BMC Genomics

BioMed Central
The Open Access Publisher

Network-based SNP meta-analysis identifies joint and disjoint genetic features across common human diseases

BMC Genomics 2012, **13**:490 doi:10.1186/1471-2164-13-490

Matthias Arnold (matthias.arnold@helmholtz-muenchen.de)
Mara L Hartsperger (mara.hartsperger@helmholtz-muenchen.de)
Hansjörg Baurecht (hbaurecht@dermatology.uni-kiel.de)
Elke Rodríguez (erodriguez@dermatology.uni-kiel.de)
Benedikt Wachinger (benedikt.wachinger@helmholtz-muenchen.de)
Andre Franke (a.franke@mucosa.de)
Michael Kabesch (Kabesch.Michael@mh-hannover.de)
Juliane Winkelmann (winkelmann@lrz.tu-muenchen.de)
Arne Pfeufer (arne.pfeufer@web.de)
Marcel Romanos (Marcel.Romanos@med.uni-muenchen.de)
Thomas Illig (illig@helmholtz-muenchen.de)
Hans-Werner Mewes (w.mewes@helmholtz-muenchen.de)
Volker Stümpflen (v.stuempflen@helmholtz-muenchen.de)
Stephan Weidinger (sweidinger@dermatology.uni-kiel.de)

ISSN	1471-2164
Article type	Research article
Submission date	29 May 2012
Acceptance date	12 September 2012
Publication date	18 September 2012
Article URL	http://www.biomedcentral.com/1471-2164/13/490

Network-based SNP meta-analysis identifies joint and disjoint genetic features across common human diseases

Matthias Arnold[1*,†]
[*] Corresponding author
Email: matthias.arnold@helmholtz-muenchen.de

Mara L Hartsperger[1,†]
Email: mara.hartsperger@helmholtz-muenchen.de

Hansjörg Baurecht[2,3,†]
Email: hbaurecht@dermatology.uni-kiel.de

Elke Rodríguez[2]
Email: erodriguez@dermatology.uni-kiel.de

Benedikt Wachinger[1]
Email: benedikt.wachinger@helmholtz-muenchen.de

Andre Franke[4]
Email: a.franke@mucosa.de

Michael Kabesch[5]
Email: Kabesch.Michael@mh-hannover.de

Juliane Winkelmann[6,7,8]
Email: winkelmann@lrz.tu-muenchen.de

Arne Pfeufer[6,8,9,10]
Email: arne.pfeufer@web.de

Marcel Romanos[10]
Email: Marcel.Romanos@med.uni-muenchen.de

Thomas Illig[11]
Email: illig@helmholtz-muenchen.de

Hans-Werner Mewes[1,12]
Email: w.mewes@helmholtz-muenchen.de

Volker Stümpflen[1]
Email: v.stuempflen@helmholtz-muenchen.de

Stephan Weidinger[2*]
[*] Corresponding author
Email: sweidinger@dermatology.uni-kiel.de

[1] Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Neuherberg, Germany

[2] Department of Dermatology and Allergy Biederstein, Technische Universität München, 80802 Munich, Germany

[3] TUM Graduate School of Information Science in Health (GSISH), Technische Universität München, 85748 Garching, Germany

[4] Institute for Clinical Molecular Biology, University of Kiel, 24105 Kiel, Germany

[5] Clinic for Pneumology and Neonatology, Hannover Medical School, 30625 Hannover, Germany

[6] Institute for Human Genetics, Technische Universität, München 81675, Munich, Germany

[7] Department of Neurology, Technische Universität München, 81675 Munich, Germany

[8] Institute of Human Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Neuherberg, Germany

[9] Institute of Genetic Medicine, European Academy Bozen/Bolzano (EURAC), Bolzano, Italy – Affiliated Institute of the University, 39100 Lübeck, Italy

[10] Department of Child and Adolescent Psychiatry, University Clinic of Munich, 80336 Munich, Germany

[11] Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Neuherberg, Germany

[12] Chair of Genome Oriented Bioinformatics, Center of Life and Food Science, Freising-Weihenstephan, Technische Universität München, 80333 Munich, Germany

[†] Equal contributors.

Abstract

Background

Genome-wide association studies (GWAS) have provided a large set of genetic loci influencing the risk for many common diseases. Association studies typically analyze one specific trait in single populations in an isolated fashion without taking into account the potential phenotypic and genetic correlation between traits. However, GWA data can be efficiently used to identify overlapping loci with analogous or contrasting effects on different diseases.

Results

Here, we describe a new approach to systematically prioritize and interpret available GWA data. We focus on the analysis of joint and disjoint genetic determinants across diseases. Using network analysis, we show that variant-based approaches are superior to locus-based analyses. In addition, we provide a prioritization of disease loci based on network properties and discuss the roles of hub loci across several diseases. We demonstrate that, in general, agonistic associations appear to reflect current disease classifications, and present the potential use of effect sizes in refining and revising these agonistic signals. We further identify potential branching points in disease etiologies based on antagonistic variants and describe plausible small-scale models of the underlying molecular switches.

Conclusions

The observation that a surprisingly high fraction (>15%) of the SNPs considered in our study are associated both agonistically and antagonistically with related as well as unrelated disorders indicates that the molecular mechanisms influencing causes and progress of human diseases are in part interrelated. Genetic overlaps between two diseases also suggest the importance of the affected entities in the specific pathogenic pathways and should be investigated further.

Keywords

Genome-wide association study, Genetic overlap, Shared variant network, Disease comorbidity

Background

In the past years enormous progress has been made in the identification of complex trait susceptibility loci. The application of genome-wide association studies (GWAS) created a still growing set of genetic markers associated with increased risk for a multitude of different diseases [1]. In contrast to few single loci exerting large effects on some phenotypes – mostly immune-related traits – the majority of traits was only associated with loci displaying small effects of odds ratios ranging from 1.1 to 1.5 [2]. Meta-analyses of several GWA studies further extended the set of known disease-related associations with even lower-effect variants. Despite the impressive progress in the field, for most traits only a small proportion of the total heritability is yet explained by known risk variants [3]. A notable exception is type 1 diabetes (T1D) where validated risk loci explain a large proportion of the total heritability [4]. In contrast, for most traits a considerably larger number of variants was reported to be associated, but typically these explain less than 50% of the total heritability [5].

Intriguingly, although published individual GWAS are usually carried out for one trait at a time, a significant overlap in the associations of several complex diseases becomes apparent [6]. Besides effects on a specific phenotype, loci and single SNPs thus may also exert pleiotropic effects by contributing to a variety of traits. While it is not surprising that susceptibility genes for closely related traits should be shared, multi-functionality of a gene in phenotype presentation, i.e. pleiotropy, *sensu stricto* refers to seemingly unrelated and distinct traits [7]. Loci or variants affecting several traits might have small effects on each specific trait, but may be of major biological interest while indicating shared or branching etiological mechanisms. In principle, the influence of such loci can be agonistic or antagonistic, i.e. involve concurrent similar or opposite effects of the same variant for different traits. So far, few studies attempted to study such loci in a systemic fashion and rather focused on shared risk variants in closely related traits like autoimmune diseases [8-10], heart diseases [11] or cancer [12].

In order to identify shared or branching pathways of related as well as diverse (i.e. medically and phenotypically distinct) diseases, we performed a systematic comparative analysis of genetic commonalities and differences across traditionally defined traits using the available repository of GWAS results. In the context of network medicine [13], we utilized an approach based on the diseasome concept [14] and investigated high-significance associations beyond conventional single-marker analysis in a hypothesis-free and comprehensive way. In former studies we found differing approaches of gene and locus assignment to association markers which partially led to controversial results (e.g. [15]). We therefore developed a more sophisticated locus assignment method and evaluate its reliability by utilizing the information contained directly in the reported markers. For this variant-based approach we manually curated a high-quality data set to construct a network extending the knowledge on genetic overlaps between diseases as provided by GWA studies.

Results and discussion

Considerable discrepancies across GWAS through differing genotyping platforms, varying sample sizes and diverging measures of statistical significance demand accurate data selection. Therefore, to sustain the genuine variant-linked information provided by GWAS, we combined several steps of data curation and filtering. To provide a comprehensive base for the analysis of potentially multi-functional loci and variants, respectively, we compiled two network representations of the information made available by GWA studies: the locus-based "shared locus network" (SLN, Figure 1B) and the variant-based "shared variant network" (SVN, Figure 1C). To be able to cluster diseases by their "genetic relatedness", we additionally created a disease-centric projection of the SVN (Additional file 1: Figure S1).

Figure 1 Illustration of the different disease networks based on genome-wide association data. A: The bipartite graph constructed from all association data. The two disjoint node sets are diseases (n = 111) and loci (n = 734; 508 gene loci and 226 intergenic loci), connected to each other by an edge if a variant (n = 1,120) within the respective locus is associated with the corresponding trait. B: The SLN (shared locus network) consisting of 84 traits and 157 loci, retrieved by removing isolated traits and loci that are associated with a single trait only. C: The SVN (shared variant network) that corresponds to a variant-based representation of the data. Here, a trait and a locus are linked if the locus contains a variant comprising associations with this and at least one other trait. The network consists of 175 SNPs located in 94 loci that are associated with 55 diseases (see also Additional file 2: Table S1). The colors of the disease nodes correspond to disease classes according to the MeSH ontology, multi-colored nodes indicate an association with different disease classes; loci are depicted as transparent, diamond-shaped nodes. The node size reflects the number of loci a disease is associated with. In C, the edge color reflects the allelic information: gray indicates agonistic variant(s), red corresponds to antagonistic variant(s), and blue mark both agonistic and antagonistic signals

We defined the associated loci over the variant-based linkage disequilibrium (LD) measure r^2 and, accordingly, expected the SLN and the SVN (Figure 1B and C) to be of similar shape. However, when visually comparing the networks, significant differences in size and structure became apparent. Therefore, we performed further analyses to compare established property measures of the networks in detail to investigate potential reasons for this divergence.

Shared locus vs. shared association analysis

Despite the size difference, the SLN shows greater network heterogeneity ($SLN = 1.30$, $SVN = 1.17$) and lower centralization ($SLN = 0.175$, $SVN = 0.205$) and density ($SLN = 0.014$, $SVN = 0.021$) values than the SVN. Furthermore, the intersection between the SVN and the SLN lacks not only 5% of the nodes but also 10% of the edges of the SVN. These numbers imply that the process of translating LD data into locus information is at least partly inconsistent. Analysis of the structure of the assigned LD blocks showed two error sources in shared locus analysis (illustrated in Figure 2). First, variants in two independent LD blocks are assigned the same locus but are not in LD (assignment error I). Thus, shared loci are found that are not reflected in the variant based data. Second, if two SNPs are in strong LD but the individual long range LD patterns of the SNPs diverge (e.g. the LD block of one SNP covers a greater area at the given r^2-threshold), an assignment error II occurs. In this case the two SNPs are assigned to different loci (in the example above, this is due to the different sizes of their LD blocks which may contain distinct gene sets) and their LD connection is lost. These observations suggest that (*i*), the SLN contains loci which overlap between traits but the associated markers are not in strong LD, (*ii*), there are several traits which are connected to the SLN via a single, potentially misleading link (as not mirrored in the variant-based data), and, (*iii*), even a LD-based locus assignment is unable to identify all shared associations ($n = 25$ of unidentified loci, based on the assignment error II). Due to this limited sensitivity and specificity in detecting LD-based correlations between the reported markers on locus scale, we used the smaller but more accurate variant-based SVN for further analyses. Moreover, the risk allele data in the SVN allows for the inclusion of the direction of the effects, i.e. agonistic and antagonistic, on different traits. Diseases that share antagonistic or agonistic, respectively, associated variants are listed in Tables 1 and 2. Another advantage is that the SVN can be compared to other variant-based approaches assessing the genetic overlap between traits. Recently, a statistic to identify SNPs with effects across phenotypes (the cross-phenotype meta-analysis statistic, CPMA) was proposed by Cotsapas et al. [8]. It compares the distribution of association P-values of a SNP across seven GWAS on distinct autoimmune diseases to the exponential(1)-distribution representing the expected decay rate of association P-values. As in our approach we use pre-filtered associations, this method cannot readily be employed on our data. However, using the data provided by Cotsapas and colleagues on autoimmune loci in the SVN, we retrieved CPMA P-values on 30 SNPs ($\sim 17\%$) corresponding to 28 loci ($\sim 30\%$) in our data. The CPMA classified all SNPs as significantly effective across diseases ($P < 0.05$, see Additional file 3: Table S2). Thus, we were able to validate nearly one third of the loci contained in the SVN by an independent approach, which underlines the suitability of the network.

Figure 2 LD based locus assignment and its error sources. At the example of chromosome 8q21.11, LD-based locus assignment is given for 6 exemplary SNPs (blue box). LD information is given by a color scale displaying the LD-measure r^2 with red depicting strong LD, blue low LD and white no LD. Example SNPs in LD are connected with black dashed lines. In the gray boxes, the two error sources of automated locus assignment are given. An assignment error I occurs if two variants not in LD, i.e. in two independent LD blocks, are located in the same gene, intergenic region or gene desert and thus are assigned the same locus. Here, this is the case for the variants rs-A/rs-B and rs-E/rs-F, respectively. The consequence of this type of error is a shared association on the locus level not mirrored on the variant level. An assignment error II is introduced if two variants are in LD but diverge in their assigned locus. Here, this is the case for rs-C and rs-D. Due to such abnormalities in the LD data the link between both variants is lost if only the locus level is considered

Table 1 Antagonistically linked traits

Disease	Abbr.	Antagonistically Linked Traits (*Loci*)	#
CROHN DISEASE	CD	AST (*17q12*); T1D (*1p13, 16p11*); VIT (*1p13*); T2D (*2p21*); RA (*1p13*); SLE (*1p13*); PS (*19p13*); MS (*17q21*)	8
ASTHMA	AST	CD (*17q12*); UC (*17q12*); BLC (*17q12*); RA (*17q12*); PS (*5q31*)	5
GLIOMA	GLI	COR (*9p21*); T2D (*9p21*); GLA (*9p21*); IPF (*5p15*) ; TN (*5p15*)	5
ARTHRITIS, RHEUMATOID	RA	AST (*17q12*); MS (*20q13*); CD (*1p13*)	3
MULTIPLE SCLEROSIS	MS	RA (*20q13*); HCC (*1p36*); CD (*17q21*)	3
COLITIS, ULCERATIVE	UC	AST (*17q12*); CeD (*1p36*); SLE (*1q23*)	3
LUNG NEOPLASMS	LN	PN (*5p15*); IPF (*5p15*); TN (*5p15*)	3
TESTICULAR NEOPLASMS	TN	GLI (*5p15*); LN (*5p15*)	2
VITILIGO	VIT	CD (*1p13*); MEL (*11q14*)	2
DIABETES MELLITUS, TYPE 1	T1D	CD (*1p13, 16p11*); IBD (*1p13*)	2
DIABETES MELLITUS, TYPE 2	T2D	GLI (*9p21*); CD (*2p21*)	2
CELIAC DISEASE	CeD	UC (*1p36*); CRC (*3q26*)	2
IDIOPATHIC PULMONARY FIBROSIS	IPF	GLI (*5p15*); LN (*5p15*)	2
PSORIASIS	PS	AST (*5q31*); CD (*19p13*)	2
LUPUS ERYTHEMATOSUS, SYSTEMIC	SLE	CD (*1p13*); UC (*1q23*)	2
GLAUCOMA	GLA	GLI (*9p21*)	1
MELANOMA	MEL	VIT (*11q14*)	1
PANCREATIC NEOPLASMS	PN	LN (*5p15*)	1
ALCOHOLISM	ALC	HNN (*12q24*)	1
COLORECTAL NEOPLASMS	CRC	CeD (*3q26*)	1
INFLAMMATORY BOWEL DISEASES	IBD	T1D (*16p11*)	1
CARCINOMA, HEPATOCELLULAR	HCC	MS (*1p36*)	1
LIVER CIRRHOSIS, BILIARY	BLC	AST (*17q12*)	1
CORONARY DISEASE	COR	GLI (*9p21*)	1
HEAD AND NECK NEOPLASMS	HNN	ALC (*12q24*)	1

Listed are diseases and the traits that share an antagonistic variant with the respective disorder. In the first column the considered disease is given. The second column specifies the abbreviation of the disorder in the first column. The third column contains the abbreviated diseases as defined in column two which have an antagonistic link to the disorder in column one, followed by the chromosomal location of the antagonistically associated variant(s) in parentheses. The last column lists the count of traits antagonistically linked to the disorder in column one. For a more detailed listing see Additional file 2: Table S1

Table 2 Agonistically linked traits

Disease	**Abbr.**	**Agonistically Linked Traits (*Loci*)**	**#**
CROHN DISEASE	CD	UC (*1p31, 1q32, 3p21, 5p13, 9p24, 9q32, 9q34, 10q24, 21q21, 21q22*); IBD (*1p31, 9q32, 10q22, 16q12, 20q13, 22q12*); CeD (*2q12, 18p11, 22q11*); LEP (*9q32, 13q14*); MG (*22q12*); T1D (*1q32, 18p11*); SC (*3p21*); OVARIAN NEOPLASMS (*8q24*); AS (*1p31*); LL (*2q37*); HTG (*2p23*); AA (*11q13*); MS (*10p15*); AD (*11q13*)	14
ARTHRITIS, RHEUMATOID	RA	CeD (*1p36, 4q27, 6q23*); SLE (*2q32, 7q32, 8p23*); T1D (*2q33, 4p15, 21q22*); BLC (*7q32*); SS (*7q32*); FL (*6p21*); UC (*6q23*); LBL (*6p21*); SCHIZOPHRENIA (*6p21*); VIT (*21q22*)	10
CELIAC DISEASE	CeD	CD (*2q12, 18p11, 22q11*); RA (*1p36, 4q27, 6q23*); UC (*2p16, 6q23*); T1D (*18p11*); BLC (*7p14*); VIT (*3q28*); HYP (*12q24*); SLE (*6p21*); MG (*6p21*)	9
COLITIS, ULCERATIVE	UC	CD (*1p31, 1q32, 3p21, 5p13, 9p24, 9q32, 9q34, 10q24, 21q21, 21q22*); CeD (*2p16, 6q23*); AS (*1p31, 2q11*); IBD (*1p31, 21q22*); SC (*3p21*); PS (*1p31*); RA (*6q23*); T1D (*1q32*); SLE (*7q32*)	9
CORONARY DISEASE	COR	MCI (*1p13, 1p32, 1q41, 2q33, 6p24, 9p21, 10q11, 19p13, 21q22*); CAD (*1p13, 3q22, 9p21*); ICA (*10q24*); PD (*10q24*); AAA (*9p21*); HYP (*10q24*); HTG (*11q23*)	7
LUPUS ERYTHEMATOSUS, SYSTEMIC	SLE	RA (*2q32, 7q32, 8p23*); SS (*2q32, 7q32*); LN (*6p21*); BLC (*7q32*); MG (*6p21*); UC (*7q32*); CeD (*6p21*)	7
DIABETES MELLITUS, TYPE 1	T1D	RA (*2q33, 4p15, 21q22*); CD (*2q33, 4p15, 21q22*); CeD (*18p11*); AA (*12q13*); VIT (*21q22*); UC (*1q32*)	6
HYPERTENSION	HYP	ICA (*10q24*); PD (*10q24*); COR (*10q24*); KIDNEY FAILURE, CHRONIC (*16p12*); CeD (*12q24*)	5
INFLAMMATORY BOWEL DISEASES	IBD	CD (*1p31, 9q32, 10q22, 16q12, 20q13, 22q12*); LEP (*9q32*); UC (*1p31, 21q22*); MG (*22q12*); AS (*1p31*)	5
LIVER CIRRHOSIS, BILIARY	BLC	RA (*7q32*); SLE (*7q32*); SS (*7q32*); MS (*12p13*); CeD (*7p14*)	5
CORONARY ARTERY DISEASE	CAD	COR (*1p13, 3q22, 9p21*); MCI (*1p13, 9p21*); AAA (*9p21*); T2D (*2q36*)	4
GLOMERULONEPHRITIS, MEMBRANOUS	MG	IBD (*22q12*); CD (*22q12*); SLE (*6p21*); CeD (*6p21*)	4
AORTIC ANEURYSM, ABDOMINAL	AAA	COR (*9p21*); CAD (*9p21*); MCI (*9p21*)	3
INTRACRANIAL ANEURYSM	ICA	PD (*10q24*); HYP (*10q24*); COR (*10q24*)	3
LUNG NEOPLASMS	LN	SLE (*6p21*); COPD (*15q25*); PVD (*15q25*)	3
LYMPHOMA, FOLLICULAR	FL	LBL (*11q24*); RA (*6p21*); LL (*6p21*)	3
MYOCARDIAL INFARCTION	MCI	COR (*1p13, 1p32, 1q41, 2q33, 6p24, 9p21, 10q11, 19p13, 21q22*); CAD (*1p13, 9p21*); AAA (*9p21*)	3
PARKINSON DISEASE	PD	HYP (*10q24*); COR (*10q24*); ICA (*10q24*)	3
SCLERODERMA, SYSTEMIC	SS	SLE (*2q32, 7q32*); RA (*7q32*); BLC (*7q32*)	3
SPONDYLITIS, ANKYLOSING	AS	UC (*1p31, 2q11*); IBD (*1p31*); CD (*1p31*)	3
VITILIGO	VIT	CeD (*3q28*); RA (*21q22*); T1D (*21q22*)	3
ALOPECIA AREATA	AA	T1D (*12q13*); CD (*11q13*)	2
CHOLANGITIS, SCLEROSING	SC	UC (*3q21*); CD (*3q21*)	2
DERMATITIS, ATOPIC	AD	GLIOMA (*20q13*); CD (*11q13*)	2
DIABETES MELLITUS, TYPE 2	T2D	OBESITY (*16q12*); CAD (*2q36*)	2
HEAD AND NECK NEOPLASMS	HNN	GASTROINTESTINAL NEOPLASMS (*10q23*); ALCOHOLISM (*12q24*)	2
HYPERTRIGLYCERIDEMIA	HTG	CD (*2p23*); COR (*11q23*)	2
LEPROSY	LEP	CD (*9q32, 13q14*); IBD (*9q32*)	2
LEUKEMIA, LYMPHOID	LL	CD (*2q27*); FL (*11q24*)	2
LYMPHOMA, LARGE B-CELL, DIFFUSE	LBL	FL (*6q21*); RA (*6q21*)	2
MULTIPLE SCLEROSIS	MS	CD (*10p15*); BLC (*12p13*)	2
PERIPHERAL VASCULAR DISEASES	PVD	COPD (*15q25*); LN (*15q25*)	2
PROSTATIC NEOPLASMS	PRN	COLORECTAL NEOPLASMS (*8q24*); ENDOMETRIAL NEOPLASMS (*17q12*)	2
PSORIASIS	PS	UC (*1p31*); ARTHRITIS, PSORIATIC (*6q21*)	2
PULMONARY DISEASE, CHRONIC OBSTRUCTIVE	COPD	LN (*15q25*); PVD (*15q25*)	2

Listed are diseases that share agonistic associations with at least two traits. In the first column the considered disease is given. The second column specifies the abbreviation of the disorder in the first column. The third column contains the disease abbreviations (as defined in column two) of traits which have an agonistic link to the disorder in column one, followed by the chromosomal location of the agonistically associated variant(s) in parentheses. Here, the full MeSH term is given for traits for which no abbreviation was defined. The last column lists the count of traits agonistically linked to the disorder in column one. For the complete list of agonistically linked traits and more details see Additional file 2: Table S1

Topology of the SVN

Its degree distribution attributes the SVN a scale-free network, i.e. it approximates a power-law ($P(k) \sim k^{-\gamma}; \gamma = 1.32; R^2 = 0.69$) (Additional file 4: Figure S2). Interestingly, also when considering the two node types separately, disease nodes ($\gamma = 0.97; R^2 = 0.71$) as well as locus nodes ($\gamma = 2.98; R^2 = 0.93$) show scale-free degree distributions (Additional file 4: Figure S2). The scale-free property classifies the network (and its two sets of node types, respectively) as structured, i.e. non-random [13]. It has to be considered that the limited size of the SVN leads to inaccuracies in distribution fitting and thus reduces the explanatory value of this observation. However, as clinically related diseases (i.e. diseases which present similar symptoms) should present a higher genetic overlap than unrelated disorders, this finding meets expectations.

The variant-based SVN also shows no artificial character with regards to its topology. Both locus and disease node sets comprise hubs, here defined as nodes with a degree >3, which form the central elements in the network. As in each GWAS multiple markers are associated with a single disease, one would expect hubs to be constituted mostly of disease nodes. In line with that, 74% of the hubs in our network are disease nodes. The remaining 26% are loci hubs (seven gene loci and three intergenic loci). Several of these loci have been previously identified as influencing susceptibility to multiple diseases like the HLA region on chromosome 6 [16], a cancer locus at chromosome 8q24 [12], and a coronary artery disease locus at chromosome 9p21 [11]. Further hub loci are *PTPN22,* a known player across several autoimmunity disorders [17], and *IL23R*, which has been shown to direct inflammatory processes [18]. In addition, we observed hubs which have not yet been described as predisposing to a whole group of diseases, such as *TNPO3* which appears to predispose to various autoimmune diseases like systemic lupus erythematosus, systemic scleroderma, and rheumatoid arthritis [19-21], or *TNFSF15*, which shows associations with several inflammatory diseases [22-25]. As expected, in the majority of cases the traits linked to one hub can be assigned to the same disease group and, further, diseases which are not obviously related to other disorders linked to the respective hub are mostly associated with antagonistic signals. For instance, in a four-gene locus at chromosome 17q12 (*GSDML*/*IKZF3*/*ORMDL3*/*ZPBP2*, see Additional file 2: Table S1), four autoimmune diseases are associated with the same risk allele that in turn has opposite effects on asthma [20,25-27]. Thus, our results indicate that loci associated with several diseases have an effect specific to a certain disease group rather than effects on unrelated diseases, and that, if there is an effect on an unrelated disease, it can often be distinguished by the direction of the effect.

Disease clustering mirrors trait relatedness

To identify shared and branching mechanisms we split the SNP association data into agonistic and antagonistic variants. Since in most cases there is no solid and comprehensive basis of experimental data that would allow for a more sensitive classification, we suggest that the best available indication of distinct effects of a variant on two diseases is the signal itself being different. Therefore, we define a SNP to be agonistic if all disorders are associated with the same risk allele of the SNP. Conversely, we consider a SNP antagonistic if the associated risk alleles differ between diseases. Accordingly, in the analysis of genetic overlaps as a measure of trait similarity only agonistic variants were included.

As similar diseases are more likely to share associations than diseases in distinct classes, we expected the SVN to be organized in a modular fashion. This was confirmed by the decrease of the degree distribution of the topological coefficient with the number of links per node (Additional file 4: Figure S2). To retrieve these modules, we applied a hierarchical clustering approach. The SVN contains two node types (loci and traits). As we wanted to directly assess variant-based disease relatedness, we used its disease centric projection for hierarchical correlation clustering. For data normalization, we calculated the Pearson correlation coefficient (PCC) for all pairs of diseases based on their genetic agonistic overlap with all other diseases. The clustering returned 15 disease clusters (Figure 3) and six diseases which show no or only weak correlation with any other disease. With the exception of the heterogeneous cluster 5 (hypertriglyceridemia, ovarian neoplasms, lymphoid leukemia, atopic dermatitis), the clusters mostly contain related diseases. However, many clusters also contain traits unrelated to the other phenotypes like schizophrenia in the autoimmune cluster 2. This indicates that clinical disease classifications appear to be reflected on the genetic level in general. Notably, several small clusters contain diseases which are either linked through common environmental risk factors – like smoking for lung neoplasms, peripheral vascular diseases, and chronic obstructive pulmonary disease – or present high frequencies of comorbidity, e.g. type 2 diabetes (T2D) and obesity. To get an insight into the overall extent of reported comorbidities of the diseases within the 15 clusters, we used publicly available resources [28,29] and literature mining. The within-cluster fraction of disease co-occurrence ranged from 75% to 100% ($\mu = 95.89\%$, $\sigma = 8.66\%$) which provides empirical evidence of the interrelation of diseases clustered together by genetic information. Such clusters containing diseases that present high ratios of comorbidity may be potential artifacts due to "contaminated" disease cohorts including a substantial number of comorbid cases. The unbiased search for the relation of a trait marker to a disease phenotype as performed in GWAS does not distinguish between markers for a primary or related secondary (comorbid) disease. Our results suggest that this aspect may have been underestimated in some studies, albeit the presence of independently shared etiological mechanisms can naturally not be ruled out in general.

Figure 3 Clustering of diseases with respect to genetic signals. We applied complete-linkage hierarchical clustering to identify groups of traits which show homogeneous patterns of genetic overlap to other disorders. We calculated for each pair of diseases the Pearson correlation of the patterns of overlap to the other diseases. The correlation values are ranging from −1 (white) indicating complete negative correlation to +1 (black) reflecting a perfect positive correlation. As the minimal value of the correlation coefficient was > −0.1, we collapsed the range of negative correlation. In red numbers, the 15 disease clusters are denoted. The Euclidian distance threshold was chosen as the maximal distance at which the six diseases showing no or only weak correlation with any other disease (disease names in gray) remain non-clustered

Overall, the outcome of the clustering poses the question of the extent of the influence of phenotype classification and population stratification on GWAS results. Frequent comorbidities (also of seemingly unrelated diseases such as obesity and cancer [30]), diagnostic difficulties in highly related diseases like Crohn's disease (CD) and ulcerative colitis (UC) [31], and structural (genetic) differences in population subgroups are known to complicate GWAS and impact their outcomes [32]. With growing sample sizes in case–control studies, the potential of false positives produced by such phenomena also increases. As a response, manifold control procedures to handle these and other confounding factors have been developed which are widely used and well appreciated [32]. The heterogeneity of the clusters we retrieved once more highlights the need for the development and application of such methods.

Odds ratio as potential indicator of primary effects

In the context of agonistic association overlap between related diseases, we used the odds ratios (ORs) reported with the SNPs to investigate their impact on the respective traits. In general, the highest ORs are reported for associations of autoimmune diseases to the HLA locus on chromosome 6. Associations with traits where few gene variants with strong effects are reported, e.g. rs6107516 in the prion protein *PRNP* associated with Creutzfeldt-Jakob disease (OR = 38.5) or rs2071348 in the hemoglobin gene cluster at 11p15.4 associated with beta-thalassemia (OR = 4.33), are exceptions from the majority of associations displaying small ORs [33,34]. Based on the effect size of variants associated with more than one trait and the same risk allele, we identified three patterns.

First, we identified variants which are likely to present general agonistic risk factors for a group of related diseases or syndromes such as rs13015714 in an interleukine receptor gene cluster at chromosome 2q12.1. This SNP is associated with celiac disease (CeD) and Crohn's disease (CD) with equal ORs (OR = 1.19) [24,35]. In cases of frequent comorbidities, though, comparable ORs have limited informative value. The SNP rs9939609 in the *FTO* gene for instance is associated with T2D and obesity with nearly equal ORs (OR = 1.34 and OR = 1.32) and thus appears to link two coequal traits of the metabolic syndrome (we refer to the definition of the International Diabetes Foundation, 2006) [36,37]. However, for SNPs in the *FTO* gene it has been shown that adjustment for body mass index results in the loss of significance (OR∼ 1.0) of the association with T2D [38].

Second, SNPs appeared in several cases to be primarily associated with one disease, which in turn represents a risk factor for another associated trait. For instance, rs2200733 on chromosome 4q25 is linked to atrial fibrillation with a higher OR (OR = 1.72) than to stroke (OR = 1.26) [39,40]. Another example is rs964184 which is located proximal to the apolipoprotein gene cluster on chromosome 11q23 which is associated with hypertriglyceridemia with a markedly higher OR (OR = 3.28) than to coronary disease (OR = 1.13) [41,42]. The lower effects of the markers on the hypothesized "secondary sequels" may be explained by the fact that these are caused by the primary diseases, but with less than 100% penetrance.

Third, we speculate that the OR might allow conclusions with respect to the evaluation of an association in cases where similar traits are linked to the same SNP with diverging effect sizes. For instance, CD and ulcerative colitis (UC) share multiple risk loci. The two diseases are strongly related in their etiology and pathology. Thus, a clinical distinction of both diseases is difficult if based only on few criteria and might lead to inaccurate case ascertainment leading to mixed associations [31]. However, for several SNPs such as rs11209026, which is located in the *IL23R* gene, we found notably higher effects on CD (OR = 3.84) than on UC (OR = 1.74) [25,43]. Conversely, rs3024505 which lies proximal to the *IL10* gene shows a greater effect on UC (OR = 1.46) than on CD (OR = 1.12) [24,44]. Interestingly, it has been shown that *IL23* is selectively upregulated in CD while levels in UC patients are normal and *IL10* expression appears to be higher in UC as compared to CD [24,45]. Thus, the OR might – similar to the above examples – allow for identifying potentially misleading associations in closely related diseases which may result from diagnostic errors.

Identification of branching etiologies

We searched for evidence that antagonistic signals represent genetic indicators of branching points in the etiologies of two diseases or disease groups. For the assessment of potentially multifunctional variants we therefore focused on markers with inverse effects. We identified 44 such variants, which represent almost 4% of the original association data analyzed and about 25% of the SNPs associated with more than one disease. Of those 44 variants, about one fifth (n = 9) are located in the HLA region. SNP-markers in that region are known to differ in their ability to capture the classical HLA-alleles [46] and therefore were not considered further for the present analysis.

For cases where the function of the harboring genes is known, we were able to identify conclusive models. For instance, rs2736100 in the telomerase reverse transcriptase (*TERT*) gene was reported to exert antagonistic effects in idiopatic pulmonary fibrosis (IPF) and testicular germ cell tumor (TGCT) and two other cancer traits [47-53]. Whereas telomerase activity is generally upregulated in tumors sustaining proliferation and potentiating mutagenesis and transformation of cancer cells [54], in IPF limited cell division due to decreased telomerase activity is thought to contribute to the phenomenon of high percentages of apoptotic cells in fibroblasts [55]. Consistent with that observation, disturbed telomerase activity in TGCT is believed to form a distinct mechanism of cancerogenesis in this tumor type [53]. This distinction from other cancer traits is believed to be based on the fact that testicular germ cells are the only adult cell type with high telomerase expression [56]. Another example is the telomerase RNA component *TERC*, which is essential for *TERT* functioning. Opposite alleles of SNP rs10936599 are associated with CeD and colorectal cancer (CRC) [35,57]. Jones et al. showed that rs2293607, a variant tagged by rs10936599, alone is sufficient to modulate *TERC* expression [58]. While in CRC this leads to *TERC* overexpression and longer telomeres, the opposite might apply to CeD, which exhibits telomere reduction and genomic instability [58,59]. The observation that both constituents of the telomerase complex contain independent antagonistic variants is an intriguing finding. It suggests parallel, autonomous evolution of two functionally interacting loci gone to fixation at a trade-off between early cell senescence or increased apoptosis rates (as in IPF and CeD) and oncogenesis.

A further example is rs1393350 in the tyrosinase (*TYR*) gene where the opposite alleles are linked to vitiligo and melanoma [60,61], potentially mirroring the inverse correlation observed for the two traits. The phenomenon is based on the presentation of *TYR* (self-) antigens on the cell surface of melanocytes. It is hypothesized that in vitiligo the immune system is hypersensitive towards *TYR* antigens, which are overexpressed in melanoma cells [62]. A possible explanation may be that opposite alleles differentially influence the antigenicity of the *TYR* protein, thereby conferring protection from melanoma but susceptibility to vitiligo through immune surveillance and vice versa.

In cases of functionally less or uncharacterized genes and their involvement in the associated diseases, our approach can still be used to suggest potential trait-specific effects. Antagonistic effects of rs12720356 (localized in the *TYK2* gene) in CD and psoriasis, for instance, might point towards different patterns of cytokine signaling in these two diseases [24,63,64]. Likewise, rs12727642 and rs35675666, both located in the *PARK7* gene and inversely associated with CeD and UC, could indicate differential effects of oxidative stress on each trait [25,35].

Variant-based analysis of joint and disjoint genetic features

In this study, we identified overlapping genetic associations and their corresponding loci with analogous or contrasting effects on different diseases. We addressed the methodological challenges of the identification of the functional entities affected by GWAS-detected variants.

Associations formally implicate genomic regions which are captured via tagging SNPs representing haplotype blocks. By using the population-specific LD-based haplotype data provided by the HapMap project [65,66] or, more recently, the 1000 genomes project [67], SNP arrays are constructed aiming at a high coverage of the total genome variation, but without considering biologically functional aspects. The advantage of GWAS as a method is its unbiased approach to identify genomic regions compromised in a disease; a major drawback is that the association of markers without knowledge of the causal variants and their effects does not allow for a straightforward biological interpretation.

As we show, the reliability of an automated assignment of LD-based loci to the trait-associated variants is strongly context-dependent. Especially in cases of high gene density or, conversely, in intergenic regions/gene deserts, resolving GWAS signals is not possible without further knowledge. Simplifications such as more basic locus assignment approaches which neglect the LD structure of the genome (e.g. classifying a SNP as affecting only the most proximal gene) may seem more intuitive, might facilitate analyses and could be useful to identify causal disease-gene associations. These correct associations of genes which are detected through significant enrichment of a harbored tagging variant in a patient cohort may not be discovered when incorporating LD data in cases where the LD block of the respective variant spans across several genes. However, such approaches disregard a basic principle defining the current GWAS paradigm, namely the use of LD information in the design of genotyping arrays to achieve the genome-wide coverage of common SNPs. Hence, it can be problematic to project the variant-based GWAS data on genes or loci. Accordingly, we decided to use variant-based methods and concentrated on strong gene candidates identified via the gene function of single-gene loci whenever suggesting potential biological effects of the considered variants.

In the analysis of genetic overlaps we followed the hypothesis that the effects of variants shared across several diseases correspond to the reported risk alleles. If the risk allele is the same in all associated diseases, we assume the effect to be the same, i.e. that there is a common underlying etiology. For closely related diseases a positive correlation is not surprising, e.g. a GWAS on psoriatic arthritis (PSA) will also detect agonistic variants such as rs33980500 that are also associated with psoriasis (PS) [68,69]. Indeed, the vast majority of agonistic variants in our data set links groups of related diseases and thus may mark interesting target regions for therapeutic interventions. However, we also found a few agonistic signals connecting apparently unrelated diseases, e.g. rs6010620 which exerts susceptibility for both glioma and atopic dermatitis (AD) [50,51,70,71]. If our hypothesis is correct, an endophenotype influencing both diseases may be present which has yet to be identified. For antagonistic SNPs, on the other hand, we describe plausible mechanisms that may render variants protective against one trait and predisposing to another, labeling the affected genes/loci as pleiotropic. If pleiotropic effects are as frequent as evolutionary modelers postulate [2,72] and this effects can be identified by analyses based on GWAS, this might have great implications for the development and use of therapeutics because it would enable avoidance of potential side effects when targeting such loci. Already, there are more than 50 genotype/drug interactions known for which therapeutic dosing recommendations are available [73].

Conclusions

Our results present new starting points for studying the genetics of complex diseases. The observation that more than 15% of the SNPs considered in our study are associated both agonistically and antagonistically with related as well as unrelated disorders indicates that the molecular mechanisms influencing causes and progress of human diseases are in part interrelated. Genetic overlaps between two diseases also suggest the importance of the affected entities in the specific pathogenic pathways and should be investigated further. These may be secondary, such as genes involved in inflammatory responses related to T2D as well as cancer [30,38]. The findings presented also demonstrate the need to clarify the relation of any phenotype linked to an associated marker. For directly interrelated diseases such as PS and PSA often PS patients without present arthritis or arthritis in the past are used as additional control group. Associations are then interpreted as PSA-specific if not as strongly associated with PS [74,75]. Comparable procedures may proof useful in frequently co-occurring diseases genetically linked by agonistic variants. Nevertheless, the complex genetics of multifactorial diseases asks for a better understanding of the functions underlying common disorders. An improved characterization of the endophenotype, such as metabolite or protein concentrations, may enhance our understanding of identical pathomechanisms that link agonistic genetic loci to clinically distinct traits. Pleiotropic effects, on the other hand, that are harbored in the same locus may trigger different mechanisms interfering with the genetic or environmental background. The detailed examination of antagonistically associated loci may thus lead to first insight into the mechanism of the various types of pleiotropy in human diseases.

Methods

Association selection and curation

We obtained the core list of candidate sentinel SNPs from ‘A Catalog of Published Genome-Wide Association Studies’ [1] accessed on June 30, 2011 (http://www.genome.gov/gwastudies). Additional associations where retrieved from HuGE Navigator [76] and automated Text Mining [77]. New (i.e. not contained in the GWAS Catalog) association markers were manually tested on compliance with the criteria for inclusion in the GWAS Catalog before insertion in the candidate list. Associations with copy number variants (CNVs) as well as with pending SNPs were removed. For consistency, SNP identifiers were mapped to the RefSNP numbers of the same dbSNP release (build 131, http://www.ncbi.nlm.nih.gov/snp). For the same reason we semi-automatically translated trait descriptions to the official terms given in the Medical Subject Headings (MeSH). In this process, associations with quantitative and non-disease traits were eliminated. Finally, we lowered the association P-value threshold from 10^{-5} (as in the GWAS Catalog criteria) to 10^{-7} as to reduce the potential of artificial associations. The workflow of the methods of our approach is sketched in Figure 4.

Figure 4 Data prioritization and analysis workflow. We established a semi-automated curation pipeline which automatically gathers and annotates GWA data obtained from three sources (locus assignment included). Last step of the preprocessing was the manual inspection of risk alleles and odds ratios. With this data set at hand, we construct a locus-based (SLN) and a variant-based (SVN) network representation of the data. For quality reasons, we then limited analyses to the SVN and investigated the contained variants and their effects further

Construction of GWAS networks

For the construction of the locus-based data representation, we defined an associated locus as the whole genomic region captured by SNPs in strong LD, $r^2 \geq 0.8$, with the marker originally reported in a GWAS contained in our data set. The locus is then characterized as all genes located within this genomic region (referred to as "gene locus") (Figure 2). If the region contains no genes, the locus is assigned to its chromosomal location (referred to as "intergenic locus"). LD data and gene information were obtained with the SNAP tool [78]. After locus assignment, our final data set consisted of 111 different traits linked via 1,120 SNPs to 508 gene loci and 226 intergenic loci.

Based on this list we constructed a bipartite graph consisting of two disjoint sets of nodes (Figure 1A) representing the complete association data. The first node set corresponds to the traits, whereas the other set comprises the associated loci. Two nodes are connected by an edge if a variant within the respective locus is associated with the corresponding trait. By removal of isolated traits, i.e. traits which share no associated locus with another trait ($n = 27$) (Figure 1A), and cutting out loci which are associated with only one trait ($n = 577$), we retrieved the SLN (Figure 1B).

To obtain a variant-based representation of the data, we repeated the network generation on marker scale by utilizing the set of variants associated with more than one distinct trait. For this, we used the LD data to mutually assign the associated traits of sentinel SNPs in pairwise LD if not already present. In other words, each variant is, in addition to its own associated traits, assigned the traits associated with all correlated SNPs. This set consists of 175 SNPs located in 94 loci and associated with 55 diseases (Additional file 2: Table S1). In the resulting bipartite SVN, a trait and a locus are linked if the locus contains a variant which comprises associations with this and at least one other trait. Here, the allele information was included in the graph visualization by coloring of the edges (Figure 1C). Both the SLN and the SVN are provided as machine-readable files, see Additional files 5 and 6.

Network analysis

The network concepts which we used to compare the properties of the SLN and the SVN are defined as given in [79]:

$$Density = \frac{\sum_i \quad \sum_{j \neq i} \quad a_{ij}}{n(n-1)} = \frac{mean(k)}{n-1} \tag{1}$$

where $a_{ij} = 1$ if nodes i and j are connected and 0 otherwise. $mean(k)$ denotes the mean connectivity, which for a node i is defined as $k_i = \sum_{j \neq i} \quad a_{ij}$.

$$Centralization = \frac{n}{n-2}\left(\frac{max(k)}{n-1} - Density\right) \tag{2}$$

$$Heterogeneity = \frac{\sqrt{variance(k)}}{mean(k)} \tag{3}$$

To automatically distinguish the two node sets contained in the SVN, we used directed edges. Direction is always from disease (source) to locus (target). The distinct node degree distributions thus are identical to the indegree distribution and the outdegree distribution. The topological coefficient as a measure of modularity [80] T_i of a node i is defined as:

$$T_i = \begin{cases} 0, if N_i < 2 \\ \text{avg}\left(\frac{S(i,j)}{N_i}\right), else \end{cases} \tag{4}$$

where N_i is the number of neighbors of i and $S(i,j)$ is the number of shared neighbors of nodes i and j (undefined if i and j do not share a neighbor) plus one if j is a neighbor of i.

Power-law functions of the form $y = e^a x^b$ were fitted using least squares fitting where the coefficients are defined as $b = \frac{n\sum_{i=1}^{n} (lnx_i lny_i) - \sum_{i=1}^{n} (lnx_i)\sum_{i=1}^{n} (lny_i)}{n\sum_{i}^{n} (lnx_i)^2 - \left(\sum_{i=1}^{n} lnx_i\right)^2}$ and $= \frac{\sum_{i=1}^{n} (lny_i) - b\sum_{i=1}^{n} (lnx_i)}{n}$. As goodness-of-fit measure, we give the coefficient of determination

$$R^2 = \left(\frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} {x_i}^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} {y_i}^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}\right)^2 \tag{5}$$

for the linear transformation of the power-law functions, i.e. $\ln y = a + b \ln x$.

Determination of agonistic and antagonistic effects

For all variants associated with more than one trait, we manually extracted the risk alleles (OR > 1, independently of major or minor allele status) and odds ratios from the reporting studies. The alleles were mapped to the forward DNA strand according to dbSNP 131. The same procedure was applied to markers which were indirectly associated with a trait over LD. If for all traits the same associated risk allele (and corresponding allele, respectively) was reported, the SNP was classified as agonistic. If the risk alleles of a SNP were opposed in the associated diseases, the variant was classified as antagonistic.

Genetic clustering

We applied complete-linkage hierarchical clustering to identify groups of traits genetically overlapping with respect to agonistic signals. Normalization was performed using the linear PCC defined as $\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$ where the input are the vectors of the variant-based agonistic overlap of two distinct diseases X and Y to all other diseases. Thus, disorders which are clustered together show a homogeneous association overlap pattern to all other diseases, while diseases which are not clearly assigned to a cluster present a more heterogeneous pattern relatively unique in the SNP data. For cluster definition, we used a Euclidian distance threshold of 1.71. This threshold was determined as the maximal distance at which the six traits not correlating with other diseases (Figure 3) remain non-clustered.

Calculation of the CPMA statistic for autoimmune loci

We downloaded the dataset S1 from [8] and extracted the information on autoimmune-linked SNPs contained in the SVN. We used the Z-scores given in the file to compute two-sided P-values for all seven GWAS. Using the CPMA code provided on http://www.cotsapaslab.info/index.php/software/cpma/ we calculated the CPMA P-values as described in [8].

Competing interests

The authors have nothing to declare.

Authors' contributions

Conceived and designed the study, curated and analyzed the data, wrote the manuscript: MA, MLH, HB, SW. Curated the data, wrote and critically revised the manuscript: ER. Delivered and interpreted data: BW. Contributed to manuscript writing: AF, MK, AP, HWM. Interpreted the data: JW. Curated and interpreted data: MR. Conceived the study: TI, VS. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the Helmholtz Alliance on Systems Biology (project CoReNe and the Joint Technology Platforms); the German Federal Ministry of Education and Research as part of the National Genome Research Network [NGFN 01GS 0429, NGFN 01GS 0818]; the German Federal Ministry of Education and Research in its MedSys initiative (project SysMBo) [FKZ: 0315494A]; the TUM Graduate School for Information Science in Health; the Christiane Kühne Center for Allergy Research and Education (http://www.ck-care.ch/); the Schleswig-Holstein Excellence Cluster "Inflammation at Interfaces"; and a Heisenberg professorship of the DFG [WE 2678/4-1 to S.W.]. The funders had no role in study design, data collection, interpretation and analysis, decision to publish, or preparation of the manuscript.

References

1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.** *Proc Natl Acad Sci U S A* 2009, **106**(23):9362–9367.

2. Stranger BE, Stahl EA, Raj T: **Progress and promise of genome-wide association studies for human complex trait genetics.** *Genetics* 2011, **187**(2):367–383.

3. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, *et al*: **Finding the missing heritability of complex diseases.** *Nature* 2009, **461**(7265):747–753.

4. Wei Z, Wang K, Qu HQ, Zhang H, Bradfield J, Kim C, Frackleton E, Hou C, Glessner JT, Chiavacci R, *et al*: **From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes.** *PLoS Genet* 2009, **5**(10):e1000678.

5. So HC, Li MX, Sham PC: **Uncovering the total heritability explained by All true susceptibility variants in a genome-wide association study.** *Genet Epidemiol* 2011, **35**(6):447–456.

6. Frazer KA, Murray SS, Schork NJ, Topol EJ: **Human genetic variation and its contribution to complex traits.** *Nat Rev Genet* 2009, **10**(4):241–251.

7. Wagner GP, Zhang J: **The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms.** *Nat Rev Genet* 2011, **12**(3):204–213.

8. Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, Wallace C, Abecasis GR, Barrett JC, Behrens T, Cho J, *et al*: **Pervasive sharing of genetic effects in autoimmune disease.** *PLoS Genet* 2011, **7**(8):e1002254.

9. Sirota M, Schaub MA, Batzoglou S, Robinson WH, Butte AJ: **Autoimmune disease classification by inverse association with SNP alleles.** *PLoS Genet* 2009, **5**(12):e1000792.

10. Zhernakova A, van Diemen CC, Wijmenga C: **Detecting shared pathogenesis from the shared genetics of immune-related diseases.** *Nat Rev Genet* 2009, **10**(1):43–55.

11. Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, Ren B, Fu XD, Topol EJ, Rosenfeld MG, *et al*: **9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response.** *Nature* 2011, **470**(7333):264–268.

12. Meyer KB, Maia AT, O'Reilly M, Ghoussaini M, Prathalingam R, Porter-Gill P, Ambs S, Prokunina-Olsson L, Carroll J, Ponder BAJ: **A functional variant at a prostate cancer predisposition locus at 8q24 is associated with PVT1 expression.** *PLoS Genet* 2011, **7**(7):e1002165.

13. Barabasi AL, Gulbahce N, Loscalzo J: **Network medicine: a network-based approach to human disease.** *Nat Rev Genet* 2011, **12**(1):56–68.

14. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL: **The human disease network.** *Proc Natl Acad Sci U S A* 2007, **104**(21):8685–8690.

15. Barrenas F, Chavali S, Holme P, Mobini R, Benson M: **Network properties of complex human disease genes identified through genome-wide association studies.** *PLoS One* 2009, **4**(11):e8090.

16. Klein J, Sato A: **The HLA system. Second of two parts.** *N Engl J Med* 2000, **343**(11):782–786.

17. Gregersen PK: **Gaining insight into PTPN22 and autoimmunity.** *Nat Genet* 2005, **37**(12):1300–1302.

18. Di Meglio P, Di Cesare A, Laggner U, Chu CC, Napolitano L, Villanova F, Tosi I, Capon F, Trembath RC, Peris K, *et al*: **The IL23R R381Q gene variant protects against immune-mediated diseases by impairing IL-23-induced Th17 effector response in humans.** *PLoS One* 2011, **6**(2):e17160.

19. Chung SA, Taylor KE, Graham RR, Nititham J, Lee AT, Ortmann WA, Jacob CO, Alarcon-Riquelme ME, Tsao BP, Harley JB, *et al*: **Differential genetic associations for systemic lupus erythematosus based on anti-dsDNA autoantibody production.** *PLoS Genet* 2011, **7**(3):e1001323.

20. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, Li Y, Kurreeman FA, Zhernakova A, Hinks A, *et al*: **Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci.** *Nat Genet* 2010, **42**(6):508–514.

21. Radstake TR, Gorlova O, Rueda B, Martin JE, Alizadeh BZ, Palomino-Morales R, Coenen MJ, Vonk MC, Voskuyl AE, Schuerwegh AJ, *et al*: **Genome-wide association study of systemic sclerosis identifies CD247 as a new susceptibility locus.** *Nat Genet* 2010, **42**(5):426–429.

22. Kugathasan S, Baldassano RN, Bradfield JP, Sleiman PM, Imielinski M, Guthery SL, Cucchiara S, Kim CE, Frackelton EC, Annaiah K, *et al*: **Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease.** *Nat Genet* 2008, **40**(10):1211–1215.

23. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, *et al*: **Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease.** *Nat Genet* 2008, **40**(8):955–962.

24. Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R, *et al*: **Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci.** *Nat Genet* 2010, **42**(12):1118–1125.

25. Anderson CA, Boucher G, Lees CW, Franke A, D'Amato M, Taylor KD, Lee JC, Goyette P, Imielinski M, Latiano A, *et al*: **Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47.** *Nat Genet* 2011, **43**(3):246–252.

26. Liu X, Invernizzi P, Lu Y, Kosoy R, Lu Y, Bianchi I, Podda M, Xu C, Xie G, Macciardi F, *et al*: **Genome-wide meta-analyses identify three loci associated with primary biliary cirrhosis.** *Nat Genet* 2010, **42**(8):658–660.

27. Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, Heath S, von Mutius E, Farrall M, Lathrop M, Cookson WO, *et al*: **A large-scale, consortium-based genomewide association study of asthma.** *N Engl J Med* 2010, **363**(13):1211–1221.

28. Hidalgo CA, Blumm N, Barabasi AL, Christakis NA: **A dynamic network approach for the study of human phenotypes.** *PLoS Comput Biol* 2009, **5**(4):e1000353.

29. Park J, Lee DS, Christakis NA, Barabasi AL: **The impact of cellular networks on disease comorbidity.** *Mol Syst Biol* 2009, **5**:262.

30. Basen-Engquist K, Chang M: **Obesity and cancer risk: recent review and evidence.** *Curr Oncol Rep* 2011, **13**(1):71–76.

31. Guindi M, Riddell RH: **Indeterminate colitis.** *J Clin Pathol* 2004, **57**(12):1233–1244.

32. Rodriguez-Murillo L, Greenberg DA: **Genetic association analysis: a primer on how it works, its strengths and its weaknesses.** *Int J Androl* 2008, **31**(6):546–556.

33. Mead S, Poulter M, Uphill J, Beck J, Whitfield J, Webb TE, Campbell T, Adamson G, Deriziotis P, Tabrizi SJ, *et al*: **Genetic risk factors for variant Creutzfeldt-Jakob disease: a genome-wide association study.** *Lancet Neurol* 2009, **8**(1):57–66.

34. Nuinoon M, Makarasara W, Mushiroda T, Setianingsih I, Wahidiyat PA, Sripichai O, Kumasaka N, Takahashi A, Svasti S, Munkongdee T, *et al*: **A genome-wide association identified the common genetic variants influence disease severity in beta0-thalassemia/hemoglobin E.** *Hum Genet* 2010, **127**(3):303–314.

35. Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, Zhernakova A, Heap GA, Adany R, Aromaa A, *et al*: **Multiple common variants for celiac disease influencing immune gene expression.** *Nat Genet* 2010, **42**(4):295–302.

36. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ, *et al*: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447**(7145):661–678.

37. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JRB, Elliott KS, Lango H, Rayner NW, *et al*: **A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity.** *Science* 2007, **316**(5826):889–894.

38. Renstrom F, Payne F, Nordstrom A, Brito EC, Rolandsson O, Hallmans G, Barroso I, Nordstrom P, Franks PW, Consortium G: **Replication and extension of genome-wide association study results for obesity in 4923 adults from northern Sweden.** *Hum Mol Genet* 2009, **18**(8):1489–1496.

39. Gudbjartsson DF, Arnar DO, Helgadottir A, Gretarsdottir S, Holm H, Sigurdsson A, Jonasdottir A, Baker A, Thorleifsson G, Kristjansson K, *et al*: **Variants conferring risk of atrial fibrillation on chromosome 4q25.** *Nature* 2007, **448**(7151):353–357.

40. Gretarsdottir S, Thorleifsson G, Manolescu A, Styrkarsdottir U, Helgadottir A, Gschwendtner A, Kostulas K, Kuhlenbaumer G, Bevan S, Jonsdottir T, *et al*: **Risk variants for atrial fibrillation on chromosome 4q25 associate with ischemic stroke.** *Ann Neurol* 2008, **64**(4):402–409.

41. Johansen CT, Wang J, Lanktree MB, Cao H, McIntyre AD, Ban MR, Martins RA, Kennedy BA, Hassell RG, Visser ME, *et al*: **Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia.** *Nat Genet* 2010, **42**(8):684–687.

42. Schunkert H, Konig IR, Kathiresan S, Reilly MP, Assimes TL, Holm H, Preuss M, Stewart AF, Barbalic M, Gieger C, *et al*: **Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease.** *Nat Genet* 2011, **43**(4):333–338.

43. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, Abraham C, Regueiro M, Griffiths A, *et al*: **A genome-wide association study identifies IL23R as an inflammatory bowel disease gene.** *Science* 2006, **314**(5804):1461–1463.

44. Franke A, Balschun T, Karlsen TH, Sventoraityte J, Nikolaus S, Mayr G, Domingues FS, Albrecht M, Nothnagel M, Ellinghaus D, *et al*: **Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility.** *Nat Genet* 2008, **40**(11):1319–1323.

45. Madsen K: **Combining T cells and IL-10: a new therapy for Crohn's disease?** *Gastroenterology* 2002, **123**(6):2140–2144.

46. de Bakker PI, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, Ke X, Monsuur AJ, Whittaker P, Delgado M, *et al*: **A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC.** *Nat Genet* 2006, **38**(10):1166–1172.

47. Landi MT, Chatterjee N, Yu K, Goldin LR, Goldstein AM, Rotunno M, Mirabello L, Jacobs K, Wheeler W, Yeager M, *et al*: **A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma.** *Am J Hum Genet* 2009, **85**(5):679–691.

48. Miki D, Kubo M, Takahashi A, Yoon KA, Kim J, Lee GK, Zo JI, Lee JS, Hosono N, Morizono T, *et al*: **Variation in TP63 is associated with lung adenocarcinoma susceptibility in Japanese and Korean populations.** *Nat Genet* 2010, **42**(10):893.

49. Mushiroda T, Wattanapokayakit S, Takahashi A, Nukiwa T, Kudoh S, Ogura T, Taniguchi H, Kubo M, Kamatani N, Nakamura Y, *et al*: **A genome-wide association study identifies an association of a common variant in TERT with susceptibility to idiopathic pulmonary fibrosis.** *J Med Genet* 2008, **45**(10):654–656.

50. Sanson M, Hosking FJ, Shete S, Zelenika D, Dobbins SE, Ma Y, Enciso-Mora V, Idbaih A, Delattre JY, Hoang-Xuan K, *et al*: **Chromosome 7p11.2 (EGFR) variation influences glioma risk.** *Hum Mol Genet* 2011, **20**(14):2897–2904.

51. Shete S, Hosking FJ, Robertson LB, Dobbins SE, Sanson M, Malmer B, Simon M, Marie Y, Boisselier B, Delattre JY, *et al*: **Genome-wide association study identifies five susceptibility loci for glioma.** *Nat Genet* 2009, **41**(8):899–904.

52. Hsiung CA, Lan Q, Hong YC, Chen CJ, Hosgood HD, Chang IS, Chatterjee N, Brennan P, Wu C, Zheng W, *et al*: **The 5p15.33 Locus is associated with risk of lung adenocarcinoma in never-smoking females in Asia.** *PLoS Genet* 2010, **6**(8):e1001051.

53. Turnbull C, Rapley EA, Seal S, Pernet D, Renwick A, Hughes D, Ricketts M, Linger R, Nsengimana J, Deloukas P, *et al*: **Variants near DMRT1, TERT and ATF7IP are associated with testicular germ cell cancer.** *Nat Genet* 2010, **42**(7):604–607.

54. Xu Y, He K, Goldkorn A: **Telomerase targeted therapy in cancer and cancer stem cells.** *Clin Adv Hematol Oncol* 2011, **9**(6):442–455.

55. Ramos C, Montano M, Garcia-Alvarez J, Ruiz V, Uhal BD, Selman M, Pardo A: **Fibroblasts from idiopathic pulmonary fibrosis and normal lungs differ in growth rate, apoptosis, and tissue inhibitor of metalloproteinases expression.** *Am J Respir Cell Mol Biol* 2001, **24**(5):591–598.

56. Schrader M, Burger AM, Muller M, Krause H, Straub B, Smith GL, Newlands ES, Miller K: **Quantification of human telomerase reverse transcriptase mRNA in testicular germ cell tumors by quantitative fluorescence real-time RT-PCR.** *Oncol Rep* 2002, **9**(5):1097–1105.

57. Houlston RS, Cheadle J, Dobbins SE, Tenesa A, Jones AM, Howarth K, Spain SL, Broderick P, Domingo E, Farrington S, *et al*: **Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33.** *Nat Genet* 2010, **42**(11):973–U989.

58. Jones AM, Beggs AD, Carvajal-Carmona L, Farrington S, Tenesa A, Walker M, Howarth K, Ballereau S, Hodgson SV, Zauber A, *et al*: **TERC polymorphisms are associated both with susceptibility to colorectal cancer and with longer telomeres.** *Gut* 2011, **61**(2):248–254.

59. Cottliar A, Palumbo M, La Motta G, de Barrio S, Crivelli A, Viola M, Gomez JC, Slavutsky I: **Telomere length study in celiac disease.** *Am J Gastroenterol* 2003, **98**(12):2727–2731.

60. Bishop DT, Demenais F, Iles MM, Harland M, Taylor JC, Corda E, Randerson-Moor J, Aitken JF, Avril MF, Azizi E, *et al*: **Genome-wide association study identifies three loci associated with melanoma risk.** *Nat Genet* 2009, **41**(8):920–925.

61. Jin Y, Birlea SA, Fain PR, Gowan K, Riccardi SL, Holland PJ, Mailloux CM, Sufit AJ, Hutton SM, Amadi-Myers A, *et al*: **Variant of TYR and autoimmunity susceptibility loci in generalized vitiligo.** *N Engl J Med* 2010, **362**(18):1686–1697.

62. Spritz RA: **The genetics of generalized vitiligo: autoimmune pathways and an inverse relationship with malignant melanoma.** *Genome Med* 2010, **2**(10):78.

63. Strange A, Capon F, Spencer CCA, Knight J, Weale ME, Allen MH, Barton A, Band G, Bellenguez C, Bergboer JGM, *et al*: **A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1.** *Nat Genet* 2010, **42**(11):985–U106.

64. Freeman AF, Holland SM: **Clinical manifestations of hyper IgE syndromes.** *Dis Markers* 2010, **29**(3–4):123–130.

65. International HapMap C: **A haplotype map of the human genome.** *Nature* 2005, **437**(7063):1299–1320.

66. International HapMap C, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, *et al*: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449**(7164):851–861.

67. Genomes Project C: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467**(7319):1061–1073.

68. Huffmeier U, Uebe S, Ekici AB, Bowes J, Giardina E, Korendowych E, Juneblad K, Apel M, McManus R, Ho P, *et al*: **Common variants at TRAF3IP2 are associated with susceptibility to psoriatic arthritis and psoriasis.** *Nat Genet* 2010, **42**(11):996–999.

69. Ellinghaus E, Ellinghaus D, Stuart PE, Nair RP, Debrus S, Raelson JV, Belouchi M, Fournier H, Reinhard C, Ding J, *et al*: **Genome-wide association study identifies a psoriasis susceptibility locus at TRAF3IP2.** *Nat Genet* 2010, **42**(11):991–995.

70. Sun LD, Xiao FL, Li Y, Zhou WM, Tang HY, Tang XF, Zhang H, Schaarschmidt H, Zuo XB, Foelster-Holst R, *et al*: **Genome-wide association study identifies two new susceptibility loci for atopic dermatitis in the Chinese Han population.** *Nat Genet* 2011, **43**(7):690–694.

71. Wrensch M, Jenkins RB, Chang JS, Yeh RF, Xiao Y, Decker PA, Ballman KV, Berger M, Buckner JC, Chang S, *et al*: **Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility.** *Nat Genet* 2009, **41**(8):905–908.

72. Caspari E: **A synopsis of contemporary evolutionary thinking.** *Evol Int J Org Evol* 1949, **3**(4):377.

73. Klein TE, Chang JT, Cho MK, Easton KL, Fergerson R, Hewett M, Lin Z, Liu Y, Liu S, Oliver DE, *et al*: **Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics research network and knowledge base.** *Pharmacogenomics J* 2001, **1**(3):167–170.

74. Bowes J, Barton A: **The genetics of psoriatic arthritis: lessons from genome-wide association studies.** *Discov Med* 2010, **10**(52):177–183.

75. Liu Y, Helms C, Liao W, Zaba LC, Duan S, Gardner J, Wise C, Miner A, Malloy MJ, Pullinger CR, *et al*: **A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci.** *PLoS Genet* 2008, **4**(3):e1000041.

76. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ: **A navigator for human genome epidemiology.** *Nat Genet* 2008, **40**(2):124–125.

77. Barnickel T, Weston J, Collobert R, Mewes HW, Stumpflen V: **Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts.** *PLoS One* 2009, **4**(7):e6393.

78. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PIW: **SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap.** *Bioinformatics* 2008, **24**(24):2938–2939.

79. Dong J, Horvath S: **Understanding network concepts in modules.** *BMC Syst Biol* 2007, **1**:24.

80. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, *et al*: **A human protein-protein interaction network: a resource for annotating the proteome.** *Cell* 2005, **122**(6):957–968.

Additional files

Additional_file_1 as PDF

Additional file 1 Figure S1: Disease-centric projection of the SVN. The SVN (see Figure 1C) is transformed in a network consisting of diseases only. Here, two traits are connected if they are associated with the same variant. The colors of the disease nodes correspond to disease classes according to the MeSH ontology, multi-colored nodes indicate an association with different disease classes. The node size reflects the number of traits a disease has shared associations with. The direction of the shared variants is indicated by the edge color reflecting the corresponding allelic information: gray indicates agonistic variant(s), red corresponds to antagonistic variant(s), and blue mark both agonistic and antagonistic signals in the two corresponding traits.

Additional_file_2 as XLSX

Additional file 2 Table S1: List of all disease-variant associations contained in the SVN. Contained is the high-quality data set which was used for the construction of the SVN, ordered by the rs-number of the tagging SNP. The first column contains this rs-number of the tagging SNP, the second column lists the disease associations and the third column gives the PubMed ID of the GWAS publication the association was reported in. In the fourth column the (gene or intergenic) locus of the tagging SNP can be found. The sixth column gives the SNP and the risk allele reported in the GWAS. If the rs-numbers of the tagging SNP (column 1) diverges from the rs-number listed here, the association was assigned via LD. For these cases, in column seven the corresponding allele of the tagging SNP is given, followed by the P-value and the odds ratio reported with the SNP (i.e. the reported SNP in column six). Blue row-coloring identifies non-HLA located antagonistic SNPs, while rows containing agonistic SNPs are not colored. Rows in green list antagonistic SNPs in the HLA region (not considered in the manuscript). Tagging SNPs which we included in our rationale are marked in bold red font.

Additional_file_3 as XLSX

Additional file 3 Table S2: CPMA P-values for autoimmune-linked SNPs and their corresponding loci in the SVN. Listed are all SNPs contained in Supplementary Table 1 for which association data could be obtained from [8]. The second column gives the LD-based loci of the SNPs as used in the SVN. The third column contains the CPMA P-Values.

Additional_file_4 as PDF

Additional file 4 Figure S2: Network properties of the SVN. A: The log-log-plot of the degree distribution of the SVN follows a power-law ($\gamma = 1.32; \mathrm{R}^2 = 0.69$) and therefore attributes the SVN to be scale-free and, thus, non-random. B: The modular structure of the SVN was confirmed by the topological coefficient which follows a power-law distribution on a log-log-scale. When considering the two node types separately, in both cases a scale-free topology can be identified: C: disease nodes ($\gamma = 0.97$; $R^2 = 0.71$) and D: locus nodes ($\gamma = 2.98$; $R^2 = 0.93$).

Additional_file_5 as GRAPHML

Additional file 5 Graph data of the SLN in yEd graphml format. View with yEd (http://www.yworks.com/en/products_yed_about.html). Using yEd, the file can be converted to GML format which is readable by Cytoscape (http://www.cytoscape.org/).

Additional_file_6 as GRAPHML

Additional file 6 Graph data of the SVN in yEd graphml format. View with yEd (http://www.yworks.com/en/products_yed_about.html). Using yEd, the file can be converted to GML format which is readable by Cytoscape (http://www.cytoscape.org/).

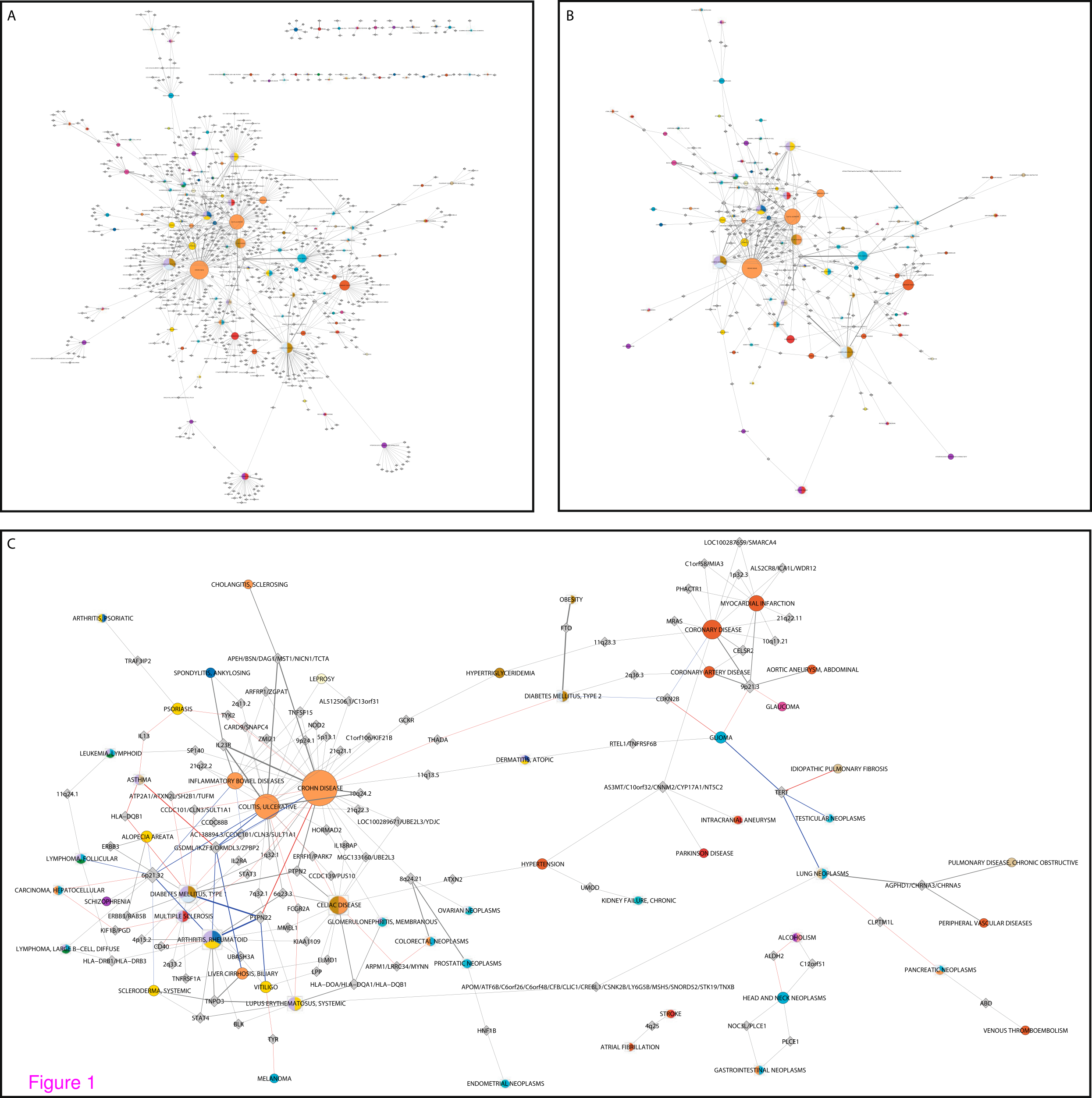
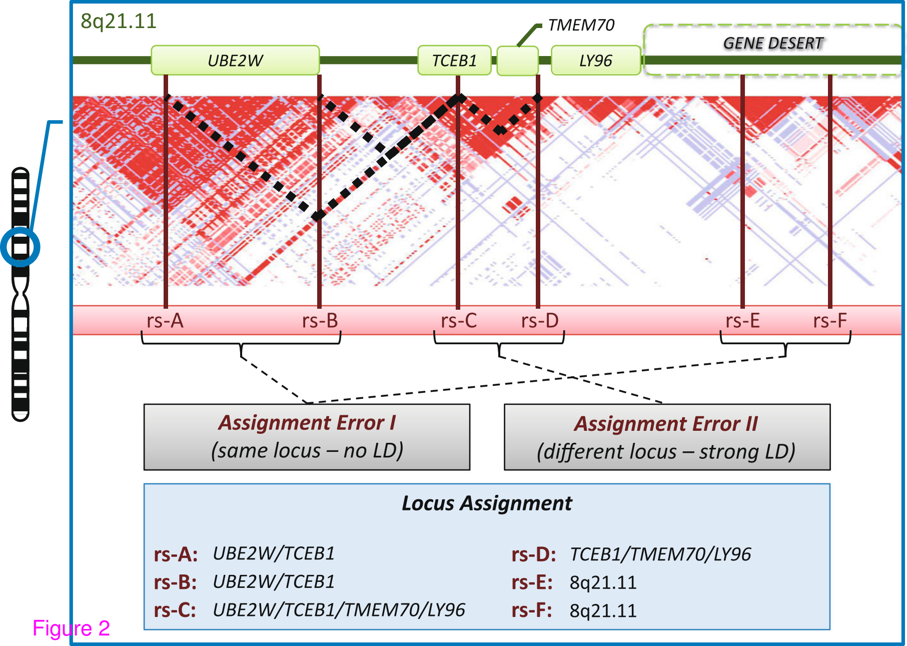

A
B
C
LOC100287659/SMARCA4
C1orf58/MIA3
ALS2CR8/ICA1L/WDR12
1p32.3
CHOLANGITIS, SCLEROSING
PHACTR1
OBESITY
MYOCARDIAL INFARCTION
ARTHRITIS, PSORIATIC
MRAS
21q22.11
CORONARY DISEASE
FTO
11q23.3
10q11.21
CELSR2
TRAF3IP2
APEH/BSN/DAG1/MST1/NICN1/TCTA
AORTIC ANEURYSM, ABDOMINAL
SPONDYLITIS, ANKYLOSING
HYPERTRIGLYCERIDEMIA
2q36.3
CORONARY ARTERY DISEASE
LEPROSY
9p21.3
ARFRP1/ZGPAT
DIABETES MELLITUS, TYPE 2
2q11.2
CDKN2B
AL512506.1/C13orf31
GLAUCOMA
PSORIASIS
TNFSF15
TYK2
GCKR
NOD2
CARD9/SNAPC4
5p13.1
IL13
ZMIZ1
C1orf106/KIF21B
THADA
9p24.1
GLIOMA
IL23R
RTEL1/TNFRSF6B
SP140
21q21.1
LEUKEMIA, LYMPHOID
DERMATITIS, ATOPIC
21q22.2
11q13.5
ASTHMA
IDIOPATHIC PULMONARY FIBROSIS
INFLAMMATORY BOWEL DISEASES
CROHN DISEASE
AS3MT/C10orf32/CNNM2/CYP17A1/NT5C2
11q24.1
ATP2A1/ATXN2L/SH2B1/TUFM
10q24.2
TERT
COLITIS, ULCERATIVE
CCDC101/CLN3/SULT1A1
21q22.3
HLA−DQB1
LOC100289671/UBE2L3/YDJC
TESTICULAR NEOPLASMS
INTRACRANIAL ANEURYSM
CCDC88B
HORMAD2
ALOPECIA AREATA
AC138894.3/CCDC101/CLN3/SULT1A1
ERBB3
IL18RAP
GSDML/IKZF3/ORMDL3/ZPBP2
PARKINSON DISEASE
1q32.1
ERRFI1/PARK7
MGC133160/UBE2L3
PULMONARY DISEASE, CHRONIC OBSTRUCTIVE
LYMPHOMA, FOLLICULAR
IL2RA
HYPERTENSION
STAT3
LUNG NEOPLASMS
PTPN2
6p21.32
CCDC139/PUS10
8q24.21
ATXN2
AGPHD1/CHRNA3/CHRNA5
UMOD
CARCINOMA, HEPATOCELLULAR
DIABETES MELLITUS, TYPE 1
7q32.1
6q23.3
KIDNEY FAILURE, CHRONIC
SCHIZOPHRENIA
CELIAC DISEASE
OVARIAN NEOPLASMS
FCGR2A
ERBB3/RAB5B
MULTIPLE SCLEROSIS
PTPN22
CLPTM1L
PERIPHERAL VASCULAR DISEASES
GLOMERULONEPHRITIS, MEMBRANOUS
MMEL1
KIF1B/PGD
ALCOHOLISM
4p15.2
ARTHRITIS, RHEUMATOID
KIAA1109
COLORECTAL NEOPLASMS
LYMPHOMA, LARGE B−CELL, DIFFUSE
CD40
ALDH2
C12orf51
HLA−DRB1/HLA−DRB3
ELMO1
UBASH3A
PROSTATIC NEOPLASMS
2q33.2
ARPM1/LRRC34/MYNN
PANCREATIC NEOPLASMS
LPP
TNFRSF1A
LIVER CIRRHOSIS, BILIARY
HLA−DOA/HLA−DQA1/HLA−DQB1
VITILIGO
SCLERODERMA, SYSTEMIC
APOM/ATF6B/C6orf26/C6orf48/CFB/CLIC1/CREBL1/CSNK2B/LY6G5B/MSH5/SNORD52/STK19/TNXB
HEAD AND NECK NEOPLASMS
TNPO3
LUPUS ERYTHEMATOSUS, SYSTEMIC
ABO
STROKE
STAT4
4q25
NOC3L/PLCE1
VENOUS THROMBOEMBOLISM
BLK
HNF1B
TYR
PLCE1
ATRIAL FIBRILLATION
GASTROINTESTINAL NEOPLASMS
MELANOMA
ENDOMETRIAL NEOPLASMS

Figure 1



Assignment Error I
(same locus – no LD)

Assignment Error II
(different locus – strong LD)

Locus Assignment

rs-A:	*UBE2W/TCEB1*	**rs-D:**	*TCEB1/TMEM70/LY96*
rs-B:	*UBE2W/TCEB1*	**rs-E:**	8q21.11
rs-C:	*UBE2W/TCEB1/TMEM70/LY96*	**rs-F:**	8q21.11

Figure 2

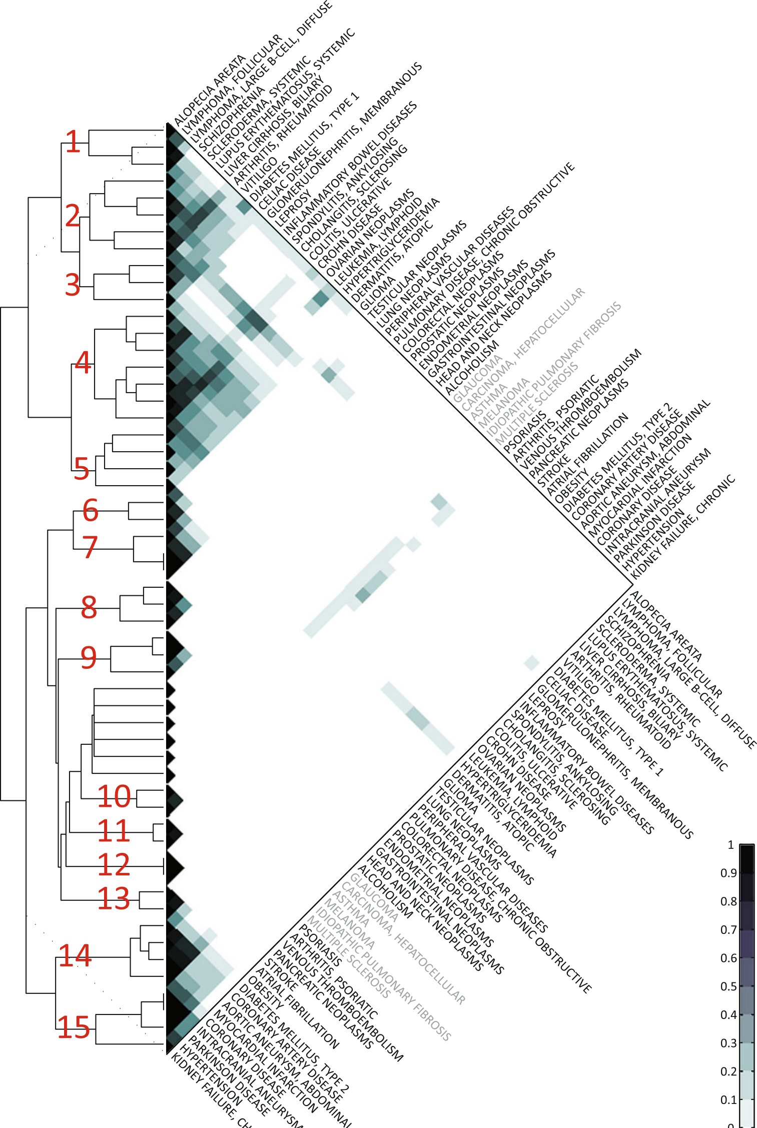

ALOPECIA AREATA
LYMPHOMA, FOLLICULAR
LYMPHOMA, LARGE B-CELL, DIFFUSE
SCHIZOPHRENIA
SCLERODERMA, SYSTEMIC
LUPUS ERYTHEMATOSUS, SYSTEMIC
LIVER CIRRHOSIS, BILIARY
ARTHRITIS, RHEUMATOID
VITILIGO
DIABETES MELLITUS, TYPE 1
CELIAC DISEASE
GLOMERULONEPHRITIS, MEMBRANOUS
LEPROSY
INFLAMMATORY BOWEL DISEASES
SPONDYLITIS, ANKYLOSING
CHOLANGITIS, SCLEROSING
COLITIS, ULCERATIVE
CROHN DISEASE
OVARIAN NEOPLASMS
LEUKEMIA, LYMPHOID
HYPERTRIGLYCERIDEMIA
DERMATITIS, ATOPIC
GLIOMA
TESTICULAR NEOPLASMS
LUNG NEOPLASMS
PERIPHERAL VASCULAR DISEASES
PULMONARY DISEASE, CHRONIC OBSTRUCTIVE
COLORECTAL NEOPLASMS
PROSTATIC NEOPLASMS
ENDOMETRIAL NEOPLASMS
GASTROINTESTINAL NEOPLASMS
HEAD AND NECK NEOPLASMS
ALCOHOLISM
GLAUCOMA
CARCINOMA, HEPATOCELLULAR
ASTHMA
MELANOMA
IDIOPATHIC PULMONARY FIBROSIS
MULTIPLE SCLEROSIS
PSORIASIS
ARTHRITIS, PSORIATIC
VENOUS THROMBOEMBOLISM
PANCREATIC NEOPLASMS
STROKE
ATRIAL FIBRILLATION
OBESITY
DIABETES MELLITUS, TYPE 2
CORONARY ARTERY DISEASE
AORTIC ANEURYSM, ABDOMINAL
MYOCARDIAL INFARCTION
CORONARY DISEASE
INTRACRANIAL ANEURYSM
PARKINSON DISEASE
HYPERTENSION
KIDNEY FAILURE, CHRONIC
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
ALOPECIA AREATA
LYMPHOMA, FOLLICULAR
LYMPHOMA, LARGE B-CELL, DIFFUSE
SCHIZOPHRENIA
SCLERODERMA, SYSTEMIC
LUPUS ERYTHEMATOSUS, SYSTEMIC
LIVER CIRRHOSIS, BILIARY
ARTHRITIS, RHEUMATOID
VITILIGO
DIABETES MELLITUS, TYPE 1
CELIAC DISEASE
GLOMERULONEPHRITIS, MEMBRANOUS
LEPROSY
INFLAMMATORY BOWEL DISEASES
SPONDYLITIS, ANKYLOSING
CHOLANGITIS, SCLEROSING
COLITIS, ULCERATIVE
CROHN DISEASE
OVARIAN NEOPLASMS
LEUKEMIA, LYMPHOID
HYPERTRIGLYCERIDEMIA
DERMATITIS, ATOPIC
GLIOMA
TESTICULAR NEOPLASMS
LUNG NEOPLASMS
PERIPHERAL VASCULAR DISEASES
PULMONARY DISEASE, CHRONIC OBSTRUCTIVE
COLORECTAL NEOPLASMS
PROSTATIC NEOPLASMS
ENDOMETRIAL NEOPLASMS
GASTROINTESTINAL NEOPLASMS
HEAD AND NECK NEOPLASMS
ALCOHOLISM
GLAUCOMA
CARCINOMA, HEPATOCELLULAR
ASTHMA
MELANOMA
IDIOPATHIC PULMONARY FIBROSIS
MULTIPLE SCLEROSIS
PSORIASIS
ARTHRITIS, PSORIATIC
VENOUS THROMBOEMBOLISM
PANCREATIC NEOPLASMS
STROKE
ATRIAL FIBRILLATION
OBESITY
DIABETES MELLITUS, TYPE 2
CORONARY ARTERY DISEASE
AORTIC ANEURYSM, ABDOMINAL
MYOCARDIAL INFARCTION
CORONARY DISEASE
INTRACRANIAL ANEURYSM
PARKINSON DISEASE
HYPERTENSION
KIDNEY FAILURE, CHRONIC
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1

Figure 3

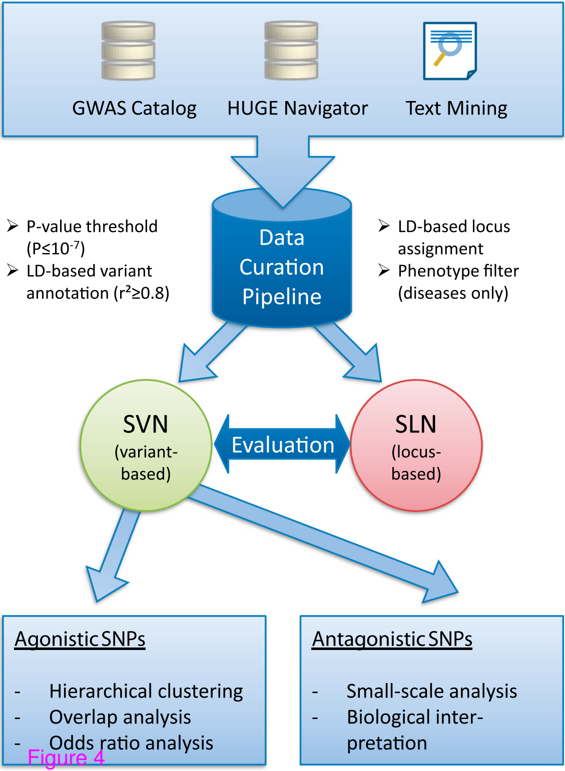

GWAS Catalog
HUGE Navigator
Text Mining
Data Curation Pipeline
➢ P-value threshold ($P \leq 10^{-7}$)
➢ LD-based variant annotation ($r^2 \geq 0.8$)
➢ LD-based locus assignment
➢ Phenotype filter (diseases only)
SVN (variant-based)
Evaluation
SLN (locus-based)
Agonistic SNPs
- Hierarchical clustering
- Overlap analysis
- Odds ratio analysis
Antagonistic SNPs
- Small-scale analysis
- Biological interpretation

Figure 4

Additional files provided with this submission:

Additional file 1: 7328979673767669_add1.pdf, 507K
http://www.biomedcentral.com/imedia/2955874668082640/supp1.pdf
Additional file 2: 7328979673767669_add2.xlsx, 54K
http://www.biomedcentral.com/imedia/1382509598808264/supp2.xlsx
Additional file 3: 7328979673767669_add3.xlsx, 12K
http://www.biomedcentral.com/imedia/1281194993808264/supp3.xlsx
Additional file 4: 7328979673767669_add4.pdf, 87K
http://www.biomedcentral.com/imedia/6772348988082639/supp4.pdf
Additional file 5: 7328979673767669_add5.graphml, 660K
http://www.biomedcentral.com/imedia/1937379993808263/supp5.graphml
Additional file 6: 7328979673767669_add6.graphml, 406K
http://www.biomedcentral.com/imedia/1006557229808263/supp6.graphml