Chromosomes and SNPs assessment

Stephen Blatti

November 21, 2017

As a result of the human genome project sequenced we have the consensus sequence of all human chromsomes, as well as several other species. We say consensus sequence because every individual has a different sequence. But well over 99% is the same.

Suppose you want to ask a questions such as: how many times does the sequence "ATG" appear on chromosome 11? Or what are the percentage of A,T,C and G on chromosome 7?

We can answer such question using Bioconductor tools. The human genome sequence is provided in the BSgenome.Hsapiens.UCSC.hg19 package. If you have not done so already please donwload and install this package. Note that it encodes 3 billion bases and is therefore a large package (over 800MB) so make time to download it especially if you have a slow internet connection.

library(BiocInstaller) biocLite("BSgenome.Hsapiens.UCSC.hg19")

Then load the package and note that you now have access to sequence information

```
library(BSgenome.Hsapiens.UCSC.hg19)
```

```
## Loading required package: BSgenome
```

```
## Loading required package: BiocGenerics
```

```
## Loading required package: parallel
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, cbind, colMeans,
##     colnames, colSums, do.call, duplicated, eval, evalq, Filter,
##     Find, get, grep, grepl, intersect, is.unsorted, lapply,
##     lengths, Map, mapply, match, mget, order, paste, pmax,
##     pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce,
##     rowMeans, rownames, rowSums, sapply, setdiff, sort, table,
##     tapply, union, unique, unsplit, which, which.max, which.min
```

```
## Loading required package: S4Vectors
```

```
## Loading required package: stats4
```

```
##
## Attaching package: 'S4Vectors'
```

```
## The following object is masked from 'package:base':
##
##     expand.grid
## Loading required package: IRanges
## Loading required package: GenomeInfoDb
## Warning: package 'GenomeInfoDb' was built under R version 3.4.2
## Loading required package: GenomicRanges
## Warning: package 'GenomicRanges' was built under R version 3.4.2
## Loading required package: Biostrings
## Loading required package: XVector
##
## Attaching package: 'Biostrings'
## The following object is masked from 'package:base':
##
##     strsplit
## Loading required package: rtracklayer
## Warning: package 'rtracklayer' was built under R version 3.4.2
```

```
BSgenome.Hsapiens.UCSC.hg19
```

```
## Human genome:
## # organism: Homo sapiens (Human)
## # provider: UCSC
## # provider version: hg19
## # release date: Feb. 2009
## # release name: Genome Reference Consortium GRCh37
## # 93 sequences:
## #   chr1                 chr2                 chr3
## #   chr4                 chr5                 chr6
## #   chr7                 chr8                 chr9
## #   chr10                chr11                chr12
## #   chr13                chr14                chr15
## #   ...                  ...                  ...
## #   chrUn_gl000235       chrUn_gl000236       chrUn_gl000237
## #   chrUn_gl000238       chrUn_gl000239       chrUn_gl000240
## #   chrUn_gl000241       chrUn_gl000242       chrUn_gl000243
## #   chrUn_gl000244       chrUn_gl000245       chrUn_gl000246
## #   chrUn_gl000247       chrUn_gl000248       chrUn_gl000249
## # (use 'seqnames()' to see all the sequence names, use the '$' or '[['
## # operator to access a given sequence)
```

Note this divided into chromosomes and includes several unmapped regions. We will learn to use this type of object.

We can access chromosome 11 like this:

chr11seq <- BSgenome.Hsapiens.UCSC.hg19[[“chr11”]]

Here, for example, is a segment of 25 bases starting at base 1 million

subseq(chr11seq,start=10^6,width=25)

Frequencies of short sequences

Read the help file for the fuction countPattern and tell us which of the following sequences is most common on chromosome 11: "ATG", "TGA", "TAA", and "TAG" Select one:

ATG
TGA
TAA
TAG

How many times does this pattern appear?:

```
?countPattern
```

```
## starting httpd help server ... done
```

```
chr11seq <- BSgenome.Hsapiens.UCSC.hg19[["chr11"]]
seqs <- c("ATG","TGA","TAA","TAG")
cp <- sapply(seqs,function(x){
  countPattern(x,chr11seq)
  })
cp
```

```
##     ATG     TGA     TAA     TAG
## 2389002 2561021 2624324 1689356
```

```
which.max(cp)
```

```
## TAA
##   3
```

Nucleotide frequencies

Now we move to a question about chromosome 7. Read the help page for the function alphabetFrequency and use it to determine what percent of chromosome 7 is T,C,G, and A. Note that we have other letters. For example N, which represents positions that are not called, appears often.

What proportion are Cs (including counts of N in the total)

```
?alphabetFrequency
chr7seq <- BSgenome.Hsapiens.UCSC.hg19[["chr7"]]
alphabetFrequency(chr7seq,as.prob=TRUE)
```

```
##          A          C          G          T          M          R
## 0.28904200 0.19901933 0.19880134 0.28935305 0.00000000 0.00000000
##          W          S          Y          K          V          H
## 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
##          D          B          N          -          +          .
## 0.00000000 0.00000000 0.02378429 0.00000000 0.00000000 0.00000000
```

Locations of SNPs in humans

As explained in the video, many of the locations on the genome that are different across individual are single nucleotide polymorphisms (SNPs). This information is not on the human genome reference sequence. Instead, this information is stored in databases such as dbSNP. Bioconductor also gives you access to these database via the

package SNPlocs.Hsapiens.dbSNP144.GRCh37. Download and install this package. This is also a large package.

```
# if (!("SNPlocs.Hsapiens.dbSNP144.GRCh37" %in% rownames(installed.packages()))) {
#    library(BiocInstaller)
#    biocLite("SNPlocs.Hsapiens.dbSNP144.GRCh37")
#    }
# library(SNPlocs.Hsapiens.dbSNP144.GRCh37)

# To see all the SNPs on, for example, chromosome 17 we can use the following commands

library(SNPlocs.Hsapiens.dbSNP144.GRCh37)
snps144 = SNPlocs.Hsapiens.dbSNP144.GRCh37
s17 = snpsBySeqname(snps144, "17")
head(s17)
```

```
## GPos object with 6 positions and 2 metadata columns:
##       seqnames       pos strand |   RefSNP_id alleles_as_ambig
##          <Rle> <integer>  <Rle> | <character>      <character>
##   [1]       17        52      * | rs556541063                M
##   [2]       17        56      * | rs145615430                Y
##   [3]       17        78      * | rs148170422                S
##   [4]       17        80      * | rs183779916                R
##   [5]       17        92      * | rs562410061                K
##   [6]       17       168      * | rs529798787                R
##   -------
##   seqinfo: 25 sequences (1 circular) from GRCh37.p13 genome
```

The first one listed is rs556541063 which is at location 52.

What is the location on chr17 of SNP rs73971683?

```
s17[which(s17$RefSNP_id == "rs73971683")]
```

```
## GPos object with 1 position and 2 metadata columns:
##       seqnames       pos strand |   RefSNP_id alleles_as_ambig
##          <Rle> <integer>  <Rle> | <character>      <character>
##   [1]       17    135246      * |  rs73971683                R
##   -------
##   seqinfo: 25 sequences (1 circular) from GRCh37.p13 genome
```

GWAS: Linking SNP genotypes to disease risk

Genome-wide association studies (GWAS) are a major tool of genetic epidemiologists. In a case-control design, individuals with a specific disease (cases) are identified and SNP chips or DNA sequencing is used to obtain individuals' genotypes for a large number of SNP. Another group of controls who are disease-free is identified and genotyped. The genotype distributions for all SNP are compared between cases and controls, and those SNP exhibiting association with disease are investigated for potential insight into disruption of gene regulation or gene function. The Bioconductor gwascat package includes information on a catalog of GWAS results assembled at EMBL-EBI (maintenance of the catalog was begun at the US NIH NHGRI and then transferred to the European institutes).

Install the gwascat package and check the version of the GWAS catalog stored in GRCh37 (hg19) coordinates.

```
library(BiocInstaller)
```

```
## Bioconductor version 3.5 (BiocInstaller 1.26.1), ?biocLite for help
```

```
## A newer version of Bioconductor is available for this version of R,
##   ?BiocUpgrade for help
```

```
biocLite("gwascat")
```

```
## BioC_mirror: https://bioconductor.org
```

```
## Using Bioconductor 3.5 (BiocInstaller 1.26.1), R 3.4.1 (2017-06-30).
```

```
## Installing package(s) 'gwascat'
```

```
## package 'gwascat' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Steve\AppData\Local\Temp\RtmpmERNp5\downloaded_packages
```

```
## installation path not writeable, unable to update packages: boot, Matrix,
##   mgcv
```

```
library(gwascat)
```

```
## Loading required package: Homo.sapiens
```

```
## Loading required package: AnnotationDbi
```

```
## Loading required package: Biobase
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
## Loading required package: OrganismDbi
```

```
## Warning: package 'OrganismDbi' was built under R version 3.4.2
```

```
## Loading required package: GenomicFeatures
```

```
## Loading required package: GO.db
```

```
##
```

```
## Loading required package: org.Hs.eg.db
```

```
##
```

```
## Loading required package: TxDb.Hsapiens.UCSC.hg19.knownGene
```

```
## gwascat loaded.  Use data(ebicat38) for hg38 coordinates;
```

```
##  data(ebicat37) for hg19 coordinates.
```

```
data(ebicat37)
ebicat37
```

```
## gwasloc instance with 22688 records and 36 attributes per record.
## Extracted:
## Genome:  GRCh37
## Excerpt:
## GRanges object with 5 ranges and 3 metadata columns:
##       seqnames                ranges strand |
##          <Rle>             <IRanges>  <Rle> |
##   [1]    chr11 [ 41820450,  41820450]      * |
##   [2]    chr15 [ 35060463,  35060463]      * |
```

```
##   [3]     chr8 [ 97512977,  97512977]      * |
##   [4]     chr9 [100983826, 100983826]      * |
##   [5]    chr15 [ 54715642,  54715642]      * |
##                        DISEASE/TRAIT        SNPS   P-VALUE
##                          <character> <character> <numeric>
##   [1] Post-traumatic stress disorder  rs10768747     5e-06
##   [2] Post-traumatic stress disorder  rs12232346     2e-06
##   [3] Post-traumatic stress disorder   rs2437772     6e-06
##   [4] Post-traumatic stress disorder   rs7866350     1e-06
##   [5] Post-traumatic stress disorder  rs73419609     6e-06
##   -------
##   seqinfo: 23 sequences from GRCh37 genome
```

You will see something like

```
ggwasloc instance with 36740 records and 37 attributes per record.
Extracted:  2017-05-20
Genome:  GRCh37
Excerpt:
...
```

The chromosome harboring the largest number of 'verified hits' can be found with

```
sort(table(ebicat37$CHR_ID),decreasing=TRUE)
```

```
##
##    6    1    2   11    3    5   10    4   12    8    7    9   16   15   19
## 2170 1983 1952 1365 1350 1231 1176 1135 1134 1104 1084  959  884  741  710
##   17   14   13   20   18   22   21   23
##  686  587  570  566  502  416  229  154
```

Which chromosome has the most GWAS hits in the catalog? Use an integer 6 has 2170

Counting traits with GWAS hits

You can use the notation mcols(ebicat37)[,"DISEASE/TRAIT"] to get a vector of names of diseases with genetic associations recorded in the gwascat. What is the disease/trait with the most associations?

```
sort(table(mcols(ebicat37)[,"DISEASE/TRAIT"]), decreasing = T)[1:6]
```

```
##
## Obesity-related traits                 Height      IgG glycosylation
##                    957                    822                    699
##        Type 2 diabetes   Rheumatoid arthritis        Crohn's disease
##                    340                    294                    249
```