Leveraging supervised learning for functionally-informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs

2020/10/17 by Qingbo Wang

A detailed description for the analysis performed is available at:
Wang. Q. S. *et al.* **Leveraging supervised learning for functionally-informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs, *biorxiv* (2020)**.

The list of additional putative causal eQTLs are available as a **supplementary file in the manuscript.**

A subset of the Expression Modifier Score (EMS), a predicted probability that a variant has a cis-regulatory effect on gene expression, trained on fine-mapped eQTLs and leveraging 6,121 features including epigenetic marks and sequence-based neural network predictions, is available for download at https://www.finucanelab.org/data .

Specifically, for each tissue, the list of top variant-gene pairs with highest EMS (= those with normalized EMS > 100) is available as *ems_top_{tissue_name}.tsv.bgz* .

Column descriptions:
v: variant in hg38
g: gene ID
rf_score_raw: Raw output from the random forest predictor (ranging from 0 to 1)
ems: the expression modifier score (a predicted probability that the variant has a cis-regulatory effect on the gene expression)
ems_normalized: Normalized EMS, corresponding to EMS divided by the probability in a random draw

Full EMS results in google cloud:

(Updated 20240521)

The full EMS results, saved as hail tables, are **not** in a public google cloud directory, but are still available upon request. Please contact qingbow@broadinstitute.org if access is needed (We expect users need access to this file ONLY IF the tsv.bgz file is insufficient for their analysis).

~~The full EMS for all the variant-gene pairs in GTEx are also available as a hail table (*gs://expression-modifier-score/public/ht/ems_{tissue_name}.ht* for each tissue).~~
~~However, due to the cost excess, the files are under requester pays bucket in google cloud.~~

~~This means that you will need to supply Google Cloud billing information when you run your pipeline, and **need to pay the computational costs**. **We recommend users to access this file ONLY IF the tsv.bgz file is insufficient for your analysis**.~~

~~For more detailed information on how to access the file and how the billing process works, please see the descriptions below:~~
~~Accessing the files:~~

- ~~When using hailctl to create Dataproc clusters, please start clusters with the flag *--requester-pays-allow-buckets expression-modifier-score*. More information can be found at https://hail.is/docs/0.2/cloud/google_cloud.html#requester-pays~~
- ~~When using *gsutil* command to download the files, add a *-u* argument with the ID of the GCP project to be billed. For example,~~
  ~~*gsutil -u my-proj cp -r gs://expression-modifier-score/public/ht/ems_Whole_Blood.ht ./*~~
  ~~More information can be found at~~
  ~~https://cloud.google.com/storage/docs/using-requester-pays#using~~

~~The cost:~~

- ~~Our data is in a US-multiregion bucket, which means that, if your cluster is also in the US, reading the file itself is for free. To export the file, you will need to pay for the export, as guided in the requester pays instructions.~~
- ~~If your cluster is outside of US-regions, you will need to pay for the read (e.g. $0.01/GB from Canada)~~

~~An example of other resources following a similar practice (gnomAD) would also be helpful.~~