Universidade Federal do Rio Grande do Sul

Centro de Biotecnologia

Programa de Pós-Graduação em Biologia Celular e Molecular

- Introdução à metagenomica: curso prático -

Bryan Tavares

**RELATÓRIO – QUALIDADE DAS SEQUÊNCIAS BRUTAS**

**Pergunta #1**

**O que é o comando *ls*?**

O comando *ls* exibe o conteúdo do diretório e informações sobre os arquivos.

**Qual o propósito da bandeira *-l*?**

A bandeira *-l* no comando *ls* serve para exibir o conteúdo do diretório e as informações sobre os arquivos em formato da lista.

**Por que utilizamos ../?**

O sinal *../* é usado para acessar o diretório parental do diretório atual.

**Sem utilizar ../, de que outra maneira você poderia listar o conteúdo dessa pasta?**

**Pergunta #2**

**O que significa ../seq\_brutas/\*.fastq? O que o asterisco (\*) está fazendo ali?**

O comando acessa o diretório parental e recruta todos arquivos com a extensão *.fastq* da pasta *seq\_brutas*. O asterisco completa o nome de arquivos com quaisquer conjuntos de caracteres.

**Qual o propósito da bandeira -o e o argumento fastqc?**

A bandeira *-o* define o nome do arquivo resultante (*output*).

**E -t 4? Qual a relação do valor 4 aqui com o comando srun que rodamos anteriormente?**

A bandeira *-t 4* no software fastqc indica o número de análises que serão realizadas referente ao número total de amostras, as quais serão computadas paralelamente ( *srun -* *Run parallel jobs*) através da bandeira com o número de CPUs (*--cpus-per-task 4*).

**Pergunta #3**

**Que tipos de problemas as sequências brutas apresentam? Quais as possíveis causas?**

As sequências brutas tem problemas em relação a quantidade de sequências duplicadas nos *reads* das amostrar (variando de 1,3 a 8,2%), além de os nucleotídeos finais das *reads* apresentarem menor qualidade de sequenciamento (valor de *phred* diminui em relação a média das amostras). A provável causa da menor qualidade das sequências brutas é a presença dos adaptadores nas extremidades das *reads*.

**RELATÓRIO – LIMPEZA DAS SEQUÊNCIAS**

**Pergunta #4**

**Olhando o manual do Cutadapt online ou rodando cutadapt --help, responda:**

- **O que significam as bandeiras -o, -p, -a, -A, -q, -m e -j?**

A bandeira *-o* significa: *output file*

A bandeira *-p* significa: *paired-output file*

A bandeira *-a* significa: *adapter*

A bandeira *-A* significa: *adapter options*

A bandeira *-q* significa: *quality-cutoff 5’ and 3’*

A bandeira *-m* significa: *minimum-length of reads*

A bandeira *-j* significa: *cores of CPUs*

*-o FILE, --output FILE*

*Write trimmed reads to FILE. FASTQ or FASTA format is chosen depending on input.*

*Summary report is sent to standard output. Use '{name}' for demultiplexing (see*

*docs). Default: write to standard output*

*-p FILE, --paired-output FILE*

*Write R2 to FILE.*

*-a ADAPTER, --adapter ADAPTER*

*Sequence of an adapter ligated to the 3' end (paired data: of the first read).*

*The adapter and subsequent bases are trimmed. If a '$' character is appended*

*('anchoring'), the adapter is only found if it is a suffix of the read.*

*-A ADAPTER 3' adapter to be removed from R2*

*-q [5'CUTOFF,]3'CUTOFF, --quality-cutoff [5'CUTOFF,] 3'CUTOFF*

*Trim low-quality bases from 5' and/or 3' ends of each read before adapter*

*removal. Applied to both reads if data is paired. If one value is given, only*

*the 3' end is trimmed. If two comma-separated cutoffs are given, the 5' end is*

*trimmed with the first cutoff, the 3' end with the second.*

*-m LEN[:LEN2], --minimum-length LEN[:LEN2]*

*Discard reads shorter than LEN. Default: 0*

*-j CORES, --cores CORES*

*Number of CPU cores to use. Use 0 to auto-detect. Default: 1*

**- Por que os argumentos para -o e -p terminam em .fastq.gz?**

Os arquivos dos argumentos *-o* e *-p* terminam com o formato das sequências no formato compactado, além disso, sendo ambos de mesma estrutura é possível verificar os resultados do Cutadapt de forma pareada, verificando os itens removidos de cada sequência.

**- Qual o propósito do redirecionamento (> seq\_limpas/${sample}.log.txt)?**

**- Além disso, qual o propósito da barra invertida (\) no comando acima?**

A barra invertida indica a existência de caractere especial na posição seguinte na linha de comando.

**Pergunta #5**

**Você conseguiria desenvolver um jeito de rodar Cutadapt para as 4 amostras utilizando um só for loop?**

Sim, seria possível fazer um *for loop* para rodar

**Ou seja, um for loop onde os argumentos passados para -a e -A dependem de qual é a amostra que está sendo analisada em cada iteração?**

**Talvez algo envolvendo declarações condicionais if/else seria uma possibilidade?**

**Pergunta #6**

**Para cada uma das amostras, responda:**

**- Quantos pares de sequências brutas tínhamos originalmente?**

ERR1713356\_1: 1x10⁷

ERR1713356\_2: 1x10⁷

ERR2683233\_1: 1x10⁷

ERR2683233\_2: 1x10⁷

ERR4998600\_1: 1x10⁷

ERR4998600\_2: 1x10⁷

ERR4998601\_1: 1x10⁷

ERR4998601\_2: 1x10⁷

**- Quantas sequências continham adaptadores?**

ERR1713356\_1: 4x10⁶

ERR1713356\_2: 4x10⁶

ERR2683233\_1: 1x10⁶

ERR2683233\_2: 1x10⁶

ERR4998600\_1: 0

ERR4998600\_2: 0

ERR4998601\_1: 0

ERR4998601\_2: 0

**- Quantos pares de sequências foram removidas porque eram muito curtas?**

ERR1713356\_1: 0

ERR1713356\_2: 0

ERR2683233\_1: 0

ERR2683233\_2: 0

ERR4998600\_1: 0

ERR4998600\_2: 0

ERR4998601\_1: 0

ERR4998601\_2: 0

**- Quantos nucleotídeos foram removidos por terem baixa qualidade?**

**- Qual a porcentagem de nucleotídeos que passou o controle de qualidade?**

**Pergunta #7**

**Quais problemas foram resolvidos?**

Os problemas resolvidos foram principalmente relacionadas a regiões duplicadas nas extremidades das *reads*, as quais estavam relacionadas com o sequenciamento de nucleotídeos dos adaptadores, os quais não são informativos para as análises.

**E quais não foram?**

Os problemas não resolvidos foram relacionados ao conteúdo GC de todas as amostras.

**Apareceram novos problemas que não existiam nas sequências limpas?**

Nas sequências limpas apareceram problemas relacionados ao menor tamanho das *reads*, visto que as sequências de adaptadores foram removidas.

**Podemos continuar nossas análises ou devemos repetir o processo de limpeza? De repente utilizando parâmetros diferentes ou até mesmo um outro programa?**

As análises podem continuar, embora limpezas adicionais possam ser realizadas a fim de melhorar a limpeza das amostras através do aumento dos valores de *cutoff* dos diversos parâmetros. Parâmetros que poderiam ser modificados são aqueles relacionados ao conteúdo GC e a sequências duplicadas nas amostras e de qualidade dos nucleotídeos através do PhredScore.

**RELATÓRIO – MONTAGEM DE GENOMAS**

**Pergunta #8**

**Você acha que se mudássemos os valores de --k-min e --k-max poderíamos obter montagens melhores? Quais seriam valores melhores para esses parâmetros?**

Sim, a alteração dos valores dos *kmers* altera consideravelmente as montagens dos *contigs*. Os fatores que influenciam os valores de *kmers* ótimos são o tamanho das sequências, volume de sequenciamento e a complexidade das sequências Os melhores valores para as amostras dos estudos seria --k-min 95 e --k-max 120.

**Existem outros programas para montagem de genomas/metagenomas? Como saber qual é o "melhor"?**

Sim, há diversos programas que realizam a montagem de genomas, como por exemplo SPAdes, Ray Meta e Meta Velvet. A própria determinação de ‘montagem ouro’ não é bem estabelecida para um padrão comparativo entre os programas disponíveis. O melhor programa depende da necessidade das análises e do trabalho computacional que se tem disponibilidade, logo, o melhor programa é o que apresenta melhor custo benefício no quesito de sensibilidade e especificidade e poder computacional disponível.

**Eu acredito em um mundo onde não precisaremos montar metagenomas para obtermos genomas completos a partir de metagenomas. O que precisa para que isso vire realidade?**

A necessidade da montagem de genomas será superada com novas tecnologias de sequenciamento de fragmentos de longo comprimento, mas principalmente com sequenciadores capazes de sequenciarem moléculas completas de DNA sem a necessidade de clivagem.

**RELATÓRIO – ANÁLISE DAS MONTAGENS DE GENOMA**

**Pergunta #9**

**Quantos contigs foram produzidos em cada montagem?**

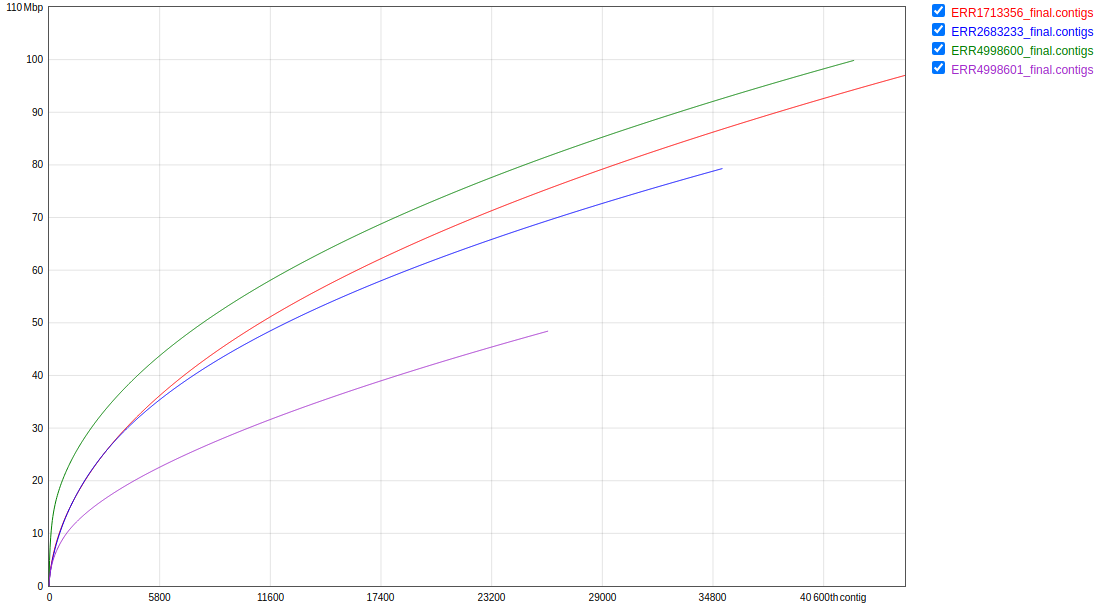
ERR1713356: 44883

ERR2683233: 35287

ERR4998600: 42185

ERR4998601: 26152

**Temos bastante contigs longos, ou maioritariamente contigs curtos?**

 A maioria dos *contigs* são curtos, próximos do tamanho de 1000 e 2000 pares de base.

**Qual o tamanho total das montagens?**

ERR1713356: 97005171

ERR2683233: 79300180

ERR4998600: 99850528

ERR4998601: 48420952

**Quão representativo do todo são as montagens; em outras palavras, qual a porcentagem das nossas comunidades que foram efetivamente montadas? Os valores que obtivemos são "bons" ou "ruins"? Se você sequenciar o mesmo número de sequências de uma amostra da microbiota intestinal humana e uma amostra de solo da Amazônia, em qual você espera uma maior taxa de mapeamento?**

A porcentagem de montagem para cada um das amostras foi de 38.16% para ERR1713356, de 31.33% para ERR2683233, de 26.62% para ERR4998600 e de 9.97% para ERR4998601. Os valores encontrados variaram de valores bons a ruins, sendo que as amostras do tratamento de esgoto tiveram melhor cobertura em relação as amostras da tundra. Isto pode ser relacionado a menor complexidade das amostras com o predomínio de um menor *set* de diversidade de genomas. Da mesma forma, exemplificando comparativamente, a maior taxa de mapeamento é da amostra da microbiota intestinal humana em relação a amostra do solo da Amazônia.

**Qual a melhor montagem dentre as quatro que fizemos e por que? Na verdade, será que faz algum sentido essa comparação?**

A melhor montagem é ERR1713356 com 38.16% da cobertura de *reads* do sequenciamento, ou seja, um melhor aproveitamento dos dados disponibilizados pelo sequenciamento. Porém a comparação realizada deve ser feita de modo crítico, visto que nem sempre uma maior cobertura indica qualidade da amostragem.

**RELATÓRIO – OBTENÇÃO DOS PERFIS TAXONÔMICOS E FUNCIONAIS**

**Pergunta #10**

**Qual a função do argumento -s 100 no comando acima e qual é a sua importância?**

O argumento *-s* serve para emparelhamento das amostras com o mesmo valor de aleatoriedade no programa para a seleção do grupo amostral. Ou seja, todas as amostras serão subamostradas com o mesmo valor *-s 100*.

**Pergunta #11**

**O que faz o subprograma blastx que estamos rodando dentro do DIAMOND?**

O subprograma blastx faz a tradução da sequência nucleotídica em sequência de aminoácidos para a pesquisa em bancos de dados para alinhamento com sequências proteicas.

**Qual a vantagem de utilizar o programa DIAMOND ao invés do bom e velho BLAST em análises de metagenomas?**

A vantagem de utilizar o programa Diamond é a maior velocidade de análise das sequências, embora para tanto não seja tão sensível e especificidade do programa Blast, o qual tem um maior custo computacional.

**Pergunta #12**

**Qual a função do programa daa-meganizer?**

O programa daa-meganizer adapta as análises provenientes do programa Diamond para que seja possível sua abertura no programa Megan.

**Pergunta #13:**

**Existem diferenças entre as amostras?**

Sim, há diferença entre as amostras.

**As amostras que vem de um mesmo ambiente são mais similares entre si?**

Sim, as amostras que vem de um mesmo ambiente compartilham maior similaridade tanto no âmbito taxonômico, quanto no âmbito funcional.

**Quais são os filos mais abundantes em cada grupo de amostra? E em relação ao metabolismo?**

**RELATÓRIO – OBTENÇÃO DE MAGS USANDO ANVI’O**

**Pergunta #14:**

**Por que calculamos frequências de *kmers* e para que são utilizadas?**

As frequências de *kmers* são calculadas nos *contigs* para uso no programa Prodigal para a identificação de genes nos *contigs*, a fim dos genes serem usados para as análises posteriores. Os gene identificados serão utilizados para o alinhamento entre as sequências das amostras e também a identificação de genes já caracterizados e depositados em bancos de dados.

**Pergunta #15:**

**Como a presença de genes de cópia única (single-copy genes) pode nos informar a respeito da completude e contaminação de um genoma/MAG?**

Os genes de cópia única são essenciais para a determinação de completude e contaminação na montagem dos genomas, visto que se em quantidade maior do que 1 indica a presença da mistura de genomas, enquanto a ausência indica que a fração do gene de cópia única não está presente no genoma montado. De forma que quantos mais genes de cópia única forem utilizados como marcadores para a montagem de genomas, melhor pode ser os valores indicativos da qualidade da montagem

**Pergunta #16:**

**Por que estamos fazendo isso e por que não mapear as sequências filtradas de cada amostra somente à montagem originária da mesma amostra? Em outras palavras, que outro aspecto importante sobre os contigs isso nos revela?**

A análise conjunta dos *contigs* entre as amostras acrescenta informações úteis para a montagem, sendo principalmente favorável para a análise do nível de representatividade dos contigs nas amostras a fim de usar o índice de cobertura diferencial entre os *contigs* para encontrar *contigs* relacionados ao mesmo genoma.

**Pergunta #17:**

**Quantos MAGs (i.e. > 50% completos e < 10% redundantes) foram obtidos no total?**

**Você espera que os MAGs sejam completamente diferentes entre si ou não?**

**E quanto aos MAGs obtidos pelos seus colegas nesse curso, eles serão diferentes aos seus?**