BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 21 2014, pages 3093–3100
doi:10.1093/bioinformatics/btu487

Systems biology

Advance Access publication July 23, 2014

Systematic analysis of gene properties influencing organ system phenotypes in mammalian perturbations

Ingo Vogt[1,2,†], Jeanette Prinz[1,2,†], Karolina Worf[3] and Monica Campillos[1,2,*]

[1]German Center for Diabetes Research (DZD), Ingolstädter Landstr. 1, 85764 Neuherberg, Germany, [2]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum Muenchen, 85764 Neuherberg, Germany and [3]Technical University Munich WZW Chair of Bioinformatics, 80333 Munich, Germany

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: Diseases and adverse drug reactions are frequently caused by disruptions in gene functionality. Gaining insight into the global system properties governing the relationships between genotype and phenotype is thus crucial to understand and interfere with perturbations in complex organisms such as diseases states.

Results: We present a systematic analysis of phenotypic information of 5047 perturbations of single genes in mice, 4766 human diseases and 1666 drugs that examines the relationships between different gene properties and the phenotypic impact at the organ system level in mammalian organisms. We observe that while single gene perturbations and alterations of nonessential, tissue-specific genes or those with low betweenness centrality in protein–protein interaction networks often show organ-specific effects, multiple gene alterations resulting e.g. from complex disorders and drug treatments have a more widespread impact. Interestingly, certain cellular localizations are distinctly associated to systemic effects in monogenic disease genes and mouse gene perturbations, such as the lumen of intracellular organelles and transcription factor complexes, respectively. In summary, we show that the broadness of the phenotypic effect is clearly related to certain gene properties and is an indicator of the severity of perturbations. This work contributes to the understanding of gene properties influencing the systemic effects of diseases and drugs.

Contact: monica.campillos@helmholtz-muenchen.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 10, 2014; revised on June 18, 2014; accepted on July 11, 2014

1 INTRODUCTION

Phenotypic traits are observable characteristic features of organisms reflecting the architecture of their biological systems. Deciphering genetic factors as well as system properties associated to phenotypic traits is still a central problem in biology, and enormous efforts have been devoted to solve it. This knowledge is fundamental for the understanding and interference with a disturbed biological system state such as a disease.

The recent explosion of 'omics' data has accelerated the discovery of novel relationships between genotypes and complex phenotypes such as multifactorial disorders and drug-induced perturbations where multiple and often unknown genes and environmental factors are involved. For example, genome-wide association studies in human populations are expanding the repertoire of genes linked to diseases (Visscher *et al*., 2012), and the systematic analyses of adverse effects of drugs and their molecular targets are elucidating novel mechanisms of drug action (Campillos, *et al*., 2008; Kuhn, *et al*., 2013).

Systematic single gene perturbation screenings in bacteria, yeast and mice have illustrated that the phenotypic responses occurring after single gene perturbations are greatly variable. While perturbations of essential genes cause lethal effects, other gene alterations show undetectable, subtle or environment-dependent phenotypes (Hillenmeyer *et al*., 2008; Nichols *et al*., 2011; White *et al*., 2013). Despite this observed phenotypic diversity, the majority of system-level analyses of perturbations has centered on gene properties linked to lethal phenotypes. For example, several topological properties of genes in protein–protein interaction networks such as betweenness centrality (hereafter named betweenness) have been associated with lethality (Goh *et al*., 2007; Jeong *et al*., 2001). Moreover, tissue gene expression and cellular localization of genes have also been related to lethality of perturbations in multicellular organisms (Goh *et al*., 2007; Liao and Zhang, 2008). In particular, genes expressed in multiple tissues tend to be essential both in human and mouse (Goh *et al*., 2007). In contrast, gene products localized in vacuoles have been found enriched among human essential genes whose orthologous genes are not essential in mouse (Liao and Zhang, 2008). Although the knowledge about the association of gene properties with lethality is relevant to understand severe perturbations such as lethal monogenic disorders, novel conceptual approaches are needed to analyze system properties of non-lethal perturbations frequently observed in humans such as complex disorders or drug treatment. In this regard, the number of side effects of a drug has been used as a measurement of the severity of drug response and has been studied in relation to the number of protein targets and to the essentiality and centrality of drug targets in human protein–protein interaction networks. However, this measurement does not take into account the diversity of impaired organ systems. For example, a drug exhibiting many side effects related to the same organ system would be considered as being as severe as a drug affecting

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first 2 authors should be regarded as Joint First Authors.

many different organ systems with the same number of side effects. Consequently, a measurement that considers the damage on the different organ systems will represent the impact of perturbations in the whole organism in a more realistic manner. To quantify the broadness of phenotypic effects across mammalian organ systems, we define the 'organ system heterogeneity [System Organ Class (SOC) heterogeneity]' measurement to analyze the influence of gene properties on the heterogeneity of organ system damage. For that, we performed a systematic analysis of the phenotypic impact of three perturbation scenarios, namely human drug treatment, human diseases and functional disruption of single genes in mice on 18 mammalian organ systems. For the three types of perturbations, we analyzed and compared the influence of the number of altered genes, tissue gene expression, betweenness and cellular localization of gene products on the overall organ system damage. We observed an expanded organ system heterogeneity for perturbations altering multiple genes, for alterations of single genes expressed in multiple tissues and for genes with high betweenness. We also find differences in the cellular localizations of proteins that are associated to a broad organ system heterogeneity in perturbations of mouse genes and human monogenic diseases, showing the variability in the molecular mechanisms leading to systemic effects for these perturbations. This analysis contributes to the elucidation of the gene properties influencing the organ system phenotypic effects of diseases and drugs.

2 MATERIAL AND METHODS

2.1 Phenotypic data

We compiled a phenotypic thesaurus from the Unified Medical Language System (UMLS) Metathesaurus based on the widely applied medical MedDRA ontology (Medical Dictionary for Regulatory Activities, Version 13.0, 2010) in a similar fashion done for the COSTART-based dictionary for the creation of the SIDER database (Kuhn *et al*., 2010a). This thesaurus was then used to extract side effect data from documents such as drug labels and monographs published by the U.S. Food and Drug Administration (FDA), the Medicines and Healthcare products Regulatory Agency (UK), BC Cancer Agency (Canada), MedEffect (only clinical report data, Canada) and the European Medicines Agency (EMA). In an analogous manner, disease signs and symptoms were collected from the clinical synopses in Online Mendelian Inheritance in Man (OMIM) as well as from disease-specific documents from CureResearch.com, the Merck Manual (home and professional edition) and the A.D.A.M. Medical Encyclopedia (content published in MedlinePlus). To extract phenotypic information from single gene perturbations in mice, we used gene–phenotype annotations provided by Mouse Genome Informatics (MGI; Blake *et al*., 2009), where the phenotypic descriptors are organized in the mammalian phenotype ontology (MPO; Smith *et al*., 2005). The terms of the MPO (from the file VOC_MammalianPhenotype.rpt, April 2012) were mapped to the UMLS with the help of MetaMap (http://mmtx.nlm.nih.gov). This application from the National Library of Medicine maps biomedical text to the UMLS Metathesaurus using natural language processing. We manually curated the high scoring matches to ensure high-quality mappings between MPO and the UMLS Metathesaurus. Finally, the most specific MedDRA terms associated to the UMLS Metathesaurus concepts annotated to each mouse gene, disease and drug comprise its final set of phenotypic features. Having all phenotypic features annotated to MedDRA enabled us to compare the three types of data sources systematically. Moreover, we could use the hierarchical structure of the ontology and analyze the phenotypic annotations on different levels of specificity like the most general level, the SOCs. Of the 26 SOCs present in MedDRA, we manually selected a subset of 18 listed in Figure 1 that can be directly linked to organ systems. In total, we collected phenotypic information for 4766 diseases, 1666 drugs and 5047 mouse genes.

2.2 Disease thesaurus

We collected all UMLS Metathesaurus (US National Library of Medicine, 2011) concepts classified as pathological function that were linked to Medical Subject Headings (2011), OMIM® (2011) or International Classification of Diseases, Ninth Revision, Clinical Modification (2010), and included all English synonyms provided by all freely accessible vocabularies included in the Metathesaurus.

2.3 Drug thesaurus

Our drug thesaurus is based on chemical synonyms provided by STITCH 2 (Kuhn *et al*., 2010b) and has been extended with information from Pubchem, RxNorm (U.S. National Library of Medicine, 2011) and KEGG as well as with the active ingredient lists provided by the Anatomical Therapeutic Chemical classification system, the electronic Medicines Compendium (www.medicines.org.uk/emc/), EMA and FDA.

2.4 Disease genes and drug targets

Information on disease genes were taken from DisGeNET (Bauer-Mehren, *et al*., 2010; Bauer-Mehren, *et al*., 2011). We only considered data from following curated sources: UniProt (Apweiler *et al*., 2004), Genetic Association Database (GAD; Becker *et al*., 2004), OMIM (Hamosh *et al*., 2005) or Comparative Toxicogenomics Database (CTD; Mattingly *et al*., 2006). Overall, we collected 9277 disease–gene associations between 2807 diseases and 3376 genes, where 2096 diseases are linked to only one gene. For 1266 of the diseases associated to only one gene we have symptom and tissue expression information available. For drugs, we extracted targets from the STITCH 3 database that have a confidence score >0.7. Moreover, we excluded indirect associations resulting in 1654 different targets for 1636 drugs.

2.5 SOC heterogeneity

In this work, we analyzed phenotypic traits at the level of organ systems as represented by MedDRA's SOCs. As a measurement of the organ system heterogeneity of a drug, mouse gene or disease, we calculated the Shannon entropy from the corresponding annotation frequencies of all SOCs and normalized by the maximum possible entropy:

$$H_{norm_{SOC}} = -\sum_{i=1}^{n} \frac{p(x_i)\log_2 p(x_i)}{\log_2(n)} \quad (1)$$

Here, p (x_i) refers to the relative annotation frequency of a SOC and n equals the number of different SOCs. The SOC heterogeneity measurement evaluates the broadness of the phenotypic effects across organ systems by accounting for the relative abundance, rather than for the number, of phenotypic traits affecting each organ system. Low heterogeneity values correspond to perturbations influencing mainly few organ systems (0 if only one organ system is affected), while high levels represent effects in multiple organ systems to a similar extent (1 if all organs are affected equally). We use the terms 'organ system heterogeneity' and 'SOC heterogeneity' synonymously.

To assess if potential annotation or study biases might alter the conclusion derived from this analysis, we repeated all the analyses presented here with a newly defined SOC heterogeneity measurement that down-weights frequently annotated classes by normalizing the counts for each SOC across all entities belonging to the same perturbation scenario (drugs, diseases, mouse genes) (Supplementary Figures S1–S4). This

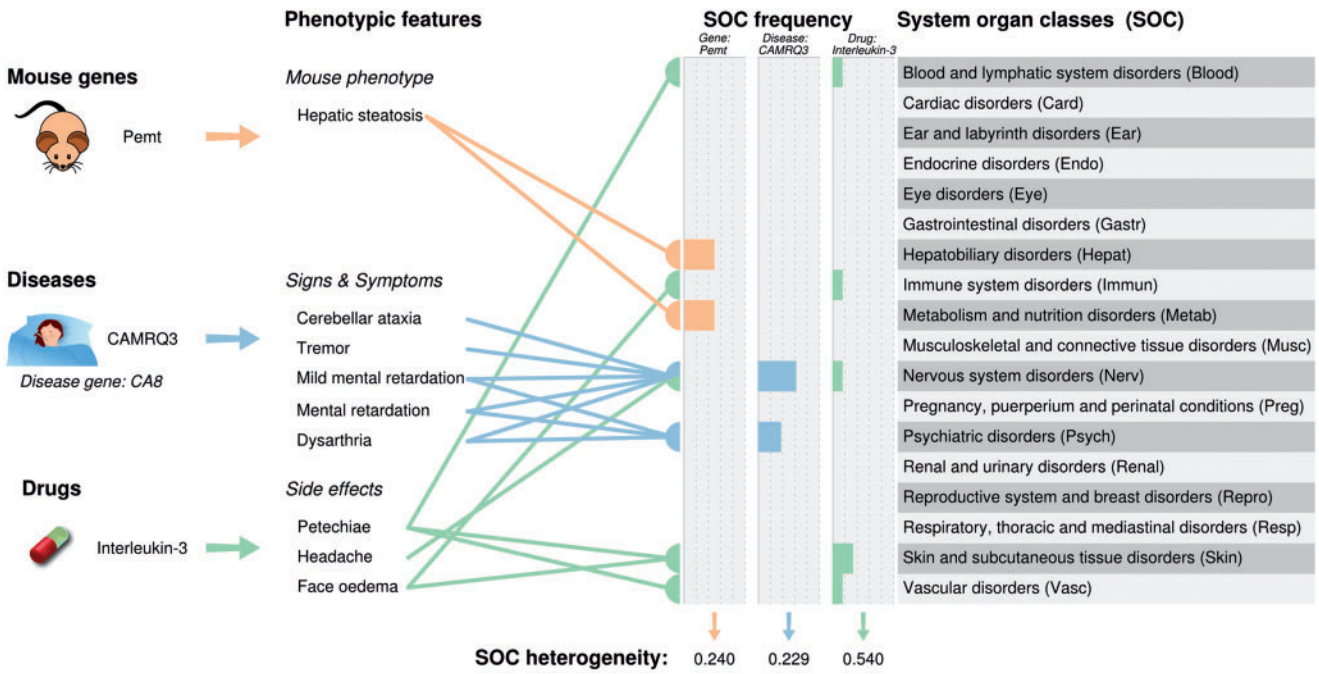

Fig. 1. SOC heterogeneity calculation. Phenotypic features of diseases, drugs and mouse genes are converted to SOC profiles, which are used to derive the SOC heterogeneity (Equation 1). The SOC heterogeneity for the mouse gene Pemt, the disease 'Cerebellar Ataxia, Mental Retardation and dyseQuilibrium syndrome 3' (CAMRQ3) and the drug Interleukin-3 is shown. For example, for Interleukin-3 the heterogeneity value is obtained as follows: $-((\log_2(1/6)\cdot 1/6))\cdot 4 + \log_2(2/6)\cdot 2/6)/\log_2(18)) = 0.54$

alternative SOC heterogeneity measurement corrects, for instance, for the annotation biased toward SOCs of interest of the biomedical community when reporting disease symptoms and drug side effects as well as for the missing phenotypic terms arising from the annotation of phenotypic sources.

2.6 Expression data

To evaluate correlations between SOC and tissue expression heterogeneity of mouse genes and human disease genes, we used the GeneAtlas GNF1H/GNF1M (Su *et al.*, 2004) dataset for human and mouse. If the data provided multiple probe sets for one gene, we retained the one with the highest mean expression across all tissues, that is, the one with the strongest signal. We normalized the tissue expression of each gene by the sum over all tissues, treating replicas independently.

2.7 Tissue expression heterogeneity

We calculate the tissue expression heterogeneity within each species for genes analogously to the organ system heterogeneity, considering the relative expression value of a gene in a particular tissue and normalizing by the total number of tissues:

$$H_{norm_{tissue}} = -\sum_{i=1}^{n} \frac{p(x_i)\log_2 p(x_i)}{\log_2(n)} \quad (2)$$

Here, p (x_i) refers to the normalized expression value in a tissue and n equals the number of different tissues.

2.8 Extraction of essential genes

We automatically scanned the phenotypic descriptions in the mammalian phenotype vocabulary provided by MGI for terms containing 'lethal' or 'death' to classify a mouse gene as essential. Genes where mutation phenotypes were available but not involved in a lethal phenotype were classified as non-essential.

2.9 Betweenness centrality

As essential genes tend to have a high betweenness, we wondered if SOC heterogeneity correlates with betweenness in the protein–protein interaction network from STRING (von Mering *et al.*, 2007). To include only high-confidence interactions, we applied a stringent cutoff of 0.7 and only kept associations resulting from experiments or extracted from curated databases. Betweenness b of a node v is a measurement for centrality in a network, which is defined as the sum over the number of shortest paths between all nodes s and t in the network running through the node v divided by the number of all shortest paths σ between s and t:

$$b(v) = \sum_{s\neq v\neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (3)$$

2.10 Cellular localizations of gene products associated with increased organ system heterogeneity

First, we obtained all Gene Ontology (GO) annotations for cellular components for genes in human monogenic diseases and mouse genes from ENSEMBL (release 73, for *Homo sapiens* and *Mus musculus*, respectively). We then excluded all annotations with IEA evidence code that identifies annotations depending directly on computation or automated transfer, which are not reviewed by a curator.

We then annotated all parent terms provided in the current GO version (http://www.geneontology.org/ontology/obo_format_1_2/gene_ontology_ext.obo, accessed November 2013) and removed 'cellular_component' as well as any terms representing unspecific subclassifications ('cytosol part', 'membrane part'), which are not intended for annotation. We thus collected 7891 annotations for 980 monogenic diseases genes

(average: ~8) and 19 860 annotations for 2816 mouse genes (average: ~7). Then, for every GO cellular component term annotated to at least 20 genes in our data in a respective gene set (human or mouse), we conducted a Mann–Whitney *U*-test with the alternative hypothesis that the SOC heterogeneity distribution of genes annotated to that term is shifted toward higher values compared with a negative set of genes not annotated to that term. To construct the negative set, we randomly picked 1000 times a number of genes equivalent to the number of genes associated to each GO term from all genes that are not annotated to that GO term and calculated their SOC heterogeneity. Lastly, the resulting *P*-values were corrected for multiple testing within each species with Benjamini–Hochberg correction, and only terms with false discovery rate (FDR) ≤ 0.05 were considered significant (Fig. 6).

3 RESULTS

3.1 SOC profiles

To study the organ systems affected by molecular perturbations, we used the MedDRA vocabulary to annotate phenotypes of mouse models, disease signs and symptoms and drug adverse effects from public sources (see Section 2 for details). These features were then mapped to a set of 18 SOCs from the most general level of the MedDRA ontology representing anatomical organ systems (Fig. 1) to allow the assessment of phenotypes at the organ system level. In total, we extracted 130 742 drug–side effect pairs for 1666 drugs, 44 125 disease–symptom pairs for 4766 diseases and 21 150 gene–phenotypic feature pairs for 5047 single gene perturbations in mice.

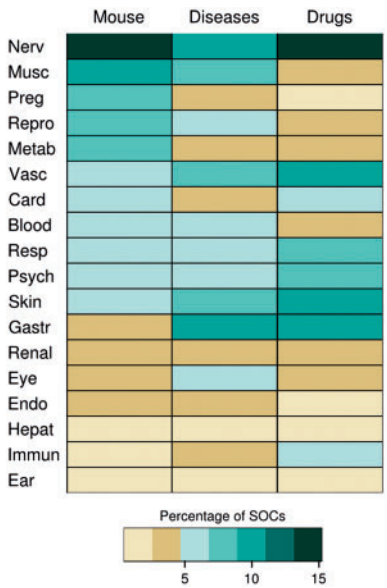

Fig. 2. Percentage of SOCs affected in mouse models of genes, human diseases and by drugs. The frequency of phenotypic annotations mapped to the most general level of the MedDRA ontology within mouse genes, diseases and drugs is shown. The nervous system appears to be most commonly affected in all three mammalian perturbations, whereas, for example, 'Pregnancy, puerperium and perinatal conditions' are more often observed in mouse perturbations, and "Skin and subcutaneous tissue disorders", "Gastrointestinal disorders" and "Vascular disorders" are more frequently seen as adverse effects of drugs

We first compared how the organ systems are affected in general in single gene perturbations in mice, diseases and by drugs by determining the distribution of SOC annotations for the phenotypes of the three perturbations types (Fig. 2). In all three types, the nervous system appears to be most commonly affected, whereas the hepatobiliary system and ear are the least frequently impaired organ systems. While reproductive and pregnancy disorders often occur in mouse genes, these conditions are rarely encountered as disease symptoms or drug side effects, probably because of the high potential to exert lethal effects. By contrast, effects on vascular, skin, respiratory and immune system are more often reported as adverse drug reactions, partially due to hypersensitivity reactions to medications that occur in certain patients (Pichler *et al.*, 2010).

3.2 Comparison of SOC heterogeneity between genes, diseases and drugs

Next, we wondered whether the three types of mammalian perturbations also differ in the individual broadness of organ system damage. To quantify the extent of organ system disruption, we define the SOC heterogeneity, a measurement that accounts for the relative abundance, rather than for the number, of phenotypic traits of a perturbation across the 18 different SOCs. The SOC heterogeneity values range from 0 to 1, where low values correspond to perturbations affecting mainly few organ systems (0 if only one organ system is affected), while high levels represent effects in multiple organ systems to a similar extent (1 if all organs are affected equally) (see Section 2 and Fig. 1).

The comparison of SOC heterogeneity distributions for drugs, diseases and single gene perturbations in mice show a clear distinction among the global phenotypic impact of the three perturbation types (Fig. 3A). While perturbations of single genes in mice exhibit the most homogeneous SOC distributions, followed by diseases, drugs yield the most heterogeneous phenotypes, indicating that drugs usually have a more widespread impact on the entire organism.

We note that the differences in SOC heterogeneity for mouse genes, diseases and drugs correlate with the number of genes associated to the three types of perturbations (Fig. 3B), suggesting that the number of genes perturbed could be an important factor influencing the diversity of organ damage in mammalian systems. To explore this hypothesis, we grouped diseases and drugs into three categories based on the number of disease genes and drug targets involved in each perturbation type and compared the SOC heterogeneity within the groups. We observed that the SOC heterogeneity increases significantly with the number of genes perturbed both in diseases and drugs (Fig. 3C), demonstrating that the number of genes altered in a perturbation has a marked impact on the diversity of organ damage.

3.3 SOC heterogeneity of single gene perturbations

The modulation of single genes in mice causes phenotypes of varying severity ranging from the absence of an observable phenotype to lethality for the organism for perturbations of essential genes.

To investigate whether the SOC heterogeneity accounts for the severity of a perturbation, we analyzed the SOC heterogeneity

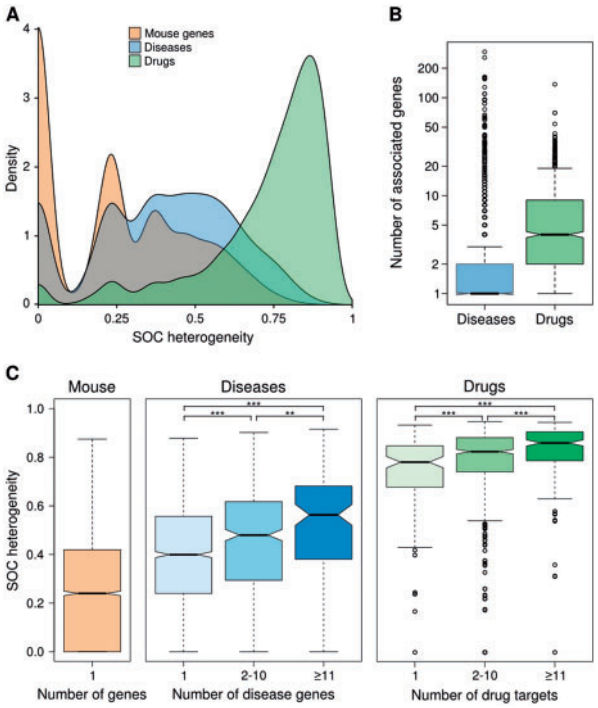

Fig. 3. SOC heterogeneity of drugs, diseases and mouse genes and the relationship to the number of associated genes in each class. (**A**) Distributions of SOC heterogeneity values for drugs, diseases and mouse genes. Drugs show the highest SOC heterogeneity, followed by diseases and then by genes. (**B**) Number of genes affected in each perturbation scenario. Drugs influence the largest number of genes, followed by diseases, while mouse models consist of single gene perturbations. (**C**) SOC heterogeneity values are plotted against the number of associated genes and targets binned into three classes. The higher the number of drug targets or disease genes the higher is also the heterogeneity of the annotated SOCs. The asterisks denote the significance of the *P*-values of the pairwise Mann–Whitney *U*-test (**$P \leq 0.01$, ***$P \leq 0.001$)

distribution of mouse and human single gene disruptions in more detail. For consistency and clarity of the analysis, we focused here on monogenic diseases and mouse genes while not taking drugs into account, where molecular causes are less well characterized (Hunter, 2005; Mestres *et al.*, 2008).

The comparison of the SOC heterogeneity distributions of 2557 essential and 2490 non-essential genes in mouse shows a markedly higher SOC heterogeneity for essential genes (Fig. 4A, *P*-value < 2.18E-198, Wilcoxon test). Interestingly, monogenic diseases associated to the human orthologs of 735 essential mouse genes also show a larger SOC heterogeneity than 343 human orthologs of non-essential mouse genes, albeit with less remarkable differences ($P < 0.007$, Wilcoxon test, Supporting Supplementary Fig. S1). The observed correlation between gene essentiality and SOC heterogeneity indicates that SOC heterogeneity is able to give insights about the severity of perturbations.

Essential genes tend to be expressed in multiple tissues, suggesting that the broad organ system effects of these genes might be caused by their activity across many tissues. We evaluated this hypothesis by testing whether a broad tissue expression

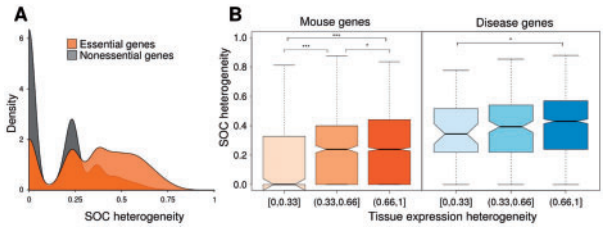

Fig. 4. Influence of perturbation severity and tissue gene expression on SOC heterogeneity. (**A**) SOC heterogeneity of essential and nonessential (in dark gray) genes in mouse. Essential genes show a significantly broader organ system heterogeneity ($P < 2.18E-198$, Wilcoxon test). (**B**) Relationship between tissue and SOC heterogeneity in mouse genes and monogenic diseases. SOC heterogeneity values are plotted against equidistantly binned tissue expression heterogeneity values perturbed in 4060 mouse models and 958 monogenic disease genes. The asterisks denote the significance of the *P*-values of the pairwise Mann–Whitney *U*-test (*$P \leq 0.05$, ***$P \leq 0.001$)

distribution of mouse and human genes correlates with an expanded organ system damage. Analogously to the SOC heterogeneity and similarly to a previous definition of overall gene tissue specificity (Schug *et al.*, 2005), we calculated the tissue expression heterogeneity for 4060 mouse and 761 human genes. For each gene, we considered its relative mRNA expression across a panel of 78 mouse and 84 human tissues, respectively, from the GeneAtlas GNF1M/H repository. The use of the relative mRNA expression levels across tissues avoids the application of an arbitrary threshold to determine the expression of genes in tissues and has the advantage of treating genes with a high or low general expression level equally.

We then classified mouse genes and those linked to human monogenic diseases in three equidistant levels of tissue expression heterogeneity and analyzed the SOC heterogeneity distribution observed within each group. Genes with low tissue expression heterogeneity show lower SOC heterogeneity (Fig. 4B) than genes expressed in many tissues, for both mouse and human genes, indicating that perturbations of genes with a specific tissue expression tend to produce damage in only few organ systems.

In protein interaction networks, betweenness has been proposed to be a significant indicator of essentiality (Yu *et al.*, 2007). Based on this, we tested whether betweenness correlates with SOC heterogeneity values in single gene perturbations. In perturbations of single genes in mice, we found that genes showing a high betweenness (third quartile) are associated with a markedly higher SOC heterogeneity (Fig. 5A). Interestingly, we observed this association also for non-essential mouse genes (Fig. 5C) and, although less striking, monogenic diseases (Fig. 5D), demonstrating that betweenness is a gene property affecting the organ system effects across mammalian organism.

We furthermore analyzed the relationship between the cellular localization of gene products and organ system heterogeneity in mouse and human single gene perturbations (Fig. 6). In mouse, we found a significantly higher SOC heterogeneity linked to proteins located in the extracellular space, membrane, plasma membrane, nucleus and transcription factor complexes. Among plasma membrane proteins associated with high SOC heterogeneity in mouse models we find transporters (SLC2A4, SLC4A2,

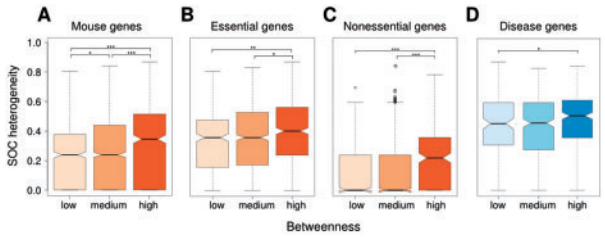

Fig. 5. SOC heterogeneity of mouse and genes of monogenic diseases in relationship to their betweenness in the protein–protein interaction network. (**A**) The organ system heterogeneity among 3302 gene perturbations in mice increases with increasing betweenness, and this trend is also observed in essential genes (**B**). (**C**) Non-essential mouse genes with a high betweenness also show a high SOC heterogeneity**. (D**) Relationship between betweenness and SOC heterogeneity for 761 genes associated with monogenic diseases. The asterisks denote the significance of the *P*-values of the pairwise Mann–Whitney *U*-test (*$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$)

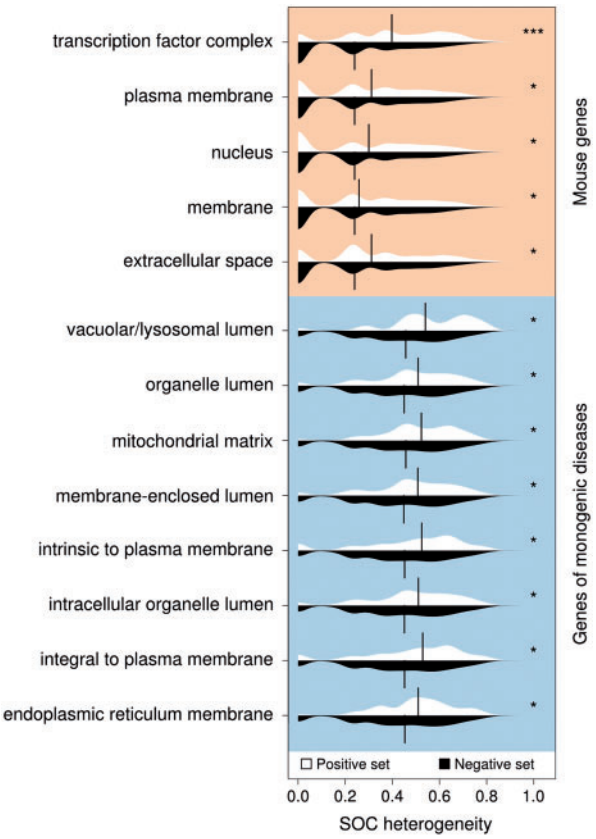

Fig. 6. Cellular localizations of gene products associated with increased organ system heterogeneity. To determine whether SOC heterogeneity values of gene products associated to GO cellular component categories are significantly higher than the rest of the genes, we performed Mann–Whitney *U*-tests. Positive set refers to genes annotated to the given GO term and Negative set to genes without that annotation extracted by randomly picking 1000 times the number of genes annotated with the GO term. The orthogonal lines represent median values for the corresponding dataset. The asterisks denote the FDR threshold after Benjamini–Hochberg correction for multiple testing applied within the datasets of the two species (*FDR ≤ 0.05, ***FDR ≤ 0.001)

SLC34A1), neurotransmitters (NOS1) and receptors (CHRNG, KIT). Genes annotated to extracellular space with broad phenotypic damage include growth factors (FGF10, IGF1, TGFB2, VEGFA) and hormones (EPO, PTHLH).

A high SOC heterogeneity in mouse genes associated to the nucleus and to transcription factor complex comprehends developmental genes such as transcription factors like ATF4, PITX2 and JUN and transcription coregulators (RB1/2, NCOR1). In human monogenic diseases, we observe that intrinsic components of the plasma membrane and gene products localized in the endoplasmic reticulum membrane and in the lumen of intracellular organelles (specifically lysosomes and mitochondria) present a significantly higher SOC heterogeneity after perturbation. Important functions of these subcellular compartments include signaling, degradation and transport. Monogenic disease genes contributing to these associations comprise for example channels (KCNJ2, CYBB), transporters (SLC7A7, SLC26A2, SLC17A5), receptors (IL2RG, MPL, NOTCH), their ligands (JAG1) and genes involved in degradation (MUT, GUSB, HEXB). In summary, a high SOC heterogeneity is distinctly linked to proteins localized in the nucleus and transcription factor complexes for mouse genes and in the lumen of different organelles for human monogenic disorders, whereas is commonly related to intrinsic membrane proteins in both mammalian systems.

Taken together, we have shown that disruptions of gene functionality resulting from mouse mutations, human diseases and drug treatments lead to significantly different SOC heterogeneity distributions influenced by the properties of the perturbed genes. These findings are not affected by potential study biases toward SOCs of interest of the biomedical community when reporting disease symptoms and drug side effects, by missing phenotypic information or by the quality of the gene–disease associations included in the analysis (Supplementary Figures S1–S4 and S6).

We thus conclude that the diversity of the phenotypic impact can be explained by the number of affected genes or gene characteristics like tissue expression, cellular localization, essentiality and betweenness.

4 DISCUSSION

In this work, we have performed a systematic analysis of system properties of genes modulating the phenotypic impact in perturbations of single genes in mice, human diseases and those caused by drug treatment. We demonstrated the influence of the number of disrupted genes, the role of tissue expression as well as topological and cellular localization gene properties on the organ system heterogeneity observed in human and mouse after chemical or gene perturbations.

Our analysis unveils a correlation between the broadness of organ system damage and the number of perturbed genes. This finding has important implications in the design of treatments for complex disorders, as these diseases are caused by the combination of the effects of multiple genes, environmental and life style factors (Hunter, 2005). Thus, as it has been proposed for schizophrenia (Roth *et al.*, 2004) and Alzheimer's disease (Espinoza-Fonseca,

2006), the treatment might require drugs with polypharmacology property, that is, the ability of binding to multiple targets, or alternatively, the controlled combination of drugs.

Although the relationship of drug polypharmacology to the increased number of side effects has been observed before (Keiser *et al*., 2007; Wang *et al*., 2013) and pointed out as the cause for the failure during clinical development of drugs (Azzaoui *et al*., 2007), we report here for the first time the broad effect of these drugs at the organ system level. We estimate that on average a drug binds to 6.7 targets, two times more than the average number of genes linked to diseases (3.3). Consistent with this, we detected a broader organ damage caused by drugs than by diseases. However, the observed effect is much greater than would be expected based on the number of known targets, as indicated by the lower SOC heterogeneity values estimated for the same number of randomly selected mouse genes (Supplementary Fig. S5). This suggests that non-genetic factors contribute to the widespread effects of drugs or that many other off-targets with a phenotypic impact remain unknown.

We identified further gene factors with a strong phenotypic influence, namely broad tissue expression, certain cellular localizations and betweenness centrality in protein networks. An example of a gene with high betweenness whose perturbation causes systemic effects in human diseases is GLB1 linked to Mucopolysaccharidosis type IV. GLB1 is involved in the degradation of molecules in lysosomes and is also part of the elastin receptor complex at the cell surface involved in cell proliferation and elastic fiber assembly (Antonicelli *et al*., 2009; Duca *et al*., 2007). These results suggest that the analyses of the phenotypic impact of diseases, in particular rare disorders with a strong genetic component, might aid in the determination of causative genes and their functions.

Regarding the phenotypic impact of cellular localization of gene products, we observed a significantly higher SOC heterogeneity for mouse proteins localized in the (plasma) membrane or extracellular space, hinting to the multiorgan effect of genes involved in signaling and transport. Likewise, disruption of proteins embedded in the plasma membrane as well as the membrane of the endoplasmic reticulum results in a significantly higher SOC heterogeneity in human diseases. Interestingly, human diseases connected to proteins localized in the lumen of intracellular organelles such as vacuoles (lysosomes, peroxisomes) and mitochondria are also linked to higher SOC heterogeneity. This suggests that the breakdown of molecules for recycling, waste disposal and detoxification in lysosomes and peroxisomes as well as energy production in mitochondria are cellular processes associated to organ system wide effects in human diseases but not in mouse models. Furthermore, in perturbations of single genes in mice transcription factor complexes and nucleus localization are associated with an increased SOC heterogeneity likely due to the relevant role of replication and transcription in early phases of development. We do not observe this association in human diseases, probably because of non-viable phenotypes caused by the disruption of transcription factor complexes. Perturbations of mouse genes comprise mainly knockout deletions (>80% of the mouse phenotypes are derived from knockouts), implying a complete absence of functional protein, whereas monogenic disorders are caused by a variety of point mutations, gene duplications and allelic gene differences that ameliorate gene function or cause a gain of function but do not necessarily lead to complete loss of function (Georgi *et al*., 2013). As a consequence, disease symptoms do not fully reflect the strong phenotypic effect associated to the essentiality, betweenness and tissue expression in perturbations of single genes in mice. These findings are in agreement with the proposition by Goh and collaborators of a selective pressure on disease genes where only those mutations compatible with survival into the reproductive years are likely to be maintained in the population. This is corroborated by the difference in subcellular localizations related to high organ system heterogeneity perturbations of single genes in mice and disease genes. The associations of human disease proteins localized in the lumen of intracellular organelles such as lysosomes and mitochondria with high SOC heterogeneity suggest that either the toxic accumulation of unprocessed molecules or impaired function of various organs due to ATP deficiency have most likely non-essential functions during embryonic development but subtle yet cumulative harmful effects over time.

The expression of a gene in multiple tissues as well as high betweenness has been linked to essentiality in mammalian organisms. We observed that the same gene properties correlate with the SOC heterogeneity values indicating that this measurement accounts for the severity of perturbations, and thus, for a high probability to be essential. Interestingly, Liao *et al*. (2008) found an enrichment of proteins localized in vacuoles among human essential genes, which are non-essential in mouse models. We have confirmed this association and extended the list of disease genes producing systemic effects. SOC heterogeneity is thus a measurement of the severity of the perturbations applicable to the analysis of non-lethal phenotypes such as diseases.

These findings imply important consequences for drug design in chronic diseases where treatments usually last over a long period. In these cases the analysis of the potential to affect off-targets expressed e.g. in lysosomes or mitochondria could help to decrease the occurrence of adverse events resulting from long-term treatment, which are difficult to detect in clinical trials.

Also, the association of betweenness, essentiality and tissue expression to the organ system heterogeneity stresses the importance of considering these characteristics for drug safety and emphasizes the crucial role of network pharmacology in rational drug design (Hopkins and Groom, 2002).

Our results demonstrate that systematic analyses relating gene attributes and associated organ system phenotypes help to elucidate the global system properties governing the relationships between genotype and phenotype. We investigated the differences and commonalities in the phenotypic impact of different perturbations in human and mouse using a new and fairly comprehensive dataset of organ system phenotypes of mammalian perturbations. Altogether, this approach contributes to the clarification of the molecular causes and phenotypic consequences of human diseases and drug treatment.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Michael Kuhn for providing the STITCH data, the support of the TUM Graduate School's Faculty Graduate Center Weihenstephan at the Technische Universität München, Germany, and Lucia Himmelein and

Bernd Streppel for their valuable assistance in compiling the phenotypic data.

Funding: This study was supported in part by a grant from the German Federal Ministry of Education and Research (BMBF) to the German Center for Diabetes Research (DZD e.V.).

Conflict of Interest: none declared.

REFERENCES

Antonicelli,F. *et al.* (2009) Role of the elastin receptor complex (S-Gal/Cath-A/Neu-1) in skin repair and regeneration. *Wound Repair Regen.*, **17**, 631–638.

Apweiler,R. *et al.* (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.

Azzaoui,K. *et al.* (2007) Modeling promiscuity based on in vitro safety pharmacology profiling data. *ChemMedChem*, **2**, 874–880.

Bauer-Mehren,A. *et al.* (2010) DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. *Bioinformatics*, **26**, 2924–2926.

Bauer-Mehren,A. *et al.* (2011) Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. *PloS One*, **6**, e20284.

Becker,K.G. *et al.* (2004) The genetic association database. *Nat. Genet.*, **36**, 431–432.

Blake,J.A. *et al.* (2009) The mouse genome database genotypes: phenotypes. *Nucleic Acids Res.*, **37**, D712–D719.

Campillos,M. *et al.* (2008) Drug target identification using side-effect similarity. *Science*, **321**, 263–266.

Duca,L. *et al.* (2007) The elastin receptor complex transduces signals through the catalytic activity of its Neu-1 subunit. *J. Biol. Chem.*, **282**, 12484–12491.

Espinoza-Fonseca,L.M. (2006) The benefits of the multi-target approach in drug design and discovery. *Bioorg. Med. Chem.*, **14**, 896–897.

Georgi,B. *et al.* (2013) From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. *PLoS Genet.*, **9**, e1003484.

Goh,K.-I. *et al.* (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.

Hamosh,A. *et al.* (2005) Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.

Hillenmeyer,M.E. *et al.* (2008) The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science*, **320**, 362–365.

Hopkins,A.L. and Groom,C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.*, **1**, 727–730.

Hunter,D.J. (2005) Gene-environment interactions in human diseases. *Nat. Rev. Genet.*, **6**, 287–298.

Jeong,H. *et al.* (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Keiser,M.J. *et al.* (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.*, **25**, 197–206.

Kuhn,M. *et al.* (2013) Systematic identification of proteins that elicit drug side effects. *Mol. Syst. Biol.*, **9**, 663.

Kuhn,M. *et al.* (2010a) A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.*, **6**, 343.

Kuhn,M. *et al.* (2010b) STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res.*, **38**, D552–D556.

Liao,B.-Y. and Zhang,J. (2008) Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc. Natl Acad. Sci. USA*, **105**, 6987–6992.

Mattingly,C.J. *et al.* (2006) The comparative toxicogenomics database: a cross-species resource for building chemical-gene interaction networks. *Toxicol. Sci.*, **92**, 587–595.

Mestres,J. *et al.* (2008) Data completeness—the Achilles heel of drug-target networks. *Nat. Biotechnol.*, **26**, 983–984.

Nichols,R.J. *et al.* (2011) Phenotypic landscape of a bacterial cell. *Cell*, **144**, 143–156.

Pichler,W.J. *et al.* (2010) Drug hypersensitivity reactions: pathomechanism and clinical symptoms. *Med. Clin. North Am.*, **94**, 645–664 xv.

Roth,B.L. *et al.* (2004) Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discov.*, **3**, 353–359.

Schug,J. *et al.* (2005) Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.*, **6**, R33.

Smith,C.L. *et al.* (2005) The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol.*, **6**, R7.

Su,A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.

Visscher,P.M. *et al.* (2012) Five years of GWAS discovery. *Am. J. Hum. Genet.*, **90**, 7–24.

von Mering,C. *et al.* (2007) STRING 7—recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, **35**, D358–D362.

Wang,X. *et al.* (2013) Target essentiality and centrality characterize drug side effects. *PLoS Comput. Biol.*, **9**, e1003119.

White,J.K. *et al.* (2013) Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. *Cell*, **154**, 452–464.

Yu,H. *et al.* (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.*, **3**, e59.