COMP90016 – Assignment 2

Haonan Li

April 30, 2018

1 Introduction

This assignment consists of four tasks.

The first task is adding a quality filter on a SNV caller. The original caller only report all heterozygotes where two bases present at the same position at a frequency between $20 - 80\%$ each, advanced caller will take account of the phred base quality (≥ 20).

The second task is to build a VCF file from a bam file and its reference. Report any position where at least one base different from the reference presents at a frequency of 20% or higher together with its genotype.

The third task is phasing. We first identify all heterozygotes from the output of task 2. Then fetch all reads contain both positions. After some necessary filter. we finally establish haplotypes if there is a consensus of at least 90% to its phasing.

In the last task. We predict three person's eye color according to (Sturm et al. [2008]) based on the detected genotype of two particular variants.

2 Results & Discussion

2.1 Task 1

Table 1 shows the result comparison of two SNV caller with pysam version 0.14.1 (on PC). From which we find that the quantity of matched variants are not directly related with the Phred base quality. For sample 1, adding the base quality filter result in the reduction of matched pairs; for sample 2, no changes, for sample 3, the results increase. Because the quantity of reported positions is directly related to bases frequency on the particular position. We analyze the frequency function 1. From this function we can find adding a base quality filter will lead to the reduce of particular base's appearances together with all bases appearances. So the frequency's trend is uncertain.

	Basic		Advanced	
	Quantity	Quality	Quantity	Quality
Sample1	114	35.6	130	36.1
Sample2	18	35.6	18	36.5
Sample3	48	30.0	32	32.6

Table 1: Results of two SNV callers with pysam version 0.14.1. (Where "Quality" is the average quality score of all bases, "Basic" and "Advanced" represent the original SNV caller and caller with quality filter separately.)

$$frequency = \frac{particular\ base\ appearances}{all\ bases\ appearances} \tag{1}$$

While Table 2 shows the results on server (digitalis.eng.unimelb.edu.au) with pysam version 0.10.0. Which shows that quantity of advanced caller reduced for all samples, the reason might be that the particular base's appearances reduce more quick than all bases'. Two tables gives the different results. Actually, we think Table 1 is more convincing, the reason already mentioned before. The algorithms for alignment and other operations maybe outdated for the pysam in older version.

	Basic		Advanced	
	Quantity	Quality	Quantity	Quality
Sample1	138	30.2	114	35.3
Sample2	52	21.9	8	34.4
Sample3	218	21.7	36	30.7

Table 2: Results of two SNV callers with pysam version 0.10.0. (Where "Quality" is the average quality score of all bases, "Basic" and "Advanced" represent the original SNV caller and caller with quality filter separately.)

As for quality, the average base quality improved for all samples on both machines.

2.2 Task 2

In this task, we only report the positions in the reference which covered by the corresponding bam file for each sample. For the item AF (Allele Frequency), we calculate it by the definition (the allele frequency of the non-reference allele) in the instructions.

For genotype. We employ two judgement methods. The first is predict a position is homozygous if one base presents at a frequency of 80% or higher. The second method is predict a position is heterozygous where two bases present at a frequency larger than 20% each. Both methods seem reasonable and they compute the same output for out task.

However, what if 'A' presents 70% on some position and the other three 'T', 'G', 'C' presents 10% separately. The first method predicts a heterozygote but the second method predicts homozygote. In my opinion, I think the second method is more reasonable and useful in practice because we must make sure

there are at least two bases who have a relatively high present, then we can ascertain a heterozygote.

2.3 Task 3

For this task. We first get all closest position pairs which both of them were identified heterozygous in task 2. For each position pair, we phase it follow the following steps:

1. We use pysam `fetch` method to get all reads contain any of the two positions. Because some positions are so far that they can not contained by one single read, but a pair of read. We use two fetch method to fetch reads contain these two positions and get intersection of their read name.

2. Parse cigar, we delete the corresponding bases for `cigar insertion` and soft-clip, and padding underscore for `cigar deletion`. We do the same process for quality scores, which is useful later.

3. Delete reads whose queried position's base quality score is less than 20. (This is a opitional filter.)

4. Filter the reads that bases on the corresponding positions is not the two variant heterozygous bases. More specifically, if one position is identified heterozygous of two bases 'A' and 'T', but the base in this position in our processing read is 'C', we ignore this read directly.

5. Calculate consensus and get the final haplotypes or rejected haplotypes. We build haplotypes of more than 2 variants by extending the phasing results. For rejected haplotypes, we only report two close variants based results as the number of combinations explode to 2^n for n varients.

Table 3 shows the phasing results of three samples on different version's pysam. Here, we only discuss results on pysam 0.14. The detected and rejected haplotypes for three samples are 8-51, 1-3, 3-3 separately. This information is very useful in further analysis. The importance of phase information was demonstrate in (Tewhey et al. [2011]).

In my opinion. phasing result can help understanding allele-specific expression. Because different gene might result in different expression. If we can find some expression that related to the specific haplotypes. It will very helpful for human genetic engineering and genetic disease detection and treatment. It is also useful in determine long distance gene relevance as consensus phasing is transitive.

Need to mentioned, we find cigars except indel and soft-clip does exist. We seek out them on sample 2 with pysam version 0.10.0 (on server).

	Pysam 0.14		Pysam 0.10	
	#Detected	#Rejected	#Detected	#Rejected
Sample1	8	51	5	48
Sample2	1	3	1	0
Sample3	3	3	3	3

Table 3: Results of phasing.

2.4 Task 4

Table 4 shows the genotype for rs1129038 and rs12913832 (their respective 1-based genomic coordinates are chr15:28356859 and chr15:28365618). The reference of these two positions are 'G' and 'T' separately (reverse strand). We predict most likely eye color for each of the sample based on (Sturm et al. [2008]) and calculate the prediction confidence use the proportion of people with the particular genotypes and have specific eye color .

For the first sample. the genotypes of two position is 'A/G' and 'C/T'. We predict the eye color of the person is brown. According the row 2,3,4 of Table 3 in (Sturm et al. [2008]), $confidence = \frac{47+81+8}{1+10+0+47+81+8} = 92.52\%$

For the second sample, two positions are not in the output VCF file. So we inferred they are homozygous and the genotype are 'G/G' and 'T/T. Predicted eye color is brown. According the row 5,6,10,11 of Table 3 in (Sturm et al. [2008]) , $confidence = \frac{18+14+1+3}{0+0+0+0+18+14+1+3} = 100\%$

For the last sample, two genotypes are 'A/A' and 'C/C' because the AF of them are both 1. The predicted eye color is blue. According the row 1,7 of Table 3 in (Sturm et al. [2008]), $confidence = \frac{169+1}{1+1+169+1} = 98.84\%$

	rs1129038	AF	rs12913832	AF	Pred-Eye color	Confidence
Sample1	A/G	0.46	C/T	0.56	Brown	92.52
Sample2*	G/G	–	T/T	–	Brown	98.84
Sample3	A/A	1	C/C	1	Blue	100

Table 4: Two specific variants detection with eye color prediction. ('*' means the result is inferred rather than directly printed.)

Actually, there are several factors may have influence on the confidence.

First, according to the paper, we know that eye color is not determined by two variants. Other two variants also have influence on it. For example, consider the case: rs1129038 is 'A/A' and rs12913832 is 'C/C', we fetch row 1 and 7 in the Table. But the number of samples of two rows are variant largely. Row 1 have 170 samples with 169 blue eyes, which means we can almost confirm that the combination of these 4 particular genotypes lead to blue eyes, however, row 7 contains only two samples with one blue eyes, it seems that 50% of this combination of genotypes have blue eyes. It seems just these condition can not lead to a correct confidence. Because small sample is not representative and we finally assume that people with this kind of haplotype combination is rare. So the calculation of confidence does not analyze the proportion of different haplotype combinations.

Besides, The investigate and research was based on a small population just as mentioned in paper. The representativeness of the result is not absolute and the paper also mentioned that strong association between eye color and rs1667394 in HERC2 was reported in Icelandic population but the author excluded this. It will also influence the confidence of eye color prediction. But we do not know the exact representativeness of this paper. So we assume it is 100%, but if we know this coefficient, the result will different.

References

Sturm, A Richard, Duffy, L David, Zhao, Zhen Zhen, P. N Fabio, Stark, and S Mitchell. A single snp in an evolutionary conserved region within intron 86 of the herc2 gene determines human blue-brown eye color. *American Journal of Human Genetics*, 82(2):424–431, 2008.

Ryan Tewhey, Vikas Bansal, Ali Torkamani, Eric Topol, and Nicholas Schork. The importance of phase information for human genomics. 12:215–23, 03 2011.