CSCI 2820 HW2: TagSNP Selection

12 March 2021 Friday

Submission Instructions: Save your write-up to a PDF document titled ***"YourLastName-CS2820-HW2.pdf"*** and e-mail it to your graduate TA at pinar_demetci@brown.edu along with the supplementary material asked for (e.g. code), if any, by the deadline indicated above. Supplementary materials should also be similarly titled (e.g. "YourLastName-CS2820-HW2-Code.py") or included in your write-up.
All deadlines are ***11:59pm Anywhere on Earth***. While this corresponds to 7am the next day in EST, we do not want you to stay up late to turn in assignments. This choice of time is to accommodate anyone currently taking classes outside the U.S.

Notes: You have a ***total of 5 free late days*** you can use however you want on written and programming assignments this semester (except for the final project) without incurring any penalties. After that, you are penalized 10% for each late day.
If there is anything unclear, please ask questions on Piazza and come to our office hours!

1 Readings & Conceptual Questions on tagSNP Selection (25 Points)

1. What are some desiderate for LD measures and do the measures we learned about in class (D, D', r^2) satisfy these? You might find page 13 of class notes on LD as well as pages marked 469-470 of this paper by Sorin helpful. Note that you do not need to read the whole paper).
2. What are tagSNPs? What are some reasons people are interested in identifying tagSNPs? Give a couple examples of when tagSNP selection algorithms might be used.
3. Take a look at the box on page 10 in "Linkage Disequilibrium in Humans: Models and Data" (also linked on the course website "Assignments" page). What is this box describing? (Note: You do not need to read the whole paper)
4. The paper below is written as a reaction to the theory described in the box from the paper above (they refer to it as the "fundamental theorem of HapMap"). The main message of the paper is that the theory summarized in that box makes some assumptions that are not likely to hold true in real-world applications. You do not need to understand the whole paper (please do not spend too much time on the details). Briefly describe a couple of the assumptions they find unrealistic and their criticism about these:
"An utter refutation of the 'Fundamental Theorem of the HapMap' " (also linked on the course website "Assignments" page)

2 TagSNP Selection Algorithms (50 points)

2.1 LD-Select Algorithm

- Describe how LD-Select algorithm works. You may use a pseudocode or just give a itemized list of steps. The original publication is here: Selecting a Maximally Informative Set of Single-Nucleotide Polymorphisms for Association Analyses Using Linkage Disequilibrium (also linked on the course website).

- Implement a greedy procedure for LD-Select. Turn in your code or add it to your write-up.

2.2 Informativeness Algorithm

- Describe how Informativeness works for tagSNP selection. You may use a pseudocode or just give a itemized list of steps. A relevant publication is here (I think the class notes will be helpful here): Optimal Haplotype Block-Free Selection of Tagging SNPs for Genome-Wide Association Studies (also linked on the course website).
- In the class, we have discussed that r^2 does not readily extend to multiple loci. How does informativeness extend to multiple loci?
- Implement the Informativeness Algorithm for SNP selection. Turn in your code or add it to your write-up.

2.3 Data Analysis

Apply both algorithms to the following data. Vary the threshold τ for r^2 (e.g. a few steps between $\tau = 0.2$ to $\tau = 0.8$. Vary the threshold ξ for informativeness, as well. Present the set of SNPs selected by each algorithm.

Sample \ SNP	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
h_1	1	1	1	1	1	1	1	1
h_2	0	1	1	1	0	1	0	1
h_3	0	1	0	1	0	1	0	1
h_4	1	1	0	1	0	1	0	0
h_5	1	1	0	1	1	1	0	1
h_6	0	0	1	1	1	1	0	0
h_7	0	0	0	0	0	0	0	0

2.4 The relationship between the algorithms and their parameters

1. Compare the tagSNPs and the number of tagSNPs picked by the two algorithms. Create plot of the number of common SNPs (intersect the sets of SNPs selected by each algorithm) when varying thresholds of r^2 and ξ. Describe any similarities or differences you observe.
2. Recall that we showed in class that the Set Cover Problem is equivalent to the Dominating Set Problem. This means that you can translate one problem into the other and solve it instead. One open question is how do these solutions relate since the two tagSNP selection algorithms have different threshold parameters (r^2 and ξ). In this question, we want you to look into the relationship between the two parameters by doing the following:
   (a) You have run the Informativeness algorithm with varying values for parameter ξ to generate different sets of tagging SNPs. At what levels of r^2 would the LD-Select algorithm select these tagSNPs and return a similar solution set?
   (b) Similarly, you have run the LD-Select algorithm with varying values for the threshold parameter r^2. At what levels of ξ would the Informativeness algorithm select these tagSNPs and return a similar solution set?

3 Working with Real Genetic Data: VCF Files (25 Points)

This course has mostly been a conceptual course. However, we would like you to gain some experience working with real genetic data, especially in case your final project requires working with such data. So, this exercise is here to introduce you to the most commonly used file format for storing genetic variants.

Many human genetic datasets are not publicly available due to privacy concerns. It is not too difficult to figure out what a person looks like, what their ethnicity is, what sort of diseases they might be at risk for etc from large enough genetic data. However, **1000 Genomes Project** makes their data publicly available. It was launched in 2008 and has become a major resource for the scientific community. At the time, their goal was to establish by far the most detailed catalogue of human genetic variation. Today, we have much larger genetic datasets, such as the UK Biobank, which we cannot access publicly.

If you need to access to human genetic data for your final projects or any other future projects, you can visit the most up-to-date data files from 1000 Genome Project here: `http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`. For this assignment, we will work with an older and much smaller pilot file.

3.1 Downloading the data

Run the following command in your terminal to get the genotype data:

```
wget ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/release/2010_07/trio/snps/experimental_validation/trio.2009_12.batch1_2_validation.genotypes.vcf.gz
```

Then get the genotype indexing file ("TABIX" format, which is a tab-delimited format index file that stores information about which variants are at what indices in the VCF file and their genomic location), run:

```
wget ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/release/2010_07/trio/snps/experimental_validation/trio.2009_12.batch1_2_validation.genotypes.vcf.gz.tbi
```

3.2 Understanding VCF and bash scripts

The most commonly used file format for storing genetic variant data is called **"variant calling format (VCF)"**. The VCF file format is documented here: http://samtools.github.io/hts-specs/VCFv4.1.pdf. Take a look at the documentation, especially "1.3 Header line syntax" section.
Use a series of piped shell (bash) commands to count the number of samples present in the sample. Bash scripts are commonly used by bioinformaticians to work with VCF files. Note that you do not need to decompress the VCF.GZ file for this task. These commands might be helpful: `gzcat`, `awk`, `sed`, `grep`, `wc`. You won't need to use them all. There are multiple different ways to write this bash command/script.

How many samples are there? Please also add the command you used.

3.3 Turning VCF into TSV

There are plenty of downstream analysis tools developed to be used with VCF files, but many (especially more recent computational biology tools) expect the genotype data to be in a matrix format. Here we will convert the data stored in VCF file into something easier to work with/understand, using `BCFtools`.
Download `BCFtools` following this page: https://samtools.github.io/bcftools/howtos/install.html. If you run into an error regarding the "libgsl" package, you might need to install that before running "make" in your BCFtools folder.

Once you have BCFtools installed, run the following command on your terminal to output a TSV file, which stores the genotype data from VCF file in a table format:

```
./path_ToBCFtools/bcftools query --print-header -f '%CHROM\t%POS\t%ID\t%TYPE\t%REF\t%ALT[\t%GT]\n'
/path_to_VCF_GZ_FILE/trio.2009_12.batch1_2_validation.genotypes.vcf.gz -o trioTable.tsv
```

Describe the structure of the resulting table saved in `trioTable.tsv`. What are "0/1" etc? What other information is saved in this table? What sort of variants are there in this sample? Do you see any multi-allelic SNPs or structural variants?
Submit the TSV file, as well.

Note: There are a number of other tools that might come in handy if you need to work with VCF files in the future, such as:

- GATK (e.g. for variants of certain types with the `SelectVariant` command)
- VCFtools, which is highly similar to `BCFtools` and has a number of commands for modifying VCF files (e.g. performing LD pruning, pruning based on MAF –minor allele frequency–, converting VCF to other file formats etc).
- PLINK, which is commonly used in GWAS related pre-processing pipelines.

It is good to be aware of the existence of these tools.

3.4 Obtaining a genotype matrix

Turn the data in the TSV file into a numerical genotype matrix (containing 0s, 1s, and 2s), where the rows correspond to samples (individuals) and the columns correspond to genomic locations. You may use a package that facilitates working with table-formatted data, such as `pandas` in `python` or `dplyr` in `R` Submit the matrix.

4 Final Projects (0 Points)

There is nothing to be turned in for this part but we would like you to start thinking about a final project. You may choose to work on your own or with a partner depending on the complexity of your proposed project. For ideas, Sorin directs you to the following page from the course website: `http://cs.brown.edu/courses/csci2820/projects.html`. We will also release an additional list. In a couple weeks or so, we will ask you to write up a project proposal/summary.

Bonus Questions

We want to make sure people don't feel pressed on time and get discouraged from attempting the bonus questions. As a result, we extend the deadline for these questions (noted below).
Note that the second bonus question is significantly more challenging than the first one and has a much later deadline (beginning of the "reading period").

Bonus 1 (10 points – Deadline: March 18)

Mathematically derive the relationship depicted in box 10 of the paper linked in Question 1.2.

Possibly useful hint: Make a table of parameters and what they correspond to. Then, similarly to the HW1 bonus question, think about how you'd create a contingency table for what's described in the box. The rest is substitutions and simplifications.

Bonus 2 (25 points – Deadline: April 12)

Take a look at the paper: Conservative Extensions of Linkage Disequilibrium Measures from Pairwise to Multi-loci and Algorithms for Optimal Tagging SNP Selection, which extends "Informativeness" to "Directed Informativeness" (Sorin's research paper). On pages marked 473-474, the authors derive a relationship between r^2 and Directed Informativeness, as well as χ^2 and Directed Informativeness.
Attempt to derive a similar relationship as above (the relationship presented in box 10 of Pritchard - Przeworski paper for r^2) for Directed Informativeness.