Course › Section 1: Biology b... › A Review of What ... › Chromosomes and ...

Chromosomes and SNPs assessment

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)
BSgenome.Hsapiens.UCSC.hg19
```

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

2/2 points (graded)
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?:

2624324 ✓

2624324

Submit You have used 2 of 5 attempts

Nucleotide frequencies

1/1 point (graded)

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)

0.19901933 ✓

0.19901933

Submit You have used 2 of 5 attempts

Locations of SNPs in humans

1/1 point (graded)

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)
```

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

What is the location on chr17 of SNP rs73971683?

135246 ✓

135246

Submit You have used 1 of 5 attempts

GWAS: Linking SNP genotypes to disease risk

1/1 point (graded)

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(gwascat)
data(ebicat37)
ebicat37
```

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)
```

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

6 ✓

6

Submit

You have used 1 of 5 attempts

Counting traits with GWAS hits

1/1 point (graded)

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?

Obesity-related traits ✓

Submit You have used 1 of 5 attempts