germany

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on October 2, 2015; revised on February 5, 2016; accepted on March 1, 2016

Abstract

Motivation: The functional mechanisms underlying disease association remain unknown for Genome-wide Association Studies (GWAS) susceptibility variants located outside coding regions. Synthesis of effects from multiple surrounding functional variants has been suggested as an explanation of hard-to-interpret findings. We define filter criteria based on linkage disequilibrium measures and allele frequencies which reflect expected properties of synthesizing variant sets. For eligible candidate sets, we search for haplotype markers that are highly correlated with associated variants.
Results: Via simulations we assess the performance of our approach and suggest parameter settings which guarantee 95% sensitivity at 20-fold reduced computational cost. We apply our method to 1000 Genomes data and confirmed Crohn's Disease (CD) and Type 2 Diabetes (T2D) variants. A proportion of 36.9% allowed explanation by three-variant-haplotypes carrying at least two functional variants, as compared to 16.4% for random variants ($P = 1.72 \times 10^{-8}$). Association could be explained by missense variants for MUC19, PER3 (CD) and HMG20A (T2D). In a CD GWAS—imputed using haplotype reference consortium data (64 976 haplotypes)—we could confirm the syntheses of MUC19 and PER3 and identified synthesis by missense variants for 6 further genes (ZGPAZ, GPR65, CLN3/NPIPB8, LOC102723878, rs2872507, GCKR). In all instances, the odds ratios of the synthesizing haplotypes were virtually identical to that of the index SNP. In summary, we demonstrate the potential of synthesis analysis to guide functional follow-up of GWAS findings.
Availability and implementation: All methods are implemented in the C/C++ toolkit GetSynth, available at http://sourceforge.net/projects/getsynth/.
Contact: tim.becker@uni-greifswald.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome-wide Association Studies (GWAS), see e.g. Bush and Moore, (2012) and Hirschhorn and Daly (2005), detected a multitude of genetic risk variants associated with complex diseases and phenotypes. For a great portion of these variants, the underlying biological mechanisms are still unknown. While in public databases the gene closest to the strongest association signal is provisionally listed as the susceptibility gene, other genes nearby might embody the true functional origin and cause the association signal via more or less complicated patterns of linkage disequilibrium (LD). As pointed

out by the authors of Dickson *et al.* (2010) and Platt *et al.* (2010) there is no guarantee that causal variants are in particular high LD with the top association signal (we will call the *tag variant* throughout this work). For instance, interaction between multiple variants may interfere with the ability to find either of them separately, but create a signal at a distantly linked marker (Atwell *et al.*, 2010).

Goldstein (2009), see also Edge *et al.* (2013), suggested that an association of a common variant with a complex disease can be synthetically created by multiple rarer functional variants from a surrounding genomic region. In this case the rarer variants occur more often or exclusively on a haplotype branch carrying a specific allele of the tag variant, generating in this way the strongest association signal at this locus. This situation is depicted in Figure 1 of Wang *et al.* (2010). The idea can be quantified by checking whether the LD measures between tag variant and candidate variants yield $|D'| \approx 1$, while R^2 need not be large (Goldstein, 2009; Takeuchi *et al.*, 2011). Examples for such *synthetic associations* have been reported from GWAS (Fellay *et al.*, 2010; Kumar *et al.*, 2013; Scherag *et al.*, 2010; Takeuchi *et al.*, 2009; Wadelius *et al.*, 2007) and sequencing studies (Wang *et al.*, 2010). A statistical method to test a given set of variants for synthetic association with a quantitative trait has been described by (Takeuchi *et al.*, 2011).

The concept of synthetic association has also given rise to some debate: some authors considered the ubiquity of synthetic associations to be unlikely (Orozco *et al.*, 2010), while others discussed whether synthetic associations should have already been detected by linkage analysis (Anderson *et al.*, 2011; Goldstein, 2011) and whether a rare-only model is applicable to a lot of GWAS findings (Goldstein, 2011; Wray *et al.*, 2011). In addition, the expected properties of synthetic associations were empirically assessed via simulation studies (Chang and Keinan, 2012; Dickson *et al.*, 2010), with partially contradictory results. The authors of Dickson *et al.* (2010) stated that rarer variants that contribute to a synthetic association might be as far as 2.5 Megabases (Mb) away, whereas the authors of Chang and Keinan (2012) reported that in 90% of their simulations at least one rare causal variant was already captured within a window of size 100 kilobases (kb). In any case, it can be stated that until now there is no complete understanding of which role synthetic associations actually play in the etiology of complex diseases, and how frequent the phenomenon of synthesis really is.

In view of the lacking consensus, we started an empirical evaluation of the frequency of the phenomenon. Until now, no methods have been provided to systematically search for sets of variants that synthesize the association of a common variant. A major reason for this is the computational load. Already within a ±100 kb region surrounding a susceptibility locus, typically 2000 variants are to be expected. With n eligible variants in an identified trait-related susceptibility region, there are $2^n - 1$ variant sets to be investigated. Even when only sets with, for instance, up to 6 variants shall be tested for potential synthesis, 8.84×10^{16} different sets have to be investigated. In view of the large number, an efficient search engine is prerequisite for the detection of synthetic associations. Identification of such variant sets is highly relevant for the follow-up of association signals that were produced by GWAS or next generation sequencing (NGS) association studies, in order to come closer disease relevant biological function (Marian, 2012).

2 Methods and data

We consider an LD region that is associated with a disease phenotype. The top association signal (tag variant) has been reported in a GWAS catalog or a consensus meta-analysis. We assume that genotype data, either from public reference data, GWAS or NGS association studies, are available for the tag variant and a sufficiently large surrounding region. We advise to include variants from a 2 or 5 Mb flanking region around the tag variant.

2.1 Filtering rules

Let a_i, A_i be the alleles of variant s_i. Let $f(a_i)$ be the allele frequency of a_i and let $h(a_i a_j)$ be the frequency of the 2-variants haplotype. The alleles a_i and a_j are said to be *in-phase* if $D = h(a_i a_j) - f(a_i)f(a_j) > 0$. From the data we calculate the allele frequencies and the LD measures $R^2 = D^2/(a_i a_j A_i A_j)$ and $D' = D/D_{\max}$, where $D_{\max} = \min(a_i a_j, A_i A_j)$ if $D < 0$ and $D_{\max} = \min(a_i A_j, A_i a_j)$ if $D > 0$. Note that D' comprises a leading sign.

We employ the following allele frequency criteria for synthetic association. Let $S = \{s_1, \ldots, s_k\}$ be a set of k variants and let $s \notin S$ be the tag variant. Let $D'(a_i, a_j)$, $R^2(a_i, a_j)$ be the pairwise LD measures for two variants s_i, s_j. Let $\hat{D}'$ and $\hat{R}^2$ be predefined intervals that quantify conditions on $|D'| \approx 1$ and R^2 and let $t \gtrsim 1$ be a tolerance parameter.

Let a be the risk allele of s, i.e. the allele with OR > 1. Let a_i be the alleles of s_i that are in-phase with a. S is said to be a *risk candidate set* if

1. $|D'(a_i, a)| \in \hat{D}' \ \wedge \ R^2(a_i, a) \in \hat{R}^2 \quad \forall\, i \in \{1 \ldots k\}$,
2. $f(a)/t \leq \sum_{i=1}^{k} f(a_i) \leq f(a) \cdot t$,

In the same way, we can regard A to be the protective allele of s, i.e. the allele with a reported OR < 1. Let now A_i be the alleles of s_i that are in-phase with A. S is then said to be a *protective candidate set* if the above criteria hold, whereupon the second condition is replaced by

2. $f(A)/t \leq \sum_{i=1}^{k} f(A_i) \leq f(A) \cdot t$.

One may also see this twofold search from a different perspective: once we seek candidate sets that are synthesized by the minor allele of the tag variant, which means $D'(a_i, a) > 0$, and another where the sets are synthesized by the major allele of the same, $D'(a_i, a) < 0$.

Note that in Eq. 2, an immediate consequence of the target SNP being an ancestor of the rarer causal variants would be $t = 1$. In practice, $t > 1$ can be caused by genotyping errors or rare recombinatorial. Furthermore, it is necessary to completely omit the frequency criterion to capture also syntheses involving frequent recombinants. Desirable choices of the parameters are evaluated in the next paragraph.

In the context of this work, we prune variants for $R^2 = 1$, while we do not remove variants that are marked to have known functional consequences. More details on our implementation are given in appendix A (Supplementary Data).

2.2 Tag variant: set haplotype correlation

The goal of our approach is to find variant sets that explain a given tag variant via synthesis, in particular to find haplotypes that are in nearly perfect LD with one of the alleles of the tag variant. For a candidate set S of tag variant s we phase and reconstruct the haplotypes from the genotypes employing the EM-algorithm using maximum-likelihood estimation of haplotype frequencies according to Excoffier and Slatkin (1995) and Long *et al.*(1995). Here, we use an implementation that improves our previous implementation in FAMHAP (Becker and Knapp, 2004). We evaluate the haplotypes in a binary storing version, which is presented in detail in appendix B

(Supplementary Data). From the reconstructed haplotypes with their estimated frequencies we compose dichotomized *haplotype markers* consisting each of one haplotype versus all others. Then we calculate the Pearson product-moment correlation coefficient of that allele x of tag variant s which is tested for being synthesized by S and each haplotype marker h existing for S,

$$r_{xh} = \frac{\sum_{i=1}^{M}(h_i - \bar{h})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{M}(h_i - \bar{h})^2 \cdot \sum_{i=1}^{M}(x_i - \bar{x})^2}}, \quad (1)$$

where M is the number of individuals, $x_i \in \{0, 1, 2\}$ is the ith individual's allele count of s and $h_i = h_i(a_1 \ldots a_k) \in [0; 2]$ is the frequency of the haplotype with set variant alleles a_j for individual i from the maximum-likelihood estimation. $\bar{x}$ ($\bar{h}$) denote the mean values of all x_i (h_i). Synthesis is established if $|r_{xh}| \approx 1$.

2.3 Data

For pilot analysis, we systematically searched for syntheses in the CEU sample of the 1000 Genomes Project phase 1 integrated release (1000 Genomes Project Consortium, 2012) reference data (accessed Mar 2012). For confirmation, we used a GWAS on Crohn's disease (CD) with 1598 individuals (480 cases and 1118 controls) genotyped with Illumina© HumanHap550 Bead array. CD patients were recruited either at the Department of General Internal Medicine, Christian-Albrechts-University, Kiel, and the Charité University Hospital, Berlin, through local outpatient services, or nationwide with the support of the German Crohn and Colitis Foundation. The 1118 German healthy control individuals were obtained from the Popgen biobank (Krawczak *et al.*, 2006).

The GWAS contributed to most relevant CD findings and is described in detail at Jostins *et al.* (2012). We adapted quality control standards described there. The GWAS was imputed the reference panel form the haplotype reference consortium (http://www.haplotype-reference-consortium.org/), first release consisting of 64 976 haplotypes at 39 235 157 SNPs (McCarthy, 2015). Imputation was performed at the Michigan Imputation Server (MIS; https://imputationserver.sph.umich.edu) using SHAPEIT2 (Deleneau *et al.*, 2013) and Minimac3

(http://genome.sph.umich.edu/wiki/Minimac3) with default parameters.

2.4 Simulation study

In order to evaluate our filter criteria, we set up simulation based on the imputed CD GWAS data. We repeatedly selected k-tuples ($k = 3$, 5) of functional SNPs with a maximum distance of 4 MB and simulated a hypothetical tag SNP with $r_{xh} = 1$ with the k-tuple (full synthesis). We then tried to retrieve the synthesis in the surrounding region of the tag SNP, both in the complete sample, as well as in a sub-sample of 300 randomly selected individuals. Under each setting, we simulated 1000 datasets and evaluated each of these under 5 filter parameter configurations (3).

3 Results

3.1 Results of simulation study

The simulated data was evaluated according to different filters, upper limit for R^2 (software option **–filter-R2-upper {decimal}**), lower limit for D' (**–filter-d-prime-lower {decimal}**) and limit for t (**–filter-tolerance {decimal}**). We did not apply a lower limit for R^2 or an upper limit for D'. The exploration of an upper limit for R^2 is motivated by the results in Goldstein (2009) and Takeuchi *et al.* (2011). We applied five settings, from strong filtering to exhaustive search, setting 1 (0.49 as upper limit for R^2, 0.7 as lower limit for D', 1.2 as upper limit for t), setting 2 (1.0, 0.7, 1.2) with no upper limit for R^2, setting 3 (1.0, 0.7, 1.5) with, in addition, relaxed limit for t, setting 4 (1.0, 1.0, 1.5) with, in addition, no limit for D' and setting 5 (1.0, 1.0, –), exhaustive search without filters active.

The results of the simulation study can be found in Table 1. We list the percentage of datasets $|r_{xh}|$ for which we retrieved the tuple that actually was used to simulate the tag SNP.

We first discuss the analysis of the entire sample. With exhaustive search (setting 5 without filters active) our software was able to retrieve all syntheses (100% sensitivity). Under filter settings 3 and 4, specificity remains high (95%), both for 3- and 5-SNP-tuples. There is no difference in performance between settings 3 and 4, indicating that a filter on D' can be applied without loss of sensitivity. Under setting 2, a substantially worse performance is observed (sensitivity $< 85\%$). Setting 1, with the most stringent filtering, has low sensitivity (around 60%) and cannot be recommended.

Interestingly, sensitivity is also very good when only a sub sample of limited size is analyzed. In particular for $k = 3$, sensitivity is very close to that in the whole sample, irrespective of the filtering parameters. For $k = 5$, the loss of sensitivity is more pronounced, under the exhaustive search, for instance, sensitivity drops to 91%.

With filters active, running time typically is reduced substantially. Comparing exhaustive search (setting 5) and setting 1 for 5-tuples, a factor of 160 for the whole and of 177 for the sub-sample can be observed. Under setting 3, for 5-SNP-tuples, we observe a reduction by a factor of 22.5 for the whole sample and by 20.0 for the sub sample. For 5-SNP-tuples, we observe reductions by factors of 18.5 and 19.3, respectively.

In summary, balancing computational cost and trade-off in specificity, filter setting 3 is recommendable when computational resources are limited.

Table 1. Sensitivity and running time

Setting	k[a]	Sample	Sensitivity	Time[b]
1	3	whole	0.621	0.9
		sub	0.611	0.3
	5	whole	0.626	90
		sub	0.612	22
2	3	whole	0.840	1.1
		sub	0.832	0.3
	5	whole	0.786	1017
		sub	0.769	258
3	3	whole	0.961	1.6
		sub	0.957	0.4
	5	whole	0.964	822
		sub	0.917	202
4	3	whole	0.961	19
		sub	0.961	5
	5	whole	0.964	9519
		sub	0.929	2492
5	3	whole	1.000	32
		sub	0.996	9
	5	whole	1.000	15213
		sub	0.957	3899

[a]Size of tuple.
[b]In minutes on average.

Table 2. Number and proportion of syntheses for $k=3$

		Random		Crohn's Disease			Type 2 Diabetes		
n[a]	$\lvert r_{xh}\rvert$	#[b]	h	#[b]	h	P[d]	#[b]	h	P[d]
3	0.995	41	0.041	9	0.129	7.91e-4	6	0.100	3.11e-2
3	0.975	63	0.063	14	0.200	**1**.80e-5	8	0.133	3.43e-1
2	0.995	105	0.105	24	0.343	**3**.46e-9	15	0.250	**5**.75e-4
2	0.975	164	0.164	30	0.429	**2**.79e-8	18	0.300	6.66e-3
1	0.995	424	0.424	42	0.600	4.09e-3	30	0.500	2.48e-1
1	0.975	556	0.556	47	0.671	5.98e-2	39	0.650	1.54e-1
0	0.995	557	0.557	48	0.686	5.21e-2	37	0.617	4.50e-1
0	0.975	694	0.694	55	0.786	1.05e-1	48	0.800	8.18e-2
total		1000		70			60		

[a]Number of 'broad-sense' functionals.
[b]Amount of syntheses.
[c]Proportion of syntheses.
[d]As compared to random.

3.2 Frequency of the synthesis phenomenon in 1000 Genomes data

Nearly perfect pairwise LD ($R^2 \approx 1$) between neighbouring variants is a common phenomenon. Likewise, perfect LD between a single variant and a haplotype marker is likely to exist in regions of strong LD. In order to assess the frequency of the phenomenon, we randomly selected 1000 variant markers from the Illumina© 550K marker panel. For these variants, we systematically searched for all three-marker syntheses in a 2 Mb surrounding interval in the 1000 Genomes Project Consortium (2012) data. In Table 2, we list the absolute numbers and proportions of syntheses for $|r_{xh}|$ of either 0.995 or 0.975. A portion of 69.4% (55.7%) of tag variant markers allowed a synthesis by three surrounding variants at an r_{xh} cut-off of 0.975 (0.995). We investigated how many of these syntheses comprised 'broad-sense' functional variants, i.e. variants classified not as 'unknown', 'intergenic' or 'intronic'. 55.6% (42.4%) of variants allowed a synthesis including at least one functional variant, 16.4% (10.5%) allowed a synthesis with at least two functional variants, and 6.3% (4.1%) could entirely be explained by functional variants. Thus, formal synthesis, also involving functional variants, is a common phenomenon.

Next, we investigated the frequency of syntheses for Crohn's Disease (CD) and Type 2 Diabetes (T2D) susceptibility loci. We took 71 consensus variants for CD (Franke *et al.*, 2010) and 62 for T2D (Mahajan *et al.*, 2014) from the Burdett,T (EBI). *et al.* (2014). Of these variants, 70 CD variants and 60 T2D variants were available in the 1000 Genomes CEU reference data. For those loci our method was able to reveal synthesizing sets, we list the respective rs-numbers, the top set and the correlation coefficient in Supplementary Table A.

The portion of syntheses, ignoring functional annotations, for CD, 78.6% (68.6%) and T2D, 80.0% (61.7%), did not differ significantly from the proportions observed for random variants ($P > 0.05$). However, a substantial increase in the proportion of synthetic associations involving functional variants was observed. After adjustment for multiple comparisons, five settings showed significance. 20.0% (12.9%) of the CD variants allowed a synthesis with three functional variants at $r_{xh}=0.975$ (0.995) as compared to a portion of 6.3% (4.1%) for random variants ($P = 1.8 \times 10^{-5}$ and $P = 7.9 \times 10^{-4}$); 42.9% (34.3%) of CD variants allowed a synthesis with at least two functional variants at $r_{xh}=0.975$ (0.995), $P = 2.01 \times 10^{-8}$ and $P = 4.1 \times 10^{-6}$) compared to random variants. Finally, T2D showed 25.0% synthesis involving at least two functional variants at $r_{xh}=0.995$ ($P = 5.8 \times 10^{-4}$). In summary, Table 2 shows a significant increase of syntheses involving functional variants for CD and also a respective tendency for T2D susceptibility loci. This suggests that a portion of these syntheses potentially reflects the actual functional causes behind the respective GWAS association signal.

3.3 Examples of associated susceptibility loci for Crohn's disease and Type 2 Diabetes

We further restricted the syntheses discovered in section 3 for 'narrow-sense' functional variants, i.e. 'missense', 'nonsense', 'stop-loss', 'frameshift' and splice variants ('splice-3','splice-5'). We discovered several complete or nearly complete syntheses made up from variants of those annotation types. In the following we will describe three examples in detail.

First, rs11564258 is an intron variant of MUC19 (chromosome 12), a CD susceptibility locus (Franke *et al.*, 2010; GWAS Catalog *et al.*, 2014). Its minor A-allele conveys an odds ratio of 1.73 [1.55;1.95] (Franke *et al.*, 2010). Synthesis analysis revealed twelve different three-variant functional syntheses for rs11564258 with $|r_{xh}| \geq 0.99$. A list of these sets can be found in Supplementary Table B. The synthesizing sets overlapped and were made up by a total of 14 different missense variants, partially reported also in Kumar *et al.* (2013). In order to disentangle the LD pattern, we determined the joint haplotype distribution of the tag variant and all synthesizing variants.

The respective 15-variant haplotypes with a total length of 138 kb can be found in Supplementary Table C.

The A-allele of the tag variant perfectly tags a single haplotype of frequency 0.024 which fits the previously reported (Franke *et al.*, 2010) minor allele frequency of 0.025 for rs11564258. While all 14 variants that are involved in the synthesis are missense variants, the tagged haplotype, marked by the red A-allele of the tag variant, does not always carry the minor allele of these variants. Under the convention that the minor allele is the missense allele this suggests that the missense alleles are protective alleles: all further haplotypes carry at least one of these 'protective' alleles and the tag haplotype is characterized by an absence of protective alleles. We note, however, that the classification into risk and protective alleles is to a large extent a matter of terminology to describe one of the two sides of a coin. In any case, it can be stated that rs11564258 risk allele carriers can fully be characterized by the allele patterns present at 14 missense variants in MUC19. As shown above, actually various subsets made of three variants are sufficient to obtain the one-to-one correspondence.

Second, rs2797685 is an intron variant of PER3 (chr 1), also a CD susceptibility locus (Franke *et al.*, 2010; GWAS Catalog *et al.*, 2014). Its minor T-allele conveys an odds ratio of 1.05 [1.01–1.10] (Franke *et al.*, 2010). Synthesis analysis revealed a manifold of 40 syntheses at $|r_{xh}| \geq 0.99$ with a cardinality between 3 and 6 variants. The synthesizing sets are given in Supplementary Table B and the 17-variant haplotypes of the joint distribution with a total length of 56 kb can be found in in Supplementary Table D (upper panel).

Third, rs7178572 is an intron variant located in HMG20A (chr 15) which has previously been identified as a T2D susceptibility locus (GWAS Catalog *et al.*, 2014; Mahajan *et al.*, 2014). Its minor G-allele conveys an odds ratio of 1.08 [1.04-1.13] (Mahajan *et al.*, 2014). Synthesis is established by a two-variant haplotype of size 371 kb comprising the missense variant rs1867780 located in PEAK1 and rs7119 in the untranslated-3′ HMG20A region. In total synthesis analysis revealed 17 synthesizing sets, which are listed in

Supplementary Table B. All sets include the aforementioned two variants and between none and two additional missense variants inside exons of the genes TBC1D2B, CHRNA5, ADAMTS7 and RASGRF1. In Supplementary Table E we list the 1877 kb joint nine-variant haplotypes of the tag and all synthesizing variants. The A-allele of the tag variant perfectly tags a single haplotype of frequency 0.324 consisting of the wildtype alleles of all contributing variants.

3.4 Analysis of CD GWAS

We applied our method to the CD GWAS in order to investigate if synthetic haplotypes capture the GWAS index association signal. Since the GWAS is only part of the original discovery samples (Jostins *et al.*, 2012), the index SNPs listed are typically not genome-wide significant in the data we present here. However, suggestive association is present and consistent with the results in Jostins *et al.* (2012).

First, we searched to confirm the syntheses for MUC19 and PER3 which we had detected using 1000 Genomes reference data. In the CD GWAS, for MUC 19, we identified a 16-SNP-haplotype, mostly defined by the same variants as before (Supplementary Table F). The haplotype is tagged by the minor allele of the GWAS index SNP rs11564258 (OR = 1.67 [1, 09; 2.54], $P = 0.017$). The effect size is in line with that reported by Jostins *et al.* (2012). The functional haplotype has a frequency of 0.040 in cases and 0.022 in controls (Supplementary Table F), (1.86 [1.21; 2.85], $P = 0.007$), i.e. its association signal is even a bit stronger than that of the index allele. For PER3, we also confirm the finding in 1000 Genomes data and establish a functional syntheses. The odds ratio of the GWAS index allele rs2797685-T is virtually identical to that of the synthetic haplotype (Supplementary Table F).

Next, we searched for further narrow-sense functional syntheses given by SNP sets with size up to 5. We applied filter settings 3 and 5, which yielded identical results, thus also confirming the usefulness of the faster setting 3. In addition to the analysis of the 1000 Genomes data, we were able to identify functional syntheses for 6 further CD loci, ZGPAZ, GPR65, CLN3/NPIPB8, LOC102723878, LD region around rs2872507 and GCKR. A full description of the underlying SNPs, haplotypes and association measures can be found in Supplementary Table F. Three of these genes, ZGPAZ, GPR65 and CLN3/NPIPB8, are also shown in Table 3 and are described in the following in more detail.

The major G-allele of rs4809330 (ZGPAZ, intron) is associated with elevated CD risk (Jostins *et al.*, 2012). In our GWAS, the G-allele has a frequency of 0.763 in cases and 0.708 in controls (Table 3), (OR = 1.32 [1.11; 1.58], $P = 0.0016$). The G-allele tags a 5-SNP-haplotype defined by missense variants in the surrounding genes HELZ2, RTEL1-TNFRSF6B, LIME1 and SLC2A4RG. The functional haplotype has a frequency of 0.762 in cases and 0.701 in controls (Table 3), (1.34 [1.13; 1.60], $P = 0.00038$), i.e. its association signal goes in the same direction as that of the index SNP, but is stronger.

Next, the minor T-allele of rs8005161 (GPR65, intron), associated with CD risk (Jostins *et al.*, 2012), has a frequency of 0.119 in cases and 0.083 in controls, (OR = 1.42 [1.11; 1.81], $P = 0.0050$). The other allele, rs8005161-C(0.70 [0.55; 0.90]), tags a 2-SNP-haplotype defined by missense variants in GALC and GPR65. The functional haplotype (Table 3) has frequencies and odds ratio virtually identical to that of rs8005161-C. In addition, the synthesizing variants define a further, rare haplotype with stronger association signal. $Rs1805078_G - rs3742704_C$ has a frequency of 0.044 in cases and 0.024 in controls (Table 3), (OR = 1.85 [1.23; 2.79], $P = 0.0040$).

Finally, rs151181-C (CLN3/NPIPB8) is associated with elevated CD risk (Jostins *et al.*, 2012). The other allele, rs151181-T, has a frequency of 0.524 in cases and 0.574 in controls (Table 3), (0.82 [0.71; 0.95], $P = 0.0114$), and tags a 2-SNP-haplotype defined by missense variants in APOBR and IL27. The functional haplotype has a frequency of 0.525 in cases and 0.575 in controls (Table 3), (0.81 [0.72; 0.95], $P = 0.0073$). Again, the synthesizing variants define an additional haplotype, $rs180743_C - rs147413292_C$ with stronger association signal ($P = 0.0024$).

In summary, the synthetic haplotypes always confirm the association signal of the GWAS index SNP. Upon analysis conditional on the synthesizing variants, all index signals vanished (Table 3), indicating that the functional variants simultaneously capture the index association.

4 Discussion

We presented methods for the search for multi-locus haplotype markers in near perfect LD with a GWAS tag variant. Such haplotype markers fulfil the formal criteria of a synthetic association (Goldstein, 2009). We suggest and evaluated filtering criteria which we deducted heuristically from typical examples of synthetic association presented in Dickson *et al.* (2010), based on marker allele frequencies and LD measure criteria. Via a simulation study, we showed that the enormously large space of potentially synthesizing variant sets can effectively be reduced while keeping sensitivity high. Our results were confirmed by the analysis of the CD GWAS, in which all results that were found with exhaustive search were also retrievable with filter setting 3 active. In practice, we recommend exhaustive search for tuples of size up to 5, and filter setting 3 when larger tuples or multitudes of index SNPs shall be investigated. As a future plan, we wish to apply our approach to all known phenotype associations listed in GWAS Catalog *et al.* (2014) and to provide a respective data base of functional syntheses of susceptibility loci.

With t1000 Genomes reference data, we found syntheses involving variants which are as far as 1 Mb, and spanning 2.1 MB in the most extreme case, away from the tag variant. In this sense, we can confirm the statement given in Dickson *et al.* (2010) that synthesizing variants can be located rather far away from the main association peak. Of note, the functional synthesis for the diabetes susceptibility locus in HMG20A we described also comprised variants more than 1 Mb away from the tag variant. Syntheses at further distances were not observed with the real data. Nevertheless, large window sizes should be chosen when non-human data or regions of extended LD shall be searched.

Our approach can be applied in a case–control setting as well as to public reference genotype data. Via analysis of 1000 Genomes reference data and confirmation in our case–control GWAS, we could demonstrate that syntheses can readily be detected in reference data, without the incorporation of phenotype information. Sensitivity in reference data, however, may be reduced, for instance when relevant markers are missing. Validation and exhaustive search in a case–control setting, therefore, is the gold standard.

Our data analysis demonstrated that formal synthesis is a very common phenomenon in regions of LD. We could further show that syntheses involving functional variants occur more frequently with known GWAS susceptibility loci (Crohn's Disease and Type 2 Diabetes) than with random variants. A limitation of this finding is that it is not trivial to select an appropriate set of random variants

Table 3. Results CD GWAS (selected)

Gene	Function	chr	SNP	bp	Index allele[a]	Second allele[b]	Functional haplotype[c]	Further haplotype[d]
HELZ2	missense	20	rs115251319	62194212			C	
RTEL1-TNFRSF6B	missense	20	rs35640778	62321128			G	
RTEL1-TNFRSF6B	missense	20	rs3208008	62326110			C	
ZGPAZ	intron	20	rs4809330	62349586	G	A	[G]	
LIME1	missense	20	rs1151625	62369997			C	
SLC2A4RG	missense	20	rs8957	62373707			T	
				Case[e]	0.763	0.237	0.762	
				Control [f]	0.708	0.292	0.701	
				OR[g]	1.32 [1.11;1.58]	0.76 [0.63;0.90]	1.34 [1.13;1.60]	
				P-val[h]	0.0016 [0.660]		0.00038	
GALC	missense	14	rs1805078	88450770			G	G
GPR65	intron	14	rs8005161	88472595	T	C	[C]	[T]
GPR65	missense	14	rs3742704	88477882			A	C
				Case	0.119	0.881	0.881	0.044
				Control	0.087	0.913	0.914	0.024
				OR	1.42 [1.11;1.81]	0.71 [0.55;0.90]	0.70 [0.55;0.90]	1.85 [1.23;2.79]
				P-val	0.0050 [0.413]		0.0051	0.0039
CLN3/NPIPB8	intron	16	rs151181	28490517	C	T	[T]	[C]
APOBR	missense	16	rs180743	28507644			C	G
IL27	missense	16	rs147413292	28511206			C	C
				Case	0.476	0.524	0.524	0.417
				Control	0.428	0.572	0.572	0.360
				OR	1.22 [1.05;1.42]	0.82[0.71;0.96]	0.82 [0.71;0.96]	1.27 [1.09;1.49]
				P-val	0.0114 [0.625]		0.0074	0.0024

[a]Risk allele of index SNP reported in GWAS catalogue.
[b]Other allele of GWAS index SNP.
[c]Functional haplotype in LD with index or second allele of index SNP.
[d]Further haplotype defined by synthesizing SNPs. Listed in case its association is stronger than that of the functional haplotype.
[e]Allele or haplotype frequency in cases in CD GWAS.
[f]Allele or haplotype frequency in controls in CD GWAS.
[g]Odds ratio of allele or haplotype in CD GWAS.
[h]Unconditional *P*-value in CD GWAS and *P*-value after conditioning on synthetic variants in brackets.

for comparison, as GWAS susceptibility loci are certainly not randomly distributed in the genome. We note, however, that the enrichment among CD loci is also markedly stronger than that in T2D loci. Thus, at least for CD the hypothesis of a functional enrichment is supported. Irrespective of this, the detection of syntheses is of substantial relevance, whether or not is the result of statistical enrichment, since index allele carriers will typically carry the functional haplotype of synthesis. Thus, this haplotype will, even in case it should not be the actual cause of the disease, expose its biological impact and should at least be acknowledged as a confounding factor when it comes to the investigation of causality. In this context, it has to be emphasized that by statistical means alone it will not possible to ultimately prove causality of a set of variants. In the presence of near perfect LD between two or more variant or haplotype markers any of these might provide causality. Still it is important to know all syntheses of a susceptibility locus, since they point to functional variants that are at least primary candidates for causality.

For, in total, 8 out 70 CD GWAS index SNPs we detected syntheses by missense variants. The respective haplotypes showed association signals very similar to that of the GWAS index SNP. In three cases (ZGPA/, CLN3/NPIPB8, MUC19), the haplotype signal was more pronounced and lead to improved significance. In the case of GPR65 and CLN3/NPIPB8, the synthesizing SNPs defined and additional, rare haplotype with an association signal that was stronger than that of the index SNP and the synthetic haplotype. Thus, syntheses may not only explain the GWAS index signal, but provide additional statistical support for the finding. In view of this, the search for 'incomplete' syntheses should be a future area of investigation: in practice, the GWAS index SNP might only partially capture surrounding signals. In this sense, GWAS index SNPs which define full syntheses must be regarded as events fortunate for gene detection.

In summary, we have shown that the inference of synthetic variants has the potential to yield additional insight into the biology underlying hard-to-interpret association signals.

Our methods are implemented in the efficient software tool GetSynth, which is freely available at http://sourceforge.net/projects/getsynth/. The software is written in C/C++ and requires the binary genotype files defined by PLINK (Chang *et al.*, 2015; Purcell *et al.*, 2007) as input format. Filter criteria, set size, function classes and the number of functional variants that shall be involved in a synthesis can be pre-specified by the user.

Conflict of Interest: none declared.

References

1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Anderson,C.A. *et al.* (2011) Synthetic associations are unlikely to account for many common disease genome-wide association signals. *PLoS Biol.*, **9**, e1000580.

Atwell,S. *et al.* (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*, **465**, 627–631.

Becker,T. and Knapp,M. (2004) Maximum-likelihood estimation of haplotype frequencies in nuclear families. *Genet. Epidemiol.*, **27**, 21–32.

Burdett,T (EBI). *et al.* (2012) The NHGRI-EBI Catalog of published genome-wide association studies. Version v1.0. Available at: http://www.ebi.ac.uk/gwas (10 January 2016, date last accessed).

Bush,W.S. and Moore,J.H. (2012) Chapter 11: genome-wide association studies. *PLoS Comput. Biol.*, **8**, e1002822.

Chang,D. and Keinan,A. (2012) Predicting signatures of "synthetic associations" and "natural associations" from empirical patterns of human genetic variation. *PLoS Comput. Biol.*, **8**, e1002600.

Chang,C.C. *et al.* (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, **4**, 7.

Deleneau,S.P. *et al.* (2013) Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods*, **10**, 5–6.

Dickson,S.P. *et al.* (2010) Rare variants create synthetic genome-wide associations. *PLoS Biol.*, **8**, e1000294.

Edge,M.D. *et al.* (2013) Windfalls and pitfalls: applications of population genetics to the search for disease genes. *Evol. Med. Public Health*, 254–272.

Excoffier,L. and Slatkin,M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.*, **12**, 921–927.

Fellay,J. *et al.* (2010) ITPA gene variants protect against anaemia in patients treated for chronic hepatitis C. *Nature*, **464**, 405–408.

Franke,A. *et al.* (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's Disease susceptibility loci. *Nat. Genet.*, **42**, 1118–1125.

Goldstein,D.B. (2009) Common genetic variation and human traits. *N. Engl. J. Med.*, **360**, 1696–1608.

Goldstein,D.B. (2011) The importance of synthetic associations will only be resolved empirically. *PLoS Biol.*, **9**, e1001008.

Hirschhorn,J.N. and Daly,M.J. (2005) Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.*, **6**, 95–108.

Jostins,L. *et al.* (2012) Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, **1**, 119–124.

Kumar,V. *et al.* (2013) Genome-wide association study signal at the 12q12 locus for Crohn's Disease may represent associations with the MUC19 gene. *Inflamm. Bowel Dis.*, **19**, 1254–1259.

Krawczak,M. *et al.* (2006) PopGen: population based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. *Commun. Genet.*, **9**, 55–61.

Long,J.C. *et al.* (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.*, **56**, 799–810.

Mahajan,A. *et al.* (2014) Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type2 diabetes susceptibility. *Nat. Genet.*, **46**, 234–244.

Marian,A.J. (2012) Molecular genetic studies of complex phenotypes. *Transl. Res.*, **159**, 64–79.

McCarthy,S. (2015) A reference panel of 64,976 haplotypes for genotype imputation. *BioRXiv: Preprint Server Biol.*, 1–24.

Orozco,G. *et al.* (2010) Synthetic associations in the context of genome-wide association scan signals. *Hum. Mol. Genet.*, **19**, R137–R144.

Platt,A. *et al.* (2010) Conditions under which genome-wide association studies will be positively misleading. *Genetics*, **186**, 1045–1052.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Scherag,A. *et al.* (2010) Investigation of a genome wide association signal for obesity: synthetic association and haplotype analyses at the melanocortin 4 receptor gene locus. *PLoS One*, **5**, e13967.

Takeuchi,F. *et al.* (2009) A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. *PLoS Genet.*, **5**, e1000433.

Takeuchi,F. *et al.* (2011) Detection of common single nucleotide polymorphisms synthesizing quantitative trait association of rarer causal variants. *Genome Res.*, **21**, 1122–1130.

Wadelius,M. *et al.* (2007) Association of warfarin dose with genes involved in its action and metabolism. *Hum. Genet.*, **121**, 23–34.

Wang,K. *et al.* (2010) Interpretation of association signals and identification of causal variants from genome-wide association studies. *Am. J. Hum. Genet.*, **86**, 730–742.

Wray,N.R. *et al.* (2011) Synthetic associations created by rare variants do not explain most GWAS results. *PLoS Biol.*, **9**, e1000579.