# PLINK/Seq

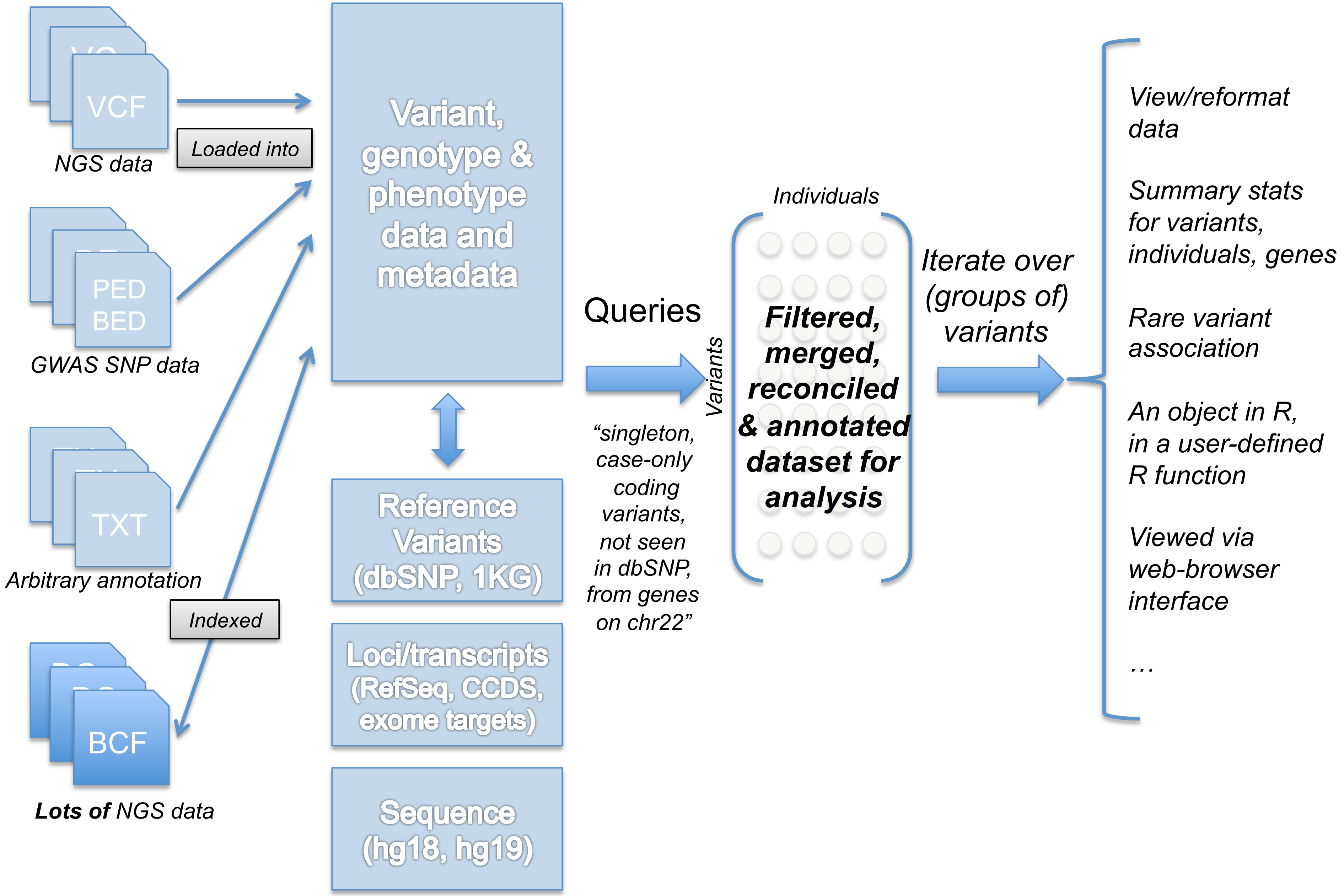
## Analysis of genetic variation data from large-scale, population-based medical sequencing studies

Shaun Purcell
shaun@pngu.mgh.harvard.edu

Analytic and Translational Genetics Unit, MGH
Center for Human Genetic Research, MGH

| PLINK |
|---|
| GWAS |
| PED file |
| Large datasets<br>(but held in RAM) |
| Common variation |
| Simple SNPs |
| Sequential tests of single variants |
| Analysis largely<br>“self-contained” |
| Command-line interface |
| Downstream of Birdsuite |

“PLINK/Seq project”



VCF

NGS data

Loaded into

PED
BED

GWAS SNP data

TXT

Arbitrary annotation

Indexed

BCF

**Lots of** *NGS data*

Variant, genotype & phenotype data and metadata

Reference Variants (dbSNP, 1KG)

Loci/transcripts (RefSeq, CCDS, exome targets)

Sequence (hg18, hg19)

Queries

*“singleton, case-only coding variants, not seen in dbSNP, from genes on chr22”*

*Individuals*

*Variants*

***Filtered, merged, reconciled & annotated dataset for analysis***

*Iterate over (groups of) variants*

*View/reformat data*

*Summary stats for variants, individuals, genes*

*Rare variant association*

*An object in R, in a user-defined R function*

*Viewed via web-browser interface*

*…*

# Command-line interface: `pseq`

Some basic commands illustrated here: a growing number of utilities for viewing, filtering, annotating, presenting summary statistics and performing various types of association analysis

*Initiate a new project*

```
./pseq project1 new-project
       --vcf /path/to/data1.vcf.gz /path/to/data2.vcf.gz
       --resources /path/to/core/databases/hg19
```

*Populate with phenotype and genotype information*

```
./pseq project1 load-vcf
./pseq project1 load-pheno --file /path/to/data.phe
```

*View variants, genotypes for a certain gene, excluding variants with non-PASS filters*

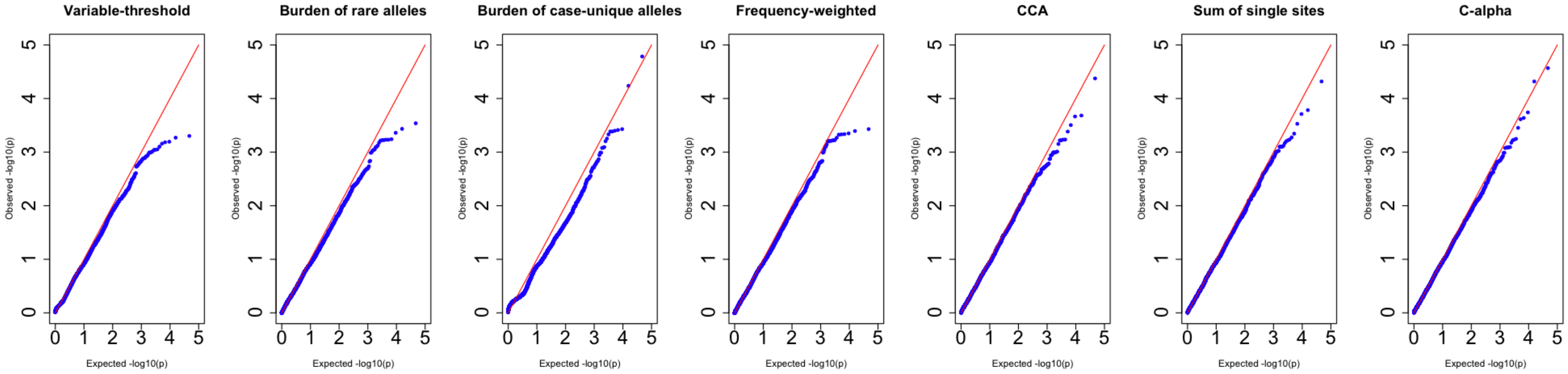
```
./pseq project1 v-view --gene ABC123  --mask any.filter.ex
```

*Gene-based association*

```
./pseq project1 assoc --phenotype dis1 --mask loc.group=refseq
```

# A growing number of gene-based rare-variant association tests

- A core set of methods, including the *variable threshold test* (Price et al, 2010) and *C-Alpha* (Neale et al, in press)
- External investigators collaborating to incorporate their methods in this framework (e.g. J. Witte, S. Leal)
- PolyPhen2 weights (courtesy of Sunyaev lab) can be applied



One-sided tests

Two-sided tests

# Filters and grouping data with *masks*

*List variants in genes in file* `mygenes.txt` *that are seen not more than twice, not in dbSNP, and are unique to cases. Exclude the MHC region. Set to missing individual genotypes with an individual read depth less than 10.*

```
./pseq project1 v-view
       --mask loc.subset=refseq,@mygenes.txt
              mac=1-2
              ref.ex=dbsnp
              case.uniq
              reg.ex=chr6:25000000..35000000
              geno=DP:ge:10
```

*Arbitrary expressions can be evaluated based on a variant’s metadata, to use as a filter for inclusion in analysis, or to produce on-the-fly new metadata*

```
--mask include=“ XX = g( GQ > 0.95 );  DB || XX > 0.8 “
```

*Here, a new tag XX is added to the variant for this particular analysis*

*The function* `g(cond)` *gives the proportion of individuals for whom GQ greater 0.95*

# Filtering against 1000 Genomes data

*(1) Obtain VCF from 1000Genomes FTP site (sites and some meta-information)*

```
wget ftp://ftp-trace.ncbi.nih.gov/1000genomes/…/ALL…sites.vcf.gz
```

*(2) Load into a "REFDB" (database of reference variants, e.g. dbSNP, HGMD, etc)*

```
./pseq - load-ref  --refdb g1k.db --group g1k
       --vcf ALL.2of4intersection.20100804.sites.vcf.gz
```

*(3) With REFDB as part of project, can filter your data for presence in G1K and G1K metadata*

```
./pseq /path/to/project v-view
       --mask ref=g1k  v-include="g1k_AF < 0.01 && g1k_DP > 100"

  chr1:865628:rs41285790  G/A g1k_DP=1955;g1k_AF=0.002
  chr1:879413:rs116279254 G/A g1k_DP=1321;g1k_AF=0.005
  chr1:879482:.           G/C g1k_DP=1366;g1k_AF=0.004
  chr1:892569:rs41285806  C/T g1k_DP=2226;g1k_AF=0.002
  chr1:901922:rs62639980  G/C g1k_DP=1171;g1k_AF=0.007
  ...
```

# Accessing data via R

*Attach an existing PLINK/Seq project*

```
pseq.project( “/path/to/my/project” )
```

*Either “apply” a user-defined function* `func1` *to the data given a mask…*

```
res <- var.iterate( func1 , “mac=1-2 any.filter.ex” )
```

*… or obtain a list of variants*

```
k <- var.fetch( “mac=1-2 any.filter.ex” )
```

*Various convenience functions to work with variant and variant group data and metadata*

```
x.consensus.altcount( k )
```

*Individuals*

*Variants*

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
[1,]    2   NA    1    2    2    2    1    1
[2,]   NA    1    1    2    2    1    2   NA
[3,]    0    0    1    1    0   NA   NA   NA
[4,]   NA    0   NA   NA   NA    1    1    0
```

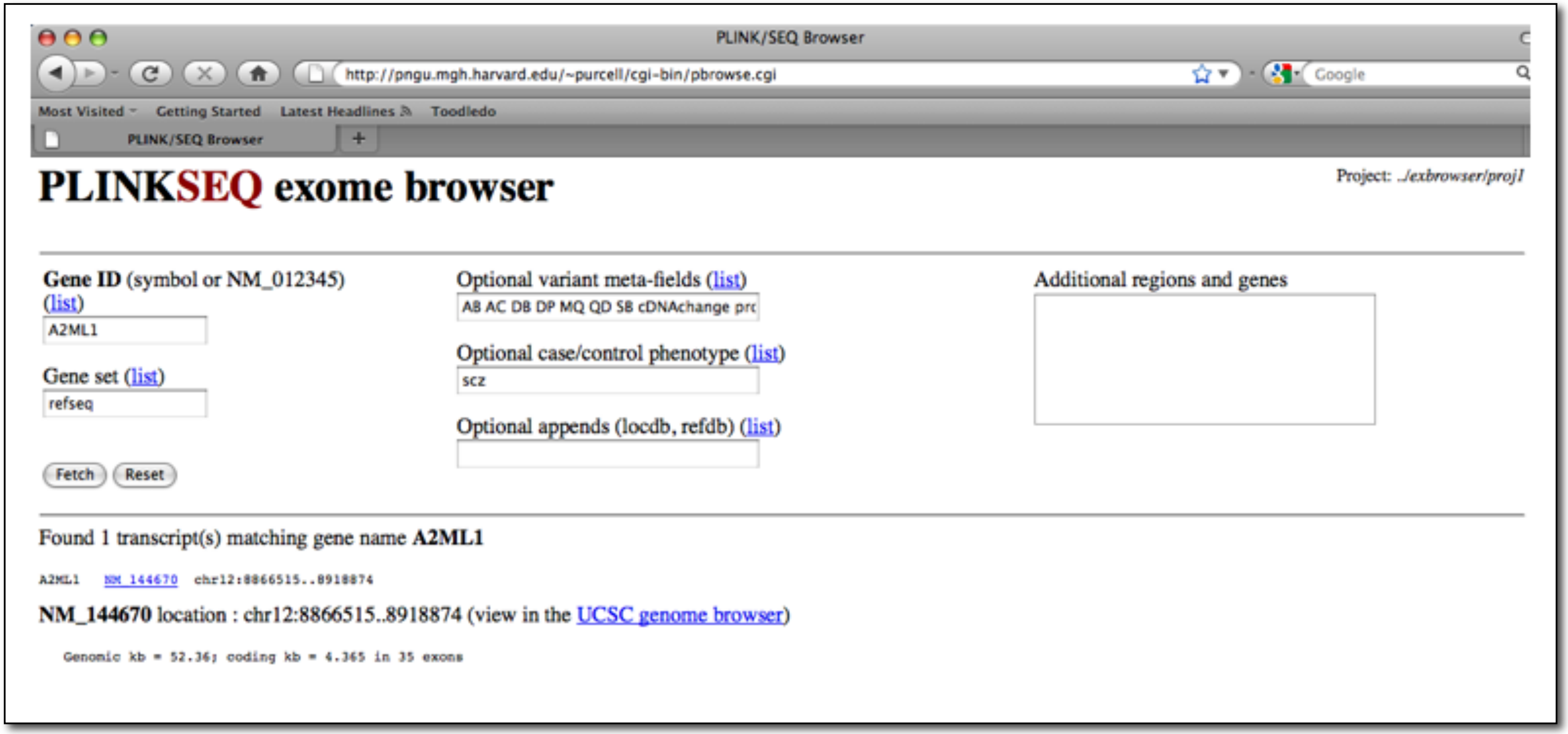
```
> str( k[[1]]$VAR[[1]] , max.level=3)

List of 7
 $ CHR: int 10
 $ BP1: int 61472483
 $ BP2: int 61472483
 $ ID : chr ".”
 $ CON:List of 8
  ..$ REF   : chr "C"
  ..$ ALT   : chr "T"
  ..$ QUAL  : num 13194
  ..$ FILTER: chr ”PASS"
  ..$ META  :List of 20
  .. ..$ cDNAchange   : chr "c.13106G>A"
  .. ..$ codonchange  : chr "c.(13105-13107)CGG>CAG"
  .. ..$ gene         : chr "ANK3"
  .. ..$ genomechange : chr "g.chr10"
  .. ..$ proteinchange: chr "p.R4369Q"
  .. ..$ strand       : chr "-"
  .. ..$ transcript   : chr "NM_020987"
  .. ..$ type         : chr "Missense"
  .. ..$ AC           : int 2
  .. ..$ AN           : int 262
  .. ..$ DB           : int 0
  .. ..$ DP           : int 49758
  .. ..$ AB           : num 0.61
  .. ..$ AF           : num 0.01
  .. ..$ MQ           : num 94.7
  .. ..$ QD           : num 17.2
  .. ..$ SB           : num -5285
  ..$ GENO  :List of 4
  .. ..$ GT: int [1:132] 0 0 0 0 0 0 0 0 0 0 ...
  .. ..$ DP: int [1:132] 355 259 504 463 451 76 499 99 420 349 ...
  .. ..$ GL: num [1:132, 1:3] -1.16 -7.27 -5.08 -4.89 -1.2 -1.15 -1.24 -3.06 -1.15 -3.37 ...
  .. ..$ GQ: num [1:132] 99 99 99 99 99 99 99 99 99 99 ...
```

PLINK/Seq library to handle the large datasets and serve up variation data according to “genomically-oriented” queries

R to provide a rich, standardised and well-documented statistical and graphical environment

Easy prototyping of methods

```
## C-alpha, implemented as a single R function

calpha.test <- function(g) {
   d <- x.consensus.genotype(g);
   y <- apply( g[ ,p==1 ], 1, sum )
   n <- apply( g, 1, sum )
   score <- sum( (y - n * ratio)^2 - n * ratio2 )
   var <- sum( sapply( lwr:upr , function(m)
               sum(length(n[n==m])
                   * ((0:m-m*ratio)^2-m*ratio2)^2
                   * dbinom(0:m,m,ratio))))
   return( score/sqrt(var) )
}

## Attach project

  pseq.project(”/path/to/my/project")

## Phenotype data from individual datastore

  p <<- inddb.phenotype("scz")

## Specify a mask: rare missense variants in CCDS genes

  mask <- “loc.group=CCDS mac=2-10 include =
           type == ‘Missense’ || type == ‘Nonsense’ “

## Run the analysis

  results <- vardb.iterate( calpha.test , mask )
```



# PLINKSEQ exome browser

Project: *../exbrowser/proj1*

**Gene ID** (symbol or NM_012345) (list): A2ML1

Gene set (list): refseq

Optional variant meta-fields (list): AB AC DB DP MQ QD SB cDNAchange prc

Optional case/control phenotype (list): scz

Optional appends (locdb, refdb) (list):

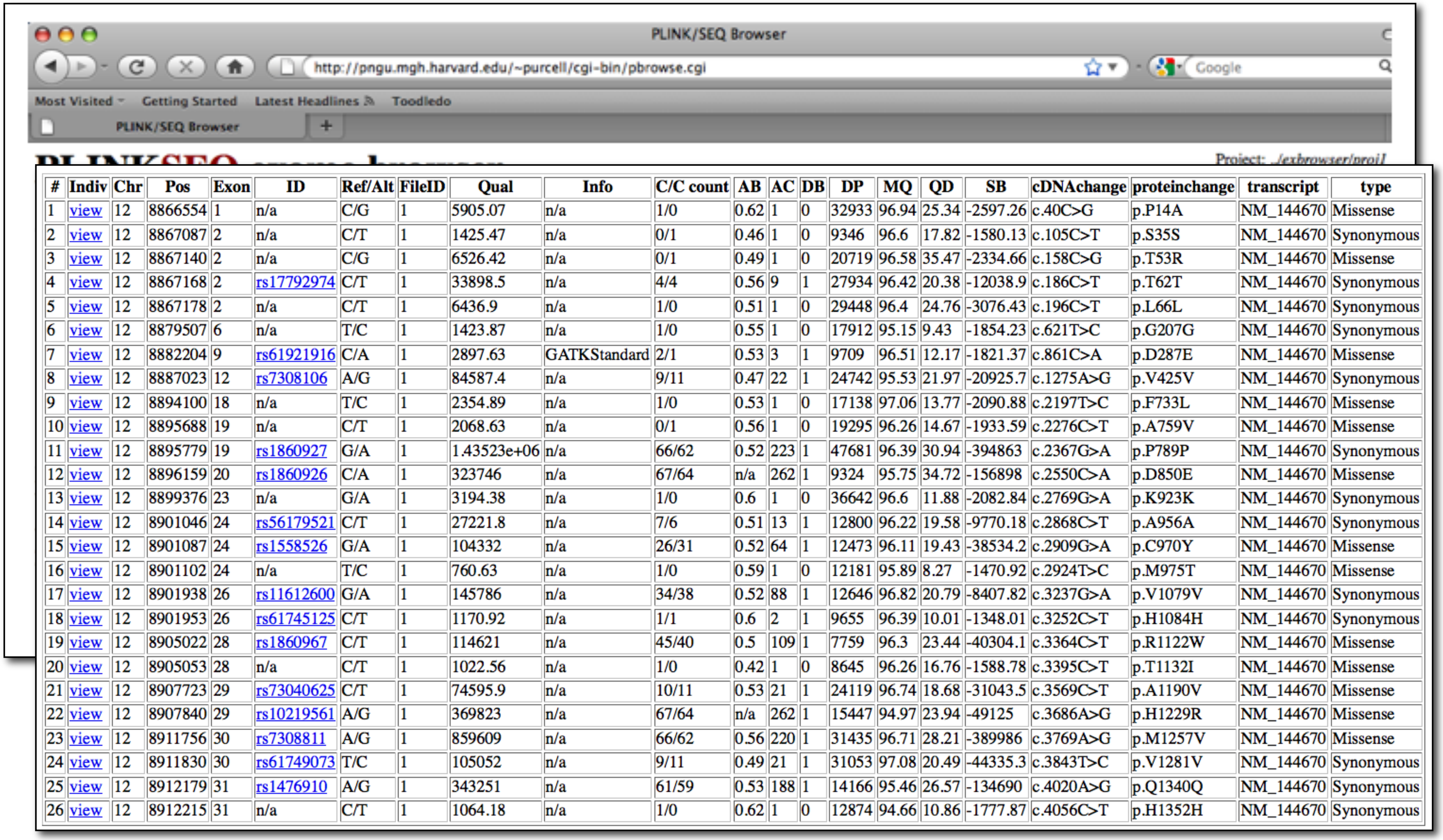
Additional regions and genes

Found 1 transcript(s) matching gene name **A2ML1**

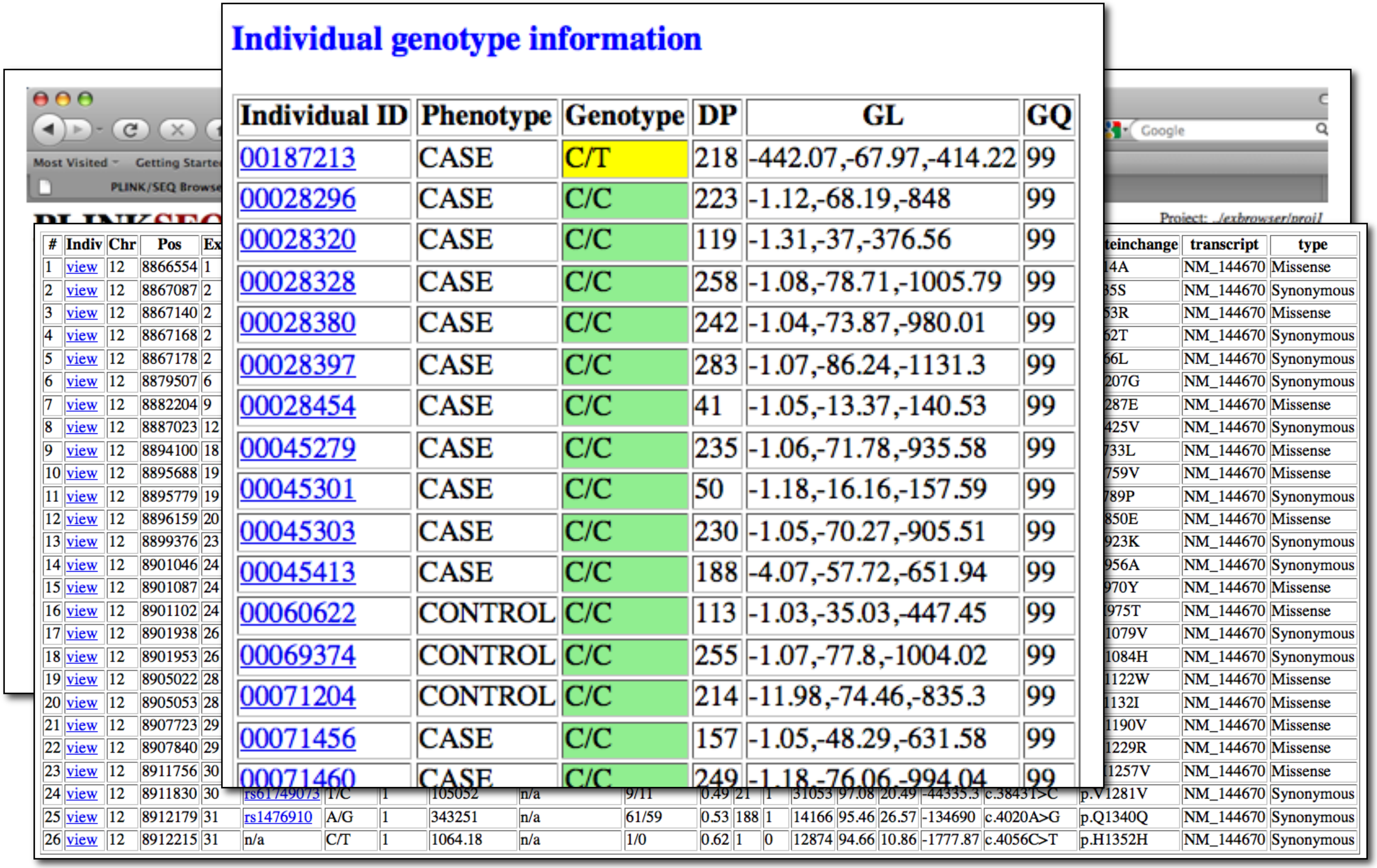
```
A2ML1   NM_144670  chr12:8866515..8918874
```

**NM_144670** location : chr12:8866515..8918874 (view in the UCSC genome browser)

```
   Genomic kb = 52.36; coding kb = 4.365 in 35 exons
```



| # | Indiv | Chr | Pos | Exon | ID | Ref/Alt | FileID | Qual | Info | C/C count | AB | AC | DB | DP | MQ | QD | SB | cDNAchange | proteinchange | transcript | type |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | view | 12 | 8866554 | 1 | n/a | C/G | 1 | 5905.07 | n/a | 1/0 | 0.62 | 1 | 0 | 32933 | 96.94 | 25.34 | -2597.26 | c.40C>G | p.P14A | NM_144670 | Missense |
| 2 | view | 12 | 8867087 | 2 | n/a | C/T | 1 | 1425.47 | n/a | 0/1 | 0.46 | 1 | 0 | 9346 | 96.6 | 17.82 | -1580.13 | c.105C>T | p.S35S | NM_144670 | Synonymous |
| 3 | view | 12 | 8867140 | 2 | n/a | C/G | 1 | 6526.42 | n/a | 0/1 | 0.49 | 1 | 0 | 20719 | 96.58 | 35.47 | -2334.66 | c.158C>G | p.T53R | NM_144670 | Missense |
| 4 | view | 12 | 8867168 | 2 | rs17792974 | C/T | 1 | 33898.5 | n/a | 4/4 | 0.56 | 9 | 1 | 27934 | 96.42 | 20.38 | -12038.9 | c.186C>T | p.T62T | NM_144670 | Synonymous |
| 5 | view | 12 | 8867178 | 2 | n/a | C/T | 1 | 6436.9 | n/a | 1/0 | 0.51 | 1 | 0 | 29448 | 96.4 | 24.76 | -3076.43 | c.196C>T | p.L66L | NM_144670 | Synonymous |
| 6 | view | 12 | 8879507 | 6 | n/a | T/C | 1 | 1423.87 | n/a | 1/0 | 0.55 | 1 | 0 | 17912 | 95.15 | 9.43 | -1854.23 | c.621T>C | p.G207G | NM_144670 | Synonymous |
| 7 | view | 12 | 8882204 | 9 | rs61921916 | C/A | 1 | 2897.63 | GATKStandard | 2/1 | 0.53 | 3 | 1 | 9709 | 96.51 | 12.17 | -1821.37 | c.861C>A | p.D287E | NM_144670 | Missense |
| 8 | view | 12 | 8887023 | 12 | rs7308106 | A/G | 1 | 84587.4 | n/a | 9/11 | 0.47 | 22 | 1 | 24742 | 95.53 | 21.97 | -20925.7 | c.1275A>G | p.V425V | NM_144670 | Synonymous |
| 9 | view | 12 | 8894100 | 18 | n/a | T/C | 1 | 2354.89 | n/a | 1/0 | 0.53 | 1 | 0 | 17138 | 97.06 | 13.77 | -2090.88 | c.2197T>C | p.F733L | NM_144670 | Missense |
| 10 | view | 12 | 8895688 | 19 | n/a | C/T | 1 | 2068.63 | n/a | 0/1 | 0.56 | 1 | 0 | 19295 | 96.26 | 14.67 | -1933.59 | c.2276C>T | p.A759V | NM_144670 | Missense |
| 11 | view | 12 | 8895779 | 19 | rs1860927 | G/A | 1 | 1.43523e+06 | n/a | 66/62 | 0.52 | 223 | 1 | 47681 | 96.39 | 30.94 | -394863 | c.2367G>A | p.P789P | NM_144670 | Synonymous |
| 12 | view | 12 | 8896159 | 20 | rs1860926 | C/A | 1 | 323746 | n/a | 67/64 | n/a | 262 | 1 | 9324 | 95.75 | 34.72 | -156898 | c.2550C>A | p.D850E | NM_144670 | Missense |
| 13 | view | 12 | 8899376 | 23 | n/a | G/A | 1 | 3194.38 | n/a | 1/0 | 0.6 | 1 | 0 | 36642 | 96.6 | 11.88 | -2082.84 | c.2769G>A | p.K923K | NM_144670 | Synonymous |
| 14 | view | 12 | 8901046 | 24 | rs56179521 | C/T | 1 | 27221.8 | n/a | 7/6 | 0.51 | 13 | 1 | 12800 | 96.22 | 19.58 | -9770.18 | c.2868C>T | p.A956A | NM_144670 | Synonymous |
| 15 | view | 12 | 8901087 | 24 | rs1558526 | G/A | 1 | 104332 | n/a | 26/31 | 0.52 | 64 | 1 | 12473 | 96.11 | 19.43 | -38534.2 | c.2909G>A | p.C970Y | NM_144670 | Missense |
| 16 | view | 12 | 8901102 | 24 | n/a | T/C | 1 | 760.63 | n/a | 1/0 | 0.59 | 1 | 0 | 12181 | 95.89 | 8.27 | -1470.92 | c.2924T>C | p.M975T | NM_144670 | Missense |
| 17 | view | 12 | 8901938 | 26 | rs11612600 | G/A | 1 | 145786 | n/a | 34/38 | 0.52 | 88 | 1 | 12646 | 96.82 | 20.79 | -8407.82 | c.3237G>A | p.V1079V | NM_144670 | Synonymous |
| 18 | view | 12 | 8901953 | 26 | rs61745125 | C/T | 1 | 1170.92 | n/a | 1/1 | 0.6 | 2 | 1 | 9655 | 96.39 | 10.01 | -1348.01 | c.3252C>T | p.H1084H | NM_144670 | Synonymous |
| 19 | view | 12 | 8905022 | 28 | rs1860967 | C/T | 1 | 114621 | n/a | 45/40 | 0.5 | 109 | 1 | 7759 | 96.3 | 23.44 | -40304.1 | c.3364C>T | p.R1122W | NM_144670 | Missense |
| 20 | view | 12 | 8905053 | 28 | n/a | C/T | 1 | 1022.56 | n/a | 1/0 | 0.42 | 1 | 0 | 8645 | 96.26 | 16.76 | -1588.78 | c.3395C>T | p.T1132I | NM_144670 | Missense |
| 21 | view | 12 | 8907723 | 29 | rs73040625 | C/T | 1 | 74595.9 | n/a | 10/11 | 0.53 | 21 | 1 | 24119 | 96.74 | 18.68 | -31043.5 | c.3569C>T | p.A1190V | NM_144670 | Missense |
| 22 | view | 12 | 8907840 | 29 | rs10219561 | A/G | 1 | 369823 | n/a | 67/64 | n/a | 262 | 1 | 15447 | 94.97 | 23.94 | -49125 | c.3686A>G | p.H1229R | NM_144670 | Missense |
| 23 | view | 12 | 8911756 | 30 | rs7308811 | A/G | 1 | 859609 | n/a | 66/62 | 0.56 | 220 | 1 | 31435 | 96.71 | 28.21 | -389986 | c.3769A>G | p.M1257V | NM_144670 | Missense |
| 24 | view | 12 | 8911830 | 30 | rs61749073 | T/C | 1 | 105052 | n/a | 9/11 | 0.49 | 21 | 1 | 31053 | 97.08 | 20.49 | -44335.3 | c.3843T>C | p.V1281V | NM_144670 | Synonymous |
| 25 | view | 12 | 8912179 | 31 | rs1476910 | A/G | 1 | 343251 | n/a | 61/59 | 0.53 | 188 | 1 | 14166 | 95.46 | 26.57 | -134690 | c.4020A>G | p.Q1340Q | NM_144670 | Synonymous |
| 26 | view | 12 | 8912215 | 31 | n/a | C/T | 1 | 1064.18 | n/a | 1/0 | 0.62 | 1 | 0 | 12874 | 94.66 | 10.86 | -1777.87 | c.4056C>T | p.H1352H | NM_144670 | Synonymous |



Individual genotype information

| Individual ID | Phenotype | Genotype | DP | GL | GQ |
|---|---|---|---|---|---|
| 00187213 | CASE | C/T | 218 | -442.07,-67.97,-414.22 | 99 |
| 00028296 | CASE | C/C | 223 | -1.12,-68.19,-848 | 99 |
| 00028320 | CASE | C/C | 119 | -1.31,-37,-376.56 | 99 |
| 00028328 | CASE | C/C | 258 | -1.08,-78.71,-1005.79 | 99 |
| 00028380 | CASE | C/C | 242 | -1.04,-73.87,-980.01 | 99 |
| 00028397 | CASE | C/C | 283 | -1.07,-86.24,-1131.3 | 99 |
| 00028454 | CASE | C/C | 41 | -1.05,-13.37,-140.53 | 99 |
| 00045279 | CASE | C/C | 235 | -1.06,-71.78,-935.58 | 99 |
| 00045301 | CASE | C/C | 50 | -1.18,-16.16,-157.59 | 99 |
| 00045303 | CASE | C/C | 230 | -1.05,-70.27,-905.51 | 99 |
| 00045413 | CASE | C/C | 188 | -4.07,-57.72,-651.94 | 99 |
| 00060622 | CONTROL | C/C | 113 | -1.03,-35.03,-447.45 | 99 |
| 00069374 | CONTROL | C/C | 255 | -1.07,-77.8,-1004.02 | 99 |
| 00071204 | CONTROL | C/C | 214 | -11.98,-74.46,-835.3 | 99 |
| 00071456 | CASE | C/C | 157 | -1.05,-48.29,-631.58 | 99 |
| 00071460 | CASE | C/C | 249 | -1.18,-76.06,-994.04 | 99 |

| # | Indiv | Chr | Pos | Ex | | | | | | | | | | | | | | | teinchange | transcript | type |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | view | 12 | 8866554 | 1 | | | | | | | | | | | | | | | 14A | NM_144670 | Missense |
| 2 | view | 12 | 8867087 | 2 | | | | | | | | | | | | | | | 35S | NM_144670 | Synonymous |
| 3 | view | 12 | 8867140 | 2 | | | | | | | | | | | | | | | 53R | NM_144670 | Missense |
| 4 | view | 12 | 8867168 | 2 | | | | | | | | | | | | | | | 62T | NM_144670 | Synonymous |
| 5 | view | 12 | 8867178 | 2 | | | | | | | | | | | | | | | 66L | NM_144670 | Synonymous |
| 6 | view | 12 | 8879507 | 6 | | | | | | | | | | | | | | | 207G | NM_144670 | Synonymous |
| 7 | view | 12 | 8882204 | 9 | | | | | | | | | | | | | | | 287E | NM_144670 | Missense |
| 8 | view | 12 | 8887023 | 12 | | | | | | | | | | | | | | | 425V | NM_144670 | Synonymous |
| 9 | view | 12 | 8894100 | 18 | | | | | | | | | | | | | | | 733L | NM_144670 | Missense |
| 10 | view | 12 | 8895688 | 19 | | | | | | | | | | | | | | | 759V | NM_144670 | Missense |
| 11 | view | 12 | 8895779 | 19 | | | | | | | | | | | | | | | 789P | NM_144670 | Synonymous |
| 12 | view | 12 | 8896159 | 20 | | | | | | | | | | | | | | | 850E | NM_144670 | Missense |
| 13 | view | 12 | 8899376 | 23 | | | | | | | | | | | | | | | 923K | NM_144670 | Synonymous |
| 14 | view | 12 | 8901046 | 24 | | | | | | | | | | | | | | | 956A | NM_144670 | Synonymous |
| 15 | view | 12 | 8901087 | 24 | | | | | | | | | | | | | | | 970Y | NM_144670 | Missense |
| 16 | view | 12 | 8901102 | 24 | | | | | | | | | | | | | | | 975T | NM_144670 | Missense |
| 17 | view | 12 | 8901938 | 26 | | | | | | | | | | | | | | | 1079V | NM_144670 | Synonymous |
| 18 | view | 12 | 8901953 | 26 | | | | | | | | | | | | | | | 1084H | NM_144670 | Synonymous |
| 19 | view | 12 | 8905022 | 28 | | | | | | | | | | | | | | | 1122W | NM_144670 | Missense |
| 20 | view | 12 | 8905053 | 28 | | | | | | | | | | | | | | | 1132I | NM_144670 | Missense |
| 21 | view | 12 | 8907723 | 29 | | | | | | | | | | | | | | | 1190V | NM_144670 | Missense |
| 22 | view | 12 | 8907840 | 29 | | | | | | | | | | | | | | | 1229R | NM_144670 | Missense |
| 23 | view | 12 | 8911756 | 30 | | | | | | | | | | | | | | | 1257V | NM_144670 | Missense |
| 24 | view | 12 | 8911830 | 30 | rs61749073 | T/C | 1 | 105052 | n/a | 9/11 | 0.49 | 21 | 1 | 31053 | 97.08 | 20.49 | -44335.3 | c.3843T>C | p.V1281V | NM_144670 | Synonymous |
| 25 | view | 12 | 8912179 | 31 | rs1476910 | A/G | 1 | 343251 | n/a | 61/59 | 0.53 | 188 | 1 | 14166 | 95.46 | 26.57 | -134690 | c.4020A>G | p.Q1340Q | NM_144670 | Synonymous |
| 26 | view | 12 | 8912215 | 31 | n/a | C/T | 1 | 1064.18 | n/a | 1/0 | 0.62 | 1 | 0 | 12874 | 94.66 | 10.86 | -1777.87 | c.4056C>T | p.H1352H | NM_144670 | Synonymous |

# Internal sharing of summary data (counts) across exome-studies

*Create a summary-level VCF of genotype counts by group, and variant meta-data*

```
./pseq /my/project counts --options vcf --name scz1 > my.vcf
```

*Upload to summary database (ACDB: “a counts database”)*

```
./acdb db1 load  --vcf my.vcf
```

*Can be queried (by qualified investigators) by position, count, disease-specificity, etc*

```
./acdb db1 lookup  --pos chr1:887188  --mask any.filter.ex
```

```
chr1:887188:

        ./.     C/C     C/G     G/G     Project|Sample
        ---     ---     ---     ---     --------------
        .       23/23   1/2     1/0     aut1|1  N=50    PASS    MQ0=0;AB=0.55
        1/0     20/20   1/2     .       aut1|2  N=44    PASS    MQ0=0;AB=0.51
        .       12/35   0/6     .       aut1|3  N=53    PASS    MQ0=0;AB=0.59
        .       19/20   5/5     1/0     aut1|4  N=50    PASS    MQ0=0;AB=0.6
        .       24/21   1/3     0/1     aut1|5  N=50    PASS    MQ0=0;AB=0.58
        .       22/22   2/3     0/1     aut1|6  N=50    PASS    MQ0=0;AB=0.59
        1/0     52/56   12/11   .       scz1|1  N=132   PASS    MQ0=0;AB=0.53
```

# GWAS 2.0

- Directly genotyping >>1 million SNPs in large samples, PLINK becomes unwieldy

- Imputation packages (e.g. BEAGLE) will output VCFs
  - calls, quality scores and posterior genotype probabilities/ dosages

- PLINK/Seq as a platform for GWAS analysis
  - Basic QC, stratification analysis (MDS), linear & logistic regression of direct and imputed genotypes, etc

- The R interface enables easier extension of methods
  - e.g. *multinomial* logistic regression for a cross-disorder GWAS

- Project has been evolving slowly but steadily over the past year

- Available internally on Broad network; public release within one month

- http://atgu.mgh.harvard.edu/plinkseq/

- Developers/collaborators:
  - Brett Thomas, Douglas Ruderfer, Jason Flannick, Jared MacGuire, Menachem Fromer, Manny Rivas, Ron Do, Ben Neale, Mark Daly
  - Adam Kiezun, Alkes Price, Paul de Bakker, LJ Wei, Shamil Sunyaev (methods grant)

- Funding: NHGRI grant R01 HG005827