Y. Ho *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415:180–183, 2002.

T. Ito, T. Chiba, R. Ozawa, *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Nat. Acad. Sci. U.S.A.*, 98:4569–4574, 2001.

Y. Jiang and J. Broach. Tor proteins and protein phosphatase 2A reciprocally regulate Tap42 in controlling cell growth in yeast. *EMBO J.*, 18:2782–2792, 1999.

N. Krogan *et al.* High-definition macromolecular composition of yeast RNA-processing complexes. *Molecular Cell*, 13(2):225–239, 2004.

N. Krogan *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, 440:637–643, 2006.

D. Scholtens and R. Gentleman. Making sense of high-throughput protein-protein interaction data. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 39, 2004.

D. Scholtens, M. Vidal, and R. Gentleman. Local modeling of global interactome networks. *Bioinformatics*, 21:3548–3557, 2005.

P. Uetz, L. Giot, G. Cagney, *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403:623–627, 2000.

S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge University Press, New York, 1999.

*Denise Scholtens*
*Northwestern University Medical School*
*Chicago, IL, USA*
dscholtens@northwestern.edu

# SNP Metadata Access and Use with Bioconductor

*by Vince Carey*

## Introduction

“Single nucleotide polymorphisms (or SNPs) ... are DNA sequence variations that occur when a single nucleotide in genomic sequence is altered”[1]. Conventionally, a given variation must be present in at least one percent of the population in order for the variant to be regarded as a SNP.

There are many uses of data on SNPs in bioinformatics. Two recent contributions which lay out aspects of the concept of “genetical genomics” are Li and Burmeister (2005) and Cheung et al. (2005). In this short contribution I review some functionality provided by Bioconductor for investigating analyses related to the Cheung *et al.* paper.

## The *RSNPper* package

The SNPper[2] web service of the Children’s Hospital (Boston) Informatics Program provides interactive access to a curated database of metadata on SNPs. Details of the system are provided in Riva and Kohane (2005). In addition to the browser-based interface, SNPper has an XML-RPC query resolution system. The *RSNPper* package provides an interface to this XML-RPC-based service. The objective of *RSNPper* is to provide a convenient high-level interface to the SNPper database contents, by providing a small number of high-level query functions with simple calling sequence, and by translating XML responses to convenient R-language objects for further use.

## Getting gene-level information

A `geneInfo` function takes a string argument with a HUGO gene symbol and returns an object of class `SNPperGeneMeta`:

```
> cpm = geneInfo("CPNE1")
> cpm
SNPper Gene metadata:
There are  8 entries.
Basic information:
  GENEID  NAME CHROM STRAND  PRODUCT NSNPS
1  12431 CPNE1 chr20      - copine I   160
  TX.START   TX.END CODSEQ.START CODSEQ.END
1 33677382 33705245     33677577   33684259
  LOCUSLINK   OMIM   UNIGENE SWISSPROT
1      8904 604205 Hs.166887    Q9NTZ6
    MRNAACC   PROTACC REFSEQACC
1 NM_003915 NP_003906      NULL
SNPper info:
     SOURCE           VERSION
[1,] "*RPCSERV-NAME*" "$Revision: 1.38 $"
```

[1] http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml
[2] snpper.chip.org

```
     GENOME DBSNP
[1,] "hg17" "123"
```

The notion of multiple “entries” mentioned in the show result concerns the multiplicity of mRNA and protein accession numbers referenced by annotation of the chosen gene. The allGeneMeta method provides access to such details.

```
> allGeneMeta(cpm)[,15:16]
    MRNAACC   PROTACC
1 NM_003915 NP_003906
2 NM_152925 NP_690902
3 NM_152926 NP_690903
4 NM_152927 NP_690904
5 NM_152928 NP_690905
6 NM_152929 NP_690906
7 NM_152930 NP_690907
8 NM_152931 NP_690908
```

Note that the show result gives a GENEID field, which is an internal SNPper-based index, which must be used for further gene-level queries. A geneLayout function provides information on the extents of the coding region and exons in a gene.

## Getting SNP-level information

The SNPinfo function takes standard dbSNP[3] identifiers (deleting the rs prefix) and returns curated metadata:

```
> mysnp = SNPinfo("rs6060535")
> mysnp
SNPper SNP metadata:
     DBSNPID     CHROMOSOME POSITION
[1,] "rs6060535" "chr20"     "33698936"
     ALLELES VALIDATED
[1,] "C/T"   "Y"
There are details on 4 populations
and 10 connections to gene features
SNPper info:
     SOURCE           VERSION
[1,] "*RPCSERV-NAME*" "$Revision: 1.38 $"
     GENOME DBSNP
[1,] "hg17" "123"
```

Information on populations in which allele frequencies were analyzed is obtained with the popDetails method:

```
> popDetails(mysnp)
             PANEL   SIZE MAJOR.ALLELE
1         Japanese sanger            C
2      Han_Chinese sanger            C
3 Yoruba-30-trios sanger            C
4   CEPH-30-trios sanger            C
  MINOR.ALLELE   majorf    minorf
1            T 0.918605 0.0813954
2            T  0.94186 0.0581395
3            T    0.925     0.075
4            T      0.9       0.1
```

The genes near this SNP are described using the geneDetails method:

```
>  geneDetails(mysnp)[8:9,]
   HUGO LOCUSLINK
8 CPNE1      8904
9 RBM12     10137
                          NAME      MRNA
8                     copine I NM_152931
9 RNA binding motif protein 12 NM_006047
    ROLE RELPOS AMINO AMINOPOS
8   Exon -14677  <NA>     <NA>
9 3' UTR   7722  <NA>     <NA>
```

Broad queries can also be handled by this system. The *itemsInRange* function allows tabulation of SNPs in specific chromosomal regions:

```
> itemsInRange("countsnps", "chr20", "36000000",
    "37000000")
 total exonic nonsyn
  3679    145     48
```

If "genes" is supplied as the first argument, a list of genes and counts of SNPs related to those genes is returned.

The *RSNPper* interface package also includes useSNPper, permitting direct communication with the XML-RPC facility, returning XML to be parsed by the R user.

# Exploring a genome-wide association study

## Data representation

A marked benefit of Bioconductor architecture for analysis of datasets arising in high-throughput biology is the capacity for unifying diverse experimental result structures in S4 objects. For this illustration of inference in genetical genomics, we made an extension of the eSet class in Biobase to house expression and allele counts along with phenotype data. This extension is the racExSet class (rac connoting rare allele count), and an exemplar, chr20GGdem, is supplied with the package:

```
> chr20GGdem
racExSet (SNP rare allele count + expression)
rare allele count assayData:
  Storage mode: environment
  featureNames: rs4814683, ..., rs6062370,
    rs6090120 (117417 total)
  Dimensions:
           racs
Features 117417
Samples      58
```

[3] www.ncbi.nlm.nih.gov/SNP

```
expression assayData
  Storage mode: environment
  featureNames: 1007_s_at, ... (8793 total)
  Dimensions:
         exprs
Features  8793
Samples     58

phenoData
  rowNames: NA06985, ..., NA12892 (58 total)
  varLabels and varMetadata:
    sample: arbitrary numbering
...
```

Information on high-density SNP genotyping (here restricted to SNPs resident on chromosome 20) is accessible with the `snps` method:

```
> snps(chr20GGdem)[1:5,1:5]
          NA06985 NA06993 NA06994
rs4814683       2       0       0
rs6076506       0       0       0
rs6139074       2       0       0
rs1418258       2       0       0
rs7274499       0       0       0
          NA07000 NA07022
rs4814683       2       1
rs6076506       0      NA
rs6139074       2       1
rs1418258       2       1
rs7274499       0      NA
```

Entries count the number of copies of the rare allele in each subject's genotype.

The data noted here were provided by Vivian Cheung and Richard Spielman in conjunction with a summer course at Cold Spring Harbor Lab. This data will be provided in a Bioconductor experimental data package in the near future.

### An association test

Figure 1 illustrates the test for association between a specific SNP (`rs6060535`) and expression measured in a probe set annotated to gene CPNE1. The $p$ value reported by Cheung and Spielman for this test was $8.35 \times 10^{-13}$, in good agreement with the finding noted here. Comprehensive computation of such tests over a chromosome or in a specific region could be conducted with a simple iteration. Some optimizations of note include the elimination of SNPs for which all subjects sampled have identical genotype, and memoization of computations that depend only on the frequency distribution of genotypes, and not on their specific connection to outcomes.

## Conclusions

Management of high-quality metadata on SNPs is a complex task. The XML document for dbSNP's data on chromosome 20 alone decompresses to 3GB. The Informatics Program at Children's Hospital Boston provides an extremely useful resource that can be queried interactively and programatically; *RSNPper* makes use of the Omegahat[4] XML interface of Duncan Temple Lang to simplify use of SNPper by the R community. More work on efficient data representation and algorithm design for genome-wide association studies is underway.

## Bibliography

V. G. Cheung, R. S. Spielman, K. G. Ewens *et al.* Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, 437 (7063):1365–9, 2005.

J. Li and M. Burmeister. Genetical genomics: combining genetics with gene expression analysis. *Human Molecular Genetics*, 14(R2):R163–R169, 2005.

A. Riva and I. Kohane. A SNP-centric database for the investigation of the human genome. *BMC Bioinformatics*, 5(33), 2005.

*Vincent J. Carey*
*Channing Laboratory*
*Brigham and Women's Hospital*
*Harvard Medical School*
*181 Longwood Ave.*
*Boston MA 02115, USA*
`stvjc@channing.harvard.edu`

[4] `www.omegahat.org`

```
Call:
lm(formula = exprs(chr20GGdem)["206918_s_at", ] ~ snps(chr20GGdem)["rs6060535",
    ])

Residuals:
     Min       1Q   Median       3Q      Max
-0.54749 -0.17590  0.02143  0.17102  0.64717

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      7.63381    0.04027  189.57  < 2e-16
snps(chr20GGdem)["rs6060535", ] -0.84324    0.08197  -10.29 1.62e-14

(Intercept)                     ***
snps(chr20GGdem)["rs6060535", ] ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.2782 on 56 degrees of freedom
Multiple R-Squared: 0.654,      Adjusted R-squared: 0.6478
F-statistic: 105.8 on 1 and 56 DF,  p-value: 1.619e-14
```

Figure 1: Call and report on a specific fit.