Human Molecular Genetics, 2010, Vol. 19, Review Issue 2 ***R137–R144***
doi:10.1093/hmg/ddq368
Advance Access published on August 30, 2010

Synthetic associations in the context of genome-wide association scan signals

Gisela Orozco[1], Jeffrey C. Barrett[2] and Eleftheria Zeggini[2,*]

[1]Arthritis Research UK Epidemiology Unit, University of Manchester, Manchester, UK and [2]Wellcome Trust Sanger Institute, The Morgan Building, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1HH, UK

Received August 12, 2010; Revised and Accepted August 26, 2010

Genome-wide association studies (GWAS) have successfully identified a large number of genetic variants associated with complex traits, but these only explain a small proportion of the total heritability. It has been recently proposed that rare variants can create 'synthetic association' signals in GWAS, by occurring more often in association with one of the alleles of a common tag single nucleotide polymorphism. While the ultimate evaluation of this hypothesis will require the completion of large-scale sequencing studies, it is informative to place it in the broader context of what is known about the genetic architecture of complex disease. In this review, we draw from empirical and theoretical data to summarize evidence showing that synthetic associations do not underlie many reported GWAS associations.

GENOME-WIDE ASSOCIATION STUDIES AND 'THE MISSING HERITABILITY'

Numerous common human diseases and phenotypic traits are believed to arise from a combination of genetic and environmental factors. The unravelling of the genetic predisposition to complex traits is a major challenge, and it could lead to better prevention, diagnosis and treatment of disease.

Recently, advances in genotyping technologies, reduction in genotyping costs and the availability of data regarding genome-wide sequence variation through the International HapMap Project and 1000 genomes project have made genome-wide association studies (GWAS) possible. GWAS have emerged as a powerful tool for identifying genetic variants associated with complex traits. In the past few years, more than 500 loci have been found to be associated with human common diseases and traits (1). GWAS have proven to be much more successful than linkage studies, which were underpowered to detect variants of modest effect (2), and candidate gene studies, which are non-systematic and biased due to our limited knowledge of the biological pathways implicated in disease pathogenesis (3).

GWAS are based on the common disease–common variant (CDCV) hypothesis (4), which states that relatively common genetic variants (MAF $>$ 5%) of relatively low penetrance are important contributors to the genetic susceptibility to common diseases. Well-powered GWAS, which capture a substantial majority of common variation in the genome, have been now conducted for many common diseases. However, for the majority of these diseases, common variants explain only a small proportion of heritability (5), due to small individual effect sizes. It has been estimated that only 13% of all identified susceptibility loci have odds ratios (OR) above 2, and only 1% have OR above 10 (6). For example, if we consider a total estimated sibling recurrence risk ratio (λ_s) of 5–10 for rheumatoid arthritis (RA) (7), 15 for type 1 diabetes (T1D) (8), 17–35 for Crohn's disease (CD) (9) and 3 for type 2 diabetes (T2D) (10), their established susceptibility loci would contribute ~33–47%, 55.6%, 10–12.6% and 11.9% of the total heritability, respectively (Table 1).

POSSIBLE CONTRIBUTORS TO THE UNEXPLAINED HERITABILITY

Explaining this 'missing heritability' of complex diseases (11–13) is an area of active research, and there are likely to be multiple contributing factors. Part of the explanation is likely to be an underestimate of the contribution made by the types of variants targeted by GWAS. For instance, it might be that there are large numbers of variants of very small effect, which early GWAS were underpowered to

*To whom correspondence should be addressed. Tel: +44 1223496868; Fax: +44 1223 496826; Email: eleftheria@sanger.ac.uk

Table 1. Established susceptibility loci for RA, T1D, CD and T2D

Chromosome	SNP	Position	Region/gene	RAF	OR	λ_s	Reference
Rheumatoid arthritis							
1p36	rs3890745	2553624	*TNFSFR14*	0.68	1.12	1.003	(16)
1p13	rs2476601	114377568	*PTPN22*	0.10	1.94	1.068	(16)
1p13	rs11586238	117263138	*CD2*, *CD58*	0.24	1.13	1.003	(16)
1q23	rs12746613	161467042	*FCGR2A*	0.12	1.13	1.002	(16)
1q31	rs10919563	198700442	*PTPRC*	0.87	1.14	1.002	(16)
2p16	rs13031237	61136129	*REL*	0.37	1.13	1.004	(16)
2p14	rs934734	65595586	*SPRED2*	0.49	1.13	1.004	(16)
2q11	rs10865035	100835734	*AFF3*	0.47	1.12	1.003	(16)
2q32	rs7574865	191964633	*STAT4*	0.22	1.16	1.004	(16)
2q33	rs1980422	204610396	*CD28*	0.24	1.12	1.002	(16)
2q33	rs3087243	204738919	*CTLA4*	0.56	1.15	1.005	(16)
3p14	rs13315591	58556841	*PXK*	0.09	1.29	1.007	(16)
4p15	rs874040	26108197	*RBPJ*	0.30	1.14	1.004	(16)
4q27	rs6822844	123509421	*IL2*, *IL21*	0.82	1.11	1.002	(16)
5q11	rs6859219	55438580	*ANKRD55*, *IL6ST*	0.79	1.28	1.009	(16)
5q21	rs26232	102596720	*C5orf13*	0.68	1.14	1.004	(16)
6p21			*HLA*			1.800	(70,71)
6q21	rs548234	106568034	*PRDM1*	0.33	1.10	1.002	(16)
6q23	rs10499194	138002637	*TNFAIP3*	0.73	1.10	1.002	(16)
6q23	rs6920220	138006504	*TNFAIP3*	0.22	1.22	1.008	(16)
6q23	rs5029937	138195151	*TNFAIP3*	0.04	1.40	1.006	(16)
6q25	rs394581	159482521	*TAGAP*	0.70	1.10	1.002	(16)
6q27	rs3093023	167534290	*CCR6*	0.43	1.13	1.004	(16)
7q32	rs10488631	128594183	*IRF5*	0.11	1.19	1.003	(16)
8p23	rs2736340	11343973	*BLK*	0.25	1.12	1.003	(16)
9p13	rs2812378	34710260	*CCL21*	0.34	1.10	1.002	(16)
9q33	rs3761847	123690239	*TRAF1*, *C5*	0.43	1.13	1.004	(16)
10p15	rs2104286	6099045	*IL2RA*	0.73	1.09	1.001	(16)
10p15	rs4750316	6393260	*PRKCQ*	0.81	1.15	1.003	(16)
11p12	rs540386	36525293	*TRAF6*	0.86	1.14	1.002	(16)
12q13	rs1678542	57968715	*KIF5A*	0.62	1.10	1.002	(16)
20q13	rs4810485	44747947	*CD40*	0.75	1.18	1.005	(16)
22q12	rs3218253	37544810	*IL2RB*	0.26	1.09	1.001	(16)
Type 1 diabetes							
1p31	rs2269241	64108771	*PGM1*	0.19	1.10	1.001	(15)
1p13	rs2476601	114377568	*PTPN22*	0.14	2.05	1.104	(15)
1q31	rs2816316	192536813	*RGS1*	0.82	1.12	1.002	(15)
1q32	rs3024505	206939904	*IL10*	0.83	1.19	1.004	(15)
2p25	rs1534422	12640741	*2p25*	0.46	1.08	1.001	(15)
2q12	rs917997	103070568	*IL18RAP*	0.78	1.20	1.005	(15)
2q24	rs1990760	163124051	*IFIH1*	0.60	1.16	1.005	(15)
2q33	rs3087243	204738919	*CTLA4*	0.55	1.14	1.004	(15)
3p21	rs333	46345611	*CCR5*	0.88	1.18	1.003	(15)
4p15	rs10517086	26085511	*4p15*	0.30	1.09	1.002	(15)
4q27	rs17388568	123132492	*IL2*	0.26	1.26	1.011	(15)
5p13	rs6897932	35874575	*IL7R*	0.73	1.12	1.002	(15)
6p21			*MHC*			3.058	(72)
6q15	rs11755527	90958231	*BACH2*	0.47	1.13	1.004	(15)
6q22	rs9388489	126698719	*C6orf173*	0.45	1.17	1.006	(15)
6q23	rs6920220	137973068	*TNFAIP3*	0.22	1.09	1.001	(15)
6q25	rs1738074	159465977	*TAGAP*	0.56	1.09	1.002	(15)
7p15	rs7804356	26891665	*7p15*	0.76	1.14	1.003	(15)
7p12	rs4948088	51027194	*COBL*	0.95	1.30	1.002	(15)
9p24	rs7020673	4291747	*GLIS3*	0.50	1.14	1.004	(15)
10p15	rs11594656	6,62015	*IL2RA*	0.75	1.19	1.005	(15)
10p15	rs12722495	6137289	*IL2RA*	0.89	1.59	1.015	(15)
10p15	rs947474	6430456	*PRKCQ*	0.81	1.10	1.001	(15)
10q23	rs10509540	90023033	*C10orf59*	0.72	1.33	1.015	(15)
11p15	rs689	2138800	*INS*	0.71	2.30	1.096	(15)
12p13	rs4763879	9910164	*CD69*	0.37	1.09	1.002	(15)
12q13	rs2292239	56482180	*ERBB3*	0.34	1.31	1.018	(15)
12q13	rs1678536	57979190	*Multiple*	0.72	1.12	1.002	(15)
12q24	rs3184504	111884608	*SH2B3*	0.49	1.28	1.015	(15)
14q24	rs1465788	69263599	*14q24*	0.71	1.16	1.004	(15)
14q32	rs4900384	98498951	*14q32*	0.29	1.09	1.002	(15)
15q25	rs3825932	79235446	*CTSH*	0.68	1.16	1.005	(15)

Continued

Table 1. *Continued*

Chromosome	SNP	Position	Region/gene	RAF	OR	λ_s	Reference
16p13	rs12708716	11179873	*CLEC16A*	0.65	1.23	1.009	(15)
16p12	rs12444268	20342572	*16p12*	0.30	1.10	1.002	(15)
16p11	rs4788084	28539848	*IL27*	0.58	1.16	1.005	(15)
16q23	rs7202877	75247245	*16q23*	0.10	1.13	1.001	(15)
17p13	rs16956936	7633692	*17p13*	0.87	1.09	1.001	(15)
17q12	rs2290400	38066240	*ORMDL3*	0.51	1.15	1.005	(15)
17q21	rs7221109	38770286	*17q21*	0.65	1.05	1.001	(15)
18p11	rs1893217	12809340	*PTPN2*	0.17	1.13	1.002	(15)
18q22	rs763361	67531642	*CD226*	0.47	1.16	1.006	(15)
19q13	rs425105	47208481	*19q13*	0.84	1.16	1.003	(15)
20p13	rs2281808	1610551	*20p13*	0.64	1.11	1.002	(15)
21q22	rs11203203	43836186	*UBASH3A*	0.43	1.13	1.004	(15)
22q12	rs5753037	30581722	*22q12*	0.39	1.10	1.002	(15)
22q13	rs229541	37591318	*C1QTNF6*	0.43	1.11	1.003	(15)
Xq28	rs2664170	153945602	*Xq28*	0.32	1.16	1.005	(15)
Crohn's disease							
1p31	rs11465804	67475114	*IL23R*	0.93	2.50	1.025	(14)
1p13	rs2476601	114179091	*PTPN22*	0.90	1.31	1.005	(14)
1q23	rs2274910	159118670	*ITLN1*	0.68	1.14	1.004	(14)
1q24	rs9286879	171128857	*1q24*	0.24	1.19	1.006	(14)
1q32	rs11584383	199202489	*1q32*	0.70	1.18	1.005	(14)
2q27	rs3828309	233845149	*ATG16L1*	0.53	1.28	1.015	(14)
3p21	rs3197999	49696536	*MST1*	0.27	1.20	1.007	(14)
5p13	rs4613763	40428485	*PTGER4*	0.13	1.32	1.010	(14)
5q31	rs2188962	131798704	*5q31*	0.43	1.25	1.013	(14)
5q33	rs11747270	150239060	*IRGM*	0.09	1.33	1.008	(14)
5q33	rs10045431	158747111	*IL12B*	0.71	1.11	1.002	(14)
6p22	rs6908425	20836710	*CDKAL1*	0.78	1.21	1.006	(14)
6q21	rs7746082	106541962	*6q21*	0.29	1.17	1.005	(14)
6q27	rs2301436	167357978	*CCR6*	0.46	1.21	1.009	(14)
7p12	rs1456893	50240218	*7p12*	0.68	1.20	1.007	(14)
8q24	rs1551398	126609233	*8q24*	0.62	1.08	1.001	(14)
9p24	rs10758669	4971602	*JAK2*	0.35	1.12	1.003	(14)
9q32	rs4263839	116606261	*TNFSF15*	0.68	1.22	1.008	(14)
10p11	rs17582416	35327656	*10p11*	0.35	1.16	1.005	(14)
10q21	rs10995271	64108492	*ZNF365*	0.39	1.25	1.012	(14)
10q24	rs11190140	101281583	*NKX2-3*	0.48	1.20	1.008	(14)
11q13	rs7927894	75978964	*C11orf30*	0.39	1.16	1.005	(14)
12q12	rs11175593	38888207	*LRRK2, MUC19*	0.02	1.54	1.005	(14)
13q14	rs3764147	43355925	*13q14*	0.22	1.25	1.010	(14)
16q12	rs2066847	49321280	*NOD2*	0.02	3.99	1.147	(14)
17q21	rs2872507	35294289	*ORMDL3*	0.47	1.12	1.003	(14)
17q21	rs744166	37767727	*STAT3*	0.57	1.18	1.007	(14)
18p11	rs2542151	12769947	*PTPN2*	0.15	1.35	1.014	(14)
21q21	rs1736135	15727091	*21q21*	0.57	1.18	1.007	(14)
21q22	rs762421	44439989	*ICOSLG*	0.39	1.13	1.004	(14)
Type 2 diabetes							
1p13–p11	rs10923931	120319482	*NOTCH2*	0.11	1.09	1.001	(73)
1q32	rs340874	212225879	*PROX1*	0.56	1.07	1.001	(18)
2p23	rs780094	27594741	*GCKR*	0.61	1.06	1.001	(18)
2p21	rs7578597	43586327	*THADA*	0.90	1.15	1.002	(73)
2p16	rs243021	60438323	*BCL11A*	0.46	1.08	1.001	(74)
2q26	rs7578326	226728897	*IRS1*	0.64	1.11	1.002	(74)
3p25	rs1801282	12368125	*PPARG*	0.85	1.23	1.005	(73)
3p14	rs4607103	64686944	*ADAMTS9*	0.76	1.10	1.002	(73)
3q13–q21	rs11708067	124548468	*ADCY5*	0.77	1.12	1.002	(18)
3q27	rs4402960	186994381	*IGF2BP2*	0.31	1.11	1.002	(73)
4p16	rs10010131	6343816	*WFS1*	0.59	1.14	1.004	(73)
5q13	rs4457053	76460705	*ZBED3*	0.26	1.08	1.001	(74)
6p22	rs10946398	20769013	*CDKAL1*	0.33	1.09	1.002	(73)
7p21	rs2191349	15030834	*DGKB/TMEM195*	0.47	1.06	1.001	(18)
7p15	rs864745	28147081	*JAZF1*	0.50	1.08	1.001	(73)
7p15	rs4607517	44202193	*GCK*	0.2	1.07	1.001	(18)
7q32	rs972283	130117394	*KLF14*	0.55	1.07	1.001	(74)
8q22	rs896854	96029687	*TP53INP1*	0.48	1.06	1.001	(74)
8q24	rs13266634	118253964	*SLC30A8*	0.68	1.12	1.003	(73)

Continued

Table 1. *Continued*

Chromosome	SNP	Position	Region/gene	RAF	OR	λ_s	Reference
9p21	rs10811661	22124094	*CDKN2A/B*	0.84	1.17	1.003	(73)
9q21	rs13292136	81141948	*CHCHD9*	0.93	1.11	1.001	(74)
10p13	rs12779790	12368016	*CDC123/CAMK1D*	0.18	1.11	1.002	(73)
10q23	rs5015480	94455539	*HHEX/IDE*	0.59	1.10	1.002	(73)
11p15	rs2334499	1653428	*DUSP8*	0.41	1.08	1.001	(32)
11p15	rs231362	2648047	*KCNQ1*	0.52	1.08	1.001	(74)
11p15	rs2237892	2796327	*KCNQ1*	0.34	1.42	1.031	(73)
11p15	rs5219	17366148	*KCNJ11*	0.39	1.15	1.005	(73)
11q13	rs1552224	72110746	*CENTD2*	0.88	1.14	1.002	(74)
11q21	rs10830963	92348358	*MTNR1B*	0.30	1.09	1.001	(73)
12q14	rs1531343	64461161	*HMGA2*	0.1	1.1	1.001	(74)
12q14–q21	rs7961581	69949369	*TSPAN8/LGR5*	0.27	1.06	1.001	(73)
12q24	rs7957197	119945069	*HNF1A*	0.85	1.07	1.001	(74)
15q25	rs11634397	78219277	*ZFAND6*	0.6	1.06	1.001	(74)
15q26	rs8042680	89322341	*PRC1*	0.22	1.07	1.001	(74)
16q12	rs8050136	52373776	*FTO*	0.38	1.21	1.009	(73)
17cen–q21.3	rs757210	33170628	*HNF1B (TCF2)*	0.38	1.10	1.002	(73)
Xq28	rs5945326	152553116	*DUSP9*	0.21	1.27	1.011	(74)

RA, rheumatoid arthritis; T1D, type 1 diabetes; CD, Crohn's disease; T2D, type 2 diabetes; RAF, risk allele frequency in controls; OR, odds ratio.
Sibling recurrence risk ratio (λ_s) was calculated using the formula:

$$\lambda_s = \left(1 + \frac{pq(\gamma - 1)^2}{2(p + \gamma q)^2}\right)^2$$

where q is the risk allele frequency, $p = 1 - q$, and γ is the genotype relative risk under the additive model.

detect, yet to be found. This idea is supported by the observation that meta-analyses of published GWAS are discovering a substantial number of new susceptibility loci (14–25). In addition, for most loci, causal variants and potential independent additional markers within the region have not been identified yet. New ways of analysing the genetic architecture of complex traits using GWAS data are suggesting that indeed a large proportion of heritability can be explained by common variants and that larger GWAS will yield many more validated loci for complex traits (26,27).

Of course, GWAS only interrogate a portion of the types of variation that could underlie disease risk. Analysis of GWAS data has been mainly focused on single nucleotide polymorphisms (SNPs), but there are other types of genetic variation, such as structural variants, that have not been studied in depth. However, recent studies of common (MAF > 5%) copy number variants (CNVs) have shown that they seem unlikely to account for a substantial proportion of the 'missing heritability' (28). Similarly, the analysis of gene–environment and gene–gene interactions (epistasis) might improve the fraction of heritability explained by loci documented thus far. Several epistatic interactions have been indentified in humans [e.g. between the RET protooncogene and endothelin receptor type B genes in Hirschsprung disease (29), the interleukin 4 receptor variants and interleukin 13 promoter variants in asthma (30) and the alpha- and beta-adrenergic receptors in congestive heart failure (31)], although they have not been replicated. However, this phenomenon has not been thoroughly explored through large-scale analysis of genome-wide SNP interactions, first due to the fact that current sample sizes are underpowered to detect modest interaction effects and secondly due to the paucity of sample collections with genetic and detailed environmental exposure data. Complex patterns of inheritance, such as parent of origin effects (32), as well as inherited epigenetic modifications of the genome, the presence of phenotype heterogeneity in the cohorts used in the first wave of GWAS, or even an initial over-estimation of the heritability of complex traits (33) can also contribute to the missing heritability.

While the above-mentioned plausible contributors seem unlikely to play a substantial role in explaining missing heritability, rare variants are increasingly thought to account for a large proportion of it (34–36). Contrary to the CDCV hypothesis, the multiple rare variant (MRV) hypothesis argues that the summation of the effects of low-frequency polymorphisms, each conferring an intermediate increase in risk (i.e. incompletely penetrant, but greater than those observed for common variants), can explain a significant proportion of the genetic susceptibility to common diseases and traits. Some studies analysing rare variants using GWAS data have been carried out, but these have proven to be underpowered to detect robust associations. Re-sequencing approaches are more suitable for rare variant analysis, and, as these are becoming more cost-effective and new analysis methods are being developed (37,38), they will soon be applied to large-scale studies of rare variants. Indeed, several targeted sequencing studies have already proven successful for the identification of associations between rare variants and some human diseases and disease-related phenotypes (39–43). The same argument can also be extended to other forms of genetic variation, and it has been recently proposed that rare CNVs may be responsible for some fraction of the missing heritability of complex traits (44,45).

SYNTHETIC ASSOCIATIONS HYPOTHESIS

It has been recently proposed that GWAS signals that have been credited to common variants could instead reflect the effect of MRVs. Dickson *et al*. (46) argue that rare variants can create 'synthetic association' signals in GWAS, by occurring more often in association with one of the alleles of a common tag SNP (Fig. 1), which would thus synthetically confer an increased risk for disease. This might also mean that the causal variants could be megabases away from the common variants detected in GWAS, and that the real effect size could be much stronger than that implied by the common tag SNP. If true, the synthetic association hypothesis would suggest that follow-up studies from GWAS hits should encompass a much larger region than the linkage disequilibrium region surrounding the detected common variant (6).

There are very few documented examples showing that MRVs may be responsible for a common variant GWAS signal (47). It therefore seems sensible to evaluate this hypothesis in the broader context of human disease genetics, including historical study designs, functional annotations of GWAS regions and experiments in human populations with diverse ancestry. While sequencing experiments currently underway or in planning will ultimately resolve the role of synthetic association, the balance of evidence available today is already illuminating.

LINKAGE EVIDENCE SUGGESTS SYNTHETIC ASSOCIATIONS ARE RARE

One line of evidence that suggests that synthetic associations do not underlie many reported GWAS associations is provided by linkage scans that have been conducted in the past. The genetic model that underpins synthetic association (allelic heterogeneity caused by several low-frequency variants with larger effects than commonly seen in GWAS) is highly tractable by linkage analysis, which combines information from all causal variants at a particular locus. This relationship is highlighted by the widely replicated linkage between the *NOD2* gene and CD, which is driven by three independent, low-frequency causal variants (48–50) which cause a synthetic association signal in GWAS of CD (Fig. 1). *NOD2* is the exception that proves the rule that, despite many attempts, very few replicable linkages to complex diseases have been discovered (51). This dearth of findings is informative when considering the likelihood of synthetic associations because it rules out a class of genetic models from playing a substantial role in complex disease.

Power calculations comparing a large-scale linkage scan (52) with the largest GWAS considered by Dickson *et al*. (46) show that only a small fraction of the genetic models which can give rise to synthetic associations would not be detected by linkage. Furthermore, the scenario where synthetic associations could have escaped linkage comprises models with a small number of causal variants with genotype relative risk <2.5 (53). While these observations do not entirely rule out synthetic associations, they seriously confine the parameter space in which they might exist. In addition, comparisons of even modest linkage signals with GWAS regions have shown only a few overlaps, and even these are largely driven by atypically large effects like the MHC in autoimmunity. In addition, attempts to explicitly use linkage information to boost the power of GWAS (54) have not been successful. This contrast between largely overlapping genetic models that linkage and synthetic association are well powered to detect and almost completely non-overlapping results from linkage and GWAS strongly suggests that synthetic associations do not underlie many GWAS signals.

PATHWAY ANALYSES IMPLY GWAS ARE POINTING TO KEY FUNCTIONAL ELEMENTS

Another prediction made by the synthetic association hypothesis is that the most significantly associated common variant identified by GWAS might be located several megabases away from the underlying low-frequency functional variants. The empirical properties of linkage disequilibrium between low-frequency and common variants are not fully understood, although the complete 1000 Genomes project (http://www.1000genomes.org/) will soon provide information necessary to evaluate this question directly. Nevertheless, two indirect pieces of evidence suggest that most GWAS hit SNPs are within a few hundred kilobases (and many within tens of kilobases) of their tagged functional alleles. First, a large number of GWAS signals across a variety of traits are nearby to genes previously established to cause Mendelian forms of the same trait (55). Secondly, genes involved in key pathways repeatedly arise in GWAS of some diseases. For example, 8 of 10 proteins involved in the Th17-differentiation signalling pathway have been associated with one or more auto-inflammatory diseases (56). As with many aspects relating to the evaluation of the prevalence of synthetic associations, deeper sequence data sets will be needed to fully answer the question of the distance between GWAS tag SNPs and causal variants, but these early patterns imply that the tag SNP often resides in the proximity of the relevant functional genomic element.

TRANS-ETHNIC ASSOCIATIONS ARE WIDESPREAD

Under the synthetic associations model, common variant signals reflecting single or multiple rare alleles are unlikely to be consistent across populations of different ancestry. This is based on the fact that many of these rare variants would have arisen recently and will therefore not be shared across diverged populations. The majority of GWAS to date have focused on populations of European descent. However, data on more diverse populations are now starting to arise. For example, a study from early 2010 clearly demonstrated that common variant signals for T2D are reproducible and have similar effect sizes across East Asian populations including Chinese, Malays and Asian-Indians in Singapore (57). In fact, T2D-associated variants have been found to be associated with disease in diverse populations (ranging from African-Americans to Chinese) by several studies (58–62). Similarly, in RA, the *STAT4* locus, as an example, has shown reproducible association with disease in the USA

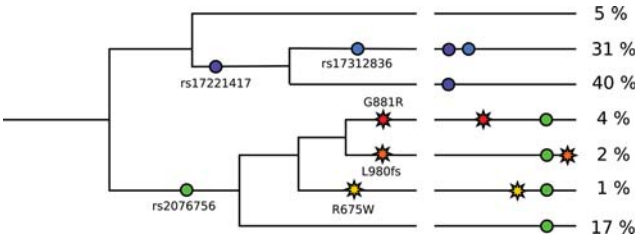

Figure 1. Simplified view of genetic variation at the *NOD2* locus, a well-documented example of a synthetic association. The left-hand side shows a genealogical tree representing six SNPs in this region after discarding rare recombinant events. The right-hand side shows the resulting haplotypes and their population frequencies (48), with coloured circles representing common GWAS SNPs, and starbursts representing previously identified low-frequency coding variants responsible for association between *NOD2* and CD. While none of the GWAS SNPs is strongly correlated with any individual causal allele, the three coding variants create a synthetic association because they cluster by chance on the side of the tree marked by the green GWAS SNP (rs2076756).

(63), UK (64), Spanish, Swedish, Dutch (65), Korean (66), Colombian (67), Japanese (68) and Greek (69) populations.

FUTURE DIRECTIONS FOR GWAS AND THE SEARCH FOR GENETIC CAUSES OF COMMON DISEASE

Although synthetic associations explaining common GWAS signals for complex polygenic traits are certainly plausible and can occur under specific circumstances (e.g. *NOD2* in CD), results from studies thus far suggest that these scenarios are actually a rarity. The idea that MRVs at a particular locus may be associated with complex traits of interest has been around for over a decade. We are now starting to accrue a growing body of empirical evidence in support of this hypothesis. The field of complex trait genetics has over the last few months engaged in discussions on the controversial topic of synthetic associations, but it transpires that there is little evidence to support this as a widespread scenario.

Empowered by advances in sequencing technologies, attention is currently shifting towards the comprehensive study of low-frequency and rare variants. Resources such as the 1000 genomes project and emerging large-scale studies like the UK10k project will undoubtedly facilitate the examination of variants at this end of the allele frequency spectrum. In parallel, improved strategies for accurate imputation and powerful analysis of low-frequency and rare variants in aggregate are being further developed and fine-tuned to the needs of these next generation truly genome-wide scans for association.

Conflict of Interest statement. None declared.

FUNDING

G.O. is funded by the European Union (Marie Curie IEF Fellowship PIEF-GA-2009-235662). E.Z. and J.C.B. are supported by the Wellcome Trust (WT088885/Z/09/Z, WT089120/Z/09/Z). Funding to pay the Open Access Charge was provided by The Wellcome Trust.

REFERENCES

1. Hindorff, L.A., Junkins, H.A., Hall, P.N., Mehta, J.P. and Manolio, T.A. (2010) A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies.
2. Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.
3. Hardy, J. and Singleton, A. (2009) Genomewide association studies and human disease. *N. Engl. J. Med.*, **360**, 1759–1768.
4. Reich, D.E. and Lander, E.S. (2001) On the allelic spectrum of human disease. *Trends Genet.*, **17**, 502–510.
5. Frazer, K.A., Murray, S.S., Schork, N.J. and Topol, E.J. (2009) Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.*, **10**, 241–251.
6. Cirulli, E.T. and Goldstein, D.B. (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.*, **11**, 415–425.
7. Wordsworth, P. and Bell, J. (1991) Polygenic susceptibility in rheumatoid arthritis. *Ann. Rheum. Dis.*, **50**, 343–346.
8. Hyttinen, V., Kaprio, J., Kinnunen, L., Koskenvuo, M. and Tuomilehto, J. (2003) Genetic liability of type 1 diabetes and the onset age among 22,650 young Finnish twin pairs: a nationwide follow-up study. *Diabetes*, **52**, 1052–1055.
9. Tysk, C., Lindberg, E., Jarnerot, G. and Floderus-Myrhed, B. (1988) Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking. *Gut*, **29**, 990–996.
10. Kobberling, J. and Tattersall, R. (1982) *The Genetics of Diabetes Mellitus*. Academic Press, London.
11. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
12. Maher, B. (2008) Personal genomes: the case of the missing heritability. *Nature*, **456**, 18–21.
13. Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H. and Nadeau, J.H. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.*, **11**, 446–450.
14. Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M. *et al.* (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.*, **40**, 955–962.
15. Barrett, J.C., Clayton, D.G., Concannon, P., Akolkar, B., Cooper, J.D., Erlich, H.A., Julier, C., Morahan, G., Nerup, J., Nierras, C. *et al.* (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.*, **41**, 703–707.
16. Stahl, E.A., Raychaudhuri, S., Remmers, E.F., Xie, G., Eyre, S., Thomson, B.P., Li, Y., Kurreeman, F.A., Zhernakova, A., Hinks, A. *et al.* (2010) Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.*, **40**, 955–962.
17. Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., de Bakker, P.I., Abecasis, G.R., Almgren, P., Andersen, G. *et al.* (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.*, **40**, 638–645.
18. Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A.U., Wheeler, E., Glazer, N.L., Bouatia-Naji, N., Gloyn, A.L. *et al.* (2010) New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.*, **42**, 105–116.
19. Tobacco and Genetics Consortium. (2010) Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat. Genet.*, **42**, 441–447.
20. Hancock, D.B., Eijgelsheim, M., Wilk, J.B., Gharib, S.A., Loehr, L.R., Marciante, K.D., Franceschini, N., van Durme, Y.M., Chen, T.H., Barr, R.G. *et al.* (2010) Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat. Genet.*, **42**, 45–52.
21. Kottgen, A., Pattaro, C., Boger, C.A., Fuchsberger, C., Olden, M., Glazer, N.L., Parsa, A., Gao, X., Yang, Q., Smith, A.V. *et al.* (2010) New loci associated with kidney function and chronic kidney disease. *Nat. Genet.*, **42**, 376–384.
22. McGovern, D.P., Gardet, A., Torkvist, L., Goyette, P., Essers, J., Taylor, K.D., Neale, B.M., Ong, R.T., Lagace, C., Li, C. *et al.* (2010)

Genome-wide association identifies multiple ulcerative colitis susceptibility loci. *Nat. Genet.*, **42**, 332–337.
23. McMahon, F.J., Akula, N., Schulze, T.G., Muglia, P., Tozzi, F., tera-Wadleigh, S.D., Steele, C.J., Breuer, R., Strohmaier, J., Wendland, J.R. *et al.* (2010) Meta-analysis of genome-wide association data identifies a risk locus for major mood disorders on 3p21.1. *Nat. Genet.*, **42**, 128–131.
24. Soranzo, N., Spector, T.D., Mangino, M., Kuhnel, B., Rendon, A., Teumer, A., Willenborg, C., Wright, B., Chen, L., Li, M. *et al.* (2009) A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat. Genet.*, **41**, 1182–1190.
25. Wang, T.J., Zhang, F., Richards, J.B., Kestenbaum, B., van Meurs, J.B., Berry, D., Kiel, D.P., Streeten, E.A., Ohlsson, C., Koller, D.L. *et al.* (2010) Common genetic determinants of vitamin D insufficiency: a genome-wide association study. *Lancet*, **376**:180–188.
26. Park, J.H., Wacholder, S., Gail, M.H., Peters, U., Jacobs, K.B., Chanock, S.J. and Chatterjee, N. (2010) Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.*, **42**, 570–575.
27. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W. *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, **42**, 565–569.
28. Craddock, N., Hurles, M.E., Cardin, N., Pearson, R.D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D.F., Giannoulatou, E. *et al.* (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, **464**, 713–720.
29. Carrasquillo, M.M., McCallion, A.S., Puffenberger, E.G., Kashuk, C.S., Nouri, N. and Chakravarti, A. (2002) Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. *Nat. Genet.*, **32**, 237–244.
30. Howard, T.D., Koppelman, G.H., Xu, J., Zheng, S.L., Postma, D.S., Meyers, D.A. and Bleecker, E.R. (2002) Gene–gene interaction in asthma: IL4RA and IL13 in a Dutch population with asthma. *Am. J. Hum. Genet.*, **70**, 230–236.
31. Smal, K.M., Wagoner, L.E., Levin, A.M., Kardia, S.L. and Liggett, S.B. (2002) Synergistic polymorphisms of beta1- and alpha2C-adrenergic receptors and the risk of congestive heart failure. *N. Engl. J. Med.*, **347**, 1135–1142.
32. Kong, A., Steinthorsdottir, V., Masson, G., Thorleifsson, G., Sulem, P., Besenbacher, S., Jonasdottir, A., Sigurdsson, A., Kristinsson, K.T., Jonasdottir, A. *et al.* (2009) Parental origin of sequence variants associated with complex diseases. *Nature*, **462**, 868–874.
33. Clarke, A.J. and Cooper, D.N. (2010) GWAS: heritability missing in action? *Eur. J. Hum. Genet.*, **18**, 859–861.
34. Pritchard, J.K. (2001) Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.*, **69**, 124–137.
35. Bodmer, W. and Bonilla, C. (2008) Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.*, **40**, 695–701.
36. Schork, N.J., Murray, S.S., Frazer, K.A. and Topol, E.J. (2009) Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.*, **19**, 212–219.
37. Li, B. and Leal, S.M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, **83**, 311–321.
38. Morris, A.P. and Zeggini, E. (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.*, **34**, 188–193.
39. Ahituv, N., Kavaslar, N., Schackwitz, W., Ustaszewska, A., Martin, J., Hebert, S., Doelle, H., Ersoy, B., Kryukov, G., Schmidt, S. *et al.* (2007) Medical sequencing at the extremes of human body mass. *Am. J. Hum. Genet.*, **80**, 779–791.
40. Cohen, J.C., Boerwinkle, E., Mosley, T.H. Jr and Hobbs, H.H. (2006) Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.*, **354**, 1264–1272.
41. Ji, W., Fo, J.N., O'Roak, B.J., Zhao, H., Larson, M.G., Simon, D.B., Newton-Cheh, C., State, M.W., Levy, D. and Lifton, R.P. (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat. Genet.*, **40**, 592–599.
42. Nejentsev, S., Walker, N., Riches, D., Egholm, M. and Todd, J.A. (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, **324**, 387–389.
43. Romeo, S., Pennacchio, L.A., Fu, Y., Boerwinkle, E., Tybjaerg-Hansen, A., Hobbs, H.H. and Cohen, J.C. (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat. Genet.*, **39**, 513–516.
44. Walsh, T., McClellan, J.M., McCarthy, S.E., Addington, A.M., Pierce, S.B., Cooper, G.M., Nord, A.S., Kusenda, M., Malhotra, D., Bhandari, A. *et al.* (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*, **320**, 539–543.
45. Zhang, F., Seeman, P., Liu, P., Weterman, M.A., Gonzaga-Jauregui, C., Towne, C.F., Batish, S.D., De Vriendt, E., De Jonghe, J.P., Rautenstraus, B. *et al.* (2010) Mechanisms for nonrecurrent genomic rearrangements associated with CMT1A or HNPP: rare CNVs as a cause for missing heritability. *Am. J. Hum. Genet.*, **86**, 892–903.
46. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. and Goldstein, D.B. (2010) Rare variants create synthetic genome-wide associations. *PLoS Biol.*, **8**, e1000294.
47. Wang, K., Dickson, S.P., Stolle, C.A., Krantz, I.D., Goldstein, D.B. and Hakonarson, H. (2010) Interpretation of association signals and identification of causal variants from genome-wide association studies. *Am. J. Hum. Genet.*, **86**, 730–742.
48. Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
49. Hugot, J.P., Chamaillard, M., Zouali, H., Lesage, S., Cezard, J.P., Belaiche, J., Almer, S., Tysk, C., O'Morain, C.A., Gassul, M. *et al.* (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, **411**, 599–603.
50. Ogura, Y., Bonen, D.K., Inohara, N., Nicolae, D.L., Chen, F.F., Ramos, R., Britton, H., Moran, T., Karaliuskas, R., Duer, R.H. *et al.* (2001) A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature*, **411**, 603–606.
51. McCarthy, M.I. (2002) Susceptibility gene discovery for common metabolic and endocrine traits. *J. Mol. Endocrinol.*, **28**, 1–17.
52. Concannon, P., Chen, W.M., Julier, C., Morahan, G., Akolkar, B., Erlich, H.A., Hilner, J.E., Nerup, J., Nierras, C., Pociot, F. *et al.* (2009) Genome-wide scan for linkage to type 1 diabetes in 2,496 multiplex families from the Type 1 Diabetes Genetics Consortium. *Diabetes*, **58**, 1018–1022.
53. Anderson, C., Barrett, J., Soranzo, N. and Zeggini, E. (2010) Synthetic associations are unlikely to account for many common disease genome-wide association signals. *PLoS Biol.*, in press.
54. Yo, Y.J., Bul, S.B., Paterson, A.D., Waggot, D. and Sun, L. (2010) Were genome-wide linkage studies a waste of time? Exploiting candidate regions within genome-wide association studies. *Genet. Epidemiol.*, **34**, 107–118.
55. O'Rahilly, S. (2009) Human genetics illuminates the paths to metabolic disease. *Nature*, **462**, 307–314.
56. Zhernakova, A., van Diemen, C.C. and Wijmenga, C. (2009) Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat. Rev. Genet.*, **10**, 43–55.
57. Tan, J.T., Ng, D.P., Nurbaya, S., Ye, S., Lim, X.L., Leong, H., Seet, L.T., Siew, W.F., Kon, W., Wong, T.Y. *et al.* (2010) Polymorphisms identified through genome-wide association studies and their associations with type 2 diabetes in Chinese, Malays, and Asian-Indians in Singapore. *J. Clin. Endocrinol. Metab.*, **95**, 390–397.
58. Han, X., Luo, Y., Ren, Q., Zhang, X., Wang, F., Sun, X., Zhou, X. and Ji, L. (2010) Implication of genetic variants near SLC30A8, HHEX, CDKAL1, CDKN2A/B, IGF2BP2, FTO, TCF2, KCNQ1, and WFS1 in type 2 diabetes in a Chinese population. *BMC Med. Genet.*, **11**, 81.
59. Lin, Y., Li, P., Cai, L., Zhang, B., Tang, X., Zhang, X., Li, Y., Xian, Y., Yang, Y., Wang, L. *et al.* (2010) Association study of genetic variants in eight genes/loci with type 2 diabetes in a Han Chinese population. *BMC Med. Genet.*, **11**, 97.
60. Takeuchi, F., Serizawa, M., Yamamoto, K., Fujisawa, T., Nakashima, E., Ohnaka, K., Ikegami, H., Sugiyama, T., Katsuya, T., Miyagishi, M. *et al.* (2009) Confirmation of multiple risk Loci and genetic impacts by a genome-wide association study of type 2 diabetes in the Japanese population. *Diabetes*, **58**, 1690–1699.
61. Yan, Y., North, K.E., Ballantyne, C.M., Brancati, F.L., Chambles, L.E., Franceschini, N., Heis, G., Kottgen, A., Pankow, J.S., Selvin, E. *et al.*

(2009) Transcription factor 7-like 2 (TCF7L2) polymorphism and context-specific risk of type 2 diabetes in African American and Caucasian adults: the Atherosclerosis Risk in Communities study. *Diabetes*, **58**, 285–289.
62. Zhou, D., Zhang, D., Liu, Y., Zhao, T., Chen, Z., Liu, Z., Yu, L., Zhang, Z., Xu, H. and He, L. (2009) The E23K variation in the KCNJ11 gene is associated with type 2 diabetes in Chinese and East Asian population. *J. Hum. Genet.*, **54**, 433–435.
63. Remmers, E.F., Plenge, R.M., Le, A.T., Graham, R.R., Hom, G., Behrens, T.W., de Bakker, P.I., Le, J.M., Le, H.S., Batliwalla, F. *et al.* (2007) STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. *N. Engl. J. Med.*, **357**, 977–986.
64. Barton, A., Thomson, W., Ke, X., Eyre, S., Hinks, A., Bowes, J., Gibbons, L., Plant, D., Wilson, A.G., Marinou, I. *et al.* (2008) Re-evaluation of putative rheumatoid arthritis susceptibility genes in the post-genome wide association study era and hypothesis of a key pathway underlying susceptibility. *Hum. Mol. Genet.*, **17**, 2274–2279.
65. Orozco, G., Alizadeh, B.Z., Delgado-Vega, A.M., Gonzalez-Gay, M.A., Balsa, A., Pascual-Salcedo, D., Fernandez-Gutierrez, B., Gonzalez-Escribano, M.F., Petersson, I.F., van Riel, P.L. *et al.* (2008) Association of STAT4 with rheumatoid arthritis: a replication study in three European populations. *Arthritis Rheum.*, **58**, 1974–1980.
66. Le, H.S., Remmers, E.F., Le, J.M., Kastner, D.L., Bae, S.C. and Gregersen, P.K. (2007) Association of STAT4 with rheumatoid arthritis in the Korean population. *Mol. Med.*, **13**, 455–460.
67. Palomino-Morales, R.J., Rojas-Villarraga, A., Gonzalez, C.I., Ramirez, G., Anaya, J.M. and Martin, J. (2008) STAT4 but not TRAF1/C5 variants influence the risk of developing rheumatoid arthritis and systemic lupus erythematosus in Colombians. *Genes Immun.*, **9**, 379–382.
68. Kobayashi, S., Ikari, K., Kaneko, H., Kochi, Y., Yamamoto, K., Shimane, K., Nakamura, Y., Toyama, Y., Mochizuki, T., Tsukahara, S. *et al.* (2008) Association of STAT4 with susceptibility to rheumatoid arthritis and systemic lupus erythematosus in the Japanese population. *Arthritis Rheum.*, **58**, 1940–1946.
69. Zervou, M.I., Sidiropoulos, P., Petraki, E., Vazgiourakis, V., Krasoudaki, E., Raptopoulou, A., Kritikos, H., Choustoulaki, E., Boumpas, D.T. and Goulielmos, G.N. (2008) Association of a TRAF1 and a STAT4 gene polymorphism with increased risk for rheumatoid arthritis in a genetically homogeneous population. *Hum. Immunol.*, **69**, 567–571.
70. Cornelis, F., Faure, S., Martinez, M., Prud'homme, J.F., Fritz, P., Dib, C., Alves, H., Barrera, P., de Vries, V.N., Balsa, A. *et al.* (1998) New susceptibility locus for rheumatoid arthritis suggested by a genome-wide linkage study. *Proc. Natl Acad. Sci. USA*, **95**, 10746–10750.
71. Jawaheer, D., Seldin, M.F., Amos, C.I., Chen, W.V., Shigeta, R., Monteiro, J., Kern, M., Criswel, L.A., Albani, S., Nelson, J.L. *et al.* (2001) A genomewide screen in multiplex rheumatoid arthritis families suggests genetic overlap with other autoimmune diseases. *Am. J. Hum. Genet.*, **68**, 927–936.
72. Tod, J.A., Walker, N.M., Cooper, J.D., Smyth, D.J., Downes, K., Plagnol, V., Bailey, R., Nejentsev, S., Field, S.F., Payne, F. *et al.* (2007) Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat. Genet.*, **39**, 857–864.
73. McCarthy, M.I. and Zeggini, E. (2009) Genome-wide association studies in type 2 diabetes. *Curr. Diab. Rep.*, **9**, 164–171.
74. Voight, B.F., Scot, L.J., Steinthorsdottir, V., Morris, A.P., Dina, C., Welch, R.P., Zeggini, E., Huth, C., Aulchenko, Y.S., Thorleifsson, G. *et al.* (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.*, **42**, 579–589.