Predicted Effector Gene Working Group meeting notes

Researchers who perform genome-wide association studies (GWAS) often aggregate and integrate multiple other evidence types in order to predict the effector (causal) genes at GWAS loci (see some examples). These predictions are a valuable outcome of GWAS, but in 2024 there were currently no community standards for organizing or sharing them, and the proliferation of multiple evidence types, methods, and presentation formats carried the risk of confusion. To spark discussion on how to address this gap, we held a community workshop on the standards, infrastructure and incentives required to promote and enable sharing, interoperability, and updating of predicted effector gene lists. Following the workshop this working group was set up to develop those standards. The working group is co-organised by the Broad Institute's Knowledge Portal Network and EMBL-EBI's NHGRI-EBI GWAS Catalog.

Materials

Material from PEG workshop (Sept 2024) https://bit.ly/pegworkshopnotes
Public webpage on workshop (includes recording of working group meetings)
PEG Standard Strawman proposal https://bit.ly/peg_strawman
Costanzo et al. Realizing the promise of genome-wide association studies for effector gene prediction
Folder containing Working Group Meeting Slides

7 Oct 2025 Working Group Meeting 5 - 202510

Agenda
Run through of ASHG ancillary session

- Laura - landscape
- Aoife - framework
- Matt - benchmarking

- Laura - discussion guiding questions

Remaining questions (20 mins) - Aoife

At ASHG, we'll be holding an Ancillary Session with outputs from our Working Group. Please register here and share with your networks, as we want to get as much community input and uptake as possible. Here are the details:
Friday, October 17, 2025
11:45AM-1:15PM ET
Thomas M. Menino Convention & Exhibition Center (MCEC, formerly the BCEC), 415 Summer St., Boston, MA 02210; Room 259A

To review the current list of proposed evidence categories, see the draft here: https://ebispot.github.io/PEGASUS/docs/peg-evidence/

2 Sept 2025 Working Group Meeting 4 - 202509

Agenda

Introduction & promoting/planning for ASHG Ancillary Session (Julie Jurgens)

Summary of benchmarking activity (Julie Jurgens)

Presentation of benchmarking results (Benchmarking Team)

Fitting PEG matrix standards to PEG lists (Laura Harris)

Participants: Julie Jurgens, Aoife McMahon, Laura Harris, Yue Ji, Ayse Demirkan, Abdurrahman Shiyanbola, Karl Heilbron, Szymon Szyszkowski, Daniel Considine, MacKenzie Brandes, Yakov Tsepilov, Kanika Kanchan, Sylvanus Toikumo, Nina Oparina,

Slides: 250902_Fourth working group meeting

Files

Revised PEG matrix metadata schema integrating feedback from our last meeting (7/9/2025) and example YAML files are available in the metadata v3.1 tab here

Doodle poll for October WG meeting (test run for PEG WG Ancillary Session at ASHG 2025); please fill out by Friday 9/5/2025:
https://doodle.com/group-poll/participate/aO9gkXLd/vote

Register for ASHG 2025 Ancillary Session:
https://docs.google.com/forms/d/e/1FAIpQLScSPEnqstCN1-4cWzZKgNX-2cWUV90F68Pjkit7UHZsEKUOIA/viewform

Notes

ASHG 2025
Friday, October 17, 2025
11:45AM-1:15PM ET
Thomas M. Menino Convention & Exhibition Center (MCEC, formerly the BCEC), 415 Summer St., Boston, MA 02210; Room 259A

Preliminary agenda for ASHG 2025 PEG Ancillary Session:
Recap of predicted effector gene landscape analysis (15 min)
Introduction to new standards for predicted effector genes (30 min)
Benchmarking presentations (15 min)
Group discussion/ review of standards (30 min)

Who from the PEG WG will be attending ASHG 2025?
Julie Jurgens, Laura Harris, Aoife McMahon, Yue Ji, Noel Burtt, Matthew Pahl, Sylvanus Toikumo, Shicheng Guo

Benchmarking Summaries

Matt Pahl
See problems/ suggestions here: 20250902_PEG_WG_benchmarkingintro

Abdurrahman Shiyanbola
Problems encountered
- Challenging when this isn't your own data
- Columns aren't consistent as the columns are used in the knowledge portals or GWAS catalog (missing columns in the knowledge portal)-should make GCAT/KPN more consistent
- YAML wasn't easy to customize to include custom columns; unsure of key values to include
- Info had to be source and validated against at least 3 sources
- Some standard yaml columns aren't c/w the types of columns that should be used in the data for migraine disorder
    - Comment from Aoife: Issue is that this is largely a data curation effort for benchmarking data that aren't your own-some of the issues encountered might not be an issue for people who apply the format to their own data

Suggestions
- Suggest that we integrate more entities for certain traits to make them more compatible

- Make python script that makes imputations for the mappings-Yue seconded this suggestion as a potential next step to make the process easier

Szymon Szyszkowski

- PEG for CAD metaanalysis
- Format is generally well defined
- Challenging curation post-analysis b/c this wasn't Szymon's data
- The publication used 2 different studies and evidence was mixed in the Knowledge Portal, so it was hard to tell what evidence came from which; mix of hg19 and hg38 data from KPN vs GWAS Catalog
    - Had to do liftover for variants to get into the desired build
- There's a mixture of different evidence categories/ had to go back to supp tables to get these actual lines of evidence
- We could get specific fields that link to the specific evidence sources and make these clearer in the knowledge portals

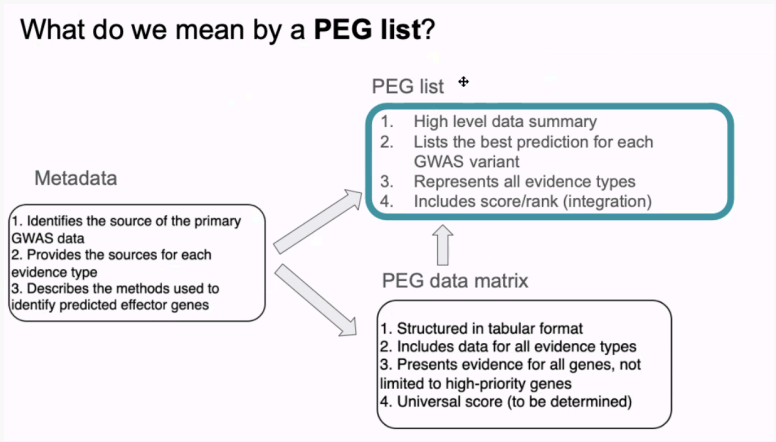
Karl Heilbron

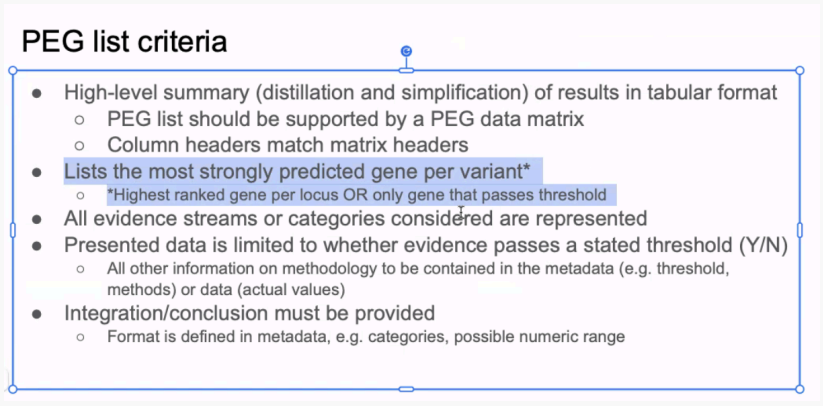
- Study that Karl led on Parkinson's Disease
- Easier to curate if it's your own data
- Think about what we can do with minimal effort (they put in rsIDs, not CPRA)
- Asked if it's the nearest gene, does it have top POP score in the locus/what's the actual POP score, credible set
- Might be good to put in medRxiv DOI in some cases
- Genome build-do we want to standardize hg19 and hg38
- Didn't see something that distinguishes sample_ancestry vs sample_ancestry_category
    - We should add more explicit guidance on ancestry
- Indented format might be challenging for users newer to yaml files
    - Response from Yue: indentations were initially made primarily for human readability here
- Might be easier to just have ontology ID attached and not human-readable phenotype
    - If we eventually make an automated submission portal, this might be something we can automate to make it easier for submitters
    - For now, keep the human-readable part included, but eventually make only machine readable
- Nature of primary variant-they ran GCTA-COJO to select primary variant in stepwise format; there wasn't a way to list this option explicitly for this particular use case
- Struggled with the fact that the Google Sheet, Google Slides, and example YAML all had slightly different names/descriptions for fields

Yue Ji

See problems/ suggestions here: Benchmark_Yue_PEGSt000004

Extending the standard to PEG lists (instead of matrices)-Aoife McMahon Slides





9 Jul 2025 Working Group Meeting 3 - 202507

Agenda
Update on benchmarking activity (Aoife McMahon)
Discuss based on review of template (Aoife McMahon)
High-level metadata intro, considerations (e.g., AI/ML readiness, ontologies) (Aoife McMahon)
Metadata standards (Marcos Casado Barbero)
Our attempts at abstraction of metadata schema (Yue Ji)

Notes

Participants: Julie Jurgens, Aoife McMahon, Laura Harris, Yue Ji, Open Targets team, Marcos Casado Barbero, Szymon Szyszkowski, Kanika Kanchan, Oliver Ruenbacker, Matt Pahl, Gabi Rinck, Lillya Kopanitsa, Oleg Borisov, Xiangyu Ge, Emrah Kacar, Wafaa Rashed, Zhanna Balkhiyarova, Quy Hoang, Nina Oparina, Sean Yao, Juliana Xavier de Miranda Cerqueira

Slides: 250709_Third working group meeting
Strawman v3 Template and Instructions
20250709_PEG metadata_strawman_V3

Laura shared EBI-EMBL survey due 7/16/2025: https://www.surveymonkey.com/r/QGFMBH8

Aoife introduced benchmarking activity that has been circulated to volunteers, final deadline in 3rd week of August. Link to template is in meeting agenda email (Strawman v3 draft template), this includes instructions, external lists that can be reformatted according to the standard, basic metadata table (details to be discussed today), folder to return files, evidence categories.

Feedback on template from WG:
Matt Pahl-template appears straightforward and clear
Szymon (Open targets)-appears generally reasonable but they have some questions on the metadata
Aoife encouraged other participants to share their feedback either during the call or afterward via email

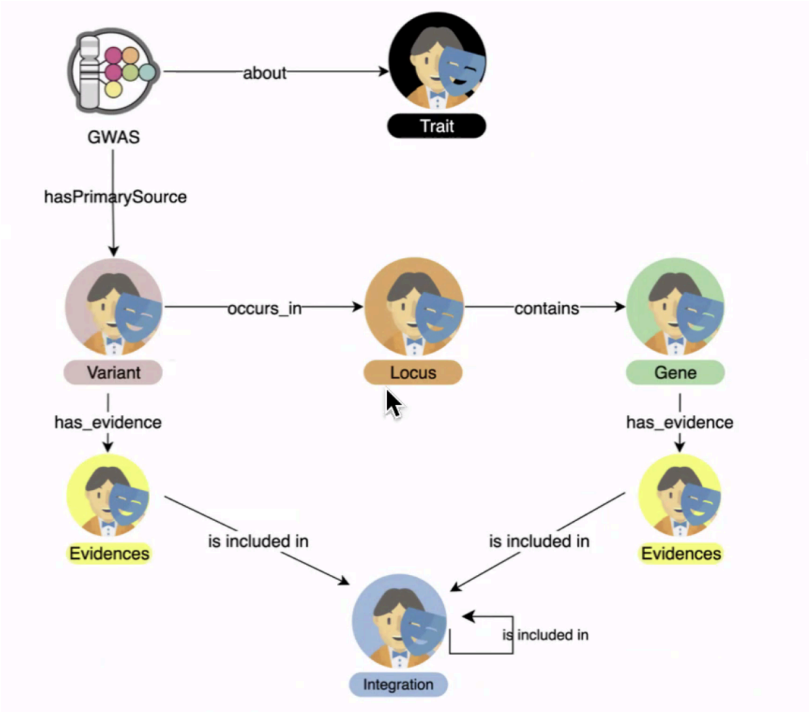
Aoife introduced the metadata component of the PEG matrices and their importance for FAIRness, and AI-readiness, reproducibility, interoperability with existing standards and ontologies (eg Croissant, Evidence and Conclusion Ontology, UBERON, EDAM, EFO)
Considerations for metadata: content, structure/schema, format

Marcos Casado Barbero (EBI): challenge in standardizing terminologies for metadata. Need to define the genre, cast the characters and outline the plot (define relationships among them) for metadata in both human- and machine-readable format. Need to be willing to revise the proposal iteratively with the community. Goal is to minimize but not eliminate distortion generated by multiple disparate standards.
Requirements for a metadata standard:

1. Human-readable format specification
2. Machine-readable schema
3. Open implementation (open, versioned code that enables adoption of the standard; must be scalable, have appropriate licensing, be easily used in existing workflows)
4. Governance (a change-protocol process for standard modification)
5. Community assets (documentation, examples, community feedback with the standard)

All 5 must exist together for a sustainable standard
Example: Federated EGA
Optional vs mandatory - needs to be decided by the community.

Yue
Defining the main cast and relationships among them for our PEG matrix metadata model:



Mandatory metadata-only entities: GWAS, trait

Including ontologies is strongly suggested

Mandatory entities represented in both data and metadata:

Variant: type, genome build, information

Locus

Gene

Evidence metadata entities:

Relevant for supporting both the variant-centric and the gene-centric categories

<u>Feedback from the WG</u>

Marcos: Re. ontologies:

- I would recommend the use of URIs, especially CURIEs (e.g., biosample:SAMEA...).
- For validation, OLS API is pretty handy, and is integrated in Biovalidator.

Marcos: Re. Evidence, Evidence_category and Evidence_streams: I would advise to revise these entities. From my uninformed POV, they may be increasing complexity without adding much information, when they could be compacted instead.

Szymon: The biosample used for QTL should be standardised with ontology - OT are using UBERON

Marcos: General comments:

- Make the model RDF-compliant.
- Integrate in the very standard links (GA4GH, DCAT-AP, Schema.org, Beacon-v2, PROV-O...) to other standards.

Open Targets team: Should “integrated” evidence be computed by the submitter? Might be better to compute it on our end after? Aoife suggests it might be easier for a broader range of submitters if they’re able to apply their own integration, but that they clearly define when they’ve done their own integration.

5 Jun 2025 Working Group Meeting 2 - 202506

AGENDA

Working Group Roadmap (ASHG & Beyond) - Julie Jurgens

Data Content Standards - Aoife McMahon

Assigning Benchmarking Activity - Yue Ji

Notes

Participants: Julie Jurgens, MacKenzie Brandes, Aoife McMahon, Laura Harris, Yue Ji, Karl Heilbron, Szymon Szyszkowski, Wafaa M. Rashed, Oleg Borisov, Yakov Tsepilov, Yong Li, Norann Zaghloul, Florence notetaker, Quy Hoang, Matt Pahl, Ayse Demirkan, Oliver Ruebenacker, Beena Akolkar

Slides: 250605_second working group meeting
Recording: https://www.youtube.com/watch?v=qSnYjfqgZe0
Scheduling for July 2025 WG Meeting: https://doodle.com/group-poll/participate/epLDm11e

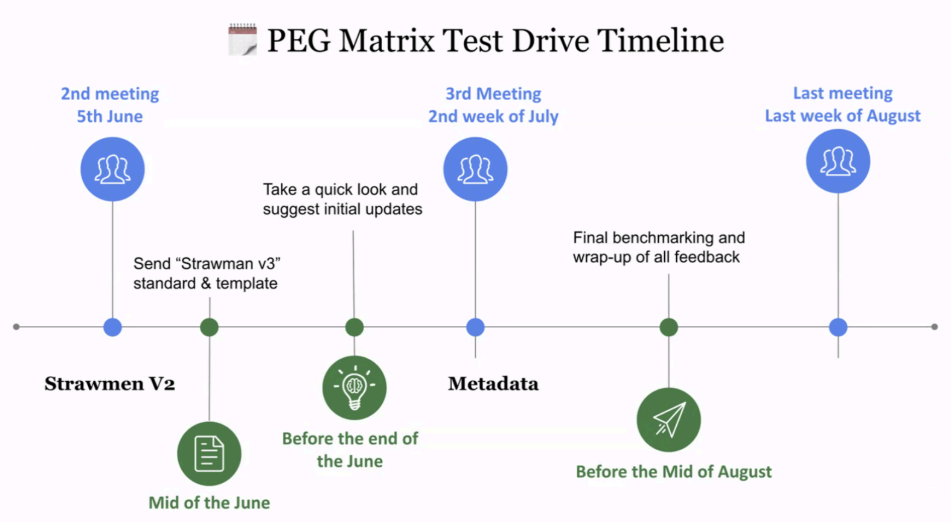
- Working Group Roadmap: ASHG & Beyond - Julie Jurgens
    - Nature Genetics Perspective paper published: https://www.nature.com/articles/s41588-025-02210-5
    - Presented roadmap for next 4 months leading up to ASHG:
        - June 2025: Data content
        - July 2025: Metadata content
        - Aug 2025: Review data content benchmarking
        - Sept 2025: Review metadata content benchmarking
        - Oct 2025: ASHG (Boston) 10/14-10/18, Ancillary Session, need presenters
    - General roadmap beyond ASHG:
        - Publication
        - PEGathon

        - AI-assisted user upload system
        - Convert existing matrices to new standard using AI-assisted tool
        - Engage with professional societies
        - Public pilot launch
        - Iteratively revise framework based on community feedback
    - Opportunities for engagement
        - WG participation
        - Benchmarking activities
        - ASHG ancillary session
- Data Content Standards - Aoife McMahon
    - Proposing submitting both metadata file and data
    - Data file
        - Post-GWAS analysis
        - Single table
        - Includes all genes analysed
        - One row per variant-gene pair
        - We provide standardized file headers (mandatory, optional, recommended), author can include additional columns
        - Evidence types are divided into gene-centric and variant-centric
        - Thoughts from the WG:
            - Szymon: It would be great to specify the offset of the position - 0-based or 1-based (expect to be 1-based)
            - Karl: I think “lead” and “sentinel” are quite common in the field and both are fine with me. Probably more for “lead”.
            - Szymon: The locus Id might be ambiguous if one would like to use multiple peg lists
            - Yakov: We use locus range and it is quite convenient. We use lead variant as locus id.
            - Szymon: We might need a standardised way to calculate it
            - Ayse: is there an instruction for the uploader on how to annotate the variant to gene?
            - Yakov: Locus range is study specific
            - Julie: consider breaking perturbation out into animal model vs cell line as 2 separate categories
            - Karl: L to G, Flames, Calder-fit these into the integration category
            - Yong: Which evidence category would a gene go if it has a high PoPS score in a locus?
            - Yakov: All of these evidence categories have their own scores. It requires the special rules for unification.
            - People like the idea of having both gene and variant-centric categories

- Yakov: For nearest gene - we use both - the nearest to TSS and nearest to footprint
- Karl: Would you also have a column with the exact distance between variant and gene?
- Yakov: I feel that variant centric and gene centric evidence should be splitted. The gene level evidence is more general and applicable to many variant-centirc studies (GWAS)
- Yakov: need to distinguish between phenotype/trait/disease

    - Locus information
        - Locus range (study specific) and variant for locus ID
    - Evidence categories
        - Nearest gene - yes or no? Maybe metadata item

- Assigning Benchmarking Activity - Yue Ji & Noel Burtt
    - We've developed a preliminary prototype for the PEG matrix data content; now we need test drivers to perform benchmarking activity
    - Recommend that each volunteer reviews 2 PEG matrices (their own data or other matrices from our PEG repository), provide preliminary input so we can revise to v3 of strawman for data content standards, log reformatting steps, provide feedback/difficulties/suggestions
        - User-supplied data must have GWAS-derived variants, gene-centric and variant-centric information, and any relevant context/metadata



    - We'll provide updated v3 standard, template to help users get started, suggested column names and mandatory column names, clear instructions for each evidence type
    - Volunteers:
        - Yakov Tsepilov: We are happy to provide our variant-centric PEG matrix calculated for all GWAS we had (mostly GWAS Catalog) but the matrix is huge. Laura asked if they could provide it for a subset of diseases
        - Karl Heilbron

        - Ayse Demirkan
        - Matthew Pahl (volunteered post-meeting)
    - Karl-Idea: one thing we could do to lower the effort required by people trying to generate a PEG matrix is to provide a tool for converting between gene names (e.g., ENSG to HGNC). Likewise for CPRA and rsID. Probably other examples exist.
    - Yakov would like to present at the next meeting ~10 min about their team's experience at open targets

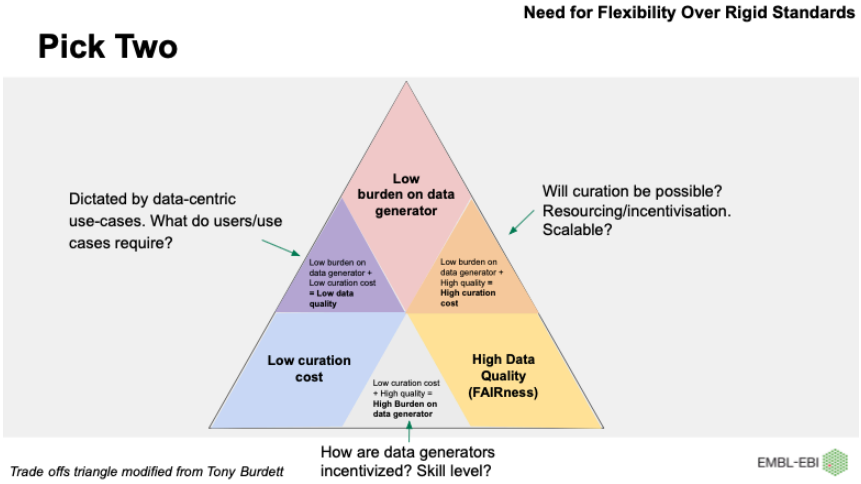
May 8, 2025 | Working Group Meeting 1 - 202505

Participants: Julie Jurgens, MacKenzie Brandes, Maria Costanzo, Noel Burtt, Aoife McMahon, Laura Harris, Yue Ji, Karl Heilbron, Wafaa M. Rashed, Adam Butterworth, Oleg Borisov, Eric Fauman, Nina Oparina, Chi Zhang, Loz Southam, Kyle Vogan, Kanika Kanchan, Ellen McDonagh *(Ellie)*, Yong Li, Nathalie Chami, Santhi Ramachandran, Norann Zaghloul, Florence, Kanika Kanchan

Notes

- Noël Burtt started the meeting with outputs from our workshop in fall 2024 and updates on the current effector list landscape
- Introductions via menti.com | 3727 6552
- Reflections from Maria Costanzo
    - Nature Genetics paper coming out soon
- Recap of September 2024 Predicted Effector Gene Workshop - Aoife McMahon
    - PEG workshop notes from previous workshop
    - In these lists, people value combination of bioinformatics, manual curation, and multiple lines of evidence
    - Learning from other efforts showed importance of FAIR metrics are critical for developing PEG lists and gene-disease relationships with standardized terminology
    - Strawman proposal: Proposed PEG standard (strawman) https://bit.ly/peg_strawman
        - Metadata file and data file for all PEGs
        - Metadata: gwas, evidence, free text
        - Data file: gene IDs, sentinel variant, evidence, other fields, score/conclusion
        - Discussion of strawman
            - Need for flexibility over rigid standards - however maximizing flexibility can include massive trade-offs
            - Evidence representation (list vs. matrix)
            - Versioning
        - Trade-offs for flexibility over rigid standards

- Low burden on data generators and high quality data (FAIRness) - high curation cost. Will curation be possible? Resourcing/incentives. Scalable?
- High data quality and low curation cost - high burden on data generators. How are data generators incentivized? Skill level?
- Low curation cost and low burden on data generators - low data quality.



- Recommendations made in Costanzo et al
- Discussion
    - Evidence types - what about missense variance or gene burden? In or near the gene category would cover missense variance.
    - Develop list to train the model for a golden standard. Stepwise approach and to train models, producers of the list must be absolutely confident and accurate.
    - Gene based annotations may change between diseases
    - Karl Helibron: I really like the idea of thinking about PEG results across colocalizing results. Probably a version 2.0 topic though, in my opinion.
        - OT coloc: https://platform-docs.opentargets.org/gentropy/colocalisation and all results are integrated here: https://platform.opentargets.org/

- Julie Jurgens
    - WG meeting structure: monthly for 1 hr, focusing on a single topic
    - Meeting topics
        - Solidifying data/metadata standards
        - Benchmarking
            - ID gaps in strawman by fitting existing PEG lists into it
        - Use cases
            - Building standards based on use cases
            - Hypothesis generation

                - reuse/computational ingest - AI/ML/KG readiness
            - Prioritization
        - Community recommendations
            - Engage with professional societies to develop endorsed recommendations
            - Public pilot launch - coordinate with community to ID new datasets and apply standards
            - Develop iterative versions of framework
        - Long term: comparing results obtained from different prioritization approaches
            - Enabling meta-analyses to define gold standards list
        - Other discussion topics?
            - Retrofitting into standards
            - Longer term goal: develop a tool to fit submitter-provided information into a standardized format?
                - Ellie McDonagh: ChatGPT may be an initial starting point to ask to format and create code for this formatting!
            - Not all the information needed in the PEGs list is in the same table - sometimes script is needed to pull the information into one
            - How much metadata do you want to capture?
- Moving from guidelines to standards for PEG data - Laura Harris
    - Variant-centric evidence matrix
        - Data content
        - Metadata content - what should be mandatory and what can be flexible?
        - Data structure - standard headers & layout
        - File formats - what can be accepted?
    - Metadata must include
        - GWAS data
            - Trait description
            - Ontology mapping
            - GWAS source
            - Genome build
        - Method
            - Free text description
        - Evidence
            - Eric Fauman: closest gene should be mandatory. Adam Butterworth: Because it's a strong predictor? How about if someone developed a PEG list that they wanted to compare against nearest gene so they deliberately didn't include nearest gene..... would that then not be a PEG list? Eric: It could have a weight of 0, but it should be in the list. Adam: I tend towards not having any single mandated evidence type but having used multiple evidence types (and some way of combining them) would be a qualifying criterion to be a PEG list

        - Main priority of WG: data content > metadata content > data structure > file format
        - Overall metadata structure?
            - Combining the evidence sources - this will be in the method
            - Ethnicity - would be in the underlying GWAS; if unpublished, would want to specify sample number, ancestry, case control
        - Metadata content evidence types that should be included - we need more discussion around this
            - Mendelian diseases
            - Rare variant associations
            - Rare variant gene burden tests
            - MR
            - Other QTL types (splicing, methylation, etc)
            - Extending the QTL category to include pQTL, sQTL, meQTL in relevant tissue type
            - Known drug target-disease list
            - PoPS
        - Mandatory evidence types to include?
            - “Tool”
            - Should be sufficiently detailed to enable someone else to reproduce the analysis
            - Cell and tissue type for genomic and molecular lines of evidence
    - Sentinel variants?
        - Genome build will be in metadata
        - Many studies consider all variants within a certain LD or the sentinel
        - Region definition/range
        - Interaction in a conditional analysis - first variant in a region, second, third
        - Rsids
    - Gene representation?
        - Agree with proposal
    - Evidence matrix vs. PEG list - confirmed it is important to define standard for PEGs list as well as evidence matrix
- Planning next WG meeting
    - Topic:
    - June 2025
- Summary and next steps
    - Develop roadmap to enable discovery
    - Continue standards discussions
- Final thoughts?
    - Look at old PEG lists that have been generated to be updated to fit our requirements - retrofitting
    - Karl agreed to review the Lange et al. Parkinson’s disease study that was mentioned at the start of the meeting for fitting to the standard

    - Step 0 or 1 - recommended best practices in generating lists then show building a model or MA on top of the gene lists
    - Generate roadmap of next steps
    - Based on this meeting we can develop roadmap
    - Noel proposes ASHG as a goal to have an initial standard

Zoom chat

Action items

- [ ] Please indicate your June availability using the following Doodle: https://doodle.com/group-poll/participate/bWoKRYnb