EMBL's European Bioinformatics Institute (EMBL-EBI)

Broad Institute

The PEGASUS framework for Predicted Effector Gene (PEG) Reporting

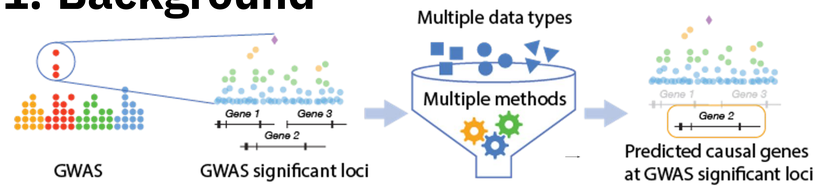
Predicted Effector Gene Aggregation, Standards and Unified Schema

A collaboration between the Knowledge Portal Network and the NHGRI-EBI GWAS Catalog

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

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

1. Background



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

- To determine which gene is the most likely the effector gene (i.e. the gene mediating the effect of the trait-associated variant), researchers aggregate and integrate multiple types of evidence
- **Effector gene prediction is a major output of post-GWAS analyses, aimed at identifying mechanistically relevant genes and potential drug targets**

2. Problem

Broad inconsistency in data content and format of PEG data impairs interpretation, integration and reusability

- 10% presented as summary images without underlying data
- 29% present evidence for top gene per locus (not all genes considered)
- 19% don't identify the locus under investigation
- 29% use a scoring system to convey a conclusion
- Data distributed throughout publications

From landscape analysis of 169 publications, Costanzo et al, 2025

3. Goals

Aim - Findable **A**ccessible **I**nteroperable **R**eusable **(FAIR)** predicted effector gene data

Ultimate long term aim - enable meta analysis to define gold standard lists of genes involved in traits

Use cases

Data integration and reuse – enable computational ingest, submission to knowledgebase, and support AI/ML/KG use.

Research and hypothesis development – generate hypotheses, confirm independent findings, prioritize genes for drug targets

Consideration balance *data quality (FAIRness) *burden on data generator *need for data curation

4. Process

Activities

Community Engagement

Workshop, Sept 2024, 80 attendees, at Broad and EBI
Working Group meetings monthly in 2025;
Developed and iterated over a 'strawman' standard, benchmarked to assess suitability.
ASHG Ancillary, Oct 2025

Working group materials in slides and recordings

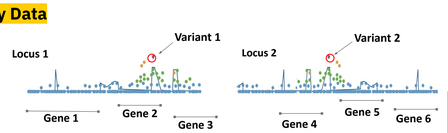
Landscape analysis Costanzo et al

Landscape analysis and general recommendations (Costanzo et al)

BOX2	
General recommendations emerging from a community workshop on standards and infrastructure for predicted effector gene lists	**Recognize** value of interpretable PEG data (journal editors, peer review). **Provide** genes, traits, loci definition, confidence measures. **Share** data in single machine-readable file. **Establish** a working group for standards and benchmarks. **Create** an open-source, FAIR-compliant catalog.

Toy Data

Trait: myocardial infarction



								Data distributed throughout publication
eQTL (aorta, GTex)								Suppl Table 4
LD								Suppl Table 4
Knock out model (mouse, IMPC)	Enlarged heart phenotype							Text
Expression (heart)	Highly expressed				Expressed			Text
PoPS score	10	3	1		7	5	4	Suppl Table 5
Author conclusion	STRONG	WEAK	WEAK		MODERATE	WEAK	WEAK	Text Suppl Table 11

Illustration of the kind of data found in predicted effector gene publications, depth of green = strength of evidence

5. Proposed Framework

PEG lists must be supported by an evidence matrix & described by metadata

Metadata

1. Trait and provenance of the original GWAS data
2. Sources for each evidence type
3. Description of methods used
4. Column header definitions

Evidence matrix

1. Includes evidence for all genes considered (not just 'top' genes)
2. Includes data for all evidence types used
3. May include integration(s)

List

1. High level data summary
2. Lists the best predictions (top genes)
3. Indicates which evidence categories were included in analysis
4. Includes an author's conclusion (integration)

Requirement

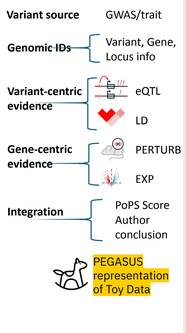
One trait -> one GWAS -> one matrix -> one list

How data is organised :

Genomic identifiers (variant, locus and gene information)

Evidence - each evidence stream assigned to categories (either variant-centric or gene-centric)

Integration - results of integrating evidence streams



PEGASUS representation of Toy Data

6. Conclusion

Proposed framework

- **facilitates submission to knowledge base** (e.g. Predicted Effector Genes Knowledge Portal pegkp.org) and **linkage to GWAS Catalog**
- enhances **interpretability** and enables **interoperability** of PEG data
- lays **foundations** for meta-analysis and generation of **gold standard** effector gene lists

Next steps

- community feedback on suitability
- uptake

Join us!

- The PEG Working Group is an open community building standards, tools, and FAIR infrastructure for PEG lists. Email: help@kp4cd.org

PEG Metadata

Content overview

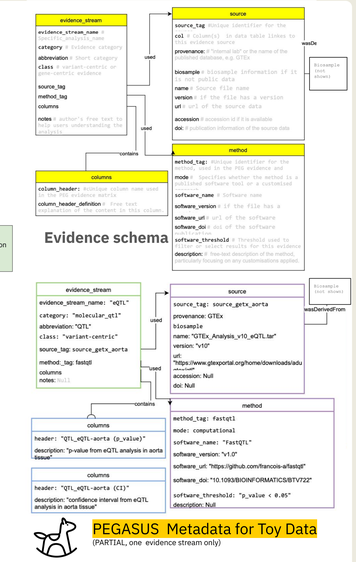
Definitions of column headers

Information on the GWAS and trait from which the variants were originally found
Metadata on genomic identifiers

Evidence streams included in the analysis (identity and provenance).
Divided into evidence categories (Evidence Categories tab).

How and which evidence streams were integrated (methods, computational tools)

Format - yaml or json



Full toy data example including yaml metadata representation (PEGASUS documentation materials)

Evidence schema

PEGASUS Metadata for Toy Data
(PARTIAL, one evidence stream only)

PEG Evidence Matrix

Content

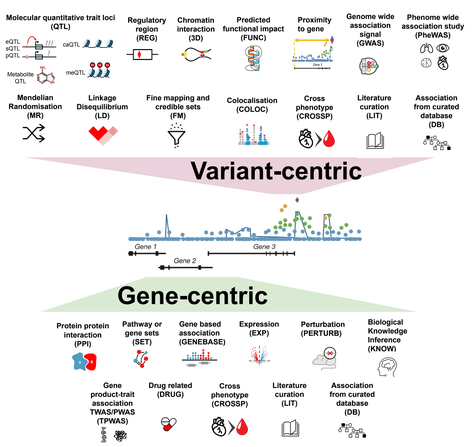
	Genomic Identifiers						Evidence		Integration
	Variant information		*Locus information*		*Gene information*		*variant-centric evidence*	*gene-centric evidence*	
Column header	**Primary Variant ID (chrbp)**	**Primary Variant ID (rsID)**	**Locus range**	**Locus ID**	**Gene ID (Ensembl Gene ID)**	**Gene symbol (HGNC)**	***Category_[xyz]***	***Category_[xyz]***	***INT_[xyz]***
Format	chr:bp:ref:alt	rs[]	chr:pos:start-end		ENSG[]		Header: Category_[xyz]		Header: INT_[xyz]
Requirement	mandatory	optional	recommended	optional	mandatory (or)	mandatory (or)	at least one evidence stream		optional; mandatory for representation as a list
Description	*Variant to which variant-centric evidence relates.*		*Range around the primary variant considered in this analysis.*	*Internal or curated ID for the region considered. The associated variant is recommended.*	*Gene to which gene-centric evidence relates.*		*Columns relating to the evidence, defined in the metadata file.*		*Columns relating to the integration, defined in the metadata file.*

PEGASUS Evidence Matrix for Toy Data

Primary Variant ID (chrbp)	Primary Variant ID (rsID)	Locus range	Locus ID	Gene ID (Ensembl Gene ID)	Gene symbol (HGNC)	Q a (p
chr10:114754071:T:C	rs[variant 1]	chr10:1000-2000	Locus 1	ENSG00000151532	Gene 1	
chr10:114754071:T:C	rs[variant 1]	chr10:1000-2000	Locus 1	ENSG00000197283	Gene 2	
chr10:114754071:T:C	rs[variant 1]	chr10:1000-2000				
chr5:224754071:A:G	rs[variant 2]	chr5:200000-205000				
chr5:224754071:A:G	rs[variant 2]	chr5:200000-205000				
chr5:224754071:A:G	rs[variant 2]	chr5:200000-205000				

QTL_eQTL-aorta (p_value)	QTL_eQTL-aorta (CI)	LD_functional_variant	LD_r2	PERTURB_mouse	EXP_heart (RPMK or TPM)	INT_pops	INT_Combined assessment (author conclusion)
0.01	[1.2, 2.5]	rs1008	0.98	MP:0000274 |	10.8	10	STRONG
0.05	[0.01, 1.9]	rs1008	0.75			3	WEAK
0.008	[0.8, 2.6]	rs1008	0.5			1	WEAK
0.0000007	[3.1, 5.0]	rs3540	0.78		7	7	MODERATE
0.008	[0.8, 2.6]	rs3540	0.3			5	WEAK
0.0000007	[3.1, 5.0]	rs3540	0.25			4	WEAK

Evidence Categories



PEG List

PEGASUS List Foundational model - records whether evidence was considered (tick = data present, blank = not assessed) and reflects the author's integrated conclusions for top genes.
Future iterations can expand on this with more detailed annotations and structured scoring.

		variant centric evidence assessed					*gene centric evidence assessed*						*integration*
Primary Variant ID (rsID)	**Gene symbol (HGNC)**	**GWAS**	**PROX**	**QTL**	**FUNC**	**LD**	**COLOC**	**TPWAS**	**EXP**	**PERTURB**	**KNOW**	**LIT**	**DRUG**	**Author conclusion**
variant 1	*Gene 1*	☑	☐	☑	☐	☑	☐	☐	☑	☑	☐	☐	☐	*STRONG*
variant 2	*Gene 4*	☑	☐	☑	☐	☑	☐	☐	☑	☐	☐	☐	☐	*MODERATE*

PEGASUS List for Toy Data

Tick = data/value present (VAL). Blank = not assessed (NA). Ticks do not imply supportive vs negative; see author interpretation & provenance

Which 'top genes' make it on to the list?
Top gene per locus (advantage -standardised)
or
author's choice (advantage - flexible, disadvantage - not standardised)

Acknowledgements

We thank working group members and participants of the community workshop for their valued input. This work was supported by NIDDK UM1DK105554, NHGRI U24HG011453 and EMB-EBI core funds.

Draft PEGASUS framework documentation: https://ebispot.github.io/PEGASUS/docs/home