Liga Brasileira de Bioinformática

Desafio Mendelics

Dia 2 - Resultados

Ana Elisa Ribeiro Orsi

Outubro/2021

No presente relatório, serão discutidos os resultados observados no segundo dia do desafio.

1. Para a obtenção do arquivo VCF no primeiro dia do desafio, foi realizada uma etapa de filtragem: apenas as variantes apresentando qualidade (QUAL) maior que 20 e profundidade (DP) maior que 10 foram selecionadas. Também foram filtradas SNPs a menos de 3 bases de indels (-g3). Assim, apenas 2254 das 4691 variantes identificadas a princípio foram incluídas no VCF final entregue (figuras 1 e 2). Não foi observada uma correlação clara entre as variáveis QUAL e DP (figura 3).

Figura 1: Distribuição da métrica QUAL antes (esquerda) e depois (direita) da filtragem realizada no primeiro dia.

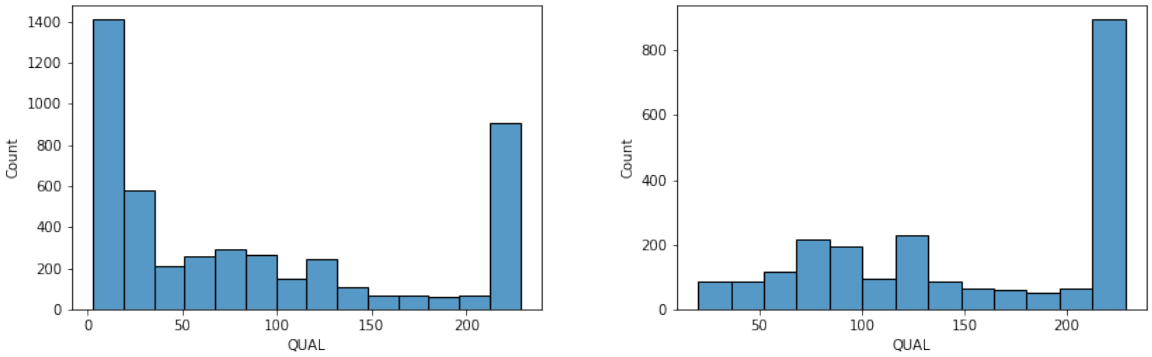
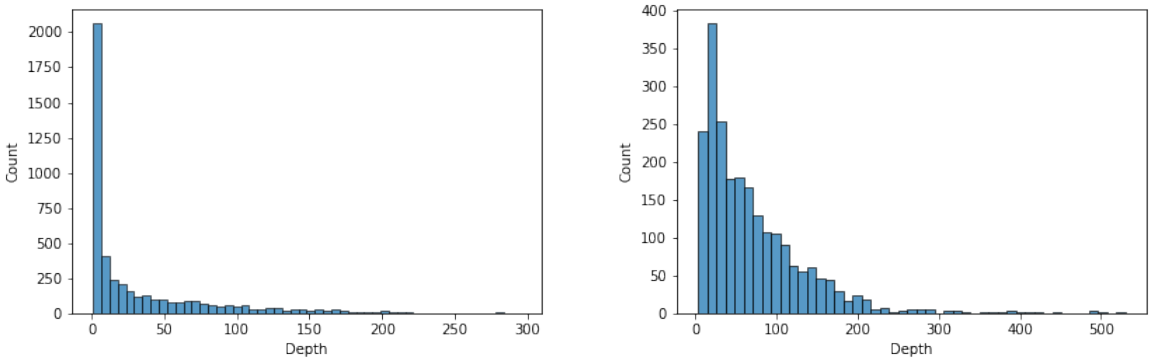


Figura 2: Distribuição da métrica DP antes (esquerda) e depois (direita) da filtragem realizada no primeiro dia.



No segundo dia, as variantes foram filtradas novamente, de forma a selecionar apenas aquelas localizadas nas regiões compreendidas pelo painel. Os histogramas correspondentes às variáveis QUAL e DP do novo arquivo foram representados na figura 4

Uma vez que a coluna QUAL apresentou picos muito altos após o valor 215, essa variável foi utilizada para uma nova filtragem do VCF. De acordo com o manual do formato VCF (https://samtools.github.io/hts-specs/VCFv4.3.pdf), a métrica QUAL pode ser definida como

Figura 3: Scatterplots das variáveis QUAL e DP antes (esquerda) e depois (direita) da filtragem realizada no primeiro dia.

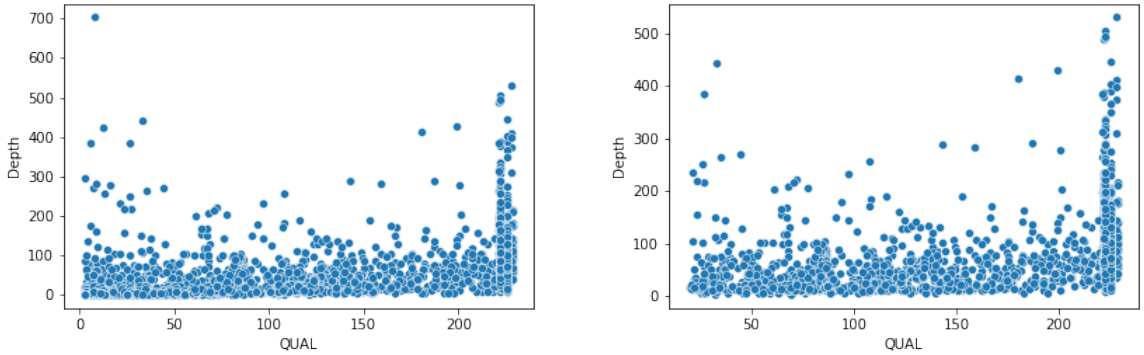
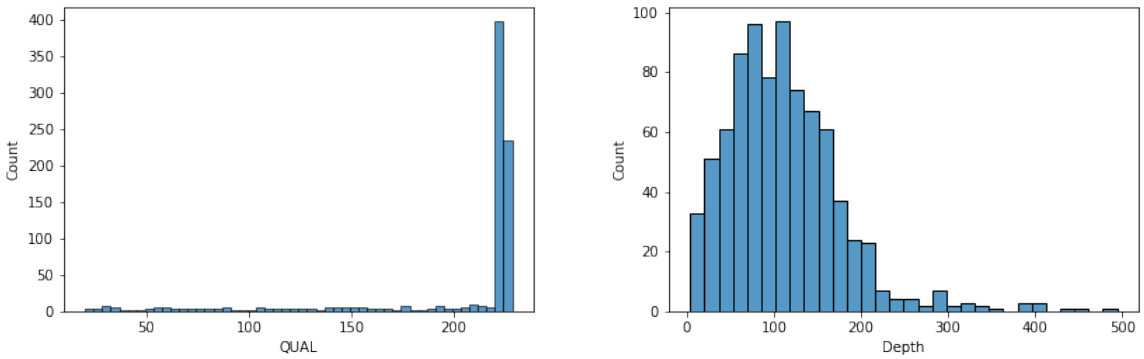


Figura 4: Distribuição das variáveis QUAL (esquerda) e DP (direita) após seleção das regiões de interesse no segundo dia.



$$-10 * log_{10} \text{ prob(chamada em ALT estar errada)}$$

Isso significa que um valor de QUAL=20 representa uma probabilidade de 1% de erro na chamada. Assim, o valor selecionado (QUAL>215) é extremamente conservador. Entretanto, uma vez que escolha foi baseada na distribuição observada, optou-se por manter essa filtragem.

2. A análise das regiões de baixa cobertura foi realizada através da aplicação bedtools. Os resultados foram visualizados utilizando os pacotes Pandas e Seaborn em Python.

O primeiro passo foi observar as métricas gerais de cobertura das regiões de interesse. A figura 5 apresenta um screenshot do output do método describe() da biblioteca Pandas aplicado ao dataframe analisado. Através da coluna Coverage, é possível observar que a média de reads correspondentes às regiões analisadas foi aproximadamente 334. Pelo menos uma região não foi coberta por nenhum read. Apesar disso, os quartis da coluna Coverage Percentage indicam que, de maneira geral, os reads sequenciados representaram bem as regiões de interesse.

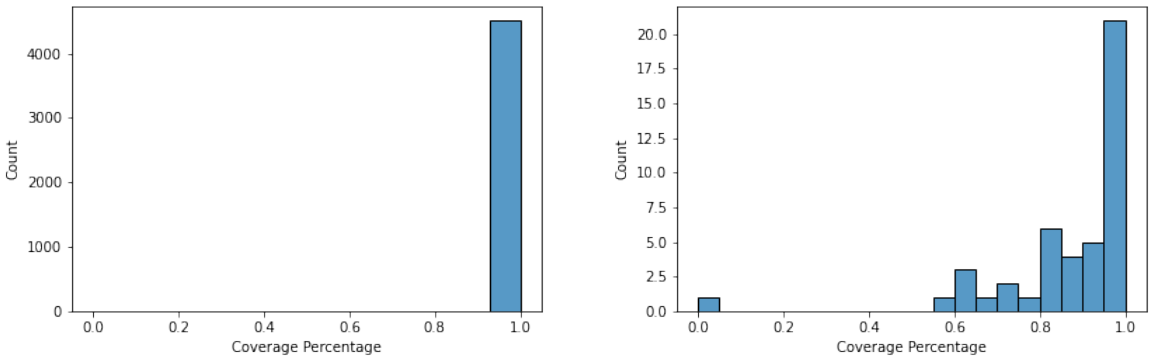
Em seguida, foram plotados histogramas das colunas Coverage e Coverage Percentage. Cor-

Figura 5: Medidas descritivas dos dataframe referentes à cobertura do alinhamento. Coverage: número de reads com algum overlap na região, Covered Length: número de pares de bases da região com cobertura >= 1, Full Length: tamanho da região, Coverage Percentage: Covered Length / Full Length

	Coverage	Covered Length	Full Length	Coverage Percentage
count	4538.000000	4538.000000	4538.000000	4538.000000
mean	333.808506	203.684442	204.142794	0.998652
std	515.903394	292.479129	294.105149	0.022070
min	0.000000	0.000000	120.000000	0.000000
25%	178.000000	120.000000	120.000000	1.000000
50%	252.000000	128.000000	128.000000	1.000000
75%	351.000000	188.000000	188.750000	1.000000
max	14905.000000	6680.000000	6762.000000	1.000000

roborando com os resultados obtidos pelo método describe(), é possível observar que a grande maioria das regiões foi totalmente representada. Apenas 44 das 4538 regiões não apresentaram cobertura completa. Dessas, somente uma não foi coberta por nenhuma read, a região RANBP1_1 (figura 6).

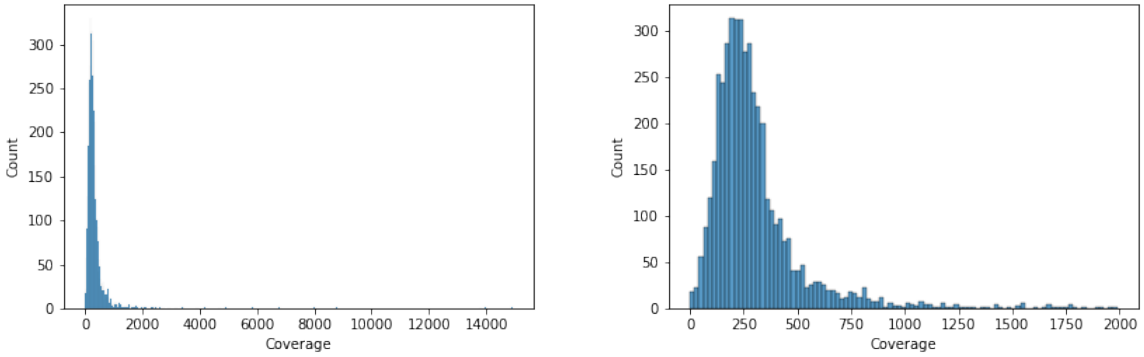
Figura 6: Distribuição da porcentagem de cobertura de todas as regiões analisadas (esquerda) e das regiões com cobertura parcial (direita).



No caso da variável Coverage, representando o número de reads alinhados a cada região, a distribuição se centrou em 250 reads por região, com uma cauda longa à direita (figura 7).

Com base nos resultados observados, decidiu-se excluir as regiões com menos de 95% de cobertura. O arquivo Dia_2.bed, contendo as regiões excluídas, encontra-se na pasta ”Outputs Esperados”.

Figura 7: Distribuição do número de reads alinhados por região analisada (esquerda), com destaque para regiões com Coverage<2000 (direita).



3. As informações relativas ao alinhamento foram obtidas através do comando *samtools view*. A seguir, os principais resultados observados:

- Total de reads: 1731879
- Total de reads alinhados à referência: 1731190
- Pares mapeados corretamente: 1728281
- Reads com mapQ = 0: 84167
- Reads com mapQ <= 20: 92011
- Alinhamentos secundários: 155

O resultados mostram que mais de 99,9% dos reads foram mapeados na referência e mais de 99,7% foram pareados corretamente. Os alinhamentos com qualidade 0 e menor que 20 corresponderam, respectivamente, a cerca de 4,9% e 5,3% do total. Por sua vez, os reads apresentando alinhamentos múltiplos totalizam menos de 0,1% do total. Essas métricas sugerem que o alinhamento realizado foi satisfatório.

Analisando as regiões esperadas de alinhamento, observou se que um total de 975598 reads (56,3%) foram mapeados a regiões não esperadas de acordo com o painel realizado, muitas vezes chegando a cobrir todas as bases desses ”gaps”.

Entretanto, na análise acima, qualquer read que cobrisse parcialmente regiões fora dos trechos delimitados pelo arquivo BED fornecido foi considerado. Dessa forma, é necessário corrigir esse valor, excluindo reads que também foram mapeados em regiões desejadas (1514823). Assim, obtemos $1731190 - 1514823 = 216367$, que corresponde a 12,5% do total. É possível perceber que uma grande quantidades de reads ultrapassou as regiões delimitadas pelo arquivo BED, mas não deixou de apresentar overlap com as regiões de interesse. Apesar da porcentagem

obtida acima ainda ser relativamente alta, o valor não deve comprometer a qualidade final da chamada de variantes após a filtragem realizada.