bioXplore: Interactive Exploration of the Gene–Disease Association Network

John Rachlin

Khoury College of Computer Sciences, Northeastern University, 360 Huntington Avenue, Boston, MA 02115, USA

Correspondence: j.rachlin@northeastern.edu

ABSTRACT

Integrative resources now assemble gene–disease associations from genetic, functional, pharmacological and literature-derived evidence, but their scale defeats visual interpretation: a single common disease returns hundreds of genes and thousands of ontology terms, and the resulting network is a hairball. bioXplore is a browser-based platform that renders such data as *contextual subnetworks* — compact, anchored views in which every element is selected by a stated rule and every view declares what it dropped. Built on Open Targets Platform release 26.06 (7,554,123 associations, 26,517 diseases, 20,097 protein-coding genes), it ranks by hypergeometric surprise rather than by degree, filters ontology terms by information content, and removes an anchor's own ontological relatives before ranking. Its distinguishing feature is an *evidence lens* that recomputes the association score under a user-selected subset of the 20 evidence sources rather than filtering rows, exactly reproducing the published scoring function. Applied to asthma, this separates disease biology from pharmacological circularity and surfaces IL6R antagonists approved for other indications as repurposing candidates. bioXplore runs entirely client-side over a 4.8 MB columnar extract and is available at https://bioxplore.org.

INTRODUCTION

Gene–disease association resources have followed a consistent trajectory: from manually curated collections such as the Genetic Association Database (1) to integrative platforms that merge genetic, functional, pharmacological and literature-derived evidence into a single score, of which DisGeNET (2) and the Open Targets Platform (3–4) are the most widely used. The integration problem is largely solved. The *interpretation* problem is not.

A researcher who asks what is known about a disease receives hundreds to thousands of genes, each carrying a score assembled from heterogeneous evidence, and each gene in turn annotated to dozens of Gene Ontology terms (5–6). Rendered as a network, the result is a hairball: a drawing that is faithful, complete and unreadable. Figure 1 shows the effect for a single common disease.

bioXplore takes the position that the useful unit of presentation is not the network but a *contextual subnetwork*: a small view assembled around one anchor — a disease, a gene or a biological process — in which every element earned its place by a stated rule, and in which the view reports what it excluded. The platform is a static web application; all queries execute in the user's browser against columnar extracts, so there is no server-side query engine to maintain and no user query leaves the client.

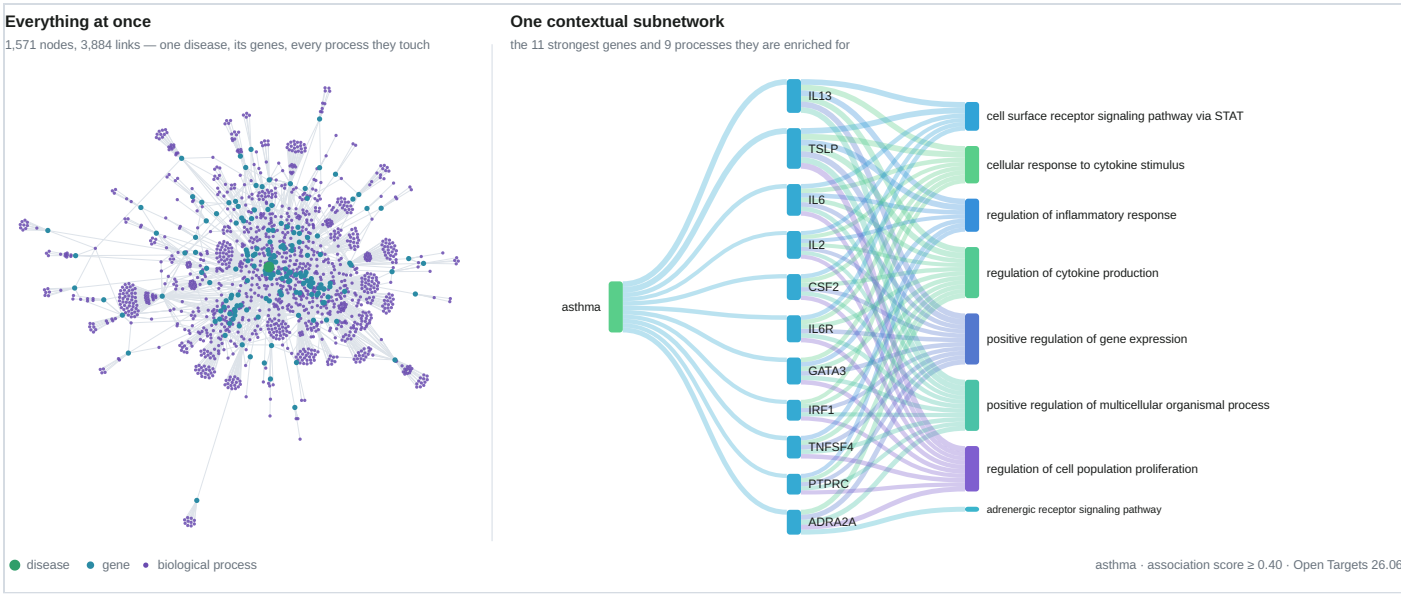

Figure 1. The same disease rendered two ways. *Left*: every node a naïve join returns for asthma — the disease, its 167 associated genes and every biological process any of them is annotated to, 1,571 nodes and 3,884 links. *Right*: the contextual subnetwork bioXplore draws from the same data at the same evidence threshold, after information-content filtering, redundancy pruning and ranking by hypergeometric surprise.

MATERIALS AND METHODS

Source data

bioXplore integrates one resource directly, the Open Targets Platform release 26.06 (3), which is distributed under CC0 and is itself an integration of approximately thirty upstream databases including Ensembl (7), the Gene Ontology (5), EFO and MONDO (8–9), ChEMBL (10), ClinVar (11), ClinGen (12), Genomics England PanelApp (13), UniProt (14), COSMIC (15), Reactome (16), Expression Atlas (17), the International Mouse Phenotyping Consortium (18) and Europe PMC (19). Licence was a hard selection criterion, since all derived data are republished; DisGeNET, on which an earlier prototype of this work was based, moved to a commercial licence and was therefore excluded.

The catalog holds 7,554,123 scored disease–gene associations over 26,517 diseases and 20,097 protein-coding genes, together with 641,624 Gene Ontology annotations, 22,407 drug molecules, 8,991 drug–target and 86,468 drug–indication links. It is built into a single 1.1 GB DuckDB file (20) by a three-stage pipeline (fetch, build, validate) that is a deterministic function of release, seeds and code.

Ranking by surprise

Three measured properties of the joined data make a naïve rendering unreadable, and each has a corresponding mechanism.

First, **generic annotations dominate**. The term `protein binding` is annotated to 13,295 of 20,097 protein-coding genes. Every Gene Ontology column is therefore filtered by *information content*, the negative logarithm of the fraction of genes carrying a term (21–22); terms true of nearly everything score near zero and are removed without a curated stop-list. Annotations are additionally restricted to experimental, high-throughput, author-statement and curator evidence codes (71,184 of 177,389 biological-process annotations, 40%), excluding electronically inferred and sequence-similarity codes, which encode expectation rather than observation.

Second, **ontological neighbours masquerade as findings**. Diseases sharing genes with an anchor are dominated by that anchor's own ancestors and descendants. bioXplore computes the transitive closure of the disease hierarchy and removes the anchor's relatives before ranking. The Gene Ontology closure is likewise recomputed over `is_a` and `part_of` edges rather than taken from the published ancestor arrays, which are not that closure: 13,370 of 48,321 terms differ, and the discrepancy inflates `bleb assembly` from 7 annotated genes to 1,541, converting a highly specific term into one that scores as generic.

Third, **counting shared genes ranks noise first**. Overlap size is confounded by set size, so every co-occurrence and enrichment result is ranked by the hypergeometric tail probability, in the manner long established for gene-set analysis (23). For asthma, *neurodegenerative disease* shares 34 genes and *chronic bronchitis* 14; the former is 0.8× chance expectation and the latter 62×. Sorting by overlap size puts the noise first.

The evidence lens

The Open Targets score is a weighted harmonic sum of per-source scores normalised by $\pi^2/6$. bioXplore reproduces this arithmetic exactly from the per-source table (mean absolute error 3.3×10^{-5} across all 7,554,123 edges), which permits an operation we call the *evidence lens*: the user disables a category of evidence and the score is **recomputed** rather than filtered. The resulting view is the one that would have obtained had that evidence never existed. This matters because 98.6% of associations rest on a single evidence type, and because the types are not interchangeable — human genetic

evidence is fixed at conception and cannot be a consequence of disease, and is the category most predictive of clinical success (24–25), whereas text-mined co-occurrence asserts only that two terms appeared near one another.

RESULTS

The web interface

The application (Figure 2) renders three-column Sankey diagrams (26) with captioned columns. Three anchor types are supported: a **disease** resolves to genes and then to biological processes or drugs; a **gene** to diseases and then therapeutic areas; a **biological process** to genes and then diseases. Double-clicking any node re-anchors the view, so a user can walk the graph without returning to a search box. Every view reports its exclusions, and any view can be exported as PNG or SVG with a caption carrying its parameters, or shared as a URL that encodes the anchor, thresholds and evidence lens.

Queries run in the browser via DuckDB-WASM (27) over Parquet extracts served through HTTP range requests; the complete data payload is 4.8 MB. The consequence is that bioXplore has no backend: it can be hosted as static files, and neither the anchor a user selects nor the thresholds they choose are observable to the server.

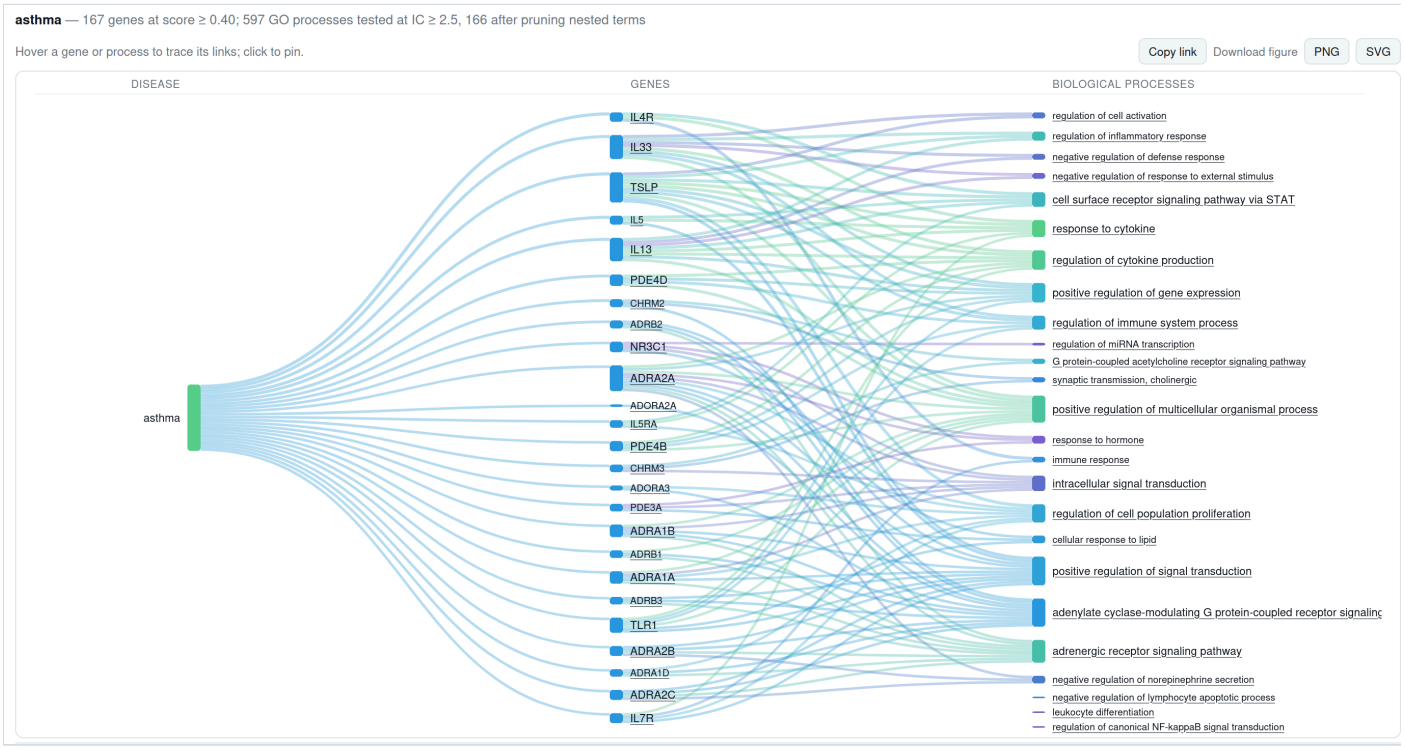

Figure 2. The bioXplore explorer, anchored on asthma. Columns are captioned; ribbon width encodes the association score; genes are ordered by score. The status line reports what was tested and what survived selection. Nodes can be pinned to trace their links, or double-clicked to become the new anchor.

Case study: asthma

Anchoring on asthma at an association score ≥ 0.4 returns 167 genes. The strongest are *FLG*, *IL4R*, *ADORA1*, *IL33*, *TSLP*, *IL5*. This is immediately interpretable: *IL4R*, *IL13*, *IL33* and *TSLP* are the type-2 inflammatory axis, and *FLG* encodes filaggrin, the epidermal barrier protein whose loss-of-function variants underlie the atopic march (28). These are the loci that large-scale asthma GWAS have converged on (29–30), recovered here without being asked for.

The process column resolves the same gene set into mechanism. The strongest enrichments are *cellular response to cytokine stimulus* (6.0×), *regulation of cytokine production* (5.2×), *positive regulation of multicellular organismal process* (3.9×), tested against a background of 8,907 disease-associated genes rather than the genome, so that the well-studied-gene bias of the underlying resources is not reported as biology. Two further terms are diagnostic of a different phenomenon: *adrenergic receptor signaling pathway* (36×) and *G protein-coupled acetylcholine receptor signaling pathway*. These are not asthma pathogenesis. They are asthma *pharmacology* — the β_2-agonist and antimuscarinic targets — arriving because drug evidence contributed to the selection of the genes.

Switching the lens to exclude drug evidence and recomputing shows the size of the effect. The gene set falls from 167 to 135, the strongest genes become *FLG*, *IL33*, *IL13*, *TLR1*, *IL7R*, *IL4R*, and the adrenergic and cholinergic terms leave the ranking entirely. The neighbourhood shifts with it: with all evidence, asthma's nearest disease neighbours are allergic rhinitis (46 shared genes, 57× expectation) and other atopic conditions; with drug evidence removed, Crohn disease enters the ranking (15 shared genes, 20×), and the picture becomes autoimmune rather than airway-specific.

Neither view is the correct one. They answer different questions — *what is known about asthma* versus *what causes asthma* — and the platform's contribution is to make the difference explicit and reversible in one click.

From subnetwork to hypothesis

The drug column makes the circularity operational. With all evidence enabled, asthma's genes reach 653 drugs of which 129 are already indicated for asthma, and the column is dominated by *NR3C1*, *ADRB2* and *CHRM3* — inhaled corticosteroids, salbutamol and ipratropium. The view restates its own premise. With drug evidence excluded the same query returns 179 drugs, of which 24 are indicated for asthma; the remainder are approved or trialled elsewhere and become candidates. Among them, four IL6R antagonists — tocilizumab, sarilumab, satralizumab and levilimab — are indicated for rheumatoid arthritis and neuromyelitis optica but not for asthma, and reach the view because *IL6R* survives at a score of 0.56 on non-pharmacological evidence alone.

This is the kind of hypothesis bioXplore is built to surface, and it is not idle: the *IL6R* Asp358Ala variant has been associated with lung function in asthma (31), and IL-6 receptor blockade has been examined experimentally in allergen challenge (32). bioXplore does not evaluate the hypothesis. It locates it, states the evidence it rests on, and reports what had to be switched off for it to become visible. Because the node budget is split between drugs already indicated for the anchor and drugs indicated elsewhere, the repurposing candidates cannot be crowded out by the 129 drugs the disease already has.

DISCUSSION

Selection does more work than the biology. Raising the association threshold from 0.1 to 0.4 leaves 493 diseases that had twenty or more genes with none at all; Huntington disease shows 111 associated genes at the looser setting and 4 at the stricter one. Any tool that silently picks one threshold is making the principal analytical decision on the user's behalf. bioXplore therefore states, on every view, how many items were dropped and why — including those that could never have appeared, such as genes with no functional annotation or no known drug.

Limitations. The platform reports associations, not mechanisms, and cannot supply the direction of an effect. It inherits the biases of the underlying resources, in which a well-studied gene appears important partly because it was studied; only 1,547 of 18,902 disease-associated genes carry any drug, and curated literature references are concentrated in rare and monogenic disease because the text-mining evidence table was not ingested. Absence of evidence here is frequently absence of attention. Finally, the platform is a reading of a public database rather than a source of new observations, and the interpretations offered above are hypotheses for experimental follow-up.

AVAILABILITY AND REQUIREMENTS

bioXplore is freely available at **https://bioxplore.org** with no registration or login. It requires only a modern browser with WebAssembly support; no software installation is needed and no data are transmitted to the server. All content is derived from the Open Targets Platform release 26.06 under CC0 1.0.

ACKNOWLEDGEMENTS

This work relies entirely on the Open Targets Platform and the upstream resources it integrates. The author thanks those projects for publishing their data under terms that permit redistribution. The software was developed with substantial assistance from a large language model; all design decisions, verification and interpretation are the author's responsibility.

CONFLICT OF INTEREST

None declared.

REFERENCES

1. Becker,K.G., Barnes,K.C., Bright,T.J. and Wang,S.A. (2004) The Genetic Association Database. *Nat. Genet.*, **36**, 431–432.
2. Piñero,J., Ramírez-Anguita,J.M., Saüch-Pitarch,J. et al. (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.*, **48**, D845–D855.
3. Buniello,A., Suveges,D., Cruz-Castillo,C. et al. (2025) Open Targets Platform: facilitating therapeutic hypotheses building in drug discovery. *Nucleic Acids Res.*, **53**, D1467–D1475.
4. Ochoa,D., Hercules,A., Carmona,M. et al. (2023) The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. *Nucleic Acids Res.*, **51**, D1353–D1359.
5. The Gene Ontology Consortium (2023) The Gene Ontology knowledgebase in 2023. *Genetics*, **224**, iyad031.
6. Ashburner,M., Ball,C.A., Blake,J.A. et al. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
7. Harrison,P.W., Amode,M.R., Austine-Orimoloye,O. et al. (2024) Ensembl 2024. *Nucleic Acids Res.*, **52**, D891–D899.
8. Malone,J., Holloway,E., Adamusiak,T. et al. (2010) Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, **26**, 1112–1118.
9. Vasilevsky,N.A., Matentzoglu,N.A., Toro,S. et al. (2022) Mondo: unifying diseases for the world, by the world. *medRxiv*, 2022.04.13.22273750.
10. Zdrazil,B., Felix,E., Hunter,F. et al. (2024) The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.*, **52**, D1180–D1192.
11. Landrum,M.J., Chitipiralla,S., Brown,G.R. et al. (2020) ClinVar: improvements to accessing data. *Nucleic Acids Res.*, **48**, D835–D844.
12. Strande,N.T., Riggs,E.R., Buchanan,A.H. et al. (2017) Evaluating the clinical validity of gene–disease associations: an evidence-based framework developed by the Clinical Genome Resource. *Am. J. Hum. Genet.*, **100**, 895–906.
13. Martin,A.R., Williams,E., Foulger,R.E. et al. (2019) PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat. Genet.*, **51**, 1560–1565.
14. The UniProt Consortium (2025) UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.*, **53**, D609–D617.
15. Sondka,Z., Bamford,S., Cole,C.G. et al. (2018) The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer*, **18**, 696–705.

16. Milacic,M., Beavers,D., Conley,P. et al. (2024) The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Res.*, **52**, D672–D678.
17. George,N., Fexova,S., Fuentes,A.M. et al. (2024) Expression Atlas update: insights from sequencing data at both bulk and single cell level. *Nucleic Acids Res.*, **52**, D107–D114.
18. Groza,T., Gomez,F.L., Mashhadi,H.H. et al. (2023) The International Mouse Phenotyping Consortium: comprehensive knockout phenotyping underpinning the study of human disease. *Nucleic Acids Res.*, **51**, D1038–D1045.
19. Ferreira,R., Rogoyski,O.M., Talavera-López,C. et al. (2023) Europe PMC in 2023. *Nucleic Acids Res.*, **51**, D1552–D1560.
20. Raasveldt,M. and Mühleisen,H. (2019) DuckDB: an embeddable analytical database. In *Proc. 2019 Int. Conf. on Management of Data (SIGMOD)*, pp. 1981–1984.
21. Resnik,P. (1995) Using information content to evaluate semantic similarity in a taxonomy. In *Proc. 14th Int. Joint Conf. on Artificial Intelligence*, pp. 448–453.
22. Lord,P.W., Stevens,R.D., Brass,A. and Goble,C.A. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.
23. Subramanian,A., Tamayo,P., Mootha,V.K. et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
24. Nelson,M.R., Tipney,H., Painter,J.L. et al. (2015) The support of human genetic evidence for approved drug indications. *Nat. Genet.*, **47**, 856–860.
25. King,E.A., Davis,J.W. and Degner,J.F. (2019) Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet.*, **15**, e1008489.
26. Riehmann,P., Hanfler,M. and Froehlich,B. (2005) Interactive Sankey diagrams. In *Proc. IEEE Symp. on Information Visualization (InfoVis)*, pp. 233–240.
27. Kohn,A., Moritz,D., Raasveldt,M., Mühleisen,H. and Neumann,T. (2022) DuckDB-Wasm: fast analytical processing for the web. *Proc. VLDB Endow.*, **15**, 3574–3577.
28. Palmer,C.N.A., Irvine,A.D., Terron-Kwiatkowski,A. et al. (2006) Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis. *Nat. Genet.*, **38**, 441–446.
29. Moffatt,M.F., Gut,I.G., Demenais,F. et al. (2010) A large-scale, consortium-based genomewide association study of asthma. *N. Engl. J. Med.*, **363**, 1211–1221.
30. Demenais,F., Margaritte-Jeannin,P., Barnes,K.C. et al. (2018) Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat. Genet.*, **50**, 42–53.
31. Hawkins,G.A., Robinson,M.B., Hastie,A.T. et al. (2012) The IL6R variation $Asp^{358}Ala$ is a potential modifier of lung function in subjects with asthma. *J. Allergy Clin. Immunol.*, **130**, 510–515.
32. Revez,J.A., Bain,L.M., Chapman,B. et al. (2019) Effects of interleukin-6 receptor blockade on allergen-induced airway responses in mild asthmatics. *Clin. Transl. Immunology*, **8**, e1044.