Pathway centrality in protein interaction networks identifies functional mediators of pulmonary disease

Authors

Jisoo Park1,*, Benjamin J. Hescott1, Donna K. Slonim1,2,*

1 Department of Computer Science, Tufts University, Medford, MA 02155, USA
2 Department of Integrative Physiology and Pathobiology, Tufts University School of Medicine, Boston, MA 02111, USA

*Correspondence

slonim@cs.tufts.edu (D.K.S.)
jisoo.park@tufts.edu (J.P.)

Summary

Identification of functional pathways mediating molecular responses may lead to better understanding of disease processes and suggest new therapeutic approaches. We introduce a method to detect such mediating functions using topological properties of protein-protein interaction networks. We introduce the concept of pathway centrality, a measure of communication between disease genes and differentially expressed genes. We find mediating pathways for three pulmonary diseases (asthma; bronchopulmonary dysplasia (BPD); and chronic obstructive pulmonary disease (COPD)) using pathway centrality. Mediating pathways shared by all three pulmonary disorders heavily favor inflammatory or immune responses and include specific pathways such as cytokine production, NF Kappa B, and JAK/STAT signaling. Disease-specific mediators, such as insulin signaling in BPD or homeostasis in COPD, are also highlighted. We support our findings, some of which suggest new treatment approaches, both with anecdotal evidence from the literature and via systematic evaluation using genetic interactions.
Introduction

It has long been noted that genes with variants implicated in disease are not necessarily differentially expressed, and that differential expression does not easily lead to the discovery of disease genes (Hudson et al., 2012). In many cases, differential expression predominantly reflects the tissue-specific consequences or downstream effects of a disease-causing process that integrates complex genetic and environmental responses. This makes differentially expressed genes useful as diagnostic markers, but often poor as therapeutic targets (Fox et al., 2011). Conversely, functional analysis of causal disease genes, whether identified through highly specific cell or animal studies or systematically via GWAS, often won’t fully explain how these downstream responses occur (Delude, 2015). However, we know that other functional pathways may mediate the expression responses we see. We hypothesize that finding these can provide new insights into disease processes and suggest novel therapeutic approaches. Here, we test this hypothesis by examining three different pulmonary disorders from different life stages to identify common and unique mediating mechanisms in airway disease.

We start by considering the roles that proteins corresponding to disease genes or differentially expressed genes play in protein-protein interaction networks. We postulate that the functional pathways mediating disease response will disproportionately reflect communication between these two sets of proteins, and conversely, that if we look for such “central” pathways we will find mediators. To identify putative mediating pathways in disease response, we first introduce a generalized notion of network centrality called pathway centrality, which combines a variation of betweenness with the concept of group centrality (Everett and Borgatti, 1999). Node betweenness characterizes how many of the shortest paths between all pairs of nodes pass through a given node (Abman et al., 1987, Yu et al., 2007). Our variation of betweenness counts only the shortest paths between disease genes and differentially expressed genes passing through a given node. The notion of group centrality averages the betweenness scores of a set of nodes to measure the centrality of the whole set. Therefore, our pathway centrality score is the average modified betweenness score of a set of genes participating in the given pathway.

There have been many other approaches to defining centrality in protein interaction networks (Mason and Verwoerd, 2007). Some have focused on modeling
communication between nodes by network or graph flows (Freeman et al., 1991). Using network flows to calculate network distance has proven helpful for protein function prediction (Nabieva et al., 2005), and has been compared favorably to betweenness centrality for integrating function prediction with contextual data (Missiuro et al., 2009). However, this approach has not yet been applied to group centrality in protein interaction networks. Viewing group centrality as an optimization problem has also been used to discover new groups of important nodes in networks (Erdos, 2015), but not for identifying functional gene sets playing a pivotal role.

Most relevant to our efforts is a collection of prior results linking expression quantitative trait loci (eQTLs) to differentially expressed genes via protein-protein, protein-DNA, and phosphorylation networks. These studies were initially intended to find the exact causal gene in a linked locus using pathway information. One early approach (Tu et al., 2006) uses a node’s accessibility in random walks to prioritize causal genes and identify regulatory paths through the network. Slightly later work (Suthram et al., 2008, Kim et al., 2011) models the integrated network as an electrical circuit, identifying putative causal genes by the flow of current through the network. Kim, et al. explicitly extended the eQTL-target work to a disease context (Kim et al., 2011). Specifically, by linking copy number variations, through eQTL mapping, to differentially expressed genes, they identified likely causal mutations in glioblastoma multiforme. Similarly, Yeger-Lotem, et al. used integrated protein-protein and protein-DNA interaction networks to identify proteins and genes on high-probability paths between “genetic hits” and transcriptionally regulated genes (Yeger-Lotem et al., 2009). In this work, which focuses on a number of cellular perturbations in yeast, the emphasis is on finding individual proteins on these paths and exploring the functions they perform.

These efforts are related to ours in the sense that they examine information flow between genes linked to disease and differentially expressed genes. Specifically, our disease genes come from a compilation of sources that includes, but is not limited to, GWAS data. However, here our focus is limited to the disease-related pathways, and our aim is to identify underlying biological functions that mediate cellular response in disease, rather than to identify causal mutations. Given a disease $d$, a set $D$ of known disease genes for $d$, and a set $E$ of genes differentially expressed in relevant tissues as a
consequence of disease \( d \), we search for gene sets \( M \) that significantly participate in passing signals from \( D \) to \( E \) (Figure 1). We calculate the pathway centrality (PC) score for a candidate pathway \( S \) considering only the paths from nodes in \( D \) to nodes in \( E \). We then use permutation tests to assess whether \( S \) has higher pathway-centrality than expected (see Methods). For each gene set \( S \), the percentage of such random sets with higher pathway centrality scores than \( S \) is reported as \( p_{\text{cen}}(S) \), a rough measure of significance.

In this initial study, we apply our novel pathway-centrality method to three pulmonary diseases that primarily affect patients at different stages of life: bronchopulmonary dysplasia (BPD), a neonatal complication of preterm birth; asthma, which is relevant across the lifespan but is becoming increasingly common in children; and chronic obstructive pulmonary disease (COPD), a term that encompasses a number of progressive lung disorders that predominantly affect the elderly (Suzuki et al., 2016). We examine mediating pathways in each disease and look for common pathways relating all three.

One caveat is that pathways with low pathway centrality significance (\( p_{\text{cen}} \)) scores for multiple diseases might be highly central in the network structure overall, and could appear to be “significant” for any disease considered. Therefore, as a further control, we applied the method to Down syndrome (DS) data, with the understanding that most functions relevant in pulmonary disorders are probably not that relevant to the etiology of DS, although we agree that there may be some exceptions. Our “disease genes” in this case are simply any genes located on chromosome 21, rather than genes that have been directly implicated in the etiology of DS symptoms.

Supplementary Figure S1 shows the overlaps between the four disease gene sets and the four differentially-expressed gene sets. Although there is some overlap between the sets of disease genes, particularly involving genes involved in inflammation and immune response, overall the disease gene sets are reasonably disjoint and the differentially-expressed gene sets even more so. Thus, common pathways across all three networks are unlikely to have arisen from shared shortest paths between identical sets of genes.

Pathway gene sets used in our experiments come from the Biological Process (BP)
terms in the Gene Ontology (Ashburner et al., 2000), pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2016), and mammalian phenotypes (MP) from the International Mouse Phenotyping Consortium in the Mouse Genome Database (Bult et al., 2016).

Our results show that pathways involved in inflammation or adaptive immunity, and several related signaling pathways including NFκB, JAK/STAT, and toll-like receptor signaling, are common to all three pulmonary disorders. Disease-specific findings include protein homeostasis pathways in COPD, leukocyte chemotaxis in asthma, and regulation of map kinase activity in BPD. While many of our findings have already been proposed as disease mediators, diagnostics, or targets in previous publications, we discovered several novel mediators that may suggest new therapeutic approaches to these pulmonary disorders.

Results and Discussion

Pathway centrality finds mediating pathways and potential drug targets

Pathways with significant pathway centrality in specific pulmonary data sets but not in the Down syndrome data include known disease mediators and potential targets. Full lists of results are available as supplemental data; Table 1 shows the top few pathways implicated in exactly one of the pulmonary disorders. We highlight some of these top pathways here.

Highly significant in asthma is the BP gene set leukocyte chemotaxis. Neutrophil chemotaxis velocity has been suggested as a biomarker for asthma and incorporated into a diagnostic test to distinguish asthma from nonasthmatic allergic rhinitis (Sackmann et al., 2014). The KEGG gene set cell adhesion molecules tops the list in COPD, pointing at similar processes, although the only genes shared between these two gene sets are ITGB2 and ITGA9. Prior work suggests that adhesion molecules also play a significant role in the pathogenesis of COPD (Ishii, 1999), and that regulation of adhesion may be a new therapeutic approach for both COPD and asthma (Chen et al., 2008, Woodside and Vanderslice, 2008). The KEGG pathway tight junction is also implicated in asthma. Although adhesion and leukocyte chemotaxis are important to all three disorders (Faura
Tellez et al., 2016, Ramsay et al., 1998, Barnes et al., 2009), the pathway centrality approach highlights different sets of genes mediating these responses. Different pathways have been implicated in recruitment of neutrophils in different contexts in COPD (Barnes et al., 2009), suggesting the plausibility of context-specific variation.

Topping the GO COPD list are two terms relating to homeostasis. Protein homeostatic imbalance has been implicated in the pathogenesis of COPD, although this approach to characterizing the disease has not yet been thoroughly explored (Bodas et al., 2012). Such imbalances have been successfully targeted through small molecules or gene therapies in other diseases such as cystic fibrosis. Addressing the imbalance of critical NFKB-regulated inflammatory proteins has thus been suggested as a possible therapeutic approach in COPD and other airways diseases (Bouchecareilh and Balch, 2012).

Mitogen-activated protein kinases (MAPKs) regulate many developmental and cellular processes (Kim and Choi, 2010), including inflammation and apoptosis. While these pathways are likely involved to some degree in all three pulmonary disorders, the top GO BP gene sets implicated in BPD are regulation of map kinase activity and ras protein signal transduction. The extracellular signal-related kinases ERK1/2 are major components of the MAPK pathway, which can be activated by the Ras GTP-ase (Kim and Choi, 2010). There is also evidence that ERK/MAPK activation can protect against the negative effects of hyperoxia on alveolar development and lung epithelial cells (Xu et al., 2006, Buckley et al., 2005). Drugs that promote this process have been suggested as potential therapies for BPD (Sakurai et al., 2013).

The KEGG insulin signaling pathway is also implicated as a mediator in BPD. The role of insulin signaling in BPD is unclear, but low neonatal serum IGF-1 levels have been shown to be associated with the later development of BPD (Lofqvist et al., 2012). This evidence is consistent with cell and animal studies showing that IGF-1 signaling, mediating by NFKB, enhances collagen synthesis in fetal lung fibroblasts (Chetty et al., 2006), and that IGF-1 and IGFR1 regulate alveologenesis in neonatal rats (Belcastro et al., 2015). Whether this finding suggests novel strategies for BPD treatment, however, is an open question.
Commonalities across all three pulmonary disorders implicate immune processes and signaling pathways

When we look for pathways that play a significant role across all three pulmonary disorders (Table 2), we are not surprised to find a preponderance of inflammatory and immune processes. GO terms topping the list implicate the adaptive immune response, cytokine production, and pro-inflammatory NF Kappa B signaling. Recent research implicates IKK-driven NFKB activation of inflammation in both COPD and asthma, but suggests that the different components of the system involved in the two diseases could explain their differing pharmacological responses (Gagliardo et al., 2011), and could suggest new avenues for therapy. NFKB-mediated inflammation has also been implicated in the pathogenesis of BPD. The increased prevalence of early-onset emphysema in BPD survivors (Wong et al., 2008) supports the hypothesis that an NFKB-related inflammatory phenotype predisposes individuals to an increased reaction to environmental airway stress (Park et al., 2014). Links between BPD and asthma through NFKB1A promoter polymorphisms have also been identified (Ali et al., 2013).

The KEGG and MP gene set collections implicate several more specific pathways, including different types of lymphocyte responses, JAK/STAT signaling, toll-like receptor signaling, and FC Epsilon RI signaling. The JAK/STAT pathway has been suggested as an asthma target through inhibitors of activating cytokines and receptors (Vale, 2016, Weltman and Karim, 1998). JAK pathway inhibitors are in development for a number of inflammatory disorders (O'Shea et al., 2015), and animal models have suggested that targeting this pathway can reduce airway hyperresponsiveness, reflecting the potential for such compounds in both COPD and asthma (Barnes, 2016). The role for JAK/STAT signaling in bronchopulmonary dysplasia is less clear, but it has been suggested that it plays a role in airway smooth muscle mitogenesis, implicated in both asthma and BPD (Simon et al., 2002), and postulated that it may be an alternative mediator of the oxidative stress response in both diseases (Zhou and Hershenson, 2003). Thus, our work suggests that the effects of JAK pathway inhibitors may be worth exploring in pre-clinical models of bronchopulmonary dysplasia.

Toll-like receptor (TLR) signaling, which activates the innate immune response, is another familiar part of the story of airway hyperreactivity and fetal lung development.
TLR polymorphisms have been linked to an increased risk of developing BPD (Carrera et al., 2015; Malash et al., 2016), and TLR agonists are already being tested for therapeutic efficacy in asthma (Biggadike et al., 2016). However, the role of this system in COPD is not as clear. Aspects of the innate immune response are often demonstrably suppressed in COPD patients (Shaykhiev and Crystal, 2013), and TLR polymorphisms play a role in disease susceptibility and severity (Apostolou et al., 2016; Yu et al., 2016). Our work therefore also suggests a role for TLR pathways in the diagnosis, stratification, and treatment of COPD.

Finally, the FC Epsilon RI signaling pathway raises interesting questions about common elements of these three diseases. Figure 2 shows a subset of this pathway and the network for BPD. The gene FCERIA is one of the primary receptors for immunoglobulin E (IgE) (Sharma et al., 2014), the key player in initiating allergic response. The role of allergy in these three disorders, however, is thought to be quite different. The significant impact of IgE response in allergic asthma is well-studied (Gould and Sutton, 2008). Its role in COPD is less clear. Many patients with COPD but no asthma diagnosis have one or more asthma-like symptoms, including atopy, and there are suggestions that allergic response plays a role in severity for a subset of those with COPD (Suzuki et al., 2016).

In contrast, most evidence suggests that bronchopulmonary dysplasia is not linked to the development of allergies or to IgE response (Kwinta et al., 2013). Indeed preterm birth has been shown to correlate with a decreased risk of atopy (Siltanen et al., 2001), although with an increased risk of asthma and ultimately of obstructive pulmonary disease (Stern et al., 2007). Further, the question of whether asthma in survivors of BPD is more likely to be non-atopic is still open, although there is some evidence supporting this hypothesis (Verhaeghe et al., 1990; Ronkainen et al., 2016). In this context, the fact that this signaling pathway is nonetheless implicated as a mediator of gene expression changes in infants with BPD is intriguing and may shed light on the mechanisms involved in the disease. Perhaps it simply reflects airway hyperreactivity in general, but there is little specific evidence supporting this possibility. Further exploration of this pathway’s role in the development of BPD is therefore warranted.
Genetic interaction data confirms the identification of mediating pathways

Finding specific examples consistent with existing knowledge provides anecdotal evidence that an approach is effective, but more systematic evaluation is needed. One indication that the proposed mediating pathways are, at least in some respect, downstream of the disease genes, would be to identify an excess of epistatic relationships between them. For example, if a mediating pathway looks like that shown in Figure 1, one might expect a higher likelihood of certain kinds of genetic interactions between a disease gene \(d\) in set \(D\) and a mediating gene \(m\) from set \(M\) than between \(d\) and genes that are not in a mediating pathway for that disease. The genetic interactions of most interest would be “alleviating” or positive genetic interactions, where the deleterious effect of the double mutant of both \(d\) and \(m\) is less severe than would be predicted by combining the independent effects of individual mutations in \(d\) or \(m\). Such relationships might arise when \(m\) is part of a pathway mediating the response of \(d\).

Because few such relationships have yet been characterized in humans or other mammals, we assembled a set of putative positive genetic relationships from orthologous genes in yeast, worm, and fly. For a mediating gene set or pathway \(S\) and disease gene set \(D\), we determine \(k\), the number of such relationships between genes in \(D\) and \(S\). We then use a separate permutation test (see Figure 3 and Methods) to assess the probability of finding \(k\) such relationships with a random gene set of size \(|S|\), which we call \(p_{\text{med}}(S)\).

Overall correlation between \(p_{\text{cent}}(S)\) and \(p_{\text{med}}(S)\) varies considerably across the data sets and annotation sets, with a range of 0.07 to 0.74. The lowest correlations come from the Down syndrome data, where this approach is more problematic due to the large number of “disease” (i.e., trisomic) genes that are likely not directly involved in the etiology of the DS phenotype. Instead of overall correlation, however, our goal is to assess the probability that a pathway \(S\) having a low \(p_{\text{cent}}(S)\) value is more likely to have a low \(p_{\text{med}}(S)\) value as well. We therefore split the range of \(p_{\text{cent}}\) values up into deciles, and plotted the probability of finding a \(p_{\text{med}}\) score below 0.05 in each decile. A sample plot of these probabilities is shown in Figure 4; the rest are available as supplemental data. If the slope of a linear regression line through the points (the blue line in Figure 4) is significantly less than zero, it suggests that the probability of being a mediator is highest for the most significantly central pathways.
Table 3 shows the p-values assessing this property (as reported by the R generalized linear model function glm()). These p-values are consistently high for the DS data, which makes sense for the reason noted above. The test is otherwise significant for the KEGG pathways and for several other pathway/disease combinations. In some of the cases where it is not, such as BPD with the MP gene sets, the first decile does in fact have the highest probability; the poor significance and relatively flat slope occur because there also happen to be many genetic relationships for $p_{cent}$ values above 0.8.

To focus on just the most-significant pathways, we can simply compare the distribution of $p_{med}$ values in the first bin to that in the remaining bins using a one-sided, non-parametric Wilcoxon test. These significance values are also shown in Table 3. This less strict test is significant in all the pulmonary disease cases and almost all the DS ones as well. Both the tests and the plots support the conclusion that there is enrichment of alleviating genetic relationships between disease genes and pathway genes for the pathways whose $p_{cent}$ values are in the most significant decile.

Although there are many issues with the data used to produce these results, the conservation of genetic interactions across species being one of the most salient, the trends shown here add systematic evidence to the anecdotal evidence above that this approach can indeed find pathways playing a mediating role in disease processes, potentially leading to the discovery of novel therapeutic interventions. Such an approach may be applied in the context of any disease or phenotype of interest.

Author Contributions
J.P. and D.K.S. conceived the pathway centrality algorithm and wrote the manuscript. J.P., B.J.H, and D.K.S. designed the statistical method to assess significance of our findings and edited the manuscript. J.P. implemented the algorithm and ran the experiments. J.P. and D.K.S. analyzed the data and provided biological interpretation of the results. All authors read and approved the final version of the manuscript.

Acknowledgments
We gratefully acknowledge the support of the Eunice Kennedy Shriver National Institute of Child Health & Human Development under award R01HD076140. We thank Diana Bianchi and Jill Maron for helpful discussions. We also thank members of the Tufts Bioinformatics and Computational Biology group for feedback and comments on earlier versions of this work.

REFERENCES


Figure Legends

Figure 1. **Hypothesized topological property of mediating pathways.**
Given that D is a set of disease genes and E is a set of differentially expressed genes, mediating pathway genes, M, will participate more significantly in passing signals from D to E.

Figure 2. **Topology of BPD-related genes and the FC Epsilon RI signaling pathway.**
One of the significant mediating pathways for Bronchopulmonary Dysplasia (Carrera et al., 2015) is the KEGG FC Epsilon RI signaling pathway. Here, pathway
genes are colored dark green, mediating pathway genes are in between BPD disease genes (red) and differentially expressed genes (blue). The VAV proteins and BTK, for example, are known to be involved in lymphocyte development and activation.

Figure 3.
**Systematic confirmation of significant mediating pathways.**
If identified pathways are truly mediating a disease response, the pathways are likely to be downstream of the corresponding disease genes. This relationship can be indicated by an excess of epistatic interactions between disease genes and pathway genes. We count alleviating genetic interactions between disease genes and our identified mediating pathways, and then assess significance by calculating the probability that a gene set of equal size has the same number of alleviating genetic interactions with the disease genes. The null distribution is learned from 10,000 random samples (R) drawn from a pool of downstream genes of any known alleviating genetic interactions.

Figure 4.
**Relationship between \( p_{\text{cent}} \) and \( p_{\text{med}} \), for KEGG pathways and COPD.**
We claim that pathways with low \( p_{\text{cent}} \) also have low \( p_{\text{med}} \) values. That is, there is a low probability (\( p_{\text{med}} \)) that a gene set of equal size has as many alleviating genetic interactions with COPD genes as a significant mediating pathway (one with low \( p_{\text{cent}} \)). This plot supports this claim because the linear regression line (in blue) has a slope significantly less than zero (\( p < 0.0006 \)), suggesting a strong relationship between \( p_{\text{cent}} \) and \( p_{\text{med}} \).

Tables

Table 1.
Most-significant GO Biological Processes and KEGG pathways in individual pulmonary disorders. These pathways are the top few with \( p_{\text{cent}} < 0.05 \) in exactly one of the pulmonary disorders and \( p_{\text{cent}} > 0.05 \) in the DS data. Cells containing p-values below 0.05 are highlighted in boldface. Pathways shown here were selected by criteria discussed in the text; full results are available as supplemental data.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>GO Biological Process Terms:</th>
<th>asthma p-value</th>
<th>BPD p-value</th>
<th>COPD p-value</th>
<th>DS p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regulation of DNA Metabolic Process</td>
<td>0.0014</td>
<td>0.1015</td>
<td>0.2034</td>
<td>0.3279</td>
<td></td>
</tr>
<tr>
<td>Leukocyte Chemotaxis</td>
<td>0.0021</td>
<td>0.3232</td>
<td>0.1634</td>
<td>0.2219</td>
<td></td>
</tr>
<tr>
<td>GI Phase</td>
<td>0.0027</td>
<td>0.0771</td>
<td>0.1375</td>
<td>0.087</td>
<td></td>
</tr>
<tr>
<td>Leukocyte Migration</td>
<td>0.0032</td>
<td>0.4966</td>
<td>0.2705</td>
<td>0.3176</td>
<td></td>
</tr>
<tr>
<td>Regulation of MAP Kinase Activity</td>
<td>0.1779</td>
<td>0.0035</td>
<td>0.0502</td>
<td>0.0874</td>
<td></td>
</tr>
<tr>
<td>Ras Protein Signal Transduction</td>
<td>0.0995</td>
<td>0.0049</td>
<td>0.0854</td>
<td>0.1305</td>
<td></td>
</tr>
<tr>
<td>Homeostatic Process</td>
<td>0.156</td>
<td>0.1313</td>
<td>0.0004</td>
<td>0.362</td>
<td></td>
</tr>
<tr>
<td>Nuclear Transport</td>
<td>0.2514</td>
<td>0.056</td>
<td>0.0018</td>
<td>0.089</td>
<td></td>
</tr>
<tr>
<td>Nucleocytoplasmic Transport</td>
<td>0.2455</td>
<td>0.0524</td>
<td>0.002</td>
<td>0.0872</td>
<td></td>
</tr>
</tbody>
</table>
Table 2.
All GO Biological Process terms and KEGG pathways, and manually selected Mammalian Phenotype terms, significant (p < 0.05) in all three pulmonary disorders and with p > 0.05 in the DS data. There were 58 such MP terms in total.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>asthma p-value</th>
<th>BPD p-value</th>
<th>COPD p-value</th>
<th>DS p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GO Biological Process Terms:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell Surface Receptor Linked Signal Transduction</td>
<td>0.0447</td>
<td>0.0091</td>
<td>0.2122</td>
<td></td>
</tr>
<tr>
<td>Protein Kinase Cascade</td>
<td>0.0058</td>
<td>0.0011</td>
<td>0.0075</td>
<td>0.2083</td>
</tr>
<tr>
<td>Positive Regulation of Response To Stimulus</td>
<td>0.02</td>
<td>0.0017</td>
<td>0.0043</td>
<td>0.1066</td>
</tr>
<tr>
<td>Positive Regulation of Immune System Process</td>
<td>0.0343</td>
<td>0.0019</td>
<td>0.0059</td>
<td>0.1561</td>
</tr>
<tr>
<td>Regulation of Immune Response</td>
<td>0.02</td>
<td>0.0025</td>
<td>0.0023</td>
<td>0.0841</td>
</tr>
<tr>
<td>Positive Regulation of Immune Response</td>
<td>0.0174</td>
<td>0.0028</td>
<td>0.0024</td>
<td>0.0711</td>
</tr>
<tr>
<td>Regulation of Immune System Process</td>
<td>0.0443</td>
<td>0.0041</td>
<td>0.0173</td>
<td>0.2604</td>
</tr>
<tr>
<td>Regulation of Response To Stimulus</td>
<td>0.0383</td>
<td>0.0043</td>
<td>0.0198</td>
<td>0.199</td>
</tr>
<tr>
<td>I KappaB Kinase NF KappaB Cascade</td>
<td>0.0021</td>
<td>0.0046</td>
<td>0.0115</td>
<td>0.1327</td>
</tr>
<tr>
<td>Positive Regulation of Multicellular Organismal Process</td>
<td>0.0487</td>
<td>0.005</td>
<td>0.0153</td>
<td>0.2606</td>
</tr>
<tr>
<td>Activation of Immune Response</td>
<td>0.017</td>
<td>0.0074</td>
<td>0.0096</td>
<td>0.1492</td>
</tr>
<tr>
<td>Regulation of Cytokine Production</td>
<td>0.0144</td>
<td>0.019</td>
<td>0.0075</td>
<td>0.0546</td>
</tr>
<tr>
<td>Adaptive Immune Response GO 0002460</td>
<td>0.0221</td>
<td>0.0283</td>
<td>0.0384</td>
<td>0.0623</td>
</tr>
<tr>
<td>Adaptive Immune Response</td>
<td>0.025</td>
<td>0.0289</td>
<td>0.0476</td>
<td>0.0687</td>
</tr>
<tr>
<td>Protein Polyubiquitination</td>
<td>0.0133</td>
<td>0.0435</td>
<td>0.0167</td>
<td>0.0683</td>
</tr>
<tr>
<td><strong>KEGG Pathways:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leishmania Infection</td>
<td>0.0002</td>
<td>0.0032</td>
<td>0.0098</td>
<td>0.1635</td>
</tr>
<tr>
<td>JAKSTAT Signaling Pathway</td>
<td>0.0003</td>
<td>0.0001</td>
<td>0.0097</td>
<td>0.1072</td>
</tr>
<tr>
<td>B Cell Receptor Signaling Pathway</td>
<td>0.0003</td>
<td>0.0004</td>
<td>0.0091</td>
<td>0.0662</td>
</tr>
<tr>
<td>FC Epsilon RI Signaling Pathway</td>
<td>0.0004</td>
<td>0.0006</td>
<td>0.0236</td>
<td>0.1027</td>
</tr>
<tr>
<td>Natural Killer Cell Mediated Cytotoxicity</td>
<td>0.0058</td>
<td>0.001</td>
<td>0.0473</td>
<td>0.5846</td>
</tr>
<tr>
<td>Chemokine Signaling Pathway</td>
<td>0.0065</td>
<td>0.0025</td>
<td>0.0476</td>
<td>0.463</td>
</tr>
<tr>
<td>Toll Like Receptor Signaling Pathway</td>
<td>0.0136</td>
<td>0.0094</td>
<td>0.0294</td>
<td>0.247</td>
</tr>
<tr>
<td>Acute Myeloid Leukemia</td>
<td>0.0175</td>
<td>0</td>
<td>0.0158</td>
<td>0.0687</td>
</tr>
<tr>
<td>T Cell Receptor Signaling Pathway</td>
<td>0.0275</td>
<td>0</td>
<td>0.0206</td>
<td>0.0574</td>
</tr>
</tbody>
</table>
Table 3.
Significance of relationships between \( p_{\text{cent}} \) and \( p_{\text{med}} \). The first set of numbers reflects the raw \( p \)-values reported by the glm() function in R showing that a regression line fit to the probabilities of having a \( p_{\text{med}} < 0.05 \) for each decile of \( p_{\text{cent}} \) values has a significantly negative slope. The second set shows the raw \( p \)-values for Wilcoxon tests comparing the \( p_{\text{med}} \) values in just the first decile to those in the remaining deciles combined. Raw \( p \)-values below 0.05 appear in bold.

<table>
<thead>
<tr>
<th>p-value of slope</th>
<th>BP</th>
<th>KEGG</th>
<th>MP</th>
</tr>
</thead>
<tbody>
<tr>
<td>asthma</td>
<td>0.034</td>
<td>0.014</td>
<td>0.002</td>
</tr>
<tr>
<td>BPD</td>
<td>0.696</td>
<td>0.032</td>
<td>0.579</td>
</tr>
<tr>
<td>COPD</td>
<td>0.607</td>
<td>0.001</td>
<td>0.017</td>
</tr>
<tr>
<td>DS</td>
<td>0.86</td>
<td>0.893</td>
<td>0.158</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wilcoxon p-value</th>
<th>BP</th>
<th>KEGG</th>
<th>MP</th>
</tr>
</thead>
<tbody>
<tr>
<td>asthma</td>
<td>4.94E-04</td>
<td>4.56E-17</td>
<td>1.41E-08</td>
</tr>
<tr>
<td>BPD</td>
<td>0.047</td>
<td>2.21E-17</td>
<td>4.10E-20</td>
</tr>
<tr>
<td>COPD</td>
<td>7.89E-07</td>
<td>1.49E-15</td>
<td>1.34E-12</td>
</tr>
<tr>
<td>DS</td>
<td>3.27E-07</td>
<td>0.632</td>
<td>1.73E-09</td>
</tr>
</tbody>
</table>

Supplemental Figure Legends

Figure S1.

**Venn diagrams for disease gene sets and differentially expressed genes.**

A) Overlap between disease genes for three pulmonary diseases and genes on chromosome 21. B) Overlap between differentially expressed genes in three pulmonary diseases and Down syndrome.

Figure S2.
R-plots showing that pathways with low \( p_{\text{cent}} \) also have low \( p_{\text{med}} \) with linear regression line. The plots comparable to that in Figure 4 for all combinations of disease and pathway collections are included here.

**Methods and Resources**

*Disease-related genes*

Disease genes for Asthma and COPD were collected from recent reviews, (Bijanzadeh et al., 2011) and (Bosse, 2012), respectively. A set of 361 Chromosome 21 genes was downloaded from the Molecular Signature DataBase (MSigDB) (Liberzon et al., 2011). We collected genes associated with BPD from Online Mendelian Inheritance in Man (OMIM) (Amberger et al., 2009) and Genopedia (Yu et al., 2010), as described in (Park et al., 2014). All datasets were downloaded on February 3, 2016.

Microarray gene expression profiles for Asthma and BPD were downloaded from the GEO database (accession numbers GSE4302 and GSE32472, respectively). The first measured differential expression in airway epithelial cells between healthy controls and asthma patients (Woodruff et al., 2007), while the second examined expression in peripheral blood cells from infants born preterm with or without BPD (Pietrzyk et al., 2013). From this study, we used only samples taken on postnatal day 5. We selected as differentially expressed genes between disease and control groups those with an adjusted Benjamini-Hochberg t-test p-value below 0.005, yielding 52 and 274 expression-related genes in asthma and BPD, respectively.

For COPD, we downloaded RNA-seq EdgeR results comparing expression in lung cells from COPD patients and controls (GSE57148) (Kim et al., 2015). We identified 266 significantly expressed genes with an EdgeR q-value below \( 10^{-10} \). For the control case, we chose a prior study from our group (GSE16176) (Slonim et al., 2009) that compared gene expression in amniotic fluid from second trimester fetuses with Down syndrome to those of age- and sex-matched controls. Using the same criteria as for the other Affymetrix microarray studies (all three of which used the U133 Plus 2.0 arrays) but paired t-tests to deal with the matched samples identified 129 differentially expressed genes in Down syndrome.
Note that there may be disease-associated genes that are also differentially expressed in that disease. To avoid confusion about how to use these in computing pathway centrality, we removed genes from the disease gene sets that also appeared in the corresponding set of differentially expressed genes. The resulting disease gene sets contained 107 (asthma), 73 (BPD), 181 (COPD), and 360 (DS). Supplementary Figure S1 shows the overlaps between these four disease gene sets and the four differentially-expressed gene sets.

Functional pathways and gene sets

Both the Gene Ontology and the KEGG gene set collections were downloaded from the Molecular Signature DataBase (MSigDB) (Liberzon et al., 2011) on August 18, 2015 (http://software.broadinstitute.org/gsea/msigdb). This GO gene set collection includes 825 biological process terms and 6,178 genes, and the KEGG collection includes 186 pathways and 5,267 genes. In addition, we created a collection of gene sets using data from the Mouse Genome Informatics site (http://www.informatics.jax.org), which includes a database characterizing mutant mouse strains with respect to the Mammalian Phenotype Ontology (Blake et al., 2017). To create gene sets with mammalian phenotype (MP) data, we downloaded the files HMD_HumanPhenotype.rpt and MGI_GenePheno.rpt on February 21, 2016. The first file includes human / mouse orthology mapping and some MP data, while the second includes more MP data characterizing mouse gene mutant strains (other than conditional mutations), but no orthology information. We therefore used the human-mouse orthologs from the first file and then combined the phenotypes from both files, excluding the two normal phenotype labels MP:0002169 ("no abnormal phenotype detected") and MP:0002873 ("normal phenotype"). We then created for each phenotype a list of genes with at least one mutant allele with that phenotype. These lists were used as functional gene sets, which we describe as the "mammalian phenotype" gene set collection. There are 8,225 phenotypes in the mammalian phenotype data and 7,778 genes.

Protein-protein and genetic interaction networks

We use two biological networks in our experiments. To measure pathway
centrality, physical protein-protein interactions were collected from the Human Integrated Protein-Protein Interaction rEference (HIPPIE) (Alanis-Lobato et al., 2017) database. HIPPIE contains experimentally verified protein interactions with confidence scores. We downloaded the interaction data (version 2) on July 18, 2016 and selected only those interactions described as “high confidence” (≥ 0.73), as these interactions are supported by more reliable evidence. We worked with the largest connected component extracted from the network, which contains 43,475 interactions between 9,379 proteins.

To compute $p_{med}$ scores based on genetic interaction data, we looked for genetic interaction data featuring alleviating (positive) genetic and phenotypic suppression interactions. Because relatively few of these types of genetic interactions are known for humans, we collected such genetic interactions from *Saccharomyces cerevisiae*, *Drosophila melanogaster*, and *Caenorhabditis elegans*. These interactions came from BioGRID (Stark et al., 2006) (version 3.4.141, downloaded October 14, 2016); the Saccharomyces Genome Database (SGD project, http://www.yeastgenome.org; downloaded October 10, 2016), and Flybase (Attrill et al., 2016) (downloaded August 15, 2016). To find human homologous interaction pairs, we use a mapping downloaded from the HomoloGene database (NCBI Resource Coordinators, 2016) on July 19, 2016. This approach gave us 4,536 pairs of putative positive genetic interactions in humans.

Identification of disease-mediating pathways

Pathway centrality measures the amount of information a set of pathway genes handles by counting paths linking disease genes and differentially expressed genes. While the classical definition of fractional betweenness (FB) for a node $v$ is the fraction of shortest paths between all pairs of nodes in a network passing through $v$, our pathway centrality (PC) score will be based on a modified FB score ($FB'$), which for node $v$, only reflects the shortest paths between disease genes, $D(d)$, and differentially expressed genes, $E(d)$, that passes through $v$. Formally, this is defined as:

$$FB'(v) = \sum_{s \in V, s \in D(d) \mid s \neq v, t \in E(d) \mid t \neq v} \frac{B_{s,t}(v)}{B_{s,t}} \quad (1)$$

where $V$ is the set of all vertices in the protein-protein interaction network, $D(d)$ is the set of genes in $V$ associated with disease $d$, $E(d)$ is the set of differentially expressed genes in
V for disease $d$, $B_{s,t}$ is the total number of shortest paths between $s$ and $t$, and $B_{s,t}(v)$ is the total number of shortest paths between $s$ and $t$ that pass through $v$.

We define **pathway centrality** as the average fractional betweenness score across all genes in a pathway, counting only the shortest paths from disease genes to differentially expressed genes. Specifically, for a pathway $k$ containing the gene set $P(k)$, pathway centrality($k$) is defined as:

$$PC(k) = \frac{\sum_{v \in P(k)} FB'(v)}{|P(k)|}$$  \hspace{1cm} (2)

**Assessing significance of identified mediating pathways: $p_{cent}$**

The significance of the observed pathway centrality score of a pathway or gene set $S$ is assessed using a null distribution learned from a permutation with 10,000 random gene sets of size $|S|$. We further require that the random gene sets be drawn from a collection of genes satisfying two conditions: they should belong to at least one pathway, and should have degree $\geq 2$. The reason for these restrictions is to reduce biases that artificially reduce the $p_{cent}$ values, making the majority of pathways appear to be significantly central. Pathway genes are known to be relatively central in protein interaction networks, and therefore, it is likely that pathway genes have higher fractional betweenness than those that are not involved in well-annotated functional processes.

Further, to the extent that protein interaction networks are scale-free, random samples will likely be enriched for genes of degree 1 if we do not somehow force the random sampling process to select higher degree genes. But this is problematic because the distribution of pathway centrality from permutation tests may then be left-skewed, given that the various centrality measurements tend to be correlated (Goh et al., 2003). Thus, to systematically exclude low-degree genes from permutation, we decided to use only genes in the 2-core of the protein-protein interaction network. A k-core of a network is a maximal group of nodes, all of which are connected to at least $k$ other nodes in the network (Bhawalkar et al., 2012). This approach improves the observed distribution of $p_{cent}$ values. It may also reduce the significance of pathways containing mostly low-degree genes, but this applies to relatively few of the gene sets in the considered collections.
Evaluation of identified disease-mediating pathways

We compute $p_{med}$ by counting the number of phenotypic alleviation or suppression interactions known to exist between disease genes and the mediators. To assess how surprising it is to see the observed number of such interactions, we compute a null distribution of such counts between the same set of disease genes and 10,000 random gene sets of equal size as the mediating pathway. Again we impose restrictions on the source of random genes to avoid over-estimating significance. Here, random gene sets are drawn from a pool of genes that belong to at least one pathway in the collection and that are downstream genes of any alleviating genetic or phenotypic suppression interactions.

Supplemental Information

We are providing one excel file containing our full experimental results, two pdf files containing two Venn diagrams and nine R plots, and three cytoscape session files.
Figure 1
Disease genes
Genes in Significant Mediating Pathway

Alleviating genetic interaction
Alleviating genetic interaction: disease genes -> pathway genes
Alleviating genetic interaction: disease genes -> randomly selected genes
Figure 4

The graph illustrates the relationship between the probability of $p_{med} < 0.05$ and the deciles of $p_{cent}$. The plot shows a clear downward trend as the deciles of $p_{cent}$ increase, indicating a higher probability of $p_{med} < 0.05$ in lower deciles of $p_{cent}$. The blue line represents a linear fit to the data, highlighting the trend.