Back to Week 3 | Lessons | Prev | Next

Review Your Peers: Application Challenge 3

Review by August 9, 11:59 PM PDT

Reviews 3 left to complete

WGS vs WES

by Manuel Breuning

August 5, 2017

like Flag this submission

PROMPT

Analyzing the Output of WGS and WES

Visit “Projects” and click on the “WGS vs. WES” results folder. Then click on “Isaac Whole Genome Sequencing”, which will take you to the WGS results page. Make sure that “*NA12878-TruSeqPCRFree-WGS*” is selected in the column on the left side of the page, and answer the following questions.

Question 1: How many WGS reads were in the dataset?

1,049,766,460 reads

Help Center

RUBRIC

1,049,766,460

○ 0 pts
The learner did not provide the correct value.

○ 1 pt
The learner provided the correct value.

PROMPT

Question 2: What was the overall coverage of the reads?

39.03 fold coverage

RUBRIC

39.03X

○ 0 pts
The learner did not provide the correct value.

○ 1 pt
The learner provided the correct value.

PROMPT

Question 3: How many total SNVs were identified?

3,553,674 SNVs

RUBRIC

3,553,674

○ 0 pts
The learner did not provide the correct value.

○ 1 pt
The learner provided the correct value.

Help Center

PROMPT

Question 4: How many SNVs were identified in exons?

49,222 SNVs in exons

RUBRIC

49,222

○ 0 pts
The learner did not provide the correct value.

○ 1 pt
The learner provided the correct value.

PROMPT

Question 5: How many total indels (insertions and deletions combined) were identified?

610,236 indels

RUBRIC

610,236

○ 0 pts
The learner did not provide the correct value.

○ 1 pt
The learner provided the correct value.

Help Center

PROMPT

Question 6: How many total indels (insertions and deletions combined) were identified in exons?

6,299 indels in exons

RUBRIC

6,299

○ 0 pts
The learner did not provide the correct value.

○ 1 pt
The learner provided the correct value.

PROMPT

We then will examine the results of the WES analysis. To do so, visit the “WGS vs. WES Results” project again in the top left of the page, and click on “Isaac Enrichment”.

Question 7: How many total WES reads were in the dataset?

65,262,238

RUBRIC

65,262,238

○ 0 pts
The learner did not provide the correct value.

○ 1 pt
The learner provided the correct value.

Help Center

PROMPT

Question 8: How many base pairs were in the reference?

45,326,818

RUBRIC

45,326,818

○ 0 pts
The learner did not provide the correct value.

○ 1 pt
The learner provided the correct value.

PROMPT

Question 9: How many total SNVs were identified?

31,497

RUBRIC

31,497

○ 0 pts
The learner did not provide the correct value.

○ 1 pt
The learner provided the correct value.

PROMPT

Question 10: How many SNVs were identified in exons?

21,972

RUBRIC

21,972

○ 0 pts
The learner did not provide the correct value.

○ 1 pt
The learner provided the correct value.

Help Center

PROMPT

Question 11: How many total indels were identified?

2,918

RUBRIC

2,918

○ 0 pts
The learner did not provide the correct value.

○ 1 pt
The learner provided the correct value.

PROMPT

Question 12: How many indels were identified in exons?

942

RUBRIC

942

○ 0 pts
The learner did not provide the correct value.

○ 1 pt
The learner provided the correct value.

Help Center

PROMPT

Question 13: What was the mean coverage identified (scroll down to “Coverage Summary”)?

102.3 fold coverage

RUBRIC

102.3 X

○ 0 pts
The learner did not provide the correct value.

○ 1 pt
The learner provided the correct value.

PROMPT

Finally, we will analyze the number of SNVs present in genome vs. exome sequencing dataset.

Question 14: Compute the ratio

Total WES SNVs / Total WGS SNVs

and compare this ratio to the ratio of the length of the exome (approximately 37 million bp) to the length of the genome (approximately 3 billion bp). Which ratio is larger?

total WES SNVs / total WGS SNVs = 0.00886
length of exome / length of genome = 0.0123
The second ratio is larger.

RUBRIC

31497 / 3553674 = 0.00886

This value is slightly smaller than 0.01233, the ratio between the length of the exome and the length of the genome.

Help Center

○ 0 pts
The learner did not provide the correct answers.

○ 1 pt
The learner provided one correct answer.

○ 2 pts
The learner provided correct answers for both questions.

PROMPT

Question 15: Provide a potential biological explanation for the answer to the previous question.

The result of the previous question shows that there are fewer SNVs in the exome than one would expect given its length. A possible explanation is that SNVs in the exome are more likely to have a negative impact on the organism (for example by reducing the functionality of the encoded proteins) and therefore have a smaller survival rate in the population than SNVs in the rest of the genome.

RUBRIC

Mutations in exons are less likely to remain fixed in a population, since they may affect transcripts and could therefore be detrimental to the survival and thus reproducibility of the organism (so the changes are not passed on). Thus, we would expect that the exome would have fewer SNVs per nucleotide, implying that

Total WES SNVs / Total WGS SNVs < |exome| / |genome|

which is what we observed.

○ 0 pts
The learner did not provide a reasonable explanation along the lines above.

○ 2 pts
The learner provided other reasonable explanation.

○ 3 pts
The learner provided a reasonable explanation along the lines above.

Help Center

PROMPT

Comparing the SNV sets revealed by WGS and WES

Open the “Projects” tab of BaseSpace, click on the “WGS vs. WES Results” project, and click on the folder whose name begins with “Variant Calling Assessment Tool”. Under “Analysis Reports” (in the left-hand panel), be sure that the “VCAT-Report” tab is selected.

We would like to determine how different the sets of SNVs are that were revealed by genome and exome sequencing. (To do so, consult the “VCF File Pairwise Intersect Stats” tab.)

Question 16: How many SNVs were identified in the target exonic regions in the genome and exome sequencing data? (Hint: This information can be found under “VCF File Stats,” in the “Total SNV Count” column).

genome seq data: 32,439 SNVs
exome seq data: 30,862 SNVs

RUBRIC

32,439 SNVs were found in the target regions using genome sequencing;

30,862 were found using exome sequencing.

○ 0 pts
The learner did not provide the correct answers.

○ 1 pt
The learner provided one correct answer.

○ 2 pts
The learner provided correct answers for both questions.

PROMPT

Question 17: How many indels were identified in the target exonic regions in the genome and exome sequencing data?

(Hint: This information can be found under “VCF File Stats,” in the “Total Indel Count” column).

genome seq data: 2,374 indels
exome seq data: 2,780 indels

Help Center

RUBRIC

2,374 indels were found in the target regions using genome sequencing;

2,780 were found using exome sequencing.

○ 0 pts
The learner did not provide the correct answers.

○ 1 pt
The learner provided one correct answer.

○ 2 pts
The learner provided correct answers for both questions.

PROMPT

Question 18: Why do you think that in the targeted region, the number of SNVs and indels in the genome sequencing data differs from the number of SNVs and indels in the exome sequencing data?

This might be because of read errors in both sets of data, which leads to errors when calling SNVs and indels.

RUBRIC

Since WES reads have higher coverage, some variants that pass the quality score threshold in WES do not pass it in WGS. On the other hand, due to the limitations of WES, such as differences in coverage across the exome, some exonic mutations captured using genome sequencing can be missed using exome sequencing.

○ 0 pts
The learner did not provide a reasonable explanation along the lines above.

○ 3 pts
The learner provided a reasonable explanation along the lines above.

Help Center

PROMPT

Comparing the accuracy of Isaac calls against Platinum Genomes

Variant calling can generate erroneous SNVs and indels and thus should be evaluated using various statistics. In order to determine the quality of the VCF files generated by Isaac app, we will use VCAT to compare our VCF files to the Platinum Genomes v8.0 dataset, a “gold standard” of genome information, including known SNVs. For both SNVs and indels, VCAT estimates two statistics for the VCF file:

- **precision:** the percentage of identified SNVs/indels that are true SNVs/indels.
- **recall**: the percentage of true SNVs/indels that are identified.

For more information about precision and recall, see FAQ: What are Precision and Recall?

Question 19: What are the SNV precision and recall as well as the indel precision and recall for the genome and exome sequencing datasets when measured against “Platinum Genomes v8.0”?

SNVs:
genome seq data: precision 99.87%, recall 93.30%
exome seq data: precision 98.92%, recall 87.43%

indels:
genome seq data: precision 96.79%, recall 86.50%
exome seq data: precision 77.91%, recall 79.28%

RUBRIC

WGS – SNV precision: 99.87%; SNV recall: 93.30%

WES – SNV precision: 98.92%; SNV recall: 87.43%

WGS – Indel precision: 96.79%; indel recall: 86.50%

WES – Indel precision: 77.91%; indel recall: 79.28%

Grading scheme: 4 points total; 1 point for each row

Help Center

○ 0 pts
See grading scheme above

○ 1 pt
See grading scheme above

○ 2 pts
See grading scheme above

○ 3 pts
See grading scheme above

○ 4 pts
See grading scheme above

PROMPT

Question 20: Which of the two datasets (WGS or WES) has higher precision and recall with respect to Platinum Genomes? Provide one explanation as to why this might be the case.

The WGS data has higher precision and recall than the WES data.
A possible explanation is that the target exonic region used for the VCAT analysis is not identical to the target that was used in the preparation of the WES data.

RUBRIC

WGS has higher precision and recall with respect to Platinum Genomes. As mentioned above, one reason why WGS may outperform WES is because of greater uniformity of coverage.

○ 0 pts
The learner did not provide a reasonable explanation along the lines above.

○ 2 pts
The learner provided a reasonable explanation along the lines above.

PROMPT

Comparing genome and exome sequencing

Question 21: Based on the results of your analyses, when would you prefer genome sequencing over exome sequencing and vice-versa?

Genome sequencing was better in identifying SNVs and indels (higher precision and higher recall) than exome sequencing. Therefore genome sequencing is preferable if we need precise results and don't necessarily know the target region in advance. Exome sequencing is preferable (because of reduced costs) if the precise target

Help Center

region is known in advance, if results are required fast, and for applications where saving money is more important than results for large parts of the genome (e.g. Genealogy testing).

RUBRIC

Genome sequencing would be preferred if we are interested in mutations occurring in non-exonic regions of the genome. It also offers more uniform coverage, and in this example was better at identifying SNVs and indels.

Exome sequencing would be preferred if we want to only look at exonic regions. For a cheaper cost, we can get higher depth/coverage than we could with genome sequencing because we are significantly limiting the amount of DNA that is being sequenced. This is useful for determining an individual's genotype for known exonic disease SNVs/indels.

○ 0 pts
The learner did not provide the correct explanation along the lines above.

○ 1 pt
The learner provided one correct explanation along the lines above.

○ 2 pts
The learner provided correct explanation for both questions along the lines above.

PROMPT

Using genome sequencing to look for CF mutations in NA12878

Next, we will see if the WGS dataset indicates any SNVs in NA128768 that are implicated in CF. Open the "Projects" tab of BaseSpace, click on the "WGS vs. WES Results" project that you joined when you ran the apps, and click on the "**WES Annotation**" folder. Next to "Save as... Result file:", click on "Annotation Result" to download a .zip file containing the complete annotation results. Expand the file, which contains a .xls file that can be opened in Microsoft Excel or OpenOffice. Each row corresponds to a single specific SNV.

Question 22: How many SNVs in this spreadsheet are related to CF? (Hint: Search for the term "cystic fibrosis" using CTRL+F/command-F). Which rows do they correspond to in the spreadsheet?

Help Center

There are 4 SNVs related to CF in the spreadsheet, in rows 1245,1246, 1247, and 15330.

RUBRIC

There are four SNVs in the WES annotation related to CF, corresponding to rows 1245, 1246, 1247, and 15330.

○ 0 pts
The learner provided more than two wrong values

○ 1 pt
The learner provided 1 or 2 wrong values

○ 2 pts
The learner provided correct answer.

PROMPT

Question 23: For each SNV related to CF, list its chromosome, position, whether it passed the quality threshold, whether it is exonic or intronic, whether it is benign or disease-causing (or unknown), and whether it is a nonsense (i.e. stop gained), missense, or silent mutation?

SNV C ->T on Chrom 1, Pos 161476204, passed quality threshold, exonic, unknown (or does the benign/disease-causing classification not exist for nonsense mutations?), nonsense mutation
SNV A -> G on Chrom 1, Pos 161476205, passed quality threshold, exonic, benign, missense mutation
SNV A -> G on Chrom 1, Pos 161479745, passed quality threshold, exonic, benign, missense mutation
SNV G -> A on Chrom 19, Pos 41858921, passed quality threshold, exonic, unknown (according to PolyPhen) and tolerated (according to SIFT), missense mutation

Help Center

RUBRIC

Chromosome 1 Position 161476204 – PASS – Exonic - Unknown – Nonsense