The Kidney-Genetics Documentation

Bernt Popp, Nina Rank, Constantin Wolff, Jan Halbritter

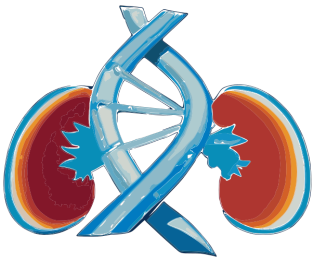
2023-10-04

Contents

Preface

This documentation is intended to describe the Kidney-Genetics[1] project.

[1] https://github.com/halbritter-lab/kidney-genetics



Objective

How can we address the lack of a unified and standardized database of kidney disease-associated genes, which hampers diagnosis, treatment, and research comparability in the field of kidney diseases?

Genetic insights are becoming increasingly influential in the understanding and treatment of various kidney diseases (KD). Hundreds of genes associated with monogenic kidney disease have been identified, providing valuable insights into their diagnosis, management, and monitoring. However, the lack of a unified and standardized database of genes assigned to kidney diseases has led to diagnostic blind spots and comparability issues among current studies of kidney genetics. To address this gap, we created the **"Kidney-Genetics"** a regularly updated, automated and publicly accessible database which aims to provide a comprehensive list of all relevant genes associated with kidney disease.

Key issues:

- Create a unified and standardized database of kidney disease-associated genes and provide a valuable resource for the diagnosis, treatment, and monitoring of those diseases
- Allow clinicians and researchers to gain a deeper understanding of the genetic factors underlying different KDs
- Compile, organize and curate important information on the genes to the identify novel candidate genes and genetic variants associated with KDs
- Group and sort the genes into different categories, for example into phenotypic groups, the onset, syndromic, etc.
- Establish genotype-phenotype correlations that can be used to assign multiple clinical entities to a single gene in order to improve understanding and treatment choices
- The information can be used to develop personalized treatment strategies and interventions, leading to more effective and targeted therapies for individuals with KD
- Researchers can freely access "Kidney-Genetics" ensuring consistency and comparability across different research projects, which can accelerate scientific progress, foster collaborations, and facilitate the development of new insights and approaches

The scientific literature highlights the need for such a database and emphasizes the importance of genetic research in kidney disease (e.g. [Boulogne et al., 2023]).

In summary, our research question and its approach have the potential to provide a deeper scientific understanding of KD genetics, improve diagnostic accuracy, guide treatment selection, advance precision medicine, and facilitate research collaboration. The establishment of the **"Kidney-Genetics"** database addresses an important gap in the field and provides a valuable resource for researchers, clinicians, and patients involved in the discovery and treatment of

KD.

Methods

To create a thorough and standardized database of kidney-related genes, we employed the following methods and compiled kidney disease-associated gene information from various sources:

1. Utilized data from Genomics England and Australia PanelApp [Martin et al., 2019]
2. Conducted a comprehensive literature review of published gene lists
3. Collected information from clinical diagnostic panels for kidney disease
4. Performed a Human Phenotype Ontology (HPO)-based [Köhler et al., 2021]) search in rare disease databases (OMIM)
5. Employed a PubTator [Wei et al., 2013] API-based automated literature extraction from PubMed

We also developed an evidence-scoring system to differentiate highly confirmed disease genes from candidate genes. We defined the presence of a certain gene in 3 or more of the 5 resources as highly evident genes. These genes were then manually curated according to predetermined criteria or, in the case of existing ClinGen curation, their data and scores were used. Genes with a score of 2 or less were accordingly more likely to be classified as candidate genes.
Furthermore, we grouped all genes into different categories to later match them in a genotype-phenotype correlation.

To get a more transparent and thus more comprehensive understanding of our several evidence source "pillars", we listed our different steps below and attached a flowchart for better visualization.

1. We retrieved all kidney disease related panels from both PanelApp UK and PanelApp Australia, meaning all panels that include "renal" or "kidney" in its name. That included xxx different lists. The access date was the xxx.
2. We identified Genes associated with kidney disease in a systematic Literature search using the following search query:
   (1) "Kidney"[Mesh] OR "Kidney Diseases"[Mesh] OR kidney OR renal AND
   (2) "Genetic Structures"[Mesh] OR "Genes"[Mesh] OR genetic test OR gene panel OR gene panels OR multigene panel OR targeted panel*
   we then screened for published lists and got xxx lists from date to date xxx.
3. We used ten common diagnostic panels that can be purchased for genome analysis and extracted the screened genes from them. Those included following panels:
   - Centogene nephrology
   - Cegat kidney diseases
   - Preventiongenetics
   - etc.
4. We used common databases (e.g. OMIM) for rare diseases and screened them for kidney disease associated Genes from a Human Phenotype Ontology (HPO) based search query. The most comprehensive HPO term used was "Abnormality of the upper urinary tract" (HP:0010935) and included all subgroup terms. We deliberately chose these to be somewhat broader in order to fully include all relevant kidney diseases such as CAKUT, among others.
5. We retrieved all kidney disease associated genes from a PubTator API-based automated literature extraction of publications available on PubMed.

Kidney-Genetics Flowchart

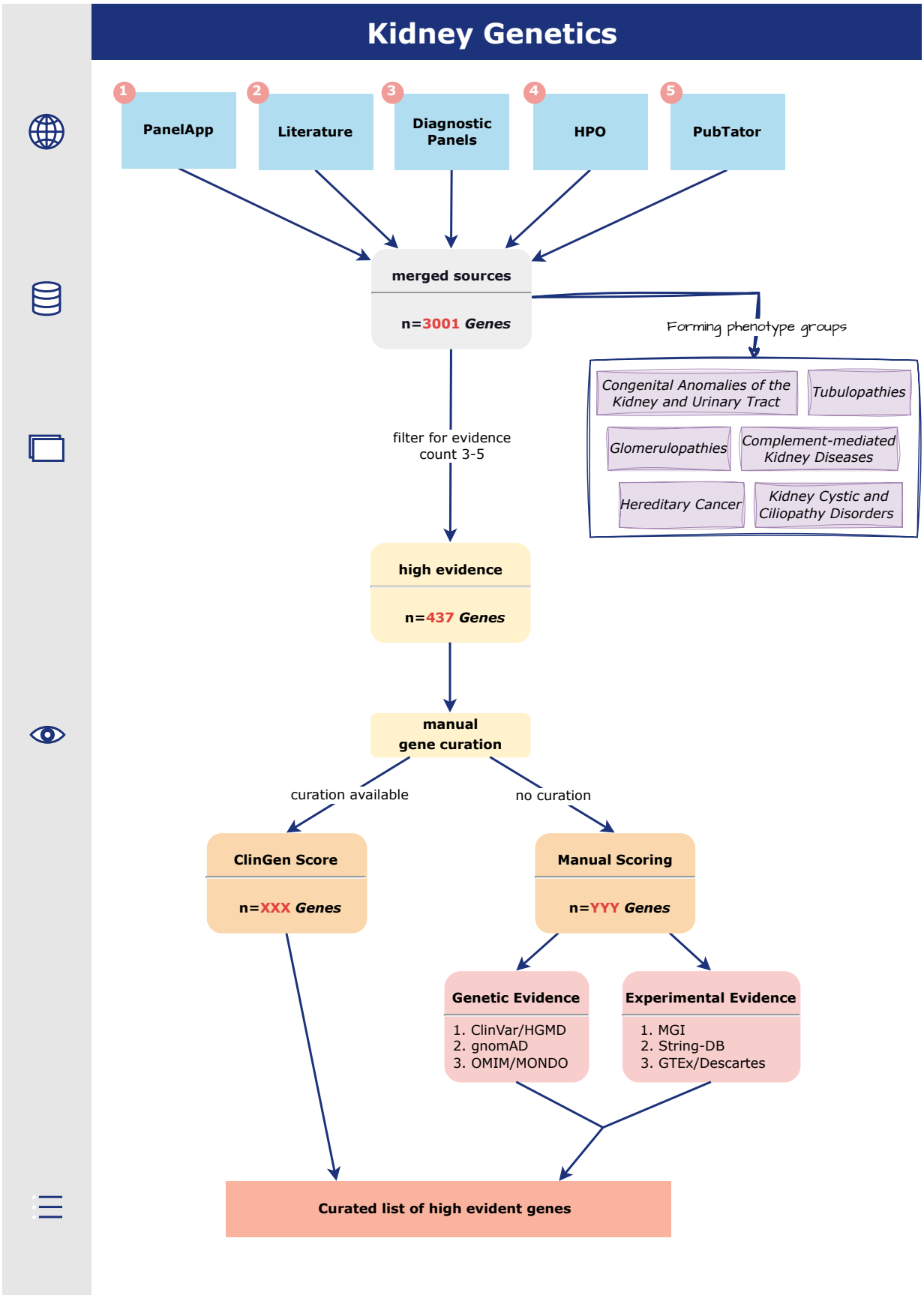


Figure 1: (#fig:curation_flow_diagram)Curation process flow diagram

Results

The "Kidney-Genetics" database currently contains detailed information on 2.906 kidney-associated genes with detailed annotations on gene function, kidney phenotype, incidence, possible syndromic disease expression and genetic variation. To automatically group the genes, we will present the results of phenotypic and functional clustering.

The number of genes extracted from the five analyzed sources of information is as follows: (1) 550, (2) 822, (3) 936, (4) 791, and (5) 2133
Notably, **437** genes (14.6%) of the **total 3001** genes are present in three or more of the analyzed information sources, thus meeting our evidence criteria, indicating high confidence and their potential for diagnostic use. Of these high evidence genes, **423** (96.8%) are present in at least one, and **56** (12.8%) are present in all 10 comprehensive diagnostic laboratory panels.

To ensure currency, Kidney-Genetics will be updated regularly and automatically at XXX week intervals. We will also provide phenotypic and functional clustering results to facilitate gene grouping.

Conclusion

Kidney-Genetics is a comprehensive, free and publicly accessible database that can be used by researchers to analyze genomic data related to KDs. The database will be routinely updated using an automated system and standardized pipeline to ensure that it is always up-to-date with the latest kidney research and diagnostics.

By utilizing Kidney-Genetics, clinicians, geneticists, and researchers can examine genomic data and improve their understanding of the genetic components of diverse KDs. The code and results are completely available on GitHub. A standardized pipeline and automated system keep our database on the cutting edge of kidney research and diagnostics. Screening efforts toward manual curation (such as through the ClinGen initiative) and assignment of diagnostic genes to nephrologic disease groups (e.g., syndromic vs. isolated; adult vs. pediatric; cystic, nephrotic, etc.) are currently in the development process and our goals for the near future.

Outlook

Future goals include the further manual curation of the high evident genes to acquire a more accurate individual assessment of each gene. For this purpose, we have developed a standardized curation process based on the ClinGen criteria, as previously discussed in the methods section. Furthermore, diagnostic genes will be assigned to certain defined nephrological disease groups, in order to obtain a phenotype-genotype correlation and gain a better clinical understanding.

1 Analyses result tables

1.1 Main table: Merged analyses sources

This table shows the merged results of all analyses files as a wide table with summarized information.

approved_symbol	hgnc_id	evidence_count	list_count	01_PanelApp	02_Literature	03_DiagnosticPanels	04_HPO	05_PubTator
All	All	All	All	All	All	All	All	All

1.2 Result table: PanelApp

This table shows results of the first analysis searching kidney disease associated genes from the PanelApp project in the UK and Australia.

approved_symbol	hgnc_id	gene_name_reported	source	source_count	source_evidence
All	All	All	All	All	All

1.3 Result table: Literature

This table shows results of the second analysis searching kidney disease associated genes from various publications.

approved_symbol	hgnc_id	gene_name_reported	source	source_count	source_evidence
All	All	All	All	All	All

1.4 Result table: Diagnostic panels

This table shows results of the third analysis searching kidney disease associated genes from clinical diagnostic panels for kidney disease.

approved_symbol	hgnc_id	gene_name_reported	source	source_count	source_evidence
All	All	All	All	All	All

1.5 Result table: HPO in rare disease databases

This table shows results of the fourth analysis searching kidney disease associated genes from a Human Phenotype Ontology (HPO)-based search in rare disease databases (OMIM, Orphanet).

approved_symbol	hgnc_id	gene_name_reported	source	source_count	source_evidence
All	All	All	All	All	All

1.6 Result table: PubTator

This table shows results of the fifth analysis searching kidney disease associated genes from a PubTator API-based automated literature extraction from PubMed.

approved_symbol	hgnc_id	gene_name_reported	source	source_count	source_evidence
All	All	All	All	All	All

2 Analyses plots

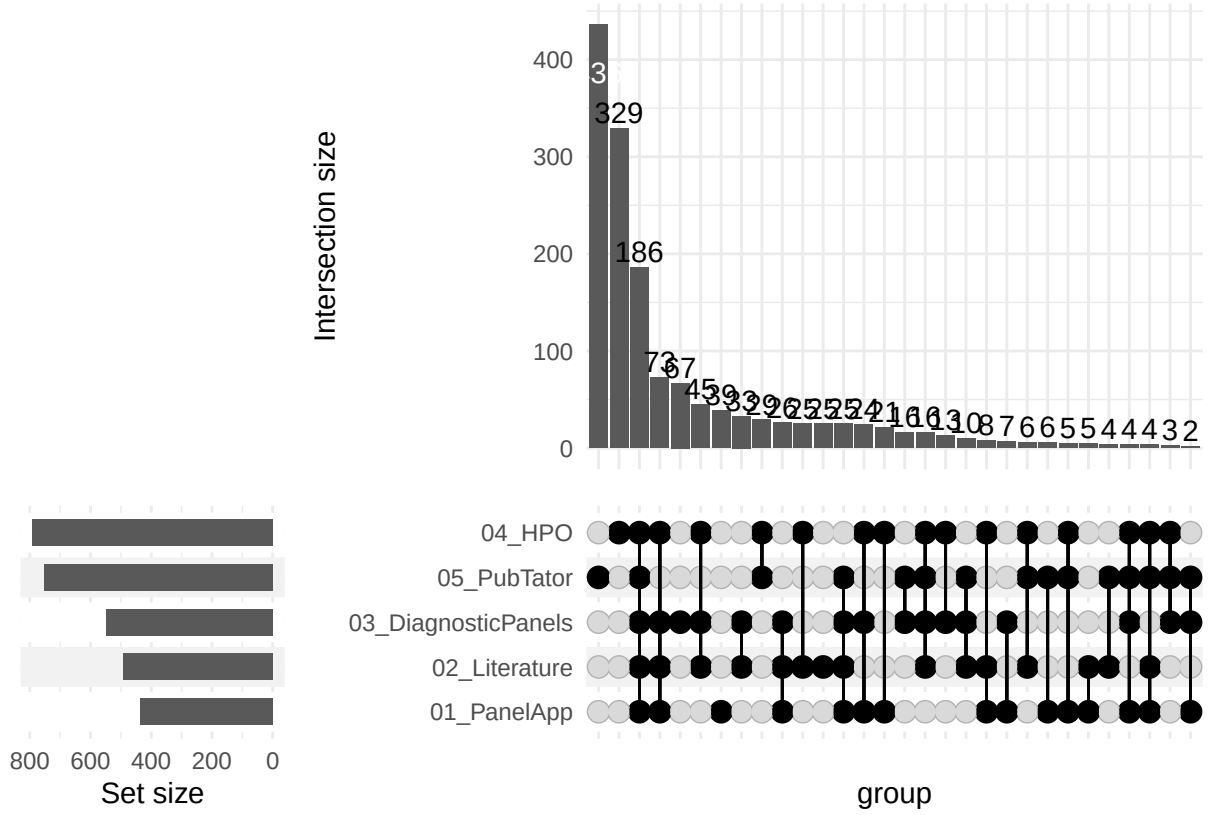
2.1 UpSet plot of merged analyses sources

Below you can see a UpSet plot of the merged analyses.

In the lower left corner you can see the number of Genes originating from each of the different resources,

after that resources are sorted on the right side. UpSet plots generally represent the intersections of a data set in the form of a matrix, as can be seen in the graph below.

- Each column corresponds to a set, and the bar graphs at the top show the size of the set.
- Each row corresponds to a possible intersection: the dark filled circles show which set is part of an intersection.
- For example, the first column shows that most of the genes found in only one of the five sources are derived from the PubTator query, and in the third column you can see that **177 Genes** are found in all five sources.

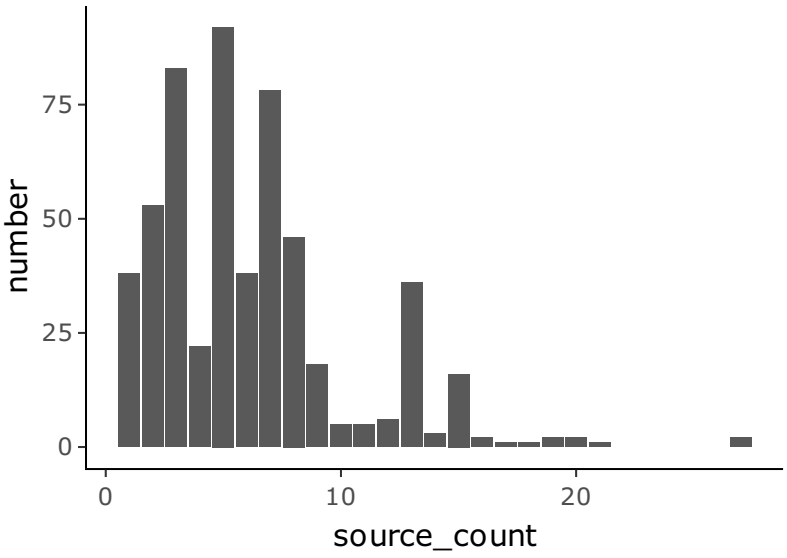


2.2 Bar plot of PanelApp results

Below you can see a Bar plot of the PanelApp analysis.

We retrieved all kidney disease related panels from both PanelApp UK and PanelApp Australia, meaning all panels that include “renal” or “kidney” in its name.

- The y axis shows the number of Genes in the different panels, which is also visualized by the height of the bars.
- The x axis displays the number of panels (source_count), i.e. in how many different panels a single Gene occurred.
- For example **38 Genes** occurred in just one panel and **2 Genes** were present in all thirty different panels.

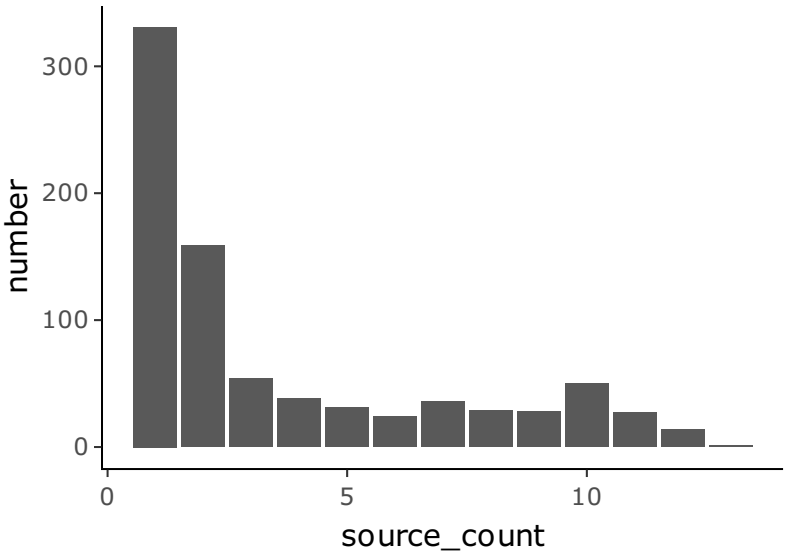


2.3 Bar plot of Literature results

Below you can see a Bar plot of the Literature analysis.

We identified Genes associated with kidney disease in a systematic Literature search using the following search query:
(1) "Kidney"[Mesh] OR "Kidney Diseases"[Mesh] OR kidney OR renal AND
(2) "Genetic Structures"[Mesh] OR "Genes"[Mesh] OR genetic test OR gene panel OR gene panels OR multigene panel OR targeted panel*

- The y axis shows the number of Genes in different publications, which is also visualized by the height of the bars.
- The x axis displays the number of publications (source_count), i.e. in how many different publications a single Gene occurred.
- For example **331 Genes** occurred in just one of the publications and **1 Gene** was present in all 13 different publications.

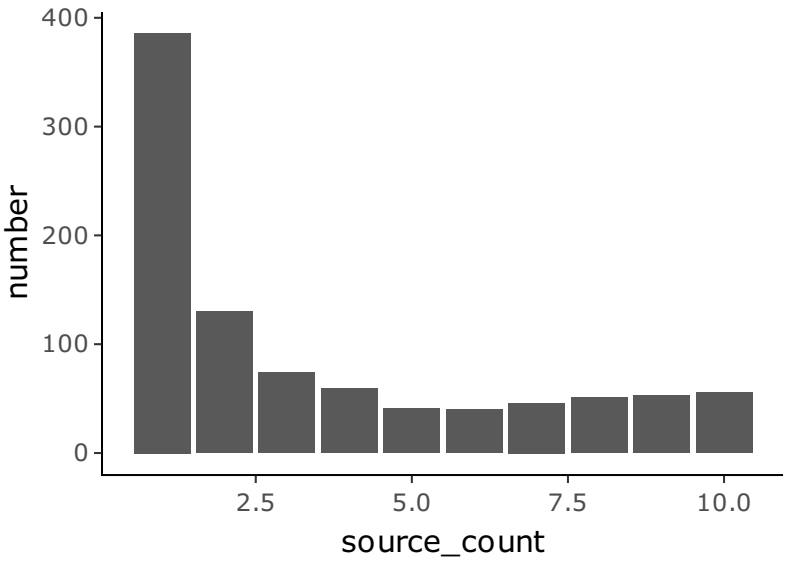


2.4 Bar plot of Diagnostic panels results

Below you can see a Bar plot of the Diagnostic panels analysis.

We used ten common diagnostic panels that can be purchased for genome analysis and extracted the screened Genes from them.

- The y axis shows the number of Genes in the different diagnostic panels, which is also visualized by the height of the bars.
- The x axis displays the number of panels (source_count), i.e. in how many different panels a single Gene occurred.
- For example **371 Genes** occurred in just one panel and **56 Genes** were present in all ten different panels.

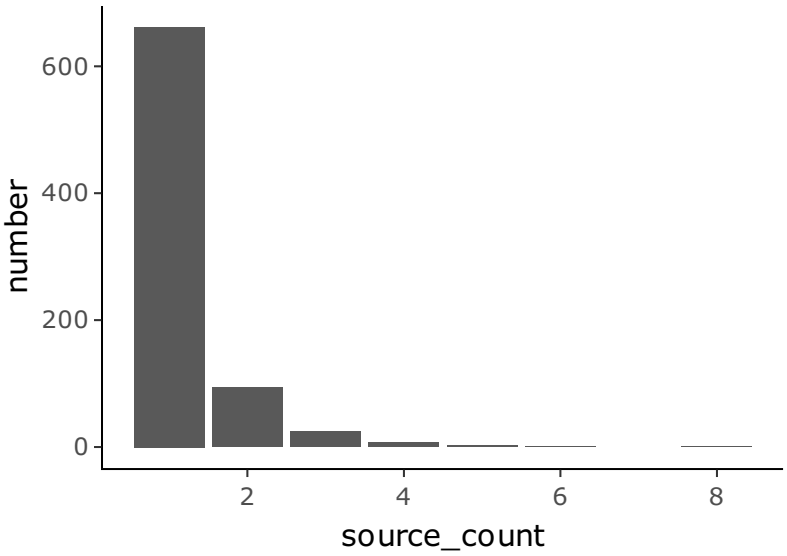


2.5 Bar plot of HPO in rare disease databases results

Below you can see a Bar plot of the HPO-term based query in rare disease databases (OMIM, Orphanet).

We used eight common databases for rare diseases and screened them for kidney disease associated Genes from a Human Phenotype Ontology (HPO) based search query. The most comprehensive HPO term used was *"Abnormality of the upper urinary tract" (HP:0010935)* and included all sub group terms. We deliberately chose these to be somewhat broader in order to fully include all relevant kidney diseases such as CAKUT, among others.

- The y axis shows the number of Genes in the different rare disease databases, which is also visualized by the height of the bars.
- The x axis displays the number of databases (source_count), i.e. in how many different databases a single Gene occurred.
- For example **652 Genes** occurred in just one database and **1 Gene** was present in all eight different databases.

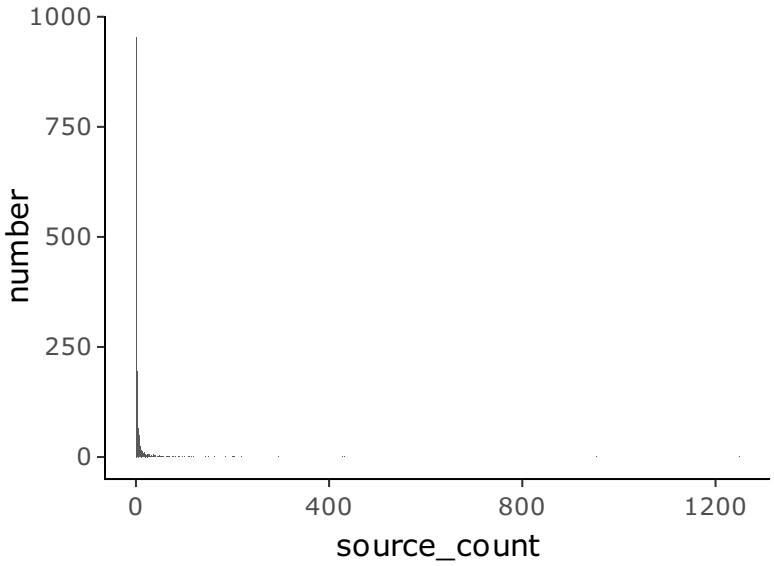


2.6 Bar plot of PubTator results

Below you can see a Bar plot of the PubTator analysis.

We retrieved all kidney disease associated Genes from a PubTator API-based automated literature extraction of publications available on PubMed.

- The y axis shows the number of Genes in the different publications, which is also visualized by the height of the bars.
- The x axis displays the number of publications (source_count), i.e. in how many different publications a single Gene occurred.
- For example **914 Genes** occurred in just one publication and **1 Gene** was present in **1221** different publications.



References

Floranne Boulogne, Laura R. Claus, Henry Wiersma, Roy Oelen, Floor Schukking, Niek de Klein, Shuang Li, Harm-Jan Westra, Bert van der Zwaag, Franka van Reekum, Genomics England Research Consortium, Dana Sierks, Ria Schönauer, Zhigui Li, Emilia K. Bijlsma, Willem Jan W. Bos, Jan Halbritter, Nine V. A. M. Knoers, Whitney Besse, Patrick Deelen, Lude Franke, and Albertien M. van Eerde. KidneyNetwork: using kidney-derived gene expression data to predict and prioritize novel genes involved in kidney disease. *European journal of human genetics: EJHG*, February 2023. ISSN 1476-5438. doi: 10.1038/s41431-023-01296-x.

Sebastian Köhler, Michael Gargano, Nicolas Matentzoglu, Leigh C Carmody, David Lewis-Smith, Nicole A Vasilevsky, Daniel Danis, Ganna Balagura, Gareth Baynam, Amy M Brower, Tiffany J Callahan, Christopher G Chute, Johanna L Est, Peter D Galer, Shiva Ganesan, Matthias Griese, Matthias Haimel, Julia Pazmandi, Marc Hanauer, Nomi L Harris, Michael J Hartnett, Maximilian Hastreiter, Fabian Hauck, Yongqun He, Tim Jeske, Hugh Kearney, Gerhard Kindle, Christoph Klein, Katrin Knoflach, Roland Krause, David Lagorce, Julie A McMurry, Jillian A Miller, Monica C Munoz-Torres, Rebecca L Peters, Christina K Rapp, Ana M Rath, Shahmir A Rind, Avi Z Rosenberg, Michael M Segal, Markus G Seidel, Damian Smedley, Tomer Talmy, Yarlalu Thomas, Samuel A Wiafe, Julie Xian, Zafer Yüksel, Ingo Helbig, Christopher J Mungall, Melissa A Haendel, and Peter N Robinson. The Human Phenotype Ontology in 2021. *Nucleic Acids Research*, 49(D1):D1207–D1217, January 2021. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkaa1043. URL https://academic.oup.com/nar/article/49/D1/D1207/6017351.

Antonio Rueda Martin, Eleanor Williams, Rebecca E. Foulger, Sarah Leigh, Louise C. Daugherty, Olivia Niblock, Ivone U. S. Leong, Katherine R. Smith, Oleg Gerasimenko, Eik Haraldsdottir, Ellen Thomas, Richard H. Scott, Emma Baple, Arianna Tucci, Helen Brittain, Anna De Burca, Kristina Ibañez, Dalia Kasperaviciute, Damian Smedley, Mark Caulfield, Augusto Rendon, and Ellen M. McDonagh. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nature Genetics*, 51(11):

1560–1565, November 2019. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-019-0528-2. URL https://www.nature.com/articles/s41588-019-0528-2.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Research*, 41(W1):W518–W522, July 2013. ISSN 1362-4962, 0305-1048. doi: 10.1093/nar/gkt441. URL http://academic.oup.com/nar/article/41/W1/W518/1105731/PubTator-a-webbased-text-mining-tool-for-assisting.