DAPPLE

I. Background

DAPPLE stands for Disease Association Protein-Protein Link Evaluator. DAPPLE looks for significant physical connectivity among proteins encoded for by genes in loci associated to disease according to protein-protein interactions reported in the literature. The hypothesis behind DAPPLE is that causal genetic variation affects a limited set of underlying mechanisms that are detectable by protein-protein interactions. Please refer to the DAPPLE publication for full details.

DAPPLE takes as input a list of genes, SNPs or genomic regions. See section II for a detailed description. It will build direct and indirect interaction networks from proteins encoded for by seed genes. It will then assess the statistical significance of a number of network connectivity parameters as well as of the connectivity of individual proteins to other seed proteins using a within-degree node-label permutation method. The individual protein scores are then used to propose candidate genes in large loci. **Please note that the most recent release uses adaptive permutation whereby networks that do not achieve P < 0.1 (for any parameter) are stopped at 100 permutations.**

II. Inputs

DAPPLE takes 4 types of input:

- A list of **genes**. These are entered as one entry per line, either in a specified file or directly via the webpage interface. Each gene should be identified with its gene symbol (ie Hugo) ID, such as "ATXN1". This mode should be used if the user does not want to group genes into regions, but rather wants each gene to stand as its own region.
- A list of **SNPs**. These are entered as one SNP per line, either in a specified file or directly via the webpage interface. Thse SNPs must be in HapMap or 1KG, because this is how DAPPLE defines the 'wingspan' region around a gene which is a function of linkage disequilibrium.
- A list of **regions**. These are entered as one region per line. Each region should be entered as 'ID chr left right' where ID is a region identifier, chr is a number from 1-23, left is the left boundary in genomic coordinates and right is the right boundary. The entries can be space or tab delimited.
- A list of **genes with region identifiers**, or "gene-regions". These are entered as one entry per line, either in a specified file or directly via the webpage interface. Each entry should be defined as 'gene ID' where gene is a gene name in gene symbol (ie Hugo) ID, such as "ATXN1", and ID refers to a region to assign the gene to. Since DAPPLE is specifically looking for connectivity between regions - and not within regions - the user can group genes based on how they want to define groups.

Here are some examples:

Gene input

```
PANK4
HES5
TNFRSF14
MMEL1
PADI4
PTPN22
```

SNP input

```
rs3890745
rs2240340
rs2476601
rs11586238
rs7528684
rs12746613
```

Region input

```
rs3890745 1    2395699   2744704
rs2240340 1    17471282  17551282
rs2476601 1    113874482 114254482
rs11586238     1    117057482 117097482
rs7528684 1    155807552 156083552
rs12746613     1    159656552 159750570
```

Gene-region input

```
PANK4     rs3890745
HES5      rs3890745
TNFRSF14  rs3890745
MMEL1     rs3890745
PADI4          rs2240340
PTPN22    rs2476601
```

Combination input

```
rs3087243
rs6822844
region1 1 159656552 159750570
region2 1 196861967 197040967
SLC26A10
PIP4K2C
```

III. Output Files

DAPPLE outputs a number of files, all of which are described here. "FILE" refers to the keyword input by the user.

FILE_summary: This file contains the parameter values for the 4 network statistics measured: (1) The number of direct connections between seen proteins from different loci, (2) the average seed protein direct connectivity (a.k.a. direct binding degree), (3) the average seed protein indirect connectivity (a.k.a. indirect binding degree) and (4) the average common indiractor binding degree (the average number of seed proteins that common interactors bind to).

FILE_NetStats: This file contains the permutation p-values for the 4 network statistics described in FILE_summary (i.e., what is the probability that I see a parameter value >= the observed value by chance?)

FILE_SeedScores: This file contains the individual p-values for seed proteins - generally, the probability that by chance the seed protein would be as connected to other seed proteins (directly or indirectly) as is observed. Please refer to the publication's supplementary materials for exact details of p-value calculation. The file contains 4 columns: gene ID, region ID, uncorrected p-value, corrected p-value.

FILE_GenesToPrioritize: This file contains genes that achieved a corrected p-value less than 0.05.

FILE_CIscores: This file contains the p-values for common interactors that describe the probability that by chance individual common interactors would be as connected to seed proteins as was observed.
FILE_directConnections: This file contains a list of the direct connections in the network.

FILE_plot: If the user chose plot=true, this is the visualization of the network. Page 1 shows the direct network and pages 2-3 show the indirect network. Colors of seeds correspond to region.

FILE_MissingGenes: This file is important to pay attention to. If the input is SNPs or regions, this describes the genes in those input regions that are in the InWeb database in contrast to those that aren't. If too many input proteins are not in the InWeb database (less than 60% average inclusion), one should be careful about interpreting DAPPLE results.

FILE_permuted*: Values of permuted parameters

IV. FAQ

What is DAPPLE testing? The hypothesis behind DAPPLE is that causal genetic variants affect common mechanisms and that these mechanisms can be inferred by looking for physical connections between proteins encoded in disease-associated regions. DAPPLE is therefore testing whether the networks built from seed regions - both direct networks and indirect networks - are more connected than chance expectation. Chance expectation is defined by the connectivity expected if connectivity were purely a function of the binding degree of participating proteins.

If I input a SNP, how is the region defined around that SNP? The region is defined using LD according to the HapMap. For a given SNP, we extend out to the region defined by SNPs in r^2>=0.5 and then extend out to the nearest hotspots.

For a region, how are overlapping genes defined? The hg18 gene list was downloaded from UCSC using Ensemble transcripts. Splice isoforms were then collapsed to define the largest gene footprint from transcription start to transcription stop. Gene footprints were then extended on either end to include 50kb of regulatory sequence by default, though the user can specify a different regulatory region. Any gene footprint that overlaps a region is included in that region. If a gene overlaps 2 regions, those regions are merged. If the user would like to keep the regions seperate, they should input genes and explicitly assign them to regions (option #3 on "What type of input does DAPPLE take?").

Where does the protein-protein interaction data come from? We use the InWeb databased, published by Kasper Lage in 2007. This database contains 428,430 reported interactions, 169,810 of which are deemed high-confidence, non-self interactions across 12,793 proteins. High-confidence is defined by a rigorously tested signal to noise threshold as determined by comparison to well-established interactions. Briefly, InWeb combines reported protein interactions from MINT, BIND, IntAct, KEGG annotated protein-protein interactions (PPrel), KEGG Enzymes involved in neighboring steps (ECrel), Reactome and others as described elsewhere in detail. All human interactions were pooled and interactions in orthologous protein pairs passing a strict threshold for orthology were included. Each interaction was assigned a probabilistic score based on the neighborhood of the interaction, the scale of the experiment in which the interaction was reported and the number of different publications in which the interaction had been sited.

V. References

Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tartar D, IIBDGC, Cotsapas C, Daly MJ. 2011 Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. PLoS Genetics 7(1): e1001273

Lundby, A. & Rossin EJ et al. Annotation of loci from genome-wide association studies using tissue-specific quantitative interaction proteomics. Nat. Methods 11, 868–874 (2014).

VI. Contact

Please email dapple@broadinstitute.org with any questions.

This version of the DAPPLE algorithm is public server only, validated and tested on the Broad CentOS 5 environment. It is not ready to be installed on other GP servers