

Power Analyses for SEQLinkage

Wed Apr 2 10:45:41 CDT 2014 | Let's rock and roll!

Basics

We simulate and analyze two autosomal recessive genes GJB2 and SLC26A4 and two autosomal dominant genes MYO7A and MYH9. The program takes two genes source data as input.

As we call seqlink from inside of the program multiple times we'd like to speed up the execution by using ramdisk of 8000MB.

```
sudo mkdir -p /ramcache
sudo mount -t tmpfs ramfs -o size=8000m /ramcache
cd /ramcache
ln -s ~/SVN/SEQLinco/simulations/LinkagePowerCalc.py pcal.py
```

Hierarchy of simulation parameters

I want to draw one figure with 4 panels. Each panel is one mode of inheritance, namely A&B are recessive and compound recessive using recessive genes; C&D are dominant and compound dominant using recessive genes. On each panel, the X-axis is sample size and Y-axis is allelic heterogeneity. Powers are contour plots of two colors, one color for new method, one for old method. The only parameter not scanned here is --offsprings. For now I set it in between 3 to 8. It can be varied for additional simulations in supplemental. In this figure only one of the gene pair will be reported and the complementary figure can be found in supplemental.

Linkage analysis parameters are always -K 0.01 -W 0 -M 1. See seqlink -h for details.

Example command

```
python pcal.py -g MYH9.tsv MYO7A.tsv -m dominant -r 10 --sample-size 20 \
--blueprint blueprint.txt --vanilla --tempdir /ramcache \
--run-linkage -K 0.01 --moi AD -W 0 -M 1 --output-entries 0
```

Power calculation script

Script generator:

PYTHON

```
name = 'pcal.sh'
rep = 500
batch = 100000000
count = 0
gdict = {'recessive': 'SLC26A4.tsv GJB2.tsv', 'compound_recessive': 'SLC26A4.tsv GJB2.tsv',
         'dominant': 'MYO7A.tsv MYH9.tsv'}
mdict = {'recessive': 'AR', 'compound_recessive': 'AR', 'dominant': 'AD'}
fdict = {'': 'CHPResult.csv', '--single-markers': 'SNVResult.csv'}
with open(name, 'w') as f:
```

```

for m in ['recessive', 'dominant', 'compound_recessive']:
    for a in [' ', '-a']:
        for p in [i * 0.01 for i in range(0, 110, 10)]:
            for n in range(5, 81, 1):
                for method in [' ', '--single-markers']:
                    count += 1
                    f.write("./pcal.py --debug -s 3 8 -g {0} -m {1} -n {2} -p {3} {4} -r {5} "\
                        "--ofile {9} {10} --blueprint blueprint.txt --vanilla {6} --jobs 1 "\
                        "--run-linkage -K 0.01 -W 0 -M 1 --moi {7} --output-entries 0 "\
                        "--tempdir /ramcache/tmp > {8}.{9}\n".\
                        format(gdict[m], m, n, p, 1-p, rep, method, mdict[m],\
                            fdict[method], "RUN%s" % count, a))
if count % batch == 0:
    f.write("echo {}\n".format(count))

```

Use command below to run the generated script on desktop:

```
cat pcal.sh | gw-parallel -j 7
```

Or on cluster:

```
cat pcal.sh | gw-qsub -s 50
```

Power calculation result

Consolidate all result data into database PowerCalc

```

cat CHPResult.csv.* > CHPResult.csv
cat SNVResult.csv.* > SNVResult.csv
./pcal.py --print-header | sqlite PowerCalc -i CHPResult.csv --as CHP --header - -d",
./pcal.py --print-header | sqlite PowerCalc -i SNVResult.csv --as SNV --header - -d",

```

To view attributes of the database

```

sqlite PowerCalc -s
sqlite PowerCalc -s CHP
sqlite PowerCalc -s SNV

```

All output data and results are backed up to 040*.tar.gz for future reference. Now let's have a rough feeling on power comparison between the methods

BASH

```

for k in 0 1; do
    for i in recessive dominant compound_recessive; do
        for j in 1 2; do
            echo $i Gene$j, allelic heterogeneity $k
            echo CHP=`sqlite PowerCalc "select avg(plod$j) from CHP where moi = '$i' and ahet = $k" \
            SNV=`sqlite PowerCalc "select avg(plod$j) from SNV where moi = '$i' and ahet = $k"
        done
    done
done

```

Results below are based on data from 0403.tar.gz:

OUTPUT

When there is no allelic heterogeneity the traditional SNV based method should be more powerful as it fits the underlying model better. Indeed SNV method is slightly more powerful for dominant mode (since creating many patterns instead of using one alternative original SNV introduces noise and lowers LOD score as observed by Di). However surprisingly there is little difference in power for recessive. When there is allelic heterogeneity there is significant power gain for compound recessive. We observed little power gain for dominant, as the improvement was offset by the use of multiple patterns than it might be necessary.

For compound recessive the CHP method is always more powerful than SNV method.

Thu Apr 3 14:51:32 CDT 2014 | A sleepy afternoon with voltaren all over my hands ...

Power plot

Script see `PowerContour.py` under this folder. Graphs below are based on data from `0403.tar.gz`: