Mining for gold standard gene sets: effector gene prediction from GWAS results

Aoife McMahon[2], Yue Ji[2], Laura W. Harris[2], Julie Jurgens,, Jason Flannick[1,3,4], Noël P. Burtt[1]

1. Programs in Metabolism and Medical & Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, MA, USA. 2. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. 3. Department of Pediatrics, Boston Children's Hospital, Boston, MA, USA. 4. Department of Pediatrics, Harvard Medical School, Boston, MA, USA.

OVERVIEW

Identifying the genes that impact disease risk is the ultimate goal of genome-wide association studies (GWAS), since the genes and their products offer the most direct clues into biological mechanisms and are the targets of most therapies. Increasingly, as the final step of a GWAS researchers now integrate multiple kinds of genetic and genomic evidence to prioritize genes near each genetic association signal and predict which is likely to be the causal, or "effector," gene.

These **predicted effector gene (PEG)** lists have the potential to greatly increase the biological utility of the GWAS, helping researchers to formulate hypotheses about disease mechanisms and serving as "gold standard" training sets for bioinformatic methods. However, in a review of published PEG lists we found that the evidence types, methods for integrating them with GWAS, and presentation formats are so varied as to risk causing more confusion than clarity.

In an effort to make effector gene predictions more widely accessible, we have developed an interactive table format to display the lists and supporting evidence. We curate these lists and display them in the open-access Predicted Effector Genes Knowledge Portal (PEGKP; pegkp.org), which is part of the Association to Function Knowledge Portal (a2fkp.org). To promote discussion on standards within the research community, we conducted a survey and convened an open workshop in September 2024 to to gather community input on standards, infrastructure, and incentives for improving the utility of PEG lists and making them FAIR (Findable, Accessible, Interoperable, and Reusable).

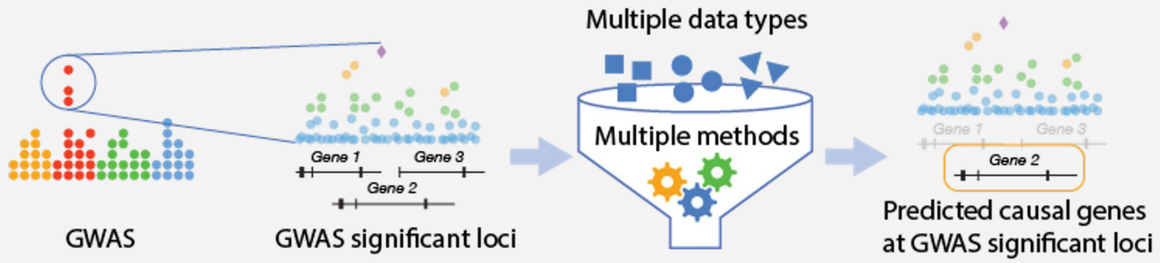
This work was supported by NIDDK UM1DK105554 and NHGRI U24HG011453.

LEARN MORE

Predicted Effector Genes Knowledge Portal (pegkp.org)

Access workshop materials

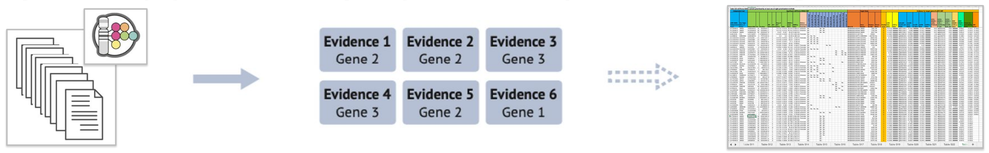
WHAT IS EFFECTOR GENE PREDICTION?



- Genome-wide association studies (GWAS) identify genomic regions (loci) where genetic variation is significantly associated with risk of a disease or magnitude of a trait
- Most GWAS variants are outside of protein-coding regions and impact regulation of nearby genes
- To predict which gene near a GWAS locus is the most likely effector gene, researchers aggregate and integrate multiple types of evidence
- **Effector gene prediction is a major output of post-GWAS analyses**

DIVERSITY OF PREDICTED EFFECTOR GENE LISTS

We surveyed GWAS papers from a 10-year period to find post-GWAS effector gene prediction efforts.

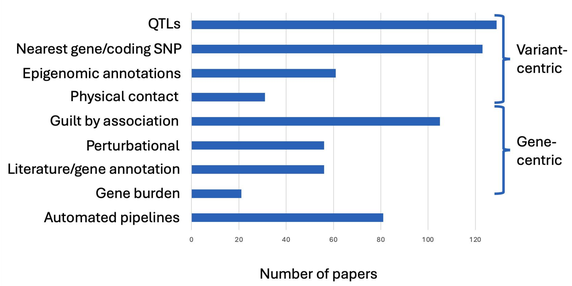


5,140 GWAS papers from GWAS Catalog

169 papers reported systematic gene prioritization

128 (76%) of the gene prioritization papers synthesized the evidence in one table

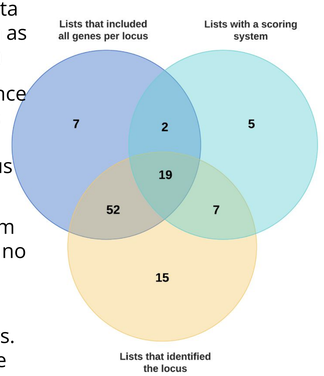
Diverse evidence types



- Most studies used 4 evidence types
- No trends observed in usage of evidence types over time
- No trends observed in types of evidence used together

Diverse content and format

- Presentation as images without underlying data (10%) vs. presentation as re-usable tables (90%)
- Presentation of evidence for all genes per locus (71%) vs. evidence for only top gene per locus (29%)
- Use of a scoring system for evidence (29%) vs. no scoring system (71%)
- Identification of the genomic locus (81%) vs. no identification of the locus (19%)



WORKSHOP ON STANDARDS

Held September 16-17, 2024 at the Broad Institute, European Bioinformatics Institute, and virtually

Agenda

Landscape of effector gene prediction studies
- Approaches and methodologies
- Data representations

Learning from other efforts
- ClinGen
- GWAS Catalog

Discussion of standards

Outcome

We agreed on some basic metadata standards:
- Reference a specific GWAS
- Use standard terminology for evidence types
- Report criteria for significance of evidence
- Cite provenance of evidence
- Document the prioritization method

We agreed on some basic data standards:
- Present results in a plain text file
- Combine all evidence in one file
- Present evidence for all genes considered at each locus
- Use standard identifiers for genes and variants
- Define coordinates and sentinel SNP for each locus

We developed a useful distinction: **PEG list** vs. **PEG evidence matrix**
- Both are valuable

Locus	Top gene	Evidence
Locus 1	Gene A	...
Locus 2	Gene B	...

*A **PEG list** displays the most likely effector gene for each locus, giving an overview of the set of predicted causal genes for a trait.*

Locus	Top gene	Evidence
Locus 1	Gene A	...
Locus 1	Gene B	...
Locus 1	Gene C	...

*A **PEG evidence matrix** displays evidence for all genes considered at each locus, allowing researchers to evaluate the details and draw their own conclusions.*

Moving forward

We need your insights to make PEG lists and PEG evidence matrices accessible, interpretable, and useful! Please let us know if you're interested in participating in future discussions: email us at **help@kp4cd.org.**