



A Bayesian analysis of the chromosome architecture of human disorders by integrating reductionist data

SUBJECT AREAS:

COMPUTATIONAL
BIOLOGY

SYSTEMS BIOLOGY

STATISTICS

NETWORKS AND SYSTEMS
BIOLOGY

Frank Emmert-Streib¹, Ricardo de Matos Simoes¹, Shailesh Tripathi¹, Galina V. Glazko²
& Matthias Dehmer³

¹Computational Biology and Machine Learning Lab, Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, 97 Lisburn Road, Belfast, UK, ²Division of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, USA, ³Institute for Bioinformatics and Translational Research, UMIT, Hall in Tyrol, Austria.

Received
29 May 2012

Accepted
27 June 2012

Published
20 July 2012

In this paper, we present a Bayesian approach to estimate a chromosome and a disorder network from the *Online Mendelian Inheritance in Man* (OMIM) database. In contrast to other approaches, we obtain statistic rather than deterministic networks enabling a parametric control in the uncertainty of the underlying disorder-disease gene associations contained in the OMIM, on which the networks are based. From a structural investigation of the chromosome network, we identify three chromosome subgroups that reflect architectural differences in chromosome-disorder associations that are predictively exploitable for a functional analysis of diseases.

Correspondence and requests for materials should be addressed to F.E.-S. (v@bio-complexity.com)

Within the last few years a new approach to biomedical problems has emerged called *network medicine* or *systems biomedicine*^{1–4}. In contrast to classic approaches to biological or medical problems emphasizing individual genes or proteins⁵, these attempts emphasize the integration of molecular and cellular data and the consideration of interactions and their hierarchy among key components on and between these levels^{6–8}. In general, systems biology approaches based on *networks* are among the most innovative contributions to the recent progress in biology and medicine^{9,10}. One reason for their success is the fact that a graphical visualization of interacting genes or gene products leads naturally to a network representation which is easily amenable for a theoretical analysis by statistical and computational means, because networks form also data structures¹¹. Hence, the graphical visualization of networks is not only appealing but the same networks pave the way for a quantitative investigation of biological information represented by the network structure itself.

A crucial part of any sensible systems approaches, especially if aiming for novel insights into biomedical problems, is the availability of data that provide information of the connection between the genetic or molecular level with phenotypes. Here, by phenotype information, we mean either information about different grades or subtypes of a disease or different stages of its pathogenesis or information about different types of disorders. An interesting example for such an approach that is based on information provided by the *Online Mendelian Inheritance in Man* (OMIM) database is a study conducted by van Driel et al.¹². They constructed feature vectors to represent phenotypes and the components of these feature vectors provide information about the occurrence frequency of medical subject headings (MeSH) in the OMIM database. A correlation analysis revealed that similar phenotypes correspond to modules of genes with a similar biological function. In addition, by comparing these modules with information about protein-protein interactions (PPI) they found that these modules contain many known protein interactions.

Probably the first study that casts a similar problem as studied by van Driel et al.¹² explicitly into a network context was proposed in a seminal work by Goh et al.¹³. There, a bipartite network, called the DISEASOME, was constructed from disorder-disease gene associations based on the OMIM database. The DISEASOME network consists of two different types of nodes. One type of nodes correspond to disorders, and the other type to disease genes. Links can only occur between nodes of different types connecting disease causing genes with disorders. The interesting role of the DISEASOME is that it allows to derive two further network from it, namely, the disease gene network and the disease network. In the former, the nodes in the network consist of disease genes and two nodes are connected if there is at least one disease that is co-associated with both disease genes. In the disease network



nodes correspond to disorders and two disorders are connected if there is at least one disease gene that is co-associated with both disorders. Formally, both networks can be easily constructed from the DISEASOME. In the meanwhile there are various applications of the DISEASOME that studied in detail the modular structure of the disease network¹⁴, improved algorithmic methods for predicting disease-genes and modules¹⁵ or integrated additional data, e.g., in the form of PPI networks or metabolic networks^{16–21}. Also, it has been shown that a DISEASOME can be constructed from various other data types, e.g., from genome-wide association studies (GWAS)^{22–24}. For a detailed review of related approaches see²⁵.

In this paper we tie in to these previous studies by further investigating the connection of associations between disease genes and disorders and the exploitation of the obtained results. More precisely, the purpose of the present paper is three-fold. First, we study the enrichment of disease genes on chromosomes to investigate if there are specific chromosomes that have a significant association with particular disorders. Second, we introduce a Bayesian framework to estimate a chromosome and a disorder network from the OMIM database. This is a methodological advancement over previous approaches^{13–15,22,24,26}, because our approach is statistic in nature and not deterministic allowing for a quantification of the uncertainty contained in the OMIM database. Third, we investigate the resulting chromosome and disorder network estimated from our Bayesian approach. Briefly, we define a chromosome network as a graph where nodes correspond to chromosomes and a link connects two chromosomes if there is at least one disease showing a statistical association with disease genes on the two chromosomes. Similarly, in the disorder network nodes correspond to ‘disorder categories’ and two nodes are connected by a link if there is at least one chromosome that is statistical associated with the two disorder categories. Due to the fact that we estimate both networks from our Bayesian analysis of the OMIM database, these networks form statistic rather than deterministic networks. The advantage of this is that the level of uncertainty in the estimated networks is controllable by a parameter, F , reflecting the degree of statistical importance. Figuratively, this is similar to a frequentist significance level allowing to control the tolerable error rate for a Type I error^{27,28}.

The underlying rationale of our investigations is based on the assumption that chromosomes can be seen as a higher organizational level, above genes. In this role, chromosomes can be perceived as disease causing variables. Evidence in support of this for which experimental results are available include studies about copy numbers variation (gain or loss of an entire chromosome), chromosomal aberrations (gain or loss of a fragment of a genetic material) and Single Nucleotide Polymorphisms (SNPs). So far, there are only a few known genetic disorders, associated with an extra copy of genetic material (duplication of the entire chromosome), such as Edwards syndrome (trisomy of 18 chromosome), or Down syndrome (trisomy of 21 chromosome), probably because the majority of fetuses with an increased number of other chromosomes are not viable. In turn, copy number variations and changes in a gene dosage resulting from the deletion or the amplification of a gene and its genomic context are overwhelming in tumors, and can be even tumor-specific. For example, a gain of 8q21.3–q24.3 and losses of 8p23.1–p21.1, 13q14.13–q22.1 and 6q14.1–q21 are considered as ‘characteristic’ aberrations in prostate tumors. That means, these aberrations were observed in more than 20% of the examined tumors²⁹. Similarly, in colorectal cancer most studies reported frequent gains of chromosome 7, 8q, 13q, 20q and losses of 4 and 18q³⁰. That is, as an example, chromosomes 13 and 6 can be associated with prostate tumors and the chromosomes 4 and 18 with colorectal tumors.

Despite that fact that the most straightforward way of associating chromosomes with disorders is via disease genes, the cases when a disease is the result of a single mutated gene are rare. In contrast, it is more common that genes responsible for diseases reside on

Table 1 | Disorders with at least 5 known disease genes which lead to statistically enriched chromosomes for a false discovery rate of $FDR = 0.01$. The second column provides information about wider disorder categories to which a disease belongs

disorder (# genes)	category	enriched Chr
Thalassemia (5)	Hematological	11, 16
Schizophrenia (9)	Psychiatric	22
Asthma (13)	Respiratory	5
Factor VII deficiency (8)	Hematological	13
Long QT syndrome (7)	Cardiovascular	21
Mental retardation (24)	Neurological	X
Pancreatic cancer (9)	Cancer	18

different chromosomes. Importantly, genome-wide association studies (GWAS) are providing more and more cases of previously unsuspected associations between genes and disorders³¹. For example, in genome-wide studies the susceptibility of age-related macular degeneration was found to be associated with allelic variants in the complement factor H. However, there can be other causal variants too³¹.

Based on such findings, we investigate the architecture of an estimated chromosome network to identify disorder associated chromosomal subgroups. That means, e.g., similar to biological pathways which represent an interacting group of genes essential to maintain particular biological functions, we aim to identify subsets of chromosomes because the genes located on these are associated with the *malfunctioning* of biological functions, which manifest in disorders. Hence, our underlying rationale is inspired by pathway-based studies utilizing significant modifications in the interaction structure among genes³².

This paper is organized as follows. In the next section we, first, present results about the enrichment of disease genes on chromosomes. Then, we present two Bayesian approaches that allow to study disorder-chromosome associations statistically. The results from these analyses are used to estimate a chromosome and a disorder network. The results section finishes with a structural analysis of the chromosome network and a discussion of the obtained results. Finally, the paper finishes with concluding remarks and by highlighting differences of our approach and previous studies.

Results

Disease gene enriched chromosomes. The first analysis we perform tests the enrichment of general disease genes on the chromosomes by a hypergeometric test; also called Fisher’s exact test³³. That means, we categorize all genes in exactly two categories. The first category consists of all 1722 known disease genes and the second consists of 18548 = (20270–1722) genes, which are all other genes. Then, we test for each chromosome separately the enrichment of the disease genes (d-genes) among all protein coding genes (p-genes) on the chromosome. This results in 24 p-values simultaneously obtained by a hypergeometric test. Due to the fact that we are testing multiple hypotheses we correct these p-values controlling the *false-discovery rate* (FDR) with the Benjamini-Hochberg procedure³⁴. For our following analysis we use always the stringent value of $FDR = 0.01$, if not stated otherwise, to declare only results as significant which are uncontroversial. The results of this analysis leads only to the X chromosome as significant. All other chromosomes test not significant.

In order to obtain disease specific results for the enrichment on the chromosomes, we repeat the above analysis for each of the 1284 disorders in the OMIM database separately. The results of this analysis are shown in table 1. We find a total of 7 different disorders that lead to at least one enriched chromosome (listed in column three in the table). It is interesting to note that only one disease, namely,



Table 2 | Disorder categories leading to statistically enriched chromosomes for a false discovery rate of $FDR = 0.01$. The first column shows the name of the disorder category, the second column gives the number of enriched chromosomes and the third column lists these chromosomes

disorder category (# genes)	# enriched Chr	enriched Chr
Bone (62)	-	-
Cancer (372)	4	10, 13, 17, 22
Cardiovascular (125)	-	-
Connective tissue (1)	1	5
Connective tissue disorder (63)	-	-
Dermatological (123)	2	12, 17
Developmental (59)	-	-
Ear, Nose, Throat (57)	-	-
Endocrine (134)	-	-
Gastrointestinal (45)	1	5
Hematological (212)	2	4, X
Immunological (137)	-	-
Metabolic (345)	-	-
multiple (252)	1	X
Muscular (104)	-	-
Neurological (344)	1	X
Nutritional (26)	-	-
Ophthalmological (196)	1	X
Psychiatric (36)	-	-
Renal (70)	-	-
Respiratory (38)	-	-
Skeletal (96)	1	4
Unclassified (32)	1	7

Thalassemia a hematological disorder leads to two enriched chromosomes. An interpretation of these results is that each of these disorders could be localized on specific, individual chromosomes only. However, given that the number of known disease genes for these disorders (shown in the first column in table 1 in bracket) is quite low there might just not be enough disease genes to enrich more than one or two chromosomes, despite the fact that a disease is not just localized on one or two chromosomes. Given the low number of disease genes, the latter reason seems more likely.

As conjectured above, the number of available disease genes is in general too low to lead to an enrichment of multiple chromosomes. In order to compensate for this lack of information, we repeat an enrichment analysis for 23 disorder categories, instead of individual disorders. Due to the fact that each category is the sum of a larger number of individual diseases, the available number of disease genes in these categories is significantly enlarged. The results of this analysis are shown in table 2.

For the 23 disorder categories we find indeed several categories that lead to an enrichment of multiple chromosomes. The category with the highest number of enriched chromosomes is ‘Cancer’ (4). This is not unexpected because cancer is known to be a very heterogeneous disease having many subtypes and subgrades and also individual tumors itself are composed of heterogeneous cells. Also the negative results in table 2, showing no enriched chromosomes at all, are plausible. For example, the disorders underlying the categories ‘Psychiatric’, ‘Nutritional’ or ‘Endocrine’ are only poorly understood on the genetic level. Further, it is of interest to note that the overlap between the 23 disorder categories is only marginal. That means, there is no category consisting of more than one enriched chromosome which is part of another category. This is also plausible because otherwise such a category would be better classified as a subcategory of some other disease class rather than to establish its own.

Two Bayesian approaches to disorder-chromosome associations. Principally, the OMIM database provides directly information about

the associations between genes, their chromosomal locations and disorders. Formally, this can be written as the following mapping:

$$g \rightarrow C \rightarrow D \quad (1)$$

Here, g corresponds to a gene, C is the chromosome the gene is located, and D to a disorder associated with g . In the following, we continue our investigation started in the previous section by studying the connection of (enriched) chromosomes for disease genes and their effect on a disorder. That means the mapping we are considering in the following is limited to:

$$C \rightarrow D \quad (2)$$

However, this limitation makes such a mapping inherently probabilistic because we neglect information about genes, but consider only the information about the chromosome a gene is located on. Interestingly, this makes such a mapping more realistic because biologically it would not be sensible to predict with certainty that an enriched chromosome with mutations in some disease genes will *for sure* lead to a certain disorder. For this reason, the mapping in Eqn. 2 corresponds to a conditional probability, $p(D|C)$. We would like to remark that similar gene-disorder relations correspond also to information typically provided by genome-wide association studies (GWAS) on which the OMIM database is partially based on.

Interestingly, if one would like to investigate the chromosomes associated with a particular disorder one would be interested in the inverse probability, i.e., $p(C|D)$. However, this information is not provided by the OMIM database but needs to be inferred. In order to obtain practical estimates for $p(C_i|D_j)$ we use a Bayesian approach in the form:

$$p(C_i|D_j) = \frac{p(D_j|C_i)p(C_i)}{p(D_j)} \quad (3)$$

with

$$p(D_j) = \sum_{i \in \{\text{all chromosomes}\}} p(D_j|C_i)p(C_i). \quad (4)$$

Here, D_j with $j \in \{1, \dots, 24\}$ correspond to the broad disease categories listed in table 2 and C_i correspond to the chromosomes, i.e., $\{C_1, C_2, \dots, C_{22}, C_X, C_Y\}$. Statistically, the conditional probabilities $p(D_j|C_i)$ are estimated from the OMIM database for all disorder (D_j) chromosome (C_i) pairs. For the prior $p(C_i)$ we study in the following two different assumptions, namely a non-informative prior and a prior that is empirically estimated from the OMIM database. The latter will lead to an Empirical Bayes approach whereas the former is a full Bayesian approach³⁵. Specifically, we estimate the empirical prior from the frequencies of chromosome appearances in OMIM. Overall, the marginal probability in Eqn. 4 can be simply obtained by integration over the likelihood and the prior probabilities.

The results of our Bayesian analysis are shown in Fig. 1. The color code we used is as follows. Dark green and light green symbols corresponds to the non-informative (NI) and empirical (E) prior. These probabilities are included for reasons of reference. The blue stars and red pluses correspond to the posterior probabilities obtained for the non-informative respectively empirical prior. Further, we add to each plot vertical lines if the *fraction of probabilities* (FOP), given by the posterior probability divided by the prior probability, is larger than a factor of F , i.e., we add a vertical line if

$$FOP_{ij}(FB) = \frac{p_{FB}(C_i|D_j)}{p_{NI}(C_i)} > F \Rightarrow \text{add blue line} \quad (5)$$

$$FOP_{ij}(EB) = \frac{p_{EB}(C_i|D_j)}{p_{EB}(C_i)} > F \Rightarrow \text{add red line} \quad (6)$$

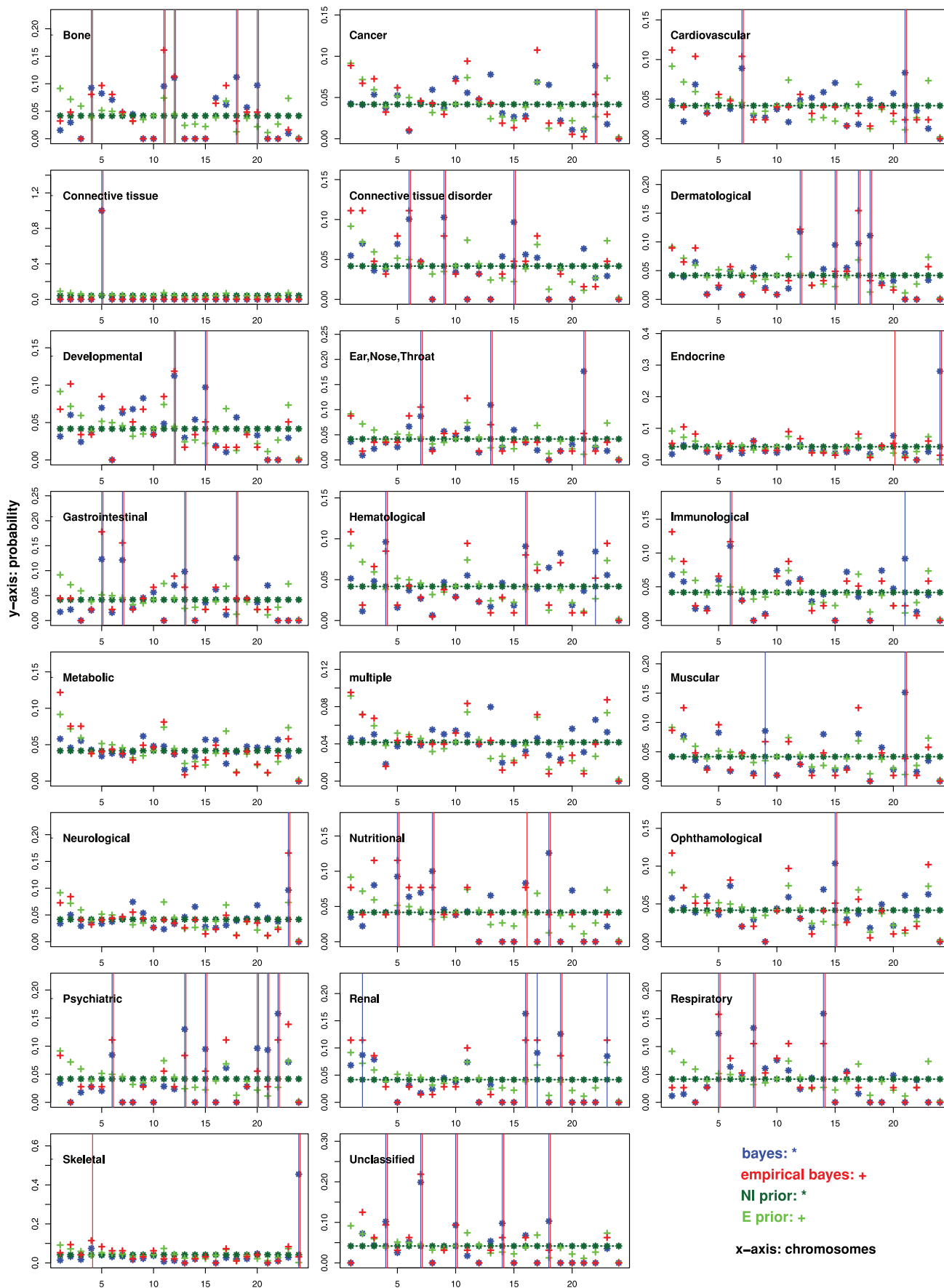


Figure 1 | Shown are the posterior and prior probabilities for the full Bayes and empirical Bayes analysis.



For reasons of a better visibility, in cases where both lines should be added we introduce a slight jitter to prevent these lines from overlapping.

For our analysis we used $F = 2.0$. That means, we require the posterior probability to be twice as large as the prior probability in order to be considered as statistically relevant. However, different choices are possible and larger values of F lead to more conservative and lower values of F to more liberal results. The reason for selecting $F = 2.0$ is partly based on our data. Estimating all values for $FOP(FB)$ and $FOP(EB)$, we find that only about 10% in each case of these values are larger than $F = 2.0$. Further, from randomizations of the data, we find all values below this threshold. Taken together, the associations that pass our criterion are unlikely to happen by chance and in addition represent only the top findings. Further, the need for a posterior probability to be twice as large as the prior probability to be considered as statistically relevant appears a sensible choice for the studied problem.

The main result of our analysis shown in Fig. 1 is that the full Bayes and the empirical Bayes method result in very similar results with respect to the modification of the posterior probabilities. This can be visually seen by the strong coincidence of the vertical lines in blue (full Bayes) and red (empirical Bayes). There are only 8 exceptions (five for FB and three for EB) compared to 52 matches. More specifically, the full Bayesian method identifies chromosome 21 for immunological, chromosome 9 for muscular and chromosome 2, 17 and X for renal disorders, whereas for the empirical Bayesian method chromosome 20 for endocrine and 16 for nutritional disorder are identified. This large correspondence is also an indicator that in this case the influence of the prior is not sensitive on the outcome, but the data, as estimated in form of the conditional probabilities of $p(D_j|C_i)$, are sufficiently strong to dominate the posterior probabilities. Further, the only disorder categories for which our method does not lead to any increased posterior chromosomal probabilities are from the categories ‘metabolic’ and ‘multiple’. In contrast, the disorder categories with the largest number of increased posterior chromosomal probabilities are ‘psychiatric’, ‘unclassified’, ‘bone’ and ‘renal’.

As a methodological alternative to the above analysis, the log-odds (LOD) is also frequently used to obtain an indicator of the statistical importance of an event in a Bayesian analysis. In order to demonstrate the similarity of both approaches for our data, we repeat the above analysis. More precisely, we estimate the log-odds (LOD) of the involvement of chromosome C_i in disease D_j compared to the none involvement of this chromosome by

$$LOD_{ij} = \log \left(\frac{p(C_i|D_j)}{p(\text{not } C_i|D_j)} \right). \quad (7)$$

The difference is that in this case, the outcome possibilities are limited to a binary case $\{C_i, \text{not } C_i\}$, instead of using all chromosomes $\{C_1, C_2, \dots, C_X, C_Y\}$. Hence, the posterior probabilities in Eqn. 7 are calculated by

$$p(C_i|D_j) = \frac{p(D_j|C_i)p(C_i)}{p(D_j)} \quad (8)$$

$$\text{with } p(D_j|C_i) = \frac{\#\text{disease genes on } C_i \rightarrow D_j}{\sum_k (\#\text{disease genes on } C_i \rightarrow D_k)} \quad (9)$$

and

$$p(\text{not } C_i|D_j) = \frac{p(D_j|\text{not } C_i)p(\text{not } C_i)}{p(D_j)} \quad (10)$$

$$\text{with } p(D_j|C_i \text{ not}) = \frac{\#\text{disease genes not on } C_i \rightarrow D_j}{\sum_k (\#\text{disease genes not on } C_i \rightarrow D_k)} \quad (11)$$

The results of the log-odds for non-informative and empirical priors are shown in Fig. 2. Similar to Fig. 1, we include vertical lines if the log-odds are larger than a factor F :

$$LOD_{ij}(FB) > \log(F) \Rightarrow \text{add blue line} \quad (12)$$

$$LOD_{ij}(EB) > \log(F) \Rightarrow \text{add red line} \quad (13)$$

This time there is a strong difference between the results for the two priors. In order to explain this discrepancy we, first, explain why the odds and the FOP are two different measures and then discuss why priors have a stronger influence on the odds.

Regarding the former point we, first, want to note that the odds in Eqn. 7 is comparing a hypothesis ($p(C_i|D_j)$) with its alternative ($p(\text{not } C_i|D_j)$), instead of the gain of a belief in a hypothesis compared to our prior information about the same hypothesis for the FOP, see Eqn. 5 and 6. Hence, the odds is generally smaller than the FOP whenever the number of alternatives is large. In our case, there is a total number of 24 different hypotheses, corresponding to the 22 autosome and the two sex chromosomes, which constitutes a large number of categories. Second, the calculation of the odds involves two posterior probabilities, whereas in Eqn. 5 and 6 a comparison between a posterior and a prior probability is conducted. That means the denominator of the odds ($p(\text{not } C_i|D_j)$) can be larger than the associated prior making the odds smaller than the FOP. Taken together, these two issues make the odds in general smaller than the FOP.

Regarding the influence of the prior information, for the odds the empirically estimated prior probabilities, i.e., $p(C_i)$ and $p(\text{not } C_i)$, give a very strong preference for the alternative, because the number of disease genes is broadly distributed over all chromosomes, see table 7, and not concentrated on a very few of these. Numerically, we find for the empirical priors from our data

$$0.905 = \min_i \{p(\text{not } C_i)\} \quad (14)$$

$$0.999 = \max_i \{p(C_i)\} \quad (15)$$

This indicates that the empirical priors are too conservative, because it would require an extremely strong signal in the data to compensate for such a strong penalty biasing the analysis. This explains the fact that only one LOD is larger than F , namely for connective tissue.

Interestingly, in addition to this result we observe that the LODs for the non-informative prior match in 54 cases with the thresholded FOPs, shown in Fig. 1. Considering that we find in total 61 cases of the LODs and 58 cases for the FOPs larger than F , this corresponds to a hit rate of 89% for the LODs and 93% for the FOPs. This large correspondence provides a justification that a non-informative prior used for the LODs is a reasonable choice, especially since for the FOPs the results for the non-informative and the empirical prior are almost identical. A summary of the results for the full Bayesian analysis shown in Fig. 1 and 2 is given in table 3.

Estimation of a chromosome network. By using consensus information from the two Bayesian analyses from the previous section, we estimate now a chromosome network (CNet). Specifically, we define the CNet in the following way. Nodes correspond to chromosomes and two chromosomes C_i and C_k are connected by an undirected link when their posterior probability for a common disease D_j is larger than a factor of F compared to either the non-informative prior or the complementary posterior probability, i.e.,

$$FOP_{ij}(FB) > F \text{ AND } LOD_{kj}(FB) > \log(F). \quad (16)$$

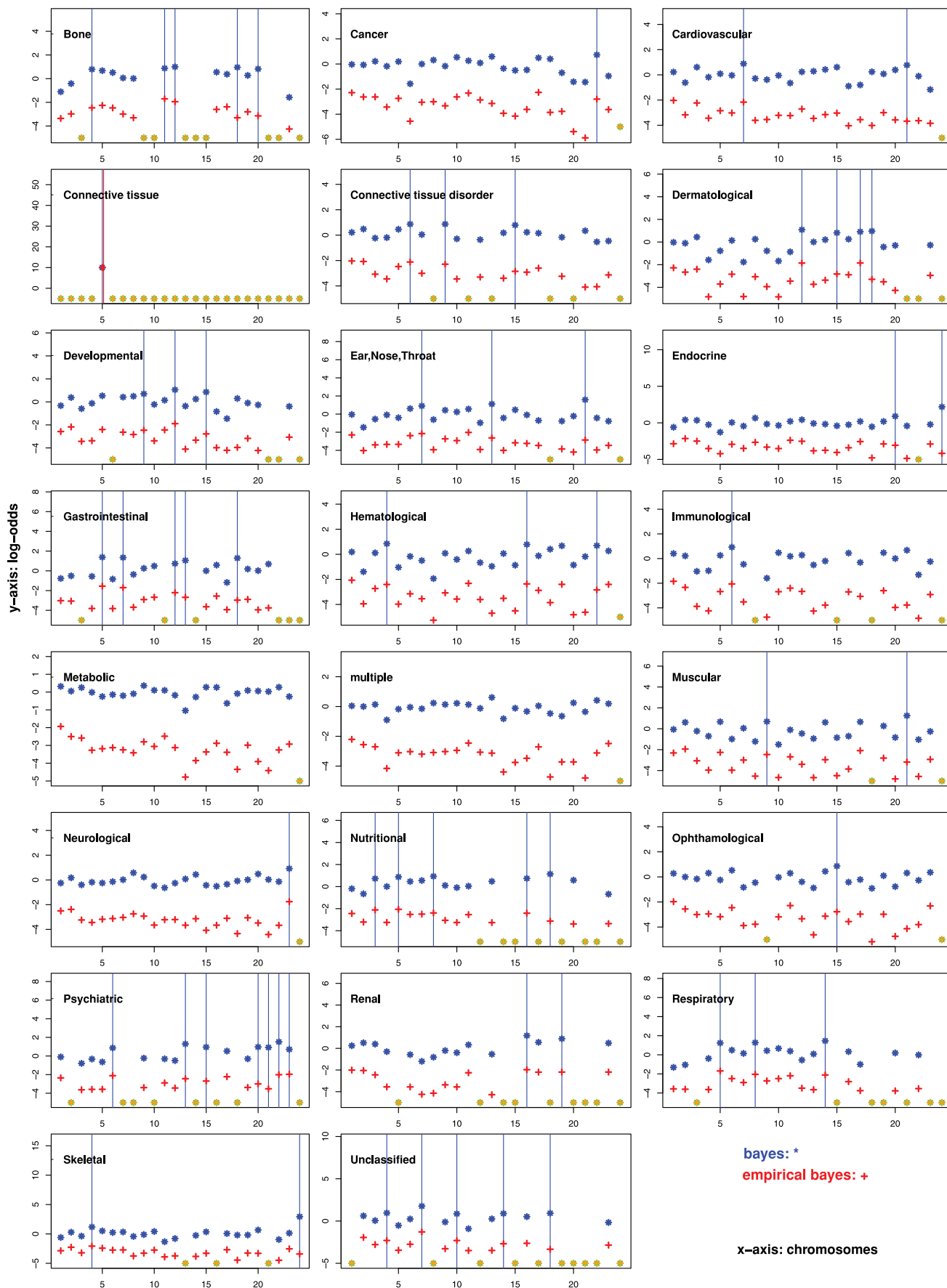


Figure 2 | Shown are the log-odds for the full Bayes and empirical Bayes analysis.



Table 3 | Summary of our full Bayesian analysis shown in Fig. 1 and 2. Column two gives the chromosomes for which $FOP_{ij}(FB) > F$ holds and column three shows results for $LOD_{ij}(FB) > \log(F)$

disorder category	Chr	Chr
Bone	4,11,12,18,20	4,11,12,18,20
Cancer	22	22
Cardiovascular	7,21	7,21
Connective tissue	5	5
Connective tissue disorder	6,9,15	6,9,15
Dermatological	12,15,17,18	12,15,17,18
Developmental	12,15	9,12,15
Ear,Nose,Throat	7,13,21	7,13,21
Endocrine	20,Y	20,Y
Gastrointestinal	5,7,13,18	5,7,12,13,18
Hematological	4,16,22	4,16,22
Immunological	6,21	6
Metabolic	-	-
multiple	-	-
Muscular	9,21	9,21
Neurological	X	X
Nutritional	5,8,16,18	3,5,8,16,18
Ophthalmological	15	15
Psychiatric	6,13,15,20,21,22	6,13,15,20,21,22,X
Renal	2,16,17,19,X	16,19
Respiratory	5,8,14	5,8,14
Skeletal	Y	4,Y
Unclassified	4,7,10,14,18	4,7,10,14,18

That means the chromosome network is constructed from consensus information of the two Bayesian analyses from the previous section.

In Fig. 3 we show a visualization of the estimated CNet obtained by application of the NetBioV package. In this figure, there are two link colors. The green links correspond to the *minimum spanning tree* of the network that connects all connectable chromosomes with each other using a minimal number of links³⁶. The orange links correspond to all remaining connections. Overall, 19 of the 22 autosome chromosomes are connected with each other, forming the *giant connected component*³⁷ of the chromosome network. Only the chromosomes 1, 2, 3 and the two sex chromosomes are unconnected. The color of the nodes corresponds to three different chromosome categories, explained below.

From a structural analysis of the chromosome network, we find that the chromosome with the largest degree is chromosome 18 which is connected to 12 other chromosomes. Interestingly, the chromosomes with the largest degree within the *minimum spanning tree* are chromosome 4, 15, 20 (degree of 9) which all have a direct link to chromosome 18. In the chromosome network, a high degree of a chromosome indicates that this chromosome is involved in similar disorders as the chromosomes it is connected with. Hence, it indicates a *co-involvement* in shared disorders. In order to assess the ‘importance’ of the chromosomes within the CNet with respect to their involvement in many different disorders, we calculate the (vertex) *betweenness centrality* (bc) index, which is a frequently used measure to assess the role of nodes within networks^{38–40}. The betweenness centrality index evaluates the fraction of shortest paths that pass through a node in a network, connecting all other chro-

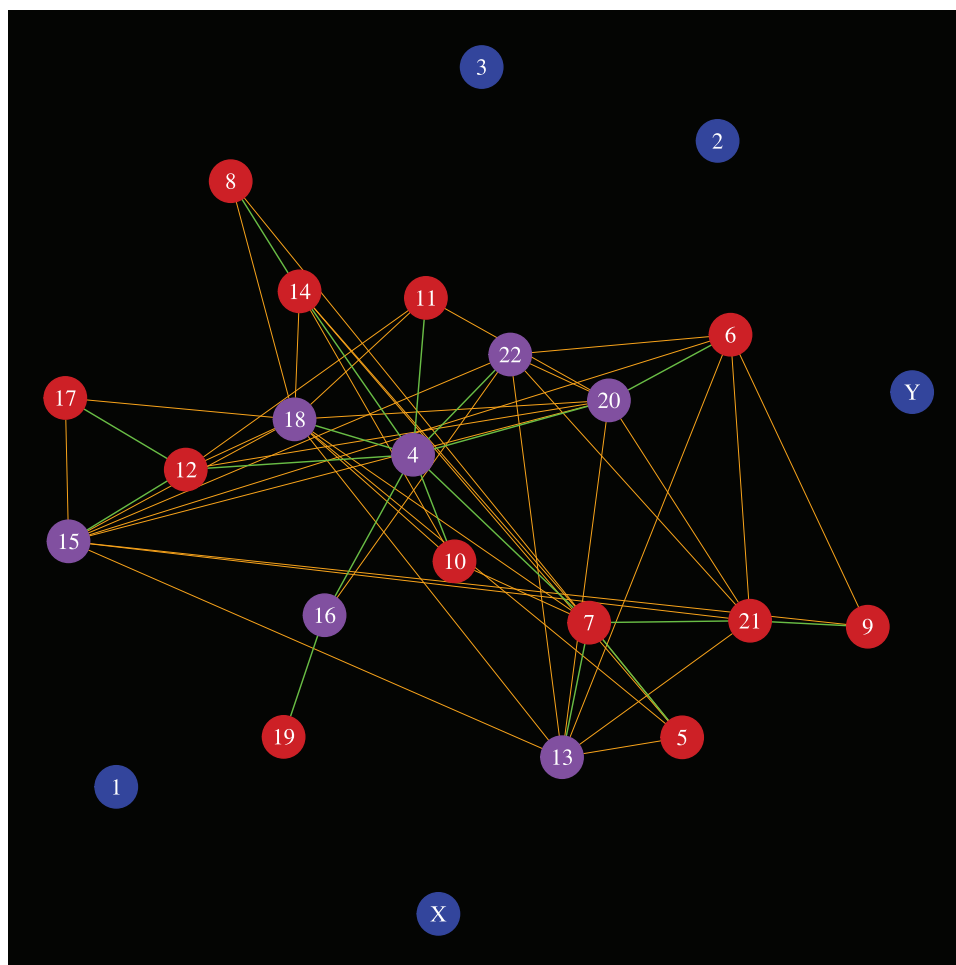


Figure 3 | The human chromosome network (CNet) where nodes correspond to chromosomes and two chromosomes C_i and C_k are connected if the joint consensus condition in Eqn. 16 is fulfilled.



Table 4 | Summary statistics of the chromosome network shown in Fig. 3. Listed are values of the degree, betweenness centrality (bc) and the FDR adjusted p-values of the betweenness centrality values for each of the chromosomes

Chr	1	2	3	4	5	6	7	8	9
degree	0	0	0	9	5	6	7	3	3
bc-value	0.0	0.0	0.0	26.4	1.9	2.3	7.6	0.0	0.0
p-value	1.0	1.0	1.0	$<10^{-5}$	1.0	1.0	0.13	1.0	1.0
Chr	10	11	12	13	14	15	16	17	18
degree	4	4	6	8	6	9	3	3	12
bc-value	0.0	0.0	2.9	8.6	3.8	15.9	17.0	0.0	32.9
p-value	1.0	1.0	1.0	0.005	1.0	$<10^{-5}$	$<10^{-5}$	1.0	$<10^{-5}$
Chr	19	20	21	22	X	Y			
degree	1	9	7	7	0	0			
bc-value	0.0	9.2	6.2	14.1	0.0	0.0			
p-value	1.0	0.0009	1.0	$<10^{-5}$	1.0	1.0			

mosome pairs with each other. A summary of our results is given in table 4. We find that the chromosome 18 ($bc = 32.9$) has the largest betweenness centrality index followed by chromosome 4 ($bc = 26.4$) and chromosome 16 (17.0). At first, the appearance of chromosome 16 among the top three nodes with a high betweenness centrality index may surprise since its degree is only 3. However, chromosome 16 is the bottleneck to connect chromosome 19 with the rest of the giant connected component and, hence, needs to be on all shortest paths to and from chromosome 19. For all other chromosomes, there are always alternative paths connecting two chromosomes with each other and, hence, the number of shortest paths is reduced leading to lower betweenness centrality values. From the analysis of the degrees and the betweenness centrality, we conclude that chromosome 4, 16 and 18 assume a *central role* within the chromosome network. This implies that these chromosomes are least disorder specific, but co-appear in many different disorders.

In order to categorize all chromosomes statistically with respect to their importance, we randomize the chromosome network shown in Fig. 3, $B = 10^5$ times by conserving the number of edges, and estimate from these networks the null distribution of mean betweenness centrality values. From this, we obtain the betweenness centrality p-values listed in table 4. We want to remark that these values have been adjusted by the Benjamini & Hochberg procedure because we are testing multiple hypotheses. The meaning of an adjusted p-value, p_i , that belongs to an observed betweenness centrality value, bc_i , is that the probability to observe bc-values in the randomized networks that are larger than bc_i is given by p_i , i.e., $p_i = \text{Prob}(bc > bc_i | bc \text{ observed in randomized networks})$.

The above analysis allows a categorization of the chromosomes in three mutually exclusive subgroups. The first category consists of chromosomes that have significant p-values for $FDR = 0.001$. The chromosomes 4, 13, 15, 16, 18, 20, 22 are in this category. In Fig. 3, we highlight these chromosomes in purple. The second category consists of chromosomes with the lowest betweenness centrality index of zero, namely, 1, 2, 3, X, Y. These correspond to the unconnected chromosomes, shown in blue, in Fig. 3. The interpretation of their isolated role is that these chromosomes correspond to disorder specific chromosomes. That means they are only involved in specific, individual disorders rather than in many different ones. Finally, chromosomes with none vanishing betweenness centrality values but none significant p-values are selectively informative for a *small number* of particular disorders. The 12 chromosomes that fall within this third category are 5, 6, 7, 8, 9, 10, 11, 12, 14, 17, 19, 21. These chromosomes are in Fig. 3 shown in red.

As an indicator that the obtained chromosome categories are not biased by, e.g., the dimension of the chromosomes, we estimate the enrichment of 'large' and 'small' chromosomes, as measured by the number of protein coding genes, see table 7, within the three categories. We find that for $FDR = 0.001$, none of the three categories is

enriched by large or small chromosomes, according to a hypergeometric test. For this analysis, we defined as 'large' and 'small' chromosomes the top/bottom 10%–30% (including intermediate sizes) of all chromosomes, without noticing any difference for any conducted hypothesis test. This result is reassuring that the estimated chromosome network and the derived three chromosome categories do not reflect simple chromosome properties that would allow to obtain such a categorization in a straight forward manner.

In our opinion the last category of chromosomes is from a practical point of view the most interesting one. The reason for this is the selective nature of these chromosomes which are only involved in a very limited number of different disorders. This enables a guided search and a potential information transfer from the knowledge available about one disease to use for another.

We would like to emphasize that the chromosome network provides information that is not directly contained in a, e.g., protein interaction network. In order to demonstrate this, we use the human protein interaction network available from the BioGrid database⁴¹. This network consists of 8429 proteins. Using this PPI network we construct another chromosome network in the following way. We start with an unconnected $C \times C$ matrix M , with $C = 24$, and increase for every protein interaction between protein A , which is on chromosome i , and protein B , on chromosome j , the edge weight between chromosome i and chromosome j by one. That means, the elements of the resulting matrix M_{ij} provide the number of protein interactions which occur between chromosome i and j , according to the human PPI network. Our analysis gives the following results. First, the obtained network is fully connected among the autosome chromosomes. Only the Y chromosome is sparsely connected to other chromosomes. Second, constructing a minimum spanning tree from the count information in the matrix M , we find that the resulting tree is actually a star graph centered around chromosome 1. This is plausible because chromosome 1 contains by far the largest number of genes. Third, by randomizing the assignment of proteins to chromosomes but keeping the topology of the PPI network fixed, we obtain cut-off values to declare count numbers between pairs of chromosomes significant. Application of these cut-off values results in a new network which has a minimum spanning tree with highly connected chromosomes 1 and 3. Both results indicate that the obtained chromosome networks are biased by the number of genes on the chromosomes, which is plausible. However, for studying the connection among diverse complex disorders this information is not immediately exploitable. Repeating a similar analysis, however, limited to the disease genes present in the PPI network, gives qualitatively very similar results.

In order to identify relevant biological process, cellular components or molecular functions, we perform a gene ontology (GO)⁴² analysis of the three chromosome categories obtained from our CNet. More precisely, we compile a list of all disease genes that are

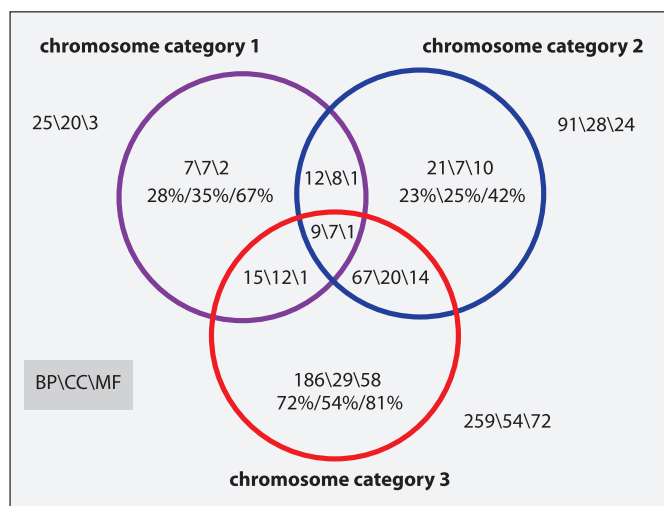


Figure 4 | Summary of the gene ontology analysis for the three categories ‘biological process’ (BP), ‘cellular component’ (CC) and ‘molecular function’ (MF). The color of the chromosomal subgroups corresponds to the color of the chromosomal subgroups in Fig. 3.

part of the three chromosome categories and test the enrichment of these gene lists for gene ontology terms from the categories: ‘biological process’ (BP), ‘cellular component’ (CC) and ‘molecular function’ (MF). A summary of our results is shown in Fig. 4. There, the color of the chromosome categories corresponds to the color of the chromosome categories in Fig. 3. The triple numbers correspond to the number of significantly enriched GO terms from the category BP/CC/MF (in that order). To adjust for multiple testing, we use a Bonferroni correction to control the family-wise error rate at 5%⁴³. The numbers outside the Venn diagram correspond to the total number of enriched GO terms for the respective chromosome categories. For example, for the three chromosome categories we find 25/91/259 enriched GO terms in the category ‘biological process’, or the number of none overlapping (unique), enriched GO terms for the three chromosome categories for BP are 7/21/186. Finally, for the number of none overlapping GO terms, we included also the percentage compared to the total number of enriched GO terms.

That means, for example, for chromosome Category 3, we find that the 186 unique GO terms in the category BP correspond to 72% (186/259) of all enriched GO categories for this chromosome category.

Our results in Fig. 4 reveal that the uniqueness of the three chromosome categories is in general quite high, ranging from about 25% to over 80%. This implies that these chromosome categories represent specific biological processes, cellular components and molecular functions. Interestingly, the highest uniqueness is obtained for chromosome Category 3. This confirms our discussion provided above, arguing that this category is the most rich from a biological perspective. In the tables 5–6, we show the top enriched GO terms for the three chromosome categories.

From a systems perspective, different chromosomal categories reflect, to some extent, biological processes operating at different levels of cellular organization (see table 5 and 6): cell-cell interactions (e.g. blood vessel remodeling, renal system process, responsible for fluid volume regulation and detoxification in an organism: Category 1), cellular processes (e.g. cell cycle and differentiation, response to substances: Category 2) and cellular metabolism (e.g. nucleotide biosynthetic processes, ATP metabolic process: Category 3). The degree of co-involvement of chromosomes in various diseases in Categories 1, 2, 3 is echoed by the level of cellular organization of biological processes, enriched in these categories. In other words, different diseases are enriched in biological processes operating at the highest levels of cellular hierarchy (cell-cell interactions), as reflected in Category 1. Unconnected chromosomes (Category 2) are disease-specific, and enriched biological processes operate at the level of a single cell. A small number of particular disorders share several biosynthetic and metabolic processes, operating at a single cell level, too (Category 3).

This information is what one would generally expect biomedically. However, it is interesting to see that this information could be deduced from our chromosomal network. This hints that by using more genes and an extended set of diseases categories one might reveal further information from the chromosomal network that could give new and important connections between different diseases and biological processes tied with their development.

Finally, we conducted a GO analysis for testing the enrichment of disorder categories within the three chromosome subgroups. From this, we find hematological disorders significantly enriched in chromosome Category 1, neurological disorders in chromosome

Table 5 | Statistically enriched GO categories for chromosome Category 1

GO.ID/BP	GO term	# of genes	p-val
GO:0051925	regulation of calcium ion transport via ...	8	1.9e-07
GO:0003014	renal system process	8	3.4e-07
GO:0009056	catabolic process	52	1.5e-06
GO:0001974	blood vessel remodeling	6	2.4e-06
GO:0050880	regulation of blood vessel size	10	2.4e-06
GO:0035150	regulation of tube size	10	2.7e-06
GO:0032412	regulation of ion transmembrane transpor...	8	4.5e-06
GO.ID/CC	GO term	# of genes	p-val
GO:0045202	synapse	20	1.6e-06
GO:0005891	voltage-gated calcium channel complex	5	8.6e-06
GO:0030424	axon	13	8.6e-06
GO:0005913	cell-cell adherens junction	6	1.2e-05
GO:0044433	cytoplasmic vesicle part	16	1.3e-05
GO:0016529	sarcoplasmic reticulum	6	2.4e-05
GO:0016528	sarcoplasm	6	3.1e-05
GO.ID/MF	GO term	# of genes	p-val
GO:0016836	hydro-lyase activity	7	1.5e-06
GO:0016835	carbon-oxygen lyase activity	8	1.5e-06



Table 6 | Statistically enriched GO categories for chromosome Category 2

GO.ID/BP	GO term	# of genes	p-val
GO:0010035	response to inorganic substance	32	1.1e-12
GO:2000026	regulation of multicellular organismal d...	42	3.5e-07
GO:0051094	positive regulation of developmental pro...	31	3.6e-07
GO:0071241	cellular response to inorganic substance	10	3.9e-07
GO:0050793	regulation of developmental process	49	4.2e-07
GO:0000080	G1 phase of mitotic cell cycle	8	5.4e-07
GO:0070482	response to oxygen levels	18	9.5e-07
GO:0045597	positive regulation of cell differentiat...	25	9.5e-07
GO:0051318	G1 phase	8	1.1e-06
GO:0001666	response to hypoxia	17	1.3e-06
GO.ID/CC	GO term	# of genes	p-val
GO:0031967	organelle envelope	38	3.6e-08
GO:0031975	envelope	38	5.7e-08
GO:0000323	lytic vacuole	20	4.7e-07
GO:0005764	lysosome	20	4.7e-07
GO:0005773	vacuole	22	5.5e-07
GO:0031090	organelle membrane	71	2.3e-06
GO:0042383	sarcolemma	10	2.8e-06
GO.ID/MF	GO term	# of genes	p-val
GO:0016705	oxidoreductase activity, acting on paire...	17	3.1e-08
GO:0008395	steroid hydroxylase activity	7	4.3e-08
GO:0042803	protein homodimerization activity	30	6.3e-08
GO:0004935	adrenergic receptor activity	5	4.6e-07
GO:0046982	protein heterodimerization activity	19	1.9e-06
GO:0004936	alpha-adrenergic receptor activity	4	2.7e-06
GO:0051400	BH domain binding	4	6.2e-06
GO:0070330	aromatase activity	6	7.8e-06
GO:0004937	alpha 1-adrenergic receptor activity	3	9.1e-06
GO:0016903	oxidoreductase activity, acting on the a...	7	9.8e-06

Category 2 and connective tissue and respiratory disorders in chromosome Category 3.

We would like to emphasize that the reason for conducting the GO analysis in this section only for the disease genes and not for all genes is that the latter would us not allow to establish a sensible connection

Table 7 | The number of protein-coding genes (p-genes) at the chromosomes according to the NCBI (accessed May 2012). Further, known disease genes (d-genes) on the chromosomes are listed

Chr	1	2	3	4	5	6	7	8	9
p-genes	2062	1266	1092	768	905	1056	939	704	808
d-genes	164	124	95	64	87	83	78	60	70
Chr	10	11	12	13	14	15	16	17	18
p-genes	768	1296	1041	338	616	613	847	1207	276
d-genes	70	121	86	40	47	44	71	103	24
Chr	19	20	21	22	X	Y			
p-genes	1421	554	226	449	833	185	total	=	20270
d-genes	76	32	20	49	111	3	total	=	1722

to the inferred chromosome associations, as represented by the structure of the estimated chromosome network, because these were also obtained from disease genes and their association with disorders, as provided by the OMIM database.

Disorder category network. Finally, we use the result from the Bayesian analysis in Fig. 2 to construct a network for the disorder categories. That means, in this network nodes correspond to disease categories, as listed in table 2 and two categories D_i and D_j are connected by an undirected link if there exists at least one chromosome C_k with a log-odds larger than $\log(F)$, i.e.,

$$\text{LOD}_{ki}(FB) > \log(F) \quad \text{AND} \quad \text{LOD}_{kj}(FB) > \log(F). \quad (17)$$

The resulting network is shown in Fig. 5.

This figure provides a visualization of the interpretation given above regarding the co-involvement of chromosomes in different disorders. For example, the connection of different cancer types and hematological disorders is well established for instance in the form of lymphocytic leukemia, e.g., *acute lymphoblastic leukemia* (ALL) or *chronic lymphocytic leukemia* (CLL)^{44,45}. Also, the connection between cancer and psychiatric disorders has been studied since decades and it is known that the prevalence of psychiatric disorders, e.g., depression or anxiety, among cancer patients increases with the severity of the patient's condition^{46–49}. Hence, the disorder category network represents the disease susceptibility among different disorders with respect to a common underlying genetic basis.

The category 'cancer' is also a good example of the above discussion about the selective informativeness of chromosomes with respect to different disorders. The implication of this can be seen in Fig. 5, because cancer is only connected to two further disorders. This allows a guided search focusing on the limited number of disorders associated with cancer instead of searching the whole medical literature randomly, which is inefficient, time consuming and costly. That means cancer, connective tissue or immunological disorders are associated with chromosomes of the third category, shown in red in Fig. 3. In contrast, the categories multiple and metabolic disorders form isolated nodes in Fig. 5 and, hence, correspond to chromosome category one with unspecific associations to other chromosomes. This can also be seen from Fig. 1 and 2 because these two disorder categories are the only disorders not leading to an enlarged posterior probability for any chromosome that would pass our threshold criterion. Lastly, bone, psychiatric, or unclassified disorders are highly connected in the disorder network in Fig. 5. This implies the involvement of many different chromosomes, which can be also seen from Fig. 1 and 2. In addition to these three disorders, there are also others with a larger number of connections to other disorder categories but not necessarily with enlarged posterior probabilities (compare Fig. 5 with Fig. 1 and 2) such as developmental, hematological or muscular disorders. These connections are the result of the Bayesian analysis we performed and the accompanied integrative inference of information from all disorders. This provides another example of the systems character of our analysis despite the fact that the data we used are of reductionist origin.

As examples for predicting common genetic causes for disorders in⁵⁰ a connection between bipolar disease (BD) and hypertension (HT), and between bipolar disease and type I diabetes (T1D) has been predicted based on GWAS data. From our disorder category network in Fig. 5, we find direct connections between psychiatric (BD) and cardiovascular (HT) disorders and endocrine (T1D) and psychiatric (BD) disorders. This provides independent support for this study, because not only the data we use, but also our methodology is different.

Discussion

In recent years the fields Network Medicine and Systems Biomedicine emerged to approach biomedical problems from a systems

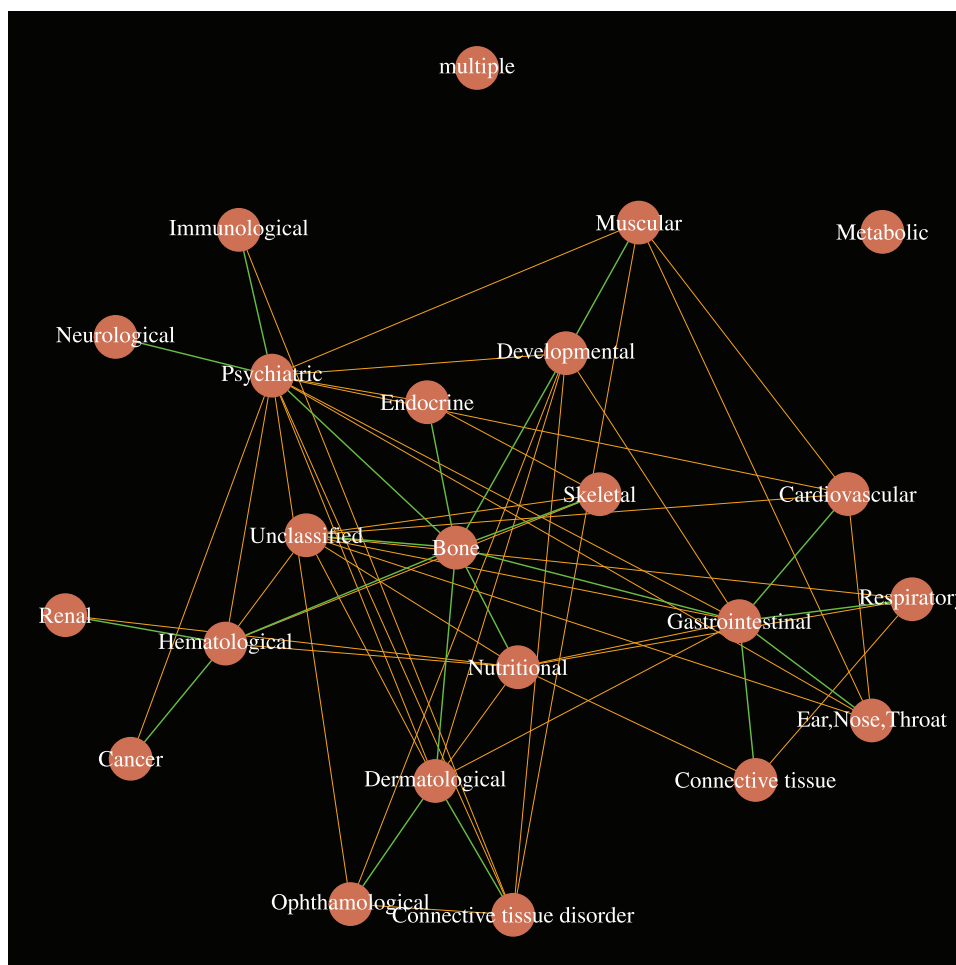


Figure 5 | The human disorder category network estimated from our Bayesian analysis.

perspective^{4,51–53}. However, it proved extremely challenging to pursue these principle ideas practically. Our analysis constitutes an example of such a practical realization.

In contrast to many previous approaches aiming to connect genetic with phenotype information in order to obtain either disorder or disease gene networks, our method for estimating such networks is based on consensus information from two Bayesian analyses instead of a deterministic method. This allows the statistical quantification of the estimated uncertainty in the underlying disorder-disease gene associations contained in the OMIM database resulting in parametric networks. The level of an acceptable uncertainty can be set by the user, similarly to the significance level in a hypotheses testing framework to control the Type I error rate of a test. In our study, we demonstrated that a sensible numerical value for such a parameter and a justification for the used priors can be obtained from the consensus information of the two Bayesian analyses. In principle, our chosen criterion could be loosened if a more exploratory estimate is desired.

In addition to this methodological difference compared to previous studies, we were not aiming to estimate a disease gene network from the OMIM database as, e.g.,^{13,15,18}, but we were interested in a higher organizational level in form of the chromosome network. This was motivated by two reasons. First, to our knowledge the systematic association among chromosomes and their co-involvement in different disorders has not been studied so far on a large scale. For this reason, our results may help in fostering a general interest due to the possibility of a practical information transfer between seemingly different disorders. This might be especially useful considering the

fact that our analysis is purely computational not requiring the direct involvement of patients. Second, if such a chromosome network should be estimated from data, instead of deterministically constructed, sufficient data are needed. This implies that the dimensionality of the problem should be balanced with the available samples. In our case this means that the number of nodes in a network, which correspond to variables, should not be too large. Specifically, when using disease genes as nodes in a network, the number of variables is 1722, because that is the number of disease genes available from OMIM. However, when using the chromosomes, we have only 22 autosomes and the two sex chromosomes as variables. Here, it is important that the amount of data we have available for our analysis is in both cases exactly the same, because our analysis is based on the OMIM database. Statistically, it is clear that the former estimation problem is more complex than the latter one. In fact, we performed also a numerical analysis for the disease genes, but the available data do not allow to perform a robust statistical analysis on this level. Similar arguments hold for disorders and disorder categories.

In summary, the major purpose of our study was the investigation of the chromosome architecture of human disorders. In contrast, for instance, to studies investigating the spatial architecture of chromosomes inside the cell nucleus⁵⁴, our focus has been on the conceptual organization and partitioning of chromosomes with respect to their involvement in human disorders. This level of abstraction implies that our obtained chromosomal subgroups cannot be experimentally observed, e.g. by microscopy. Instead, our results constitute a predictive model that can be utilized by generating novel hypotheses



about the susceptibility of either disorders or disease genes in a target pathology. Due to limitations of the data, for theoretical reasons, we had to use broad disorder categories, instead of disease specific terms as provided, e.g., by the International Classification of Diseases (ICD)⁵⁵. Hence, a future extension of our model, extending it down to the level of individual disorders, could be based on an enlarged database, possibly established from a variety of different sources in order to obtain a finer coverage of all disorders.

Methods

For our analysis we use the *Online Mendelian Inheritance in Man* (OMIM) database <http://www.ncbi.nlm.nih.gov/omim>. The OMIM contains information about known associations between disease genes and disorders. Despite the fact that this database is far from being complete, it represents currently a gold standard. Starting in the 1960s as a repository to collect information about monogenic disorders, within recent years OMIM includes more and more information about complex disorders. For our analysis, we used information about 1,284 disorders and 1,777 disease genes. Further, we utilize a classification of the 1,284 disorders into 23 broad disease categories, as compiled by¹³.

It is important to note that despite the fact that the OMIM database contains information about hundreds of different disorders and disease genes, all this information is collected from separate experiments. That means this information has been generated by experiments focusing on individual disorders and disease genes, frequently conducted in a reductionist manner.

An overview of the distribution of disease genes (d-genes) on the chromosome is given in table 7. Further, this table shows the distribution of all protein coding genes over the chromosomes. The protein coding genes have been obtained from NCBI via ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz.

- Auffray, C., Chen, Z. & Hood, L. Systems medicine: the future of medical genomics and healthcare. *Genome Med* **1**(1), 2 (2009).
- Barabási, A.-L. Network Medicine – From Obesity to the “Diseaseome” *N Engl J Med* **357**(4), 404–407 (2007).
- Liu, E. & Lauffenburger, D., editors. *Systems Biomedicine*. Elsevier/Academic Press, Boston, (2010).
- Zanzoni, A., Soler-Lopez, M. & Aloy, P. A network medicine approach to human disease. *FEBS Letters* **583**(11), 1759–1765 (2009).
- Beadle, G. W. & Tatum, E. L. Genetic Control of Biochemical Reactions in Neurospora. *Proceedings of the National Academy of Sciences of the United States of America* **27**(11), 499–506 (1941).
- Dehmer, M., Emmert-Streib, F., Graber, A. & Salvador, A., editors. *Applied Statistics for Network Biology: Methods for Systems Biology*. Wiley-Blackwell, Weinheim, (2011).
- Sales-Pardo, M., Guimera, R., Moreira, A. & Amaral, L. Extracting the hierarchical organization of complex systems. *Proceedings of the National Academy of Sciences* **104**(39), 15224–15229 (2007).
- Vidal, M. A unifying view of 21st century systems biology. *FEBS Letters* **583**(24), 3891–3894 (2009).
- Barabási, A. L. & Oltvai, Z. N. Network biology: Understanding the cell’s functional organization. *Nature Reviews* **5**, 101–113 (2004).
- Emmert-Streib, F. & Glazko, G. Network Biology: A direct approach to study biological function. *Wiley Interdiscip Rev Syst Biol Med* **3**(4), 379–391 (2011).
- Emmert-Streib, F. & Dehmer, M. Networks for Systems Biology: Conceptual Connection of Data and Function. *IET Systems Biology* **5**(3), 185 (2011).
- Van Driel, M. A., Bruggeman, J., Vriend, G., Brunner, H. G. & Leunissen, J. A. M. A text-mining analysis of the human phenome. *European journal of human genetics EJHG* **14**(5), 535–542 (2006).
- Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M. & Barabasi, A. The human disease network. *Proceedings of the National Academy of Sciences* **104**(21), 8685–8690 (2007).
- Jiang, X. *et al.* Modularity in the genetic disease-phenotype network. *FEBS Letters* **582**(17), 2549–2554 (2008).
- Xie, M., Hwang, T. & Kuang, R. Reconstructing disease phenome-genome association by bi-random walk. *Bioinformatics* **2012** (2012).
- Lee, D.-S., Park, J., Kay, K. A., Christakis, N. A., Oltvai, Z. N. & Barabasi, A.-L. The implications of human metabolic network topology for disease comorbidity. *Proceedings of the National Academy of Sciences* **105**(29), 9880–9885 (2008).
- Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology* **25**(3), 309–316 (2007).
- Yang, P., Li, X., Wu, M., Kwok, C.-K. & Ng, S.-K. Inferring Gene-Phenotype Associations via Global Protein Complex Network Propagation. *PLoS ONE* **6**(7), 07 (2011).
- Vanunu, O., Magger, O., Ruppim, E., Shlomi, T. & Sharan, R. Associating Genes and Protein Complexes with Disease via Network Propagation. *PLoS Comput Biol* **6**(1), e1000641, 01 (2010).
- Wu, X., Jiang, R., Zhang, M. Q. & Li, S. Network-based global inference of human disease genes. *Molecular Systems Biology* **4**(189), 189 (2008).
- Zhang, M., Zhu, C., Jacomy, A., Lu, L. J. & Jegga, A. G. The orphan disease networks. *The American Journal of Human Genetics* **88**(6), 755–766 (2011).
- 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* **4**, e8090, 11 (2009).
- Lee, I., Blom, U. M., Wang, P. I., Shim, J. E. & Marcotte, E. M. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research* **21**(7), 1109–1121 (2011).
- Zhang, X. *et al.* The expanded human disease network combining protein-protein interaction information. *European journal of human genetics EJHG* **19**(7), 783–788 (2011).
- Emmert-Streib, F., Tripathi, S., de Matos Simoes, R., Hawwa, A. F. & Dehmer, M. The human disease network: Opportunities for classification, diagnosis and prediction of disorders and disease genes. **1** (2012, submitted.).
- Hayasaka, S., Hugenschmidt, C. E. & Laurienti, P. J. A Network of Genes, Genetic Disorders, and Brain Areas. *PLoS ONE* **6**(6), e20907, 06 (2011).
- Casella, G. & Berger, R. *Statistical Inference*. Duxbury Press, (2002).
- Mood, A., Graybill, F. & Boes, D. *Introduction to the Theory of Statistics*. McGraw-Hill, (1974).
- Sun, J. *et al.* DNA copy number alterations in prostate cancers: A combined analysis of published CGH studies. *The Prostate* **67**(7), 692–700 (2007).
- Staub, E. *et al.* A genome-wide map of aberrantly expressed chromosomal islands in colorectal cancer. *Molecular Cancer* **5**(1), 37 (2006).
- Manolio, T. A. Genomewide Association Studies and Assessment of the Risk of Disease. *New England Journal of Medicine* **363**(2), 166–176 (2010).
- Emmert-Streib, F. The chronic fatigue syndrome: A comparative pathway analysis. *Journal of Computational Biology* **14**(7), 961–972 (2007).
- Sheskin, D. J. *Handbook of Parametric and Nonparametric Statistical Procedures*. RC Press, Boca Raton, FL, 3rd edition, (2004).
- Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)* **57**, 125–133 (1995).
- Carlin, B. & Louis, T. *Bayesian Methods for Data Analysis*. CRC Press, (2009).
- Cormen, T., Leiserson, C., Rivest, R. & Stein, C. *Introduction to Algorithms*. MIT Press, (2001).
- Dorogovtsev, S. & Mendes, J. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, (2003).
- Freeman, L. C. A set of measures of centrality based on betweenness. *Sociometry* **40** (1977).
- Freeman, L. C. Centrality in social networks: Conceptual clarification. *Social Networks* **1**, 215–239 (1979).
- Newman, M. *Networks: An Introduction*. Oxford University Press, Oxford, (2010).
- Breitkreutz, B.-J. *et al.* The BioGRID Interaction Database: 2008 update. *Nucl. Acids Res.* **36**(suppl 1), D637–640 (2008).
- Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **25**(1), 25–29, May (2000).
- Dudoit, S. & van der Laan, M. *Multiple Testing Procedures with Applications to Genomics*. Springer, New York; London, (2007).
- Chiorazzi, N., Rai, K. R. & Ferrarini, M. Chronic lymphocytic leukemia. *New England Journal of Medicine* **352**(8), 804–815 (2005).
- Pui, C.-H., Relling, M. V. & Downing, J. R. Acute lymphoblastic leukemia. *New England Journal of Medicine* **350**(15), 1535–1548 (2004).
- Akechi, T. *et al.* Psychiatric disorders in cancer patients: Descriptive analysis of 1721 psychiatric referrals at two Japanese cancer center hospitals. *Japanese Journal of Clinical Oncology* **31**(5), 188–194 (2001).
- Derogatis, L. R. *et al.* The prevalence of psychiatric disorders among cancer patients. *Jama The Journal Of The American Medical Association* **249**(6), 751–757 (1983).
- Levine, P. M., Silberfarb, P. M. & Lipowski, Z. J. Mental disorders in cancer patients. a study of 100 psychiatric referrals. *Cancer* **42**(3), 1385–1391 (1978).
- Miovic, M. & Block, S. Psychiatric disorders in advanced cancer. *Cancer* **110**(8), 1665–1676 (2007).
- Schaub, M. A., Kaplow, I. M., Sirota, M., Do, C. B., Butte, A. J. & Batzoglou, S. A classifier-based approach to identify genetic similarities between diseases. *Bioinformatics* **25**(12), i21–i29 (2009).
- von Bertalanffy, L. The theory of open systems in physics and biology. *Science* **111**, 23–29 (1950).
- Liu, E. T. Integrative biology—a strategy for systems biomedicine. *Nature Reviews Genetics* **10**(1), 64–68 (2009).
- Waddington, C. *The strategy of the genes*. Geo, Allen & Unwin, London, (1957).
- Van Steensel, B. & Dekker, J. Genomics tools for unraveling chromosome architecture. *Nature Biotechnology* **28**(10), 1089–1095 (2010).
- Organization, World Health. *International Classification of Diseases (ICD)* (2010).

Acknowledgments

Matthias Dehmer thanks the Standortagentur Tirol for financial support.



Author contributions

FES conceived the study. FES, RDMS and GVG performed the analysis. FES, RDMS, ST, GVG and MD interpreted the results and wrote the manuscript. ST prepared figures 3 and 5.

Additional information

Competing financial interests: The authors declare no competing financial interests.

License: This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/3.0/>

How to cite this article: Emmert-Streib, F., de Matos Simoes, R., Tripathi, S., Glazko, G.V. & Dehmer, M. A Bayesian analysis of the chromosome architecture of human disorders by integrating reductionist data. *Sci. Rep.* **2**, 513; DOI:10.1038/srep00513 (2012).