Community workshop on standards and infrastructure for predicted effector gene (PEG) lists

September 16-17, 2024
Hybrid meeting: at the Broad Institute, at the European Bioinformatics Institute, and on Zoom

NEW: Please fill out the post-workshop survey here.

Table of contents

Meeting Materials

Shared Google Drive folder containing:
PEG_workshop_program.pdf
Workshop notes
Breakout group discussions
Presentations

Recordings: The videos are annotated on YouTube so that you can navigate to specific presentations.

Day 1: Community workshop on standards and infrastructure for predicted effector gen…

Day 2: Community workshop on standards and infrastructure for predicted effector gen…

Workshop Aims

- Raise awareness of the diversity of approaches to creating PEG lists
- Ensure that PEG lists with their supporting evidence and related metadata are FAIR and relevant to the user community
- Establish a community standard for reporting PEG lists
- Define strategies for evaluation and collation of PEG lists
- Identify challenges in sharing, maintaining, updating PEG lists
- Identify strategies for incentivizing implementation/use of the proposed standards

Participant instructions

This document is for live note-taking during the workshop. Any attendee can contribute to these notes by suggestion or commenting. In particular, please suggest edits to the notes to correct anything you have said.

- Participants will be muted upon entering the meeting and during the presentations. Everyone will be unmuted during the question and discussion sessions. Participant videos will be off by default.
- Asking questions/making comments: 'Raise your hand' or comment in the Zoom chat. Moderators will read out questions and comments typed in the chat.
    - To raise your hand: Click on the 'React' icon (heart icon), you will then see a clickable 'Raise hand' box.

Actions to take now:

- Edit your Zoom profile to display your full name and affiliation.
    - To edit your Zoom display name: Click on Participants button, next to your name click on the 'More' drop down, you'll have the option to 'Rename' yourself.

Agenda

Monday September 16: 9am - 1pm EDT / 2pm - 6pm BST

9:00 EDT / 2:00 BST: Introduction and welcome (Noël Burtt, Knowledge Portal Network)

9:20 EDT / 2:20 BST: Who's that causal gene? Lessons learned in manually annotating 100,000 GWAS SNPs (Eric Fauman, Pfizer Inc.)

10:20 EDT / 3:20 BST: Open Targets as an end user of PEG lists (Yakov Tsepilov, Open Targets)

10:35 EDT / 3:35 BST: Open discussion

10:45 EDT / 3:45 BST: Break

10:55 EDT / 3:55 BST: Overview of current gene prioritization efforts and PEG lists (Maria Costanzo, Knowledge Portal Network)

11:25 EDT / 4:25 BST: 10-minute presentations of gene prioritizations

- Cassandra Spracklen, University of Massachusetts
- Brent Richards, McGill University
- Adam Butterworth, University of Cambridge

12:30 EDT / 5:30 BST: How ClinGen handles framework standardization with input from many invested users (Marina DiStefano, ClinGen)

12:05 EDT / 5:05 BST: Open discussion

1:00 EDT / 6:00 BST: Adjourn

Tuesday September 17: 9am - 1pm EDT / 2pm - 6pm BST

9:00 EDT / 2:00 BST: GWAS Catalog's experience developing GWAS standards (Laura Harris, GWAS Catalog)

9:15 EDT / 2:15 BST: Curation and display of PEG lists in the Knowledge Portals (Maria Costanzo, Knowledge Portal Network)

9:25 EDT / 2:25 BST: Curation and display of gold standard lists at Open Targets (Xiangyu Jack Ge, Open Targets)

9:35 EDT / 2:35 BST: Open discussion: advantages/disadvantages of PEG list presentations

9:50 EDT / 2:50 BST: Results of community survey (Maria Costanzo, Knowledge Portal Network)

10:00 EDT / 3:00 BST: Standardization requirements for use of PEG lists as input for

computational and AI methods (Jason Flannick, Boston Children's Hospital)

10:15 EDT / 3:15 BST: Break

10:25 EDT / 3:25 BST: Proposal for PEG list standards (Yue Ji and Laura Harris, GWAS Catalog)

10:40 EDT / 3:40 BST: 40-minute breakout group discussions about the proposed standards. Guiding questions will be provided.

11:20 EDT / 4:20 BST: Breakout groups report back; further discussion in the larger group

12:40 EDT / 5:40 BST: Wrap up (All organizers)

1:00 EDT / 6:00 BST: Adjourn

Notes

Notes for Monday 16 Sept 2024

Introduction (Noel Burtt) RECORDING

- Desired outputs from the workshop- engaged community, white paper, working toward deposition resource, roadmap to usable data
- Introduction to perspectives of workshop host resources -KP Network (summary representations of gene/region for non-expert; integrative analysis/meta analyses, community driven) and GWAS Catalog (repository resource, experience in setting standards with community)
- Day 1 - landscape; people and perspectives on effector genes. Perspectives from industry and resource, multiple perspectives from academia/different disease areas, rare disease field.
- Day 2 - more interactive, open discussion - what matters for contributors and consumers, input from community survey, strawman to provoke discussion on standard for breakout discussion

Who's that causal gene? Lessons learned in manually annotating 100,000 GWAS SNPs (Eric Fauman, Pfizer Inc.) RECORDING

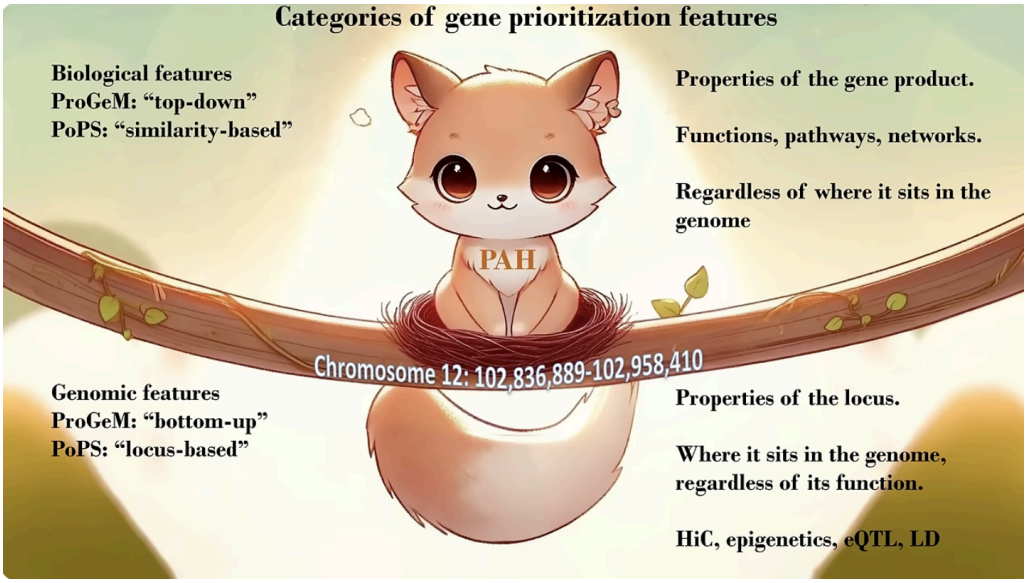
- Eric has background in biochemistry/protein structure. Introduction to GWAS in collaboration with Nicole Soranzo, circulating phenylalanine GWAS - one hit, who is that causal gene.
- When the trait is protein abundance a reasonable hypothesis is that the causal gene is the gene encoding that protein, the causal gene. Ie in protein QTLs
- Generally the distance from pQTL lead SNP to the TSS of the cognate gene is v short. But you get a bimodal distribution corresponding to cis-pQTL and trans-pQTLs

How a simple model of enhancers and promoters explains all the pQTLs in the world (linkedin); 10.1186/s12859-022-04706-x

- The cognate gene is usually the closest gene, even for intronic and intergenic variants.\

- Metabolite QTLs - phenylalanine GWAS example - Sources used for manual curation for finding possible causal genes on same chromosome as single GWAS hit on chr 12- 1) HMDB (metabolite 'known interactors'), 2) mouse knock outs (MGI) (KO has effect on phenylalanine) , Orphanet (gene containing 'disease causing germline mutation resulting in disorder of Phe metabolism). - each method identifies a single candidate on chr 12 = PAH.
- Noel B question - have you re-evaluated existing recognised causal genes for whether or not they are closest?
    - EF - should always consider closest, even just as comparator for new method
- Q - does eQTL data result in same bimodal distribution as pQTL?
    - EF - we dont' have that data.
- Why should we care about protein/metabolite QTLs when what we are really interested in is diabetes, IBD [complex disease]? Argues that mechanism must ultimately be acting through proteins/metabolites.

- [33:45] Categories of gene prioritisation features - biological features (properties of the gene product, the fox) vs genomic features (properties of the locus, where the fox's nest is). Eric's work mostly focused on biological features.
  Analogous to :ProGEM bottom up/top down- ; PoPs - similarity based/locus based



- Validate metabolite Qtl and pQTl against one another
- Correlation between protein and metabolite abundance mostly in same direction (subsetting to colocalizing pairs)
- Observation [46:22] - relative strength of a genetic association usually reflects biological proximity to the causal gene

- PheWAS

- [52:50] A framework for the reproducible curation of gwas causal genes (w Brent Richards) - using 4 types of information -mouse knockout, genes for med disease, drug target, biological mechanism (EGLN1 as causal gene for RBC traits as example)
    - ‘The criterion of biological saliance is surprisingly inclusive’ ie easy to make assertions, How to evaluate relevance of evidence? We don’t know the best answer but need to be explicit and transparent about inferences (credit to Helen Parkinson)
- Q Laura Harris - no standard for establishing salience of evidence, are there stat methods? EF - has been discussing Bayesian approach w Jason. JF - not possible without prior assumptions [couldn’t hear].. 250 genes?
- NB - the lists you show are effector gene lists - how can we apply standards to make these more interoperable and machine readable. - EF - attempts to structure and include machine readable e.g. accession IDs, but the assertion is free text
- Q Adam Butterworth - Building on the topic of the 'biological salience', this is clearly a highly predictive method for metabolite QTLs, but as you highlighted with the schizophrenia paper, biological salience is a considerably poorer predictor for complex disease GWAS signals. Does this imply that we need different rules (and possibly standards) for different kinds of GWAS phenotypes?
- EF - getting sense for types of traits this works for, haven’t focused on psych traits
- JF - mouse traits work better for metabolic traits, maybe mouse are poor psych models

Open Targets as an end user of PEG lists (Yakov Tsepilov, Open Targets) RECORDING
- OTAR - private public partnership between academic partners (EBI, Sanger) and industry partners; focused on targets to prioritised for drug development
- GWAS evidence is one stream of evidence used in OTAR in gene-drug target prioritisation effort. GWAS evidence flows through the OTAR Genetics pipeline.
- EFO trait -> GWAS -> distil to credible sets -> list of genes -> list of functional genomic features -> prioritisation score (L2G score in platform)
- Q EF -1) what maximum distance do you allow? 2) all genes or only protein coding genes?
- A Xiangyu Ge 1) 500kb 2) only protein coding
- Disease level feature matrix presents summary of evidence (inc coloc, VEP).
- Effector gene list (defined as EFO/gene pairs) and disease level feature matrix together used to train L2G model -> output prioritisation score
- Q Oleg Borisov How do you define strong evidence of colocalization in terms of posterior probabilities, e.g., do you use any threshold like PP.H4 > 0.8?
- Q Satoshi Yoshiji how are gold standard positive genes defined?
- and gold standard negative genes. Thank you!
- Q Albert Henry How do you account for overlapping samples / studies for a given EFO in the training?
- Q Adam Butterworth Do you have a recommended way of interpreting the L2G score for a credible set, or is it deliberately left to users to choose?

- Produces `GWAS genetic evidence as feature metric -estimates prioritization score

What is an effector gene lists - an established EFO-trait pair, or feature matrix with information for each gene in the region?

How would OTAR use effector gene list
- directly as evdience/additional column in association page,
- in training model

Overview of current gene prioritization efforts and PEG lists (Maria Costanzo, Knowledge Portal Network) RECORDING

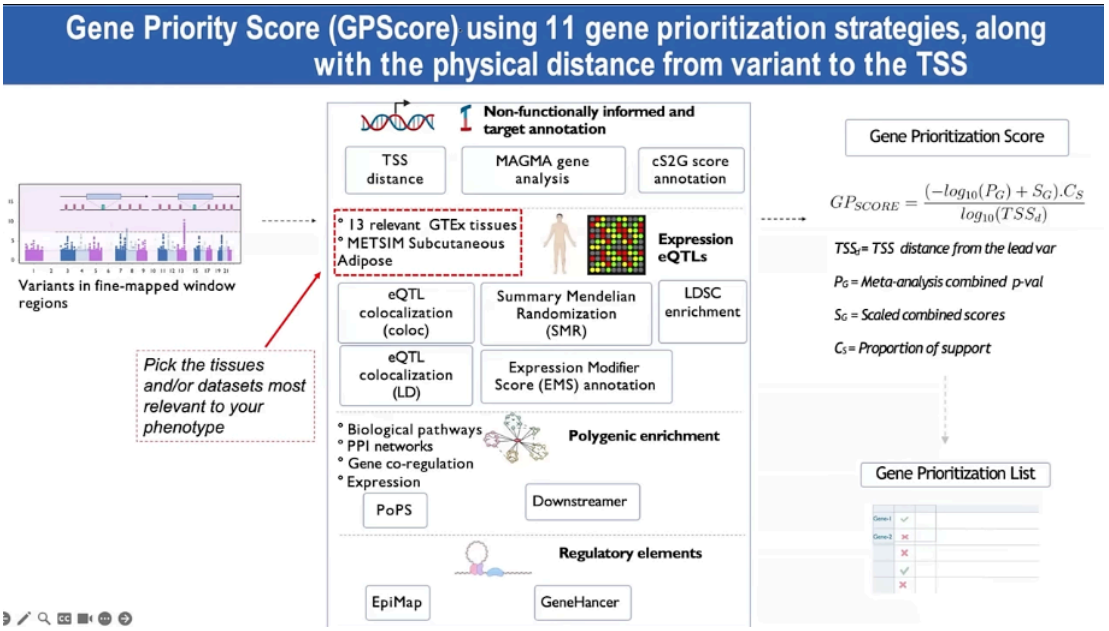
- Closest gene not always causal -> gene prioritization - evidence in each locus -> PEG list scoring system.
- History of history of effector gene lists in the context of the AMP T2D Knowledge portal. Original Anuba Mahajan effector gene list applied a heuristic to combine evidence types into a categorisation of evidence strength (causal/strong. Moderat etc). Started gathering lists for integration in the KP
- 3 categories of evidence -> overall categorisation. Curating -> rep in knowledge portals. Increasing?
- GWAS literature search for gene prioritisation. 169 papers. Number of papers increasing over time.
- What evidence used? Categories:
    - variant-centric -> nearest gene, chromatin confirmation, impact molecular properties - relevant tissues;
    - gene-centric -> within loci, “Guilt by association”, eg. proteins interact. Perturbation. Gene-burden. Literature/online resources;
    - Pipelines
- How many evidence types per study? ~4.
- Trends: QTLs, nearest gene (not universal). Perturbational not so used.
- Gene prioritisation vs. PEG lists. 75% of papers integrated all evidence in a PEG list. 25% included gene prioritisation only. Eg. Some evidence type in separate tables.
- Q: papers about evidence sources or about gene? A: More supplementary. Not objective to make overall PEG list. Specific goal in mind of author. How to make it easier for them?
- Some PEG lists only images. Tables in image format. What types of graphics could represent?
- Some present evidence for all genes in a locus. 71% of papers. The rest only give top gene per locus.
- Scoring system - only 29%.
- Comparing lists for the same trait - not too many comparable, often from same group refining GWAS and PEG list. Comparing different groups -> different formats, lots of manual work. 6/9 loci concordance.
- Conclusion: landscape = Wild West. No consensus or standardisation.
- Minimal standards suggestions:
    - Document the GWAS, boundary coordinates of loci, methods, criteria for significance, scoring system. Format - downloadable and interoperable spreadsheet. Graphical element separate. Make it obvious which gene has

most evidence. Don't count duplicate evidence types as independent observations. [More in slide.]

- Comment: Laura Harris GWAS Catalog - Author reported gene used to be a field curated. Stopped because it was time-consuming and no data behind that, what was it telling us? Use Ensembl mapping pipeline. Would like to use PEG list so data is shown behind it. How important is scoring system reporting? A: Popular in survey, think should be included. How to enforce? Standards?
- Brent Richards - experience of standardisation in mendelian randomisation, STROBE checklist 10.1001/jama.2021.18236. Resulted in improvement, need to convince journal editors to get on board
- Q Eric Fauman: Found authors weren't clear on how they arrived at mouse phenotype column? A: Yes, details not always provided.
- Q Yakov Tsepilov - data that doesn't go in publications, generated by large scale efforts like OTAR. Jason F - are these effector gene lists or different outputs, these large scale effort may not have disease specific expertise

10-minute presentations of gene prioritizations

- **Cassandra Spracklen, University of Massachusetts RECORDING**
  - Presented one method GPScore: a method to take list of genes from GWAS hit and narrow down. Combine multiple approaches to 1 prioritisation scheme. Eg. adiponectin as a proof of concept trait.
  - A lot of methods disagree with each other or have different strengths of evidence. Wanted to focus on datasets which give weight to some evidence over others (eg. eQTL data - adipose tissue as most relevant tissue so should be upweighted).
  - Score integrates eQTL, polygenic enrichment, non-functionally informed target annotation, regulatory elements, distance to TSS, MAGMA gene analysis, biological pathways, PoPs, …
  - Detail in paper/ Github https://github.com/vsarsani/GPScore , 10.1016/j.xhgg.2023.100252



- Performance for adiponectin - identified well known (IRS1, ADIPOQ, CCDC92 etc) and novel target genes (CSF1, RGS17, ADRB1, CYP2R1 - concludes these are valid targets based on literature based evidence for involvement in adiponectin)
- Can view relative contribution of terms from the GPSscore
- Also looked at other Finngen traits, beyond PoC. e.g. pulmonary fibrosis, autoimmune hypothyroidism - also find expected strongest prioritised genes
- Adiponectin list in the the Knowledge Portal

- **Brent Richards, McGill University**
  - The Effector Index - prioritising genes at GWAS loci.
  - Can we develop a method which provides the probability of causality of each gene at a GWAS locus? Would the method recover positive control genes?
  - Looked at 12 traits / diseases (e.g. Type 2 diabetes, LDL, height, Ca levels,etc)
    - But positive control genes were lacking
  - Arrived at Positive control genes - 381 from Mendelian diseases + 113 drug targets
  -
  - Major problem(s) in effector gene lists.
  - Can use knowledge from clinical medicine to assess if PEG seem like sensible targets.
    - E.g. IBD drug targets in Open Targets, assessing the 35 genes given using clinical knowledge. OPRM1 (target for Fentanyl) is given as a possible drug target gene - using clinical knowledge this seems unlikely. VDR (Vitamin D)
    - SCN4A (Lidocaine target) appears as possible target because often given in surgery.
    - Potassium channels often appears as drug targets for MS, mainly because there is one rarely use KCN drug for MS.
    - Conclusion many target genes (or predicted effector genes) are erroneous
  -

- MR - loose standards of reporting. STROBE-MR statement has improved MR reporting (JAMA), hasn't cleaned up field, required by many journals.
-
- Presented method:
-
- Effector Index - includes SNPEff, trait matched cell/tissue type, snv-gene relationship, also locus specific features, multiple models
- How does Effector Index perform? Beat locus-level features?
- Results can change if positive control genes are changed
    - Switched to expertly curated T2D PEG list (McCarthy/Mahajan), changed performance
- Every new algorithm seem to beat every other algorithm… Biased, based on data trained on. Hard to find positive controls. Metrics depend on which positive controls are used and validated against.
- Accuracy estimates deflated because most causal genes not known. Can clean up list, use different thresholds for inclusion of genes, but vast majority of causal genes are unknown?
- Believes should only be interrogating with GWAS signals. Thinks a positive outcome would be agreed upon positive control lists for a limited number of traits.

- **Adam Butterworth, University of Cambridge RECORDING**
    - Predicting GWAS effector genes for CAD.
    - Many different routes / places to start. 8 different predictors/metrics (e.g. nearest gene, monogenic disorder, drug trial, PoPs, eQTL, KO mouse, protein altering variant, rare variant assoc with CAD), taking 200 GWAS hits, annotating against each metric -> simplistic approach, summing. All info in sup tables of paper [ADD HERE]. Most likely for region. ~220 causal genes could prioritise -> can prioritise further based on number of concordant predictors (lines of evidence)
    - Discussed if they should use scoring methodology, but have lack of positive control genes. Some info sparse. Simplistic to understand. Decided to keep it simple and pragmatic.
    - Acknowledges Robert Plenge's RA GWAS gene prioritisation approach (Okada et al, Nature, 2014)
    - Metrics: Nearest gene, protein altering, rare variant (subjective, how to define an association, phenotype specific, and sparse), gene causing relevant monogenic disorder (which disorders relevant?, challenging to expand beyond small numbers of phenotypes), eQTL in CAD-relevant tissue (STARNET vs GTEx), gene implicated by drug trial or MR (MR somewhat subjective, literature based), relevant phenotype in mouse knockout (challenge to restrict to phenotypes that are disease relevant), Polygenic Priority Score priority (PoPS) (how can score be incorporated, ranked?).
    - Validation: came up with gold standard of known causal genes beforehand (community expert set). 25 were the single highest ranked gene in locus. 3 joint highest ranked, 2 'wrong'. Acknowledges that this list is not unbiased because some of the predictors provided the evidence used to define the 'known' genes.

- What learnt: Tailoring for phenotype is useful; could we expand this to related diseases/traits
- Multi-disease weighting. Multiple datasets - challenges integrating - which better / sum. Gold standards - how best validate?
- Largely ignoring 'literature' which provides evidence of causality for the gold standard list - doesn't get picked up by data sources used.

How ClinGen handles framework standardization with input from many invested users (Marina DiStefano, ClinGen)

- These discussions are reminiscent of early days of their standardization journey. Emphasises the importance of buy in, no point in developing standards if no one uses them.
- Standardising terminology
    - Eg.1. Delphi Survey: Iterative survey methodology.
    - ClinVar part of consortia, eg. Gene Curation Coalition (GenCC) - like ClinVar for genes. Assertions of gene-disease relationships
    - All compared approaches. Pilot set of gene curations, all curated and discussed. Helpful to understand what each resource did. Identified discordance - definitions, high/low, curating for differnet reasons - gene panel, gene disease relationship, different diseases, older. Compared, resolved and harmonised. Harmonising definitions = Phase1.
    - Phase 2 = prioritisation. All terms used, all GenCC members, scale of agreement (including neutrality). Prompted to add additional terms for next round.
    - Phase 3 = refinement. Terms scored worst removed.
    - Phase 4 = consensus and implementation. International genetics community. 93% term agreement. -> Terms used to standardise submissions. Way to display thousands of curations on website using same terms.
    - Objective and transparent metrics.
- Engaging experts - Gene curation framework standardisation
    - Develop, test (33 gene disease pairs), 2 curator scores - independent, same amount of evidence -> conference with expert / review -> measure concordance. 94% between curators. Between curator and expert 88%. Flexibility in framework, can tailor to disease area, points gap eg. .5 to discuss.
      Basic framework:

Assertion criteria	Genetic Evidence (0-12 points)	Experimental Evidence (0-6 points)	Total Points (0-18)	Replication Over Time (Y/N)
Description	Case-level, family segregation, or case-control data that support the gene-disease association	Gene-level experimental evidence that support the gene-disease association	Sum of Genetic & Experimental Evidence	> 2 pubs w/ convincing evidence over time (>3 yrs)
Assigned Points				
CALCULATED CLASSIFICATION		LIMITED	1-6	
		MODERATE	7-11	
		STRONG	12-18	
		DEFINITIVE	12-18 AND replication over time	
Valid contradictory evidence? (Y/N)	List PMIDs and describe evidence:			
CURATOR CLASSIFICATION				
FINAL CLASSIFICATION				

    - Buy-in to framework? (as it takes time.) As many curations as possible, all on website. Engage professional societies - policy statements.
    - Uptake? Eg. Brugada syndrome - only 1 gene kept.
- Variant curation framework standardisation
    - Survey assessing adoption of ACMG-AMP guidelines - interpreting seq variants…
    - 25% of US labs implemented a “more advanced” version of the 2015 guidelines.
    - Engaged even more pro societies to get buy-in for standards. Need to engage as many as possible, and international societies if poss.
    - Need to harmonise adjacent frameworks. (eg. CNV)
    - Continuously engage the community - present, audience poll.
    - Public pilot of rules - invested users will help.
    - Flexibility: Provide a mechanism for updates to your framework, version it, be willing to receive feedback about when it works and does not work.
    - Q: Common disease GWAS, sumstats accessible... Rare disease more open to these types of things? A: rare disease community much like common disease community. Expert panel eg. craniofacial, harmonise with ClinGen standards a struggle - disease naming, went to each small community and asked what doing / naming. Deal-breakers / bring everyone in, most important things to retain in area.
    - Q: Envision GenCC replace all resources that went into making it? A: Strength of GenCC still linking to primary data sources with granular data. Good place to aggregate, link out.

Open discussion: what are the key features of each approach and how do they compare? RECORDING

- What is effector gene list? Consensus - all genes around variant with evidence - a table? Not just a list. Proposing should include all genes in a region. Plus score. Protein-coding only? Questions for tomorrow. Any genes the contributor decides? Might not be defined by the region, may be the entire GWAS results. Define ourselves or encourage list generator to define.

- So maybe there isn't full consensus on what defines an effector gene list
    - OT combining studies? Yes, different levels. Predict differently.
- List around one GWAS or around the publication? Some analyse several studies together, can't say which study used.
- Overweighting eQTL in tissue of relevance. Assumes known which tissue relevant.
- What incorporated in deposition of list.
- Difference between GWAS and rare disease (ClinGen example). Generators may not be users. What are uses, how used, what they'd want to see - short list, high-confidence gene / pharma company - big list. Need to understand use cases better. Downstream use-cases.
- Example of when PEG list used? What did, how useful.

Zoom chat

Karl Heilbron 9:53 AM
For directionality, did you subset to colocalizing pairs?

Maria Costanzo 10:26 AM
Eric's PEG list on the Knowledge Portals:
https://hugeamp.org/research.html?pageid=egea_217

Adam Butterworth 10:26 AM
Building on the topic of the 'biological salience', this is clearly a highly predictive method for metabolite QTLs, but as you highlighted with the schizophrenia paper, biological salience is a considerably poorer predictor for complex disease GWAS signals. Does this imply that we need different rules (and possibly standards) for different kinds of GWAS phenotypes?

Eric Fauman 10:40 AM
1) what maximum distance do you allow? 2) all genes or only protein coding genes?

Xiangyu Ge 10:41 AM
1) 500kb 2) only protein coding

Satoshi Yoshiji 10:46 AM
how are gold standard positive genes defined?

Oleg Borisov 10:45 AM
How do you define strong evidence of colocalization in terms of posterior probabilities, e.g., do you use any threshold like PP.H4 > 0.8?

Yakov Tsepilov 10:52 AM
We have two methods of colocalization: coloc and eCAviar. For coloc it is h4>0.8 , for eCaviar it is clpp>0.8

Satoshi Yoshiji 10:47 AM
and gold standard negative genes. Thank you!

Yakov Tsepilov 10:52 AM
Jack will give more details on this tomorrow

Albert Henry 10:47 AM
How do you account for overlapping samples / studies for a given EFO in the training?

Yakov Tsepilov 10:52 AM
We combine all features for particular genes from alll relevant studies selecting the best value

Adam Butterworth 10:50 AM
Do you have a recommended way of interpreting the L2G score for a credible set, or is it deliberately left to users to choose?

Yakov Tsepilov 10:53 AM
Usually we use threshold of 0.5

Laura Harris 11:07 AM
Just a reminder we have a shared notes doc, we are updating it as we go and feel free to add anything you want to add or clarify:
https://docs.google.com/document/d/1BDdR3PiM6vyStju4OAzfTWxo3kfJCP8FdPIRgs9zq7M/edit?usp=sharing

Laura Harris 11:16 AM
EuropePMC extract info from tables as xml

Sylvanus Toikumo 11:18 AM
Would you be happy to share a few of the papers with the PEG graphics or Table you just shared? Really look cool!

Noel Burtt 11:20 AM
Yes, we can
all the slides will be in the google drive which you should have access to, but let us know if you cannot get in

Eric Fauman 11:23 AM
ensemble geneids are more stable than HGNC gene symbols

Karl Heilbron 11:25 AM
No question, but wow, I am so impressed by this Herculean work

Marina DiStefano 11:28 AM
I'll talk a bit about scoring implementation and how we got the community to use it in my talk about ClinGen

Joannella Morales - NHGRI 11:32 AM
Very informative, Maria!

Laura Harris 11:33 AM
With respect to pipeline vs publication - I think the key is interoperability, even if the representation is not identical

Anna Kottgen 11:35 AM
I joined a bit late, so this might have been covered earlier: when assessing the question of “disease-relevant tissue”, how do you deal with sample size imbalance in eQTL data? Many tissues are not well covered in GWAS, so consortia resort to tissue-specific datasets related to their traits of interest. This leads to many “disease-relevant, tissue-specific” eQTLs, when they may pertain to other tissues too. Maybe there is some time to discuss this tomorrow. GTEx, I mean (not GWAS)

Wafaa Rashed 11:38 AM
I think you should take in consideration the difference between ethnicity while you calculate the score.. otherwise it will end into restricted application in certain population not globally

Sylvanus Toikumo 11:18 AM
Would you be happy to share a few of the papers with the PEG graphics or Table you just shared? Really look cool!

Maria Costanzo 11:39 AM
I’ll put those references and other examples into the shared meeting notes.

Sylvanus Toikumo 11:40 AM
Thanks!

Eric Fauman 12:19 PM
@Adam Butterworth - enjoyed that a lot. I wanted to link our metabolite work and your CAD work. One of the things that really helped the INTERVAL metQTL work was the concept of the metabotypes, or the full profile of phenotypic consequences of each variant. Applied to CAD, you'd find "lipid" loci, "inflammation" loci, "vascular biology" loci. This falls out of the PheWAS of each locus. Clearly the type of gene expected at a lipid loci will be different than the type of gene expected at a vascular biology locus. Rather than applying global rules for this complex disease I think it is helpful to look for rules for different causal mechanisms for the disease, like I discussed on HbA1C.

Karl Heilbron 12:44 PM
Do you envision GenCC ultimately replacing all of the resources that went into making it?

Cassie Spracklen 12:52 PM
Only protein coding genes? Or all transcripts?

Eric Fauman 12:53 PM

I am aware of 2 causal genes that are not protein coding: TERC for telomere length, and CDKN2B-AS1 for CAD

Karl Heilbron 12:59 PM
We're doing this right now for schizophrenia

Cassie Spracklen 1:00 PM
Noel--having a statement on the portal that says "this gene has shown up in these effector gene lists" feels potentially less helpful if the standard is to have all tested genes appear in the list (scored)? Unless it has some way to indicate the level of evidence (instead of just presence/absence in the list)?

Albert Henry 1:01 PM
I think it'd be useful to discuss the outputs that are expected by the end users. Is it just simply an unweighted list of genes, ordered list with score / probability, or list with some confidence labels (e.g. ClinGen definitive / strong / etc evidence label)

Notes for Tuesday Sep 17, 2024

GWAS Catalog's experience developing GWAS standards (Laura Harris, GWAS Catalog) RECORDING

- Overview of the experience of the GWAS Catalog realising a standard was required, organising a community workshop, running working groups, arriving at a standard and implementing the standard for GWAS Catalog submitters.
- The workshop brought together around 50 stakeholders, including tool developers, data generators, and funders, to discuss the problems and potential solutions. Two working groups were formed to focus on the file format and content, and higher-level issues around privacy and diversity. After three meetings and an online survey, a standard format was proposed and published in Bio Archive and on Github. The standard was then refined based on community feedback and implemented in the GWAS Catalog and other resources. The outcome has been significant, with over 85,000 summary statistics files available in the Us catalog, a 33% increase in submission rate, and all new submissions containing the mandatory content.
    - Relevant reading - MacArthur et al, 2021 Workshop proceedings: GWAS Catalog standards and sharing. Cell Genomics Vol. 1, Issue 1, 100004.
    - Materials from community workshop on GWAS summary statistics sharing and standards
    - A community driven GWAS summary statistic standard
    - Cerezo et al, 2025 https://doi.org/10.1093/nar/gkae1070 Nucleic Acids Research, Volume 53, Issue D1, Pages D998–D1005
    -
- Thinking about FAIR for PEG lists, using GWAS Catalog as a comparator for GWAS

summary statistics
- Findable
    - unique ID for each list,
    - Searchable by study metadata
    - Linked from a central repository
    - Linked to source data in GWAS Catalog
- Accessible
    - Available from repository
    - Persistent metadata (perhaps formatted/stored separately from the actual PEG lists)
- Interoperable
    - Well defined data model
    - Controlled vocab for traits and evidence types
    - Standard nomenclature for variant, gene, region
- Reusable
    - Clear data usage licence
    - Detailed provenance (incl. links to original GWAS)
    - Mandatory metadata

Question from Maria Costanzo - How long did the whole process take? A - started taking sum stats in 2018 (OTAR has been a big push for standardisation), workshop held 2020, standard implemented in 2023 - so 5 years, but could have been quicker if more resources and without a pandemic.

Comment: Noël
- GWAS summary statistics are rather simple compared to PEG lists
- PEG lists have lots of different types and sources of data, will be difficult to standardise

Need “gatekeeping” - incentives to force people to submit/share PEG lists (similar to requirements that now exist for submitting GWAS sumstats to GWAS Catalog). It is harder now to submit summary statistics now that a standard is mandated, but authors are usually required to do it by journals, need to balance the effort required by the submitter with the incentive/pay off. Mandating too much

Ellie McDonagh - PEG lists may not be static, will evolve over time as new evidence is introduced - this may make it difficult to assign stable permanent identifiers
Laura Harris - the standards for metadata really made a difference to the quality of the data. Structured vs free text in paper would be a huge step forward for PEG lists.

Curation and display of PEG lists in the Knowledge Portals (Maria Costanzo, Knowledge Portal Network) RECORDING
- Sources of PEG lists: GWAS literature, literature survey, communications from authors

- When necessary, worked with authors to: obtain data, understand and document methods, represent evidence in a logical way
- BYOR (Bring Your Own Results) platform

- Minimal data provided for all lists, more detail for some but this became too much work over time except where there was input from authors
- Spreadsheet with expandable rows to show more data
- Lists available through PEGKP.org, sortable/filterable list of all the PEG lists
- Coverage of wide range of traits
- Search by gene feature - some genes are on many lists
- PEG lists available on gene page of Knowledge Portals also available on the Tools tab in the a2fkp and cmdkp

Curation and display of gold standard lists at Open Targets (Xiangyu Jack Ge, Open Targets) RECORDING

- There is not curation itself, but a prediction done in Open Targets called a Gold Standard list
- Data sources for Gold Standard positives (ChEMBL, PrGem, Eric's list
- For each phenotype-gene gold standard, there is a min of info
    - E.g. disease/trait name, gene IDs
- Curation (study index, credible sets and colocalisation), with the credible set as essential. There is in silico prediction (VEP and Polyphen2), distance (credible set variants, genes), Chromatin interaction and QTL colocalisation. The most important factors in the prediction are distance and strongest colocalisation
    - Distance is important because many PEG lists select the nearest gene (?), but also because it is a feature that is present in ever list, while other features may be missing

Satoshi Yoshiji: Can we run the analysis with our own data? Jack, in theory is possible (Gentropy package) there could be some problems linked to the information/format of the data as input.

Open discussion: advantages/disadvantages of PEG list presentations

Results of community survey (Maria Costanzo, Knowledge Portal Network) RECORDING

- Open in March, 58 responses good distribution of different stages of academia but also industry.
- Which terminology do you think is the most appropriate for a gene that is predicted to be responsible? Causal
    - Conclusion "predicted effector gene" - generally acceptable and seems clearest
    - A lot of arguments against using the term "causal gene" - implies high burden

of evidence
- Most people interested in the list linked to their trait and/or related traits
- Barriers to using PEG lists
    - Lack of transparency and provenance
    - …
- Most people want a combination of bioinformatics and manual curation to produce PEG lists
- Most people prefer a numerical value
- Most people want to see evidence for all genes, not just the top gene
- Most people want to be able to compare multiple lists, both within and across traits
- Most want to be able to interact with lists, and assign different weights to different evidence types

As summary of the survey- People want to have access to all possible evidence for the list, sources of data

Standardization requirements for use of PEG lists as input for computational and AI methods (Jason Flannick, Boston Children's Hospital) RECORDING

- We are not talking about big data (only up to 100s, maybe 1000s of genes), not AI generating the list, but using a short number of genes as primary source

Possible uses:

- Truth data for training models to predict genes
- Benchmark data for testing models to predict genes
- Repositories of gene "features" for building additional models
- "Knowledge" of disease/gene relationships for use in applications such as knowledge graphs

Use as training data:

- Needs: not only positive labels but negative labels as well (i.e. true positives, true negatives), confidence of genes or tiers of confidence (for weighting, rank of genes not enough for this, need a quantification of confidence)

Use as input features:

- Needs:
    - input features in addition to labels
    - Ability to match features across lists
    - Common units and definitions of the feature values
    - "Squared off" matrix
    - How to handle substructure of features
- will be difficult to compare across datasets (e.g. how do you know that experimental data in one set is comparable to another), but controlled vocabularies etc. may help
- Trade-offs between standard formats (JSON vs matrixes etc.)

Knowledge graphs:

- Needs:
    - Encode predictions as edges in a knowledge graph

    - Encode supporting data as provenance
    - Contribute these to knowledge graph efforts (SmartAPI, biolink)

Comparing two lists

- Illustrates how difficult it is to use these lists
- For example, matching up loci is really difficult. How is a locus defined? Which genes were considered?

AI/ML requirements

Clean, consistent

Which features should each list have? There are different categories of standards, required or allowed values, metadata, format and repositories.

A major challenge is that some lists have ranked genes, some have scores and some have categorisations of genes (e.g. likely causal, possibly causal…)

- How to make the data interoperable? Maybe start with some of current lists. Encouraging people to submit the list when they do a GWAS, with some numerical score maybe involving some journals, funding agencies and professionals

Proposal for PEG list standards (Yue Ji and Laura Harris, GWAS Catalog)
RECORDING

- Matrix of evidences instead of a initial list of genes. It should be suitable for publication in a journal, submission to a resource and integration in a pipeline
- Metadata standards:
    - Must include the GWAS the list was derived from
    - Focusing here on single list per GWAS
    - Standard terminology for evidence types
    - Criteria for significance for each evidence type
    - Easy to see whether the list contains de novo wet lab evidence, or purely based on computational process and evidence already in databases
    - Details of methods of prioritisation
    - Focused here on a standard suitable for used in individual publications
        - Standard for these must also be interoperable with pipeline-generated data
- Data standards
    - Single list/matrix containing all evidences
    - Plain text file not including graphics
    - Present evidences for all genes considered for each locus
    - Use standard identifiers
    - Include sentinel variant
- PEG standard proposal based on most common data types identified from 56 publications
    - 1. Identifies the source of primary GWAS
    - 2. Provides the sources for each evidence file
    - 3. Describe method used to identify PEG
    - Structured in tabular format, includes data for all evidence types, presents

evidence for all genes, universal score (tbd)
    - Should be both human- and machine-readable
    - Consider ontology to structure and annotate evidence types (e.g. Evidence and Conclusion Ontology, ECO)

Suggestion to ask data generators to provide the actual credible sets

Breakout group discussion and reporting back RECORDING
40-minute breakout group discussions about the proposed standards. Guiding questions will be provided.

Breakout groups report back; further discussion in the larger group

Zoom room 1: Cassie Spracklen

- Every study will use whatever sources they want when making their decisions - tables could become huge - would be good to have a condensed summary table of genes and ranking/prioritisation, that then links to paper or some other source with all the detailed data sources
- More user friendly to provide a short identifier (e.g. lead variant) for each credible set, rather than providing all of the variants
- Querying on lead variants would be useful
- Versioning:
    - not everyone will update their lists, more likely to update the list by doing a new GWAS analysis
    - however others (e.g. Open Targets) might be more likely to repeatedly update the same list with new data
- Suggestion: the big table should be hosted on the repository and be searchable, then the author can choose which level of detail to display in the paper. It could also include a checklist about how to reach that PEG list.

- Rasika Mathias - PEG lists are very subjective - everyone uses different approaches to pick their effector genes, how do you go from that to curating and making them interoperable (e.g. standards of evidence may be wildly different between lists)

- Jason Flannick - important that lists can't be treated as data. Instead they are an investigator's inferences, conclusions. Need to use a different approach to present knowledge vs data. Recognise that there is subjectivity, and determine what need to be provided to back up the conclusions. Summary of rationale, justification for including genes in the list and assigning them to certain categories.
    - Can have submitter's judgement of the evidence, but also a curator/expert committee's judgement of the evidence, could involve community curation in this, crowdsourcing tools

- Rasika Mathias - would be important to provide authors with guidance for how to assign genes to prioritisation categories based on the evidence they do or don't

have.

- Cassie Spracklen - Important to indicate ancestry of populations in GWAS and other relevant cohorts, e.g. for tracking of diversity

Zoom room 2: Noël Burtt

- Insist on sentinel variant, credible set less essential
- Evidence from rare disease, other human curated data sources very valuable
- Like the idea of scoring
- Need submission accessions and/or versioning system - need to be able to say if one list is an updated version of a previous list
- Provide as many automating tools as possible to assist submitters - e.g. map automatically to nearest gene
- Lots of people interested in working on PEG lists (helping develop standards and also submitting their own data) beyond today

Zoom room 3: Jason Flannick

- Whether it is time to work on this problem
    - This is still under development and subjective - maybe best at this stage to provide guidelines, rather than mandate very strict standards
- Embrace that rankings are going to be subjective, so instead provide guidelines on how to explain and back up the chosen rankings
- Biggest value is to provide guidelines on explaining how the list and data sources were generated

EBI in-person group: Aoife McMahon

- Agreed that at this stage, trying to create a universal score is not possible - too early for that
- Now we need to mandate that people provide the evidence that they used to create their own score/ranking etc.
- Can suggest categories and provenance of evidence types, but should be flexible
- Way to distinguish between new experimental data and data that has been accessed/ingested from existing sources, this should be clear from provenance
- Format - Should make clear that fields should only be filled if author actually used that data to generate the list - i.e. not filled retroactively with additional annotations

Broad in-person group: Maria Costanzo

- How to deal with link between list and journal publications - is a peer reviewed paper required for a list to be included? What happens if paper is retracted?
- Don't want to make barriers too high
- These lists are also being made already for mouse, rat etc. - should standards be harmonised with efforts in model organism as well?
    - A lot of opportunities to synergise with model organism PEG lists
    - Could multiple species be brought together in the same platform?
- Gold standard data sources:
    - Mendelian disease

    - Mouse knockouts
    - Validated drug targets
- Would be nice if authors had to submit a free-text abstract to explain what they did
- Would be good to provide a tool to help authors create a list with some automated operations (e.g. mapping to nearest gene)
- Let the journals be the enforcers of compliance to standards - otherwise, lower quality data can get through
    - Placing the QC step as a requirement for publication works as an incentive for academic submitters

Wrap up (All organizers)

- A lot of discussion to synthesise
- Seems that we are at the stage where we can convene working groups
- Benchmarking of current effector lists that are already accessible
    - Work with GWAS Catalog to back-populate lists already on Knowledge Portals into the GWAS Catalog's "straw man" proposed format - see how well the current lists fit
- Talk to people who are just about to generate a list - work with them to define a standard
- Think about what can be done with controlled vocabularies vs. free text
- This process will probably take a few years, but off to a good start!

- Coming up:
    - Ancillary session at ASHG this year - will give readout of discussions from this workshop
- Follow-up workshop in 3-6 months' time?

Zoom chat

Chi Zhang 9:47 AM
To follow up on what to capture in the portal, I think PEG is more than a list. Like the wonderful presenters mentioned yesterday, there are a lot of other efforts that went into this. For example, what tissue types are relevant, what OMIM traits are related, etc. We probably want to capture all this somewhere.

Karl Heilbron 9:49 AM
When you say "87 Phase II", did these start or complete Phase II?

Daniel Considine 9:59 AM
completed

Karl Heilbron 10:00 AM
And did it need to complete "successfully"?

Daniel Considine 10:01 AM

Yes - the phase II drugs would be those that passed phase II but not further (for many potential reasons)

Karl Heilbron 10:02 AM
Great, thanks!

Chi Zhang 9:47 AM (Edited)
To follow up on what to capture in the portal, I think PEG is more than a list. Like the wonderful presenters mentioned yesterday, there are a lot of other efforts that went into this. For example, what tissue types are relevant, what OMIM traits are related, etc. We probably want to capture all these somewhere.

Laura Harris 10:02 AM
Yes, I think we need to think about this as a “matrix” rather than a list

Karl Heilbron 10:10 AM
I agree with Maria, “candidate” is

Cassie Spracklen 10:10 AM
At least in the traits I work with, we used to call them “candidate genes” before it was decided to use the phrase “effector genes”? Based on the study sections I sit on, I think its still relatively common for those to use that language...

Laura Harris 10:12 AM
To me a candidate gene is hypothesis-based ie. not from a GWAS

Mark Keller 11:04 AM
For variant id's, should Rs# be included. If coding, can consequences also be included?

Laura Harris 12:08 PM
I think the relationship between a summary table and the paper should be the other way round - the huge table should be in the resource and searchable - then we can choose what level of detail to display. In the paper you have the option to just present the “top results”

Ellie McDonagh-Open Targets, Translational Informatics Director 12:26 PM
Example of the PanelApp functionality, where registered experts/curators can provide their own evaluation of gene-disease association:
https://panelapp.genomicsengland.co.uk/panels/285/gene/CLCN4/ and provide the underlying evidence.

Ellie McDonagh-Open Targets, Translational Informatics Director 12:26 PM
(open source code that could be utilised and developed if of interest)

Ellie McDonagh-Open Targets, Translational Informatics Director 12:26 PM
Example of the PanelApp functionality, where registered experts/curators can provide their own evaluation of gene-disease association:

https://panelapp.genomicsengland.co.uk/panels/285/gene/CLCN4/ and provide the underlying evidence.

Ellie McDonagh-Open Targets, Translational Informatics Director 12:31 PM
Everything is versioned controlled too