UNIVERSIDADE FEDERAL DA BAHIA
Instituto Multidisciplinar em Saúde
Campus Anísio Teixeira

Montagem de Genomas

Bioinformática

Prof. Dr. Leandro Martins de Freitas

IMS/CAT-UFBA - Rua Rio de Contas, 58 – Quadra 17 – Lote 58 – Bairro Candeias
Vitória da Conquista – BA - CEP 45.029-094/ Fone: (77) 3429 2709. E-mail: leandromartins@ufba.br

Introdução

As novas tecnologias de sequenciamento conseguem produzir uma quantidade de dados muito grande com custos baixos. A velocidade e quantidade de informação gerada por essas novas tecnologias de sequenciamento estão revolucionando a investigação biológica e permitindo o acesso a genomas de diferentes espécies e sequenciamentos de diferentes linhagens. O NGS permite o resequenciamento de genomas inteiros aumentando confiança dos dados. Os genomas de organismos modelo *Drosophila melanogasters* e *Caenorhabditis elegans,* e os genomas de cânceres humanos já estão sendo produzidos usando NGS.

O arquivo usado para montagem de um genoma ou região de um genoma contém muitas sequências chamadas *reads*. Os *reads* são o resultado da reação de sequenciamento e posteriormente leitura das bases que estão na sequência. Os *reads* gerados por esses sequenciadores de nova geração (*Next generation DNA sequencing - NGS*) são fragmentos curtos (*short read sequence - SRS*) comparados com os fragmentos produzido pela tecnologia *Sanger*. O tamanho dos fragmentos produzidos pelos NGS representa um desafio para a bioinformática na montagem de genomas. Os SRS apresentam problemas na distinção entre regiões repetitivas, formando fragmentos genômicos. O método aplicado para análises de SRS devem ser robustos para lidar com uma grande quantidade de sequencias. O sequenciamento e montagem de genomas sem comparação com genomas previamente sequenciados para auxiliar a montagem é chamado de genoma *de novo* (figura 1).

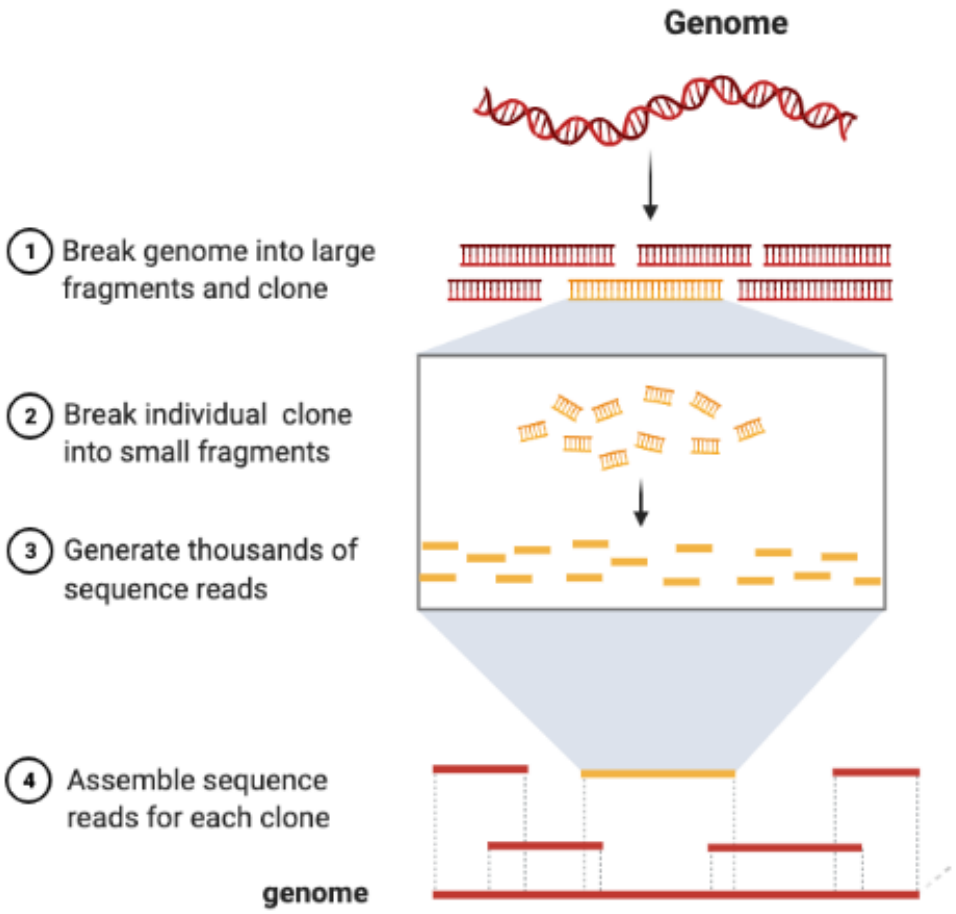

Figura 1. Representação das etapas de sequenciamento e montagem de um genoma *de novo*.

Cobertura de sequenciamento é uma média de quantas vezes cada base foi sequenciada (figura 2).

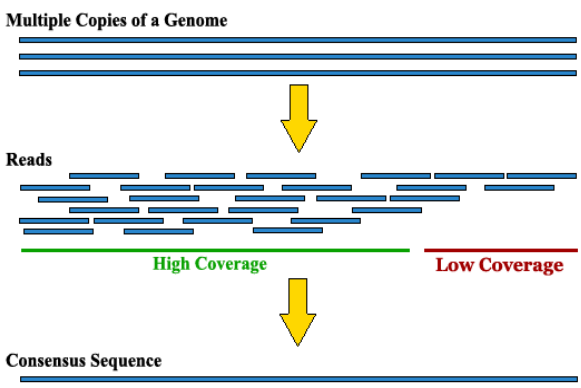

Figura 2. Montagem de uma região de um genoma. Algumas regiões foram sequenciadas várias vezes (alta cobertura, região em verde), outras regiões foram sequenciadas poucas vezes (baixa cobertura, região em vermelho).

Analisando a qualidade do sequenciamento

Devemos observar a qualidade do sequenciamento para evitar erros de montagem e erros de alinhamento. Aumentando a acurácia do genoma e dos SNPs encontradas. Podemos usar o software FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc). O FASTCQ avalia a frequência do tamanho dos fragmentos, informação útil para sequenciadores que apresentam aproximadamente o mesmo tamanho para os fragmentos gerados. O programa também avalia a qualidade dos fragmentos baseados no valore de qualidade PHRED (figura 3).

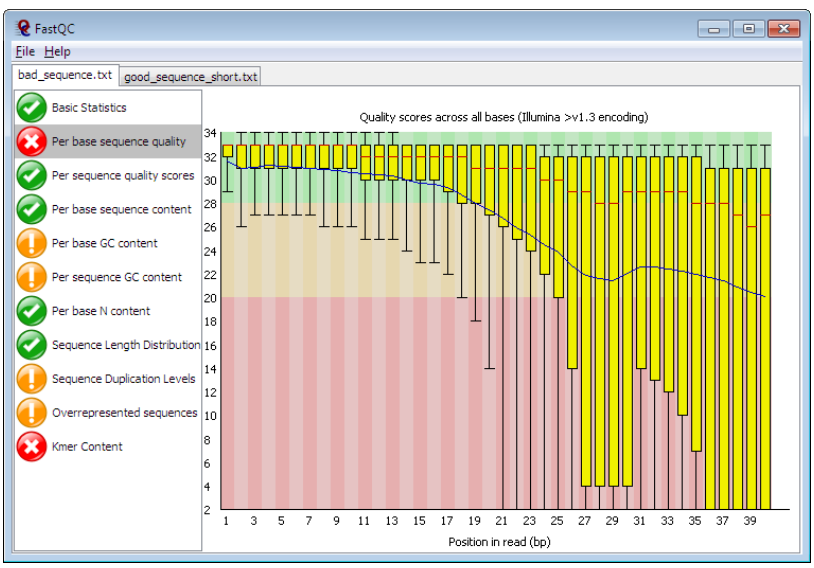

Figura 3. Relatório de qualidade de sequências gerado pelo FASTQC.

Ressequenciamento

O re-sesequeciamento de genomas já finalizados era usado para análises de genes específicos ou regiões de interesse, aumentando a confiança dos resultados e permitindo a identificação de SNPs no genoma de outros indivíduos. O NGS permite agora o re-sequenciamento de genomas inteiros, devido a produção de grande quantidade de dados. A aplicação do re-sequenciamento genômico depende do SRS serem longos o suficiente para aplicação do mapeamento no genoma referência. O mapeamento durante o sequenciamento deve ser capaz de lidar com polimorfismos e erros durante o sequenciamento.

Montagem referência.

A montagem de genomas baseados em referência utiliza de um genoma já montado como base para construção do novo genoma. Deve se utilizar o genoma de um organismo relacionado filogeneticamente para usar como referência (figura 4). A montagem referência pode ser realizada com o programa Bowtie2.

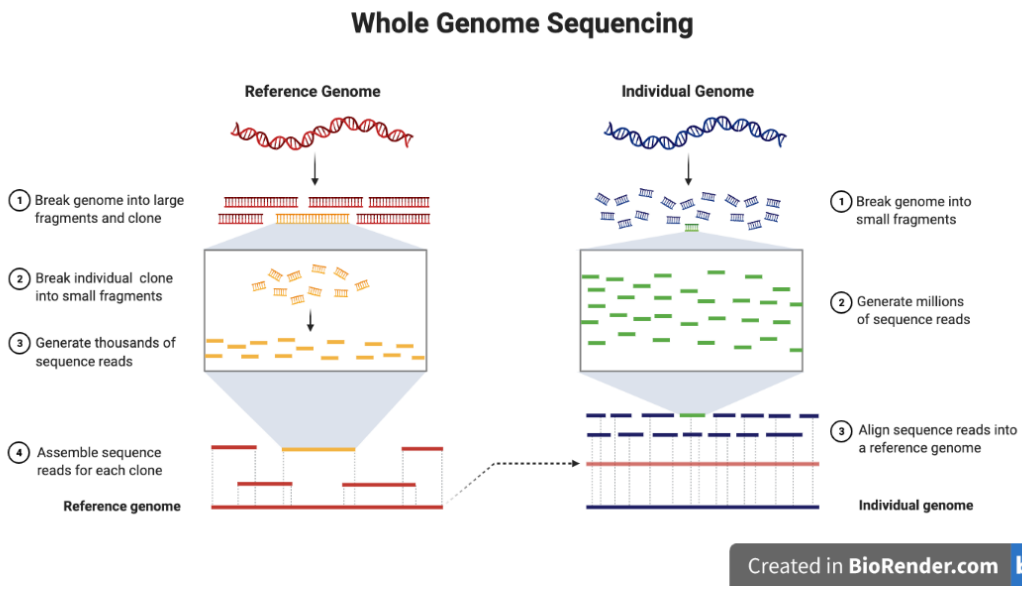

Figura 4. Representação das etapas de sequenciamento e montagem de um genoma baseado em uma referência.

Objetivos

1. Fazer análise de qualidade dos *reads* e excluir os *reads* de baixa qualidade
2. Montar genoma de procarioto usando técnicas de bioinformática e avaliar o tamanho dos *contigs*

Métodos

1. A qualidade das bases será verificada usando o programa FASTQC e Trimmomatic dentro da plataforma KBsase.
2. Usaremos o programa SPAdes para fazer a montagem dos *contigs* dentro da plataforma KBase.

Tutorial

1. Escolha do organismo - Recuperando *reads* do banco de dados do NCBI

Escolha do genoma no Sequence Read Archive (SRA) (SRA/NCBI)
https://www.ncbi.nlm.nih.gov/sra

Nesse arquivo usaremos o genoma de *Corynebacterium pseudotuberculosis*

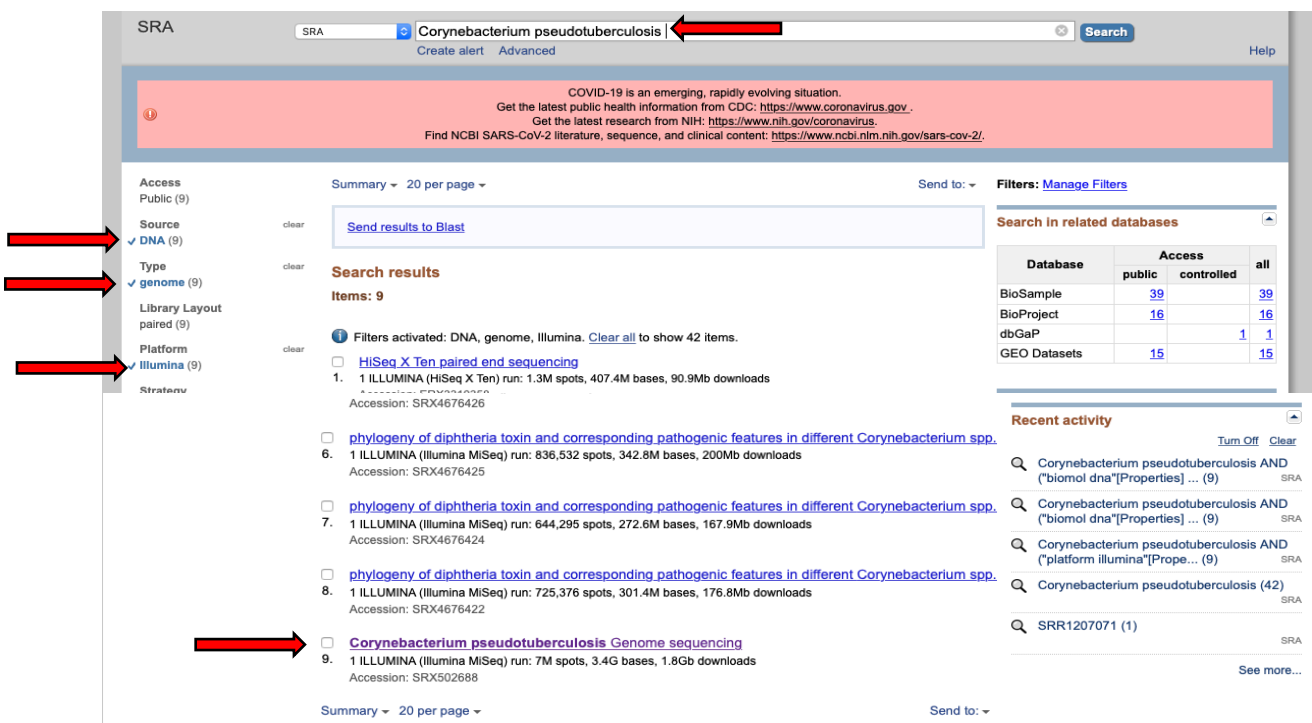

Figura 5. Pesquisa dos genomas sequenciados para o genoma da espécie *Corynebacterium pseudotuberculosis.* Foi selecionadas as opções Fonte DNA; Tipo DNA; Platform Illumina.

Foi selecionado o genoma *C. pseudotuberculosis* ID SRX502688 contendo 3.4G bases. Esse genoma foi sequenciado com o aparelho Illumina MiSeq usando sequenciamento de leituras em pares (*Paired-end sequencing*).

Figura 6. Informações básicas sobre o sequenciamento do organismo e submetido pelo Indian Council of Agricultural Research. Entrar no link indicado pela seta para obter mais informações.

A aba metadata contém informações sobre o conteúdo GC (47.7%) e o comprimento dos *reads* gerados (273 e 213). A aba *Analysis* tem informações sobre a porcentagem de *reads* que foram únicos nesse genoma e *reads* compartilhados com outros grupos taxonômicos.

Unidentified reads: 20.6%

Identified reads: 79.4%

cellular organisms: 79.4%

Bacteria: 77.83%

Eukaryota: 1.53%

Archaea: < 0.01% (2 Kbp)

Viruses: < 0.01% (20 Kbp)

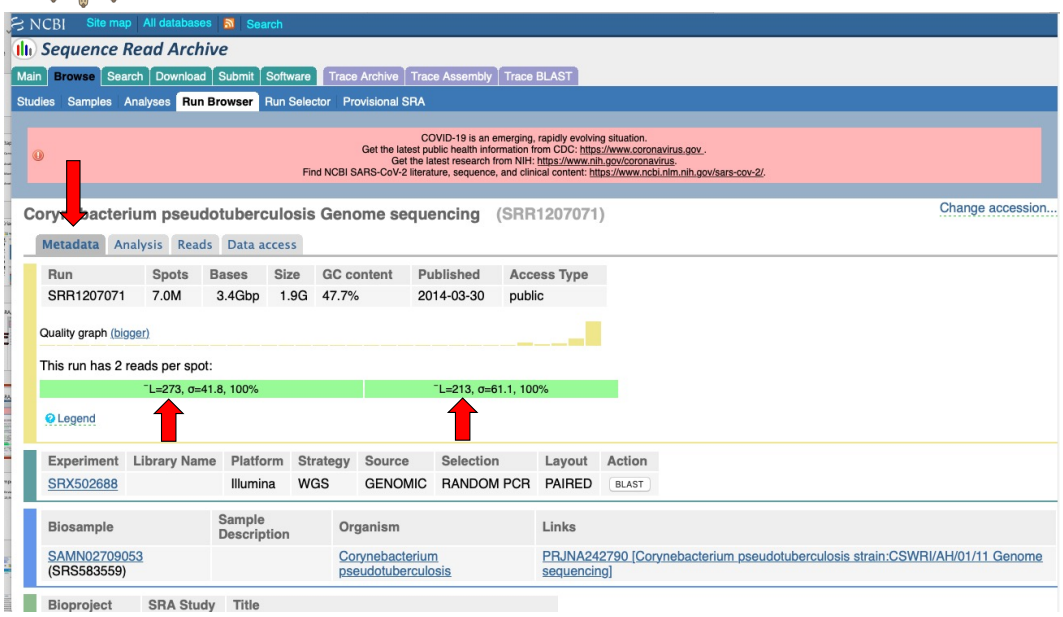

Figura 7. Informações sobre os *reads* gerados no sequenciamento. No item Links contem o link que usaremos para inserir no Kbase.

A aba data Access tem informação do link para inserir no KBase e fazer a importação dos dados. Devemos copiar o link indicado pela seta na figura 8 para inserir no KBase e fazer a montagem desse genoma.

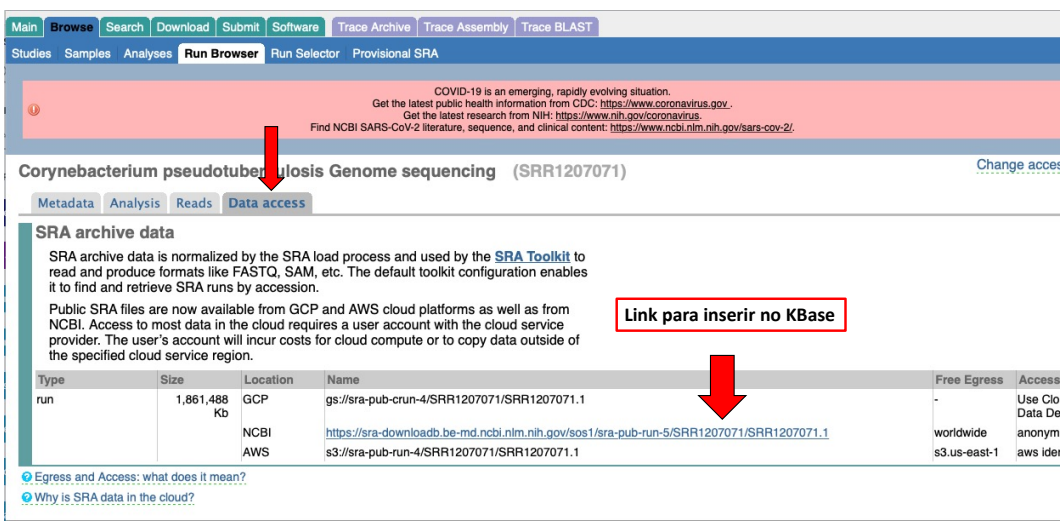

Figura 8. Aba Data access indicando o link para copiar e inserir no KBase.

Importanto os dados para o KBase

Etapas no KBase
Abrir o site do KBase e fazer o registro criando um login único de cada usuário.

Endereço: https://www.kbase.us/

Criar narrativa e carregar as leituras (reads) no KBase

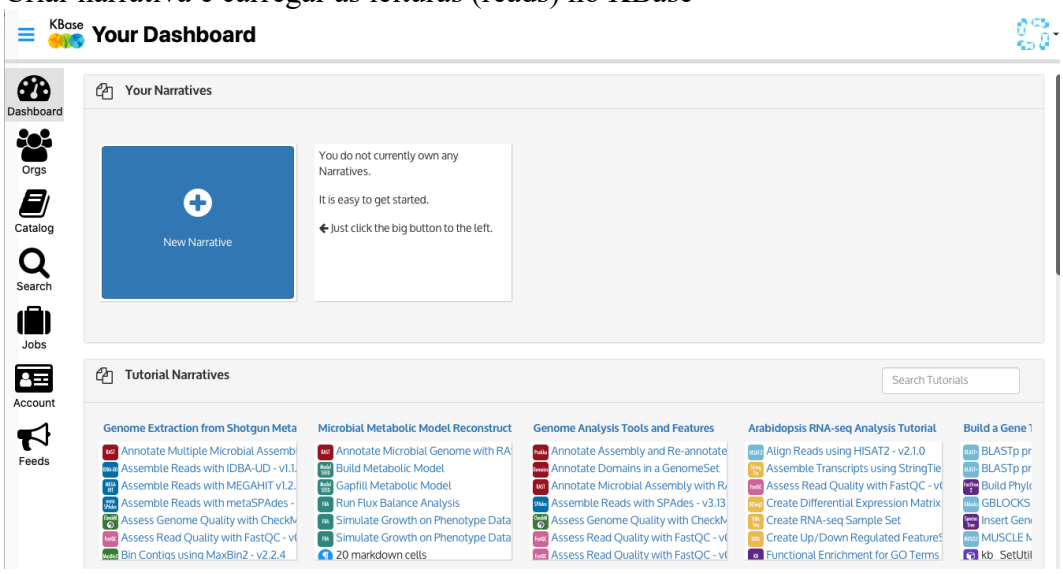

Figura 9. Página inicial do KBase após fazer o login. Vamos criar uma narrativa (New Narrative) para analisar o genoma.

Vamos fazer a importação dos *reads* do sequenciamento do genoma de *Corynebacterium pseudotuberculosis* usando um APPS do KBase. Os APPS estão no menu a esquerda. Vamos começar usando o item Upload e selecionar o APP (DATA) Import SRA File as Reads From Web - v1.0.7 (Figura 10).

Import SRA File as Reads From Web - v1.0.7

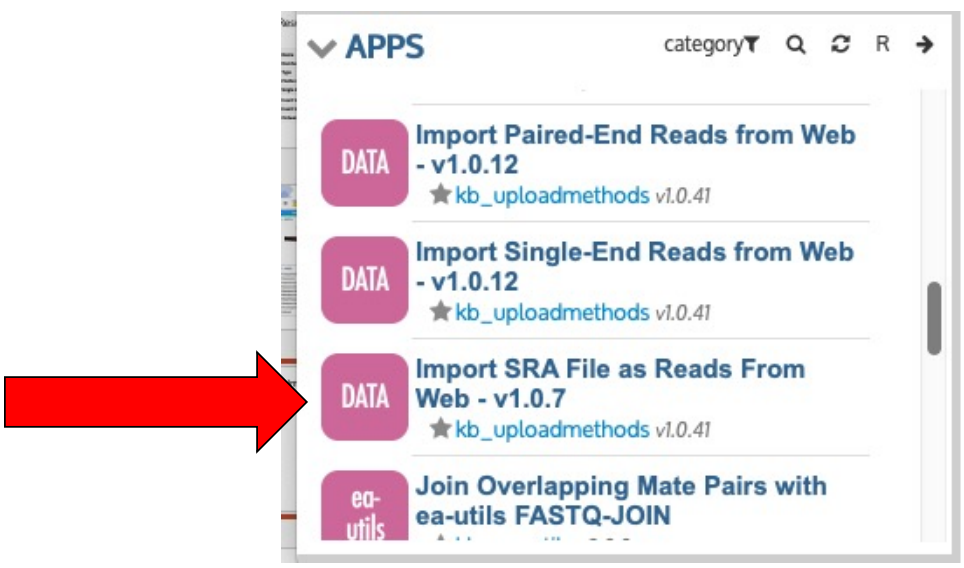

Figura 10. Menu do KBase para importação de dados de genomas sequenciados. Nesse tutorial vamos usar o APP Import SRA File as Reads From Web - v1.0.7 para importar os dados do NCBI.

Vamos inserir as três informações obrigatórias do projeto no KBase

Link: https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos1/sra-pub-run-5/SRR1207071/SRR1207071.1

Nome do projeto: Cpseudotuberculosis

Tipo de plataforma: Illumina

Tamanho médio do inserto: 273

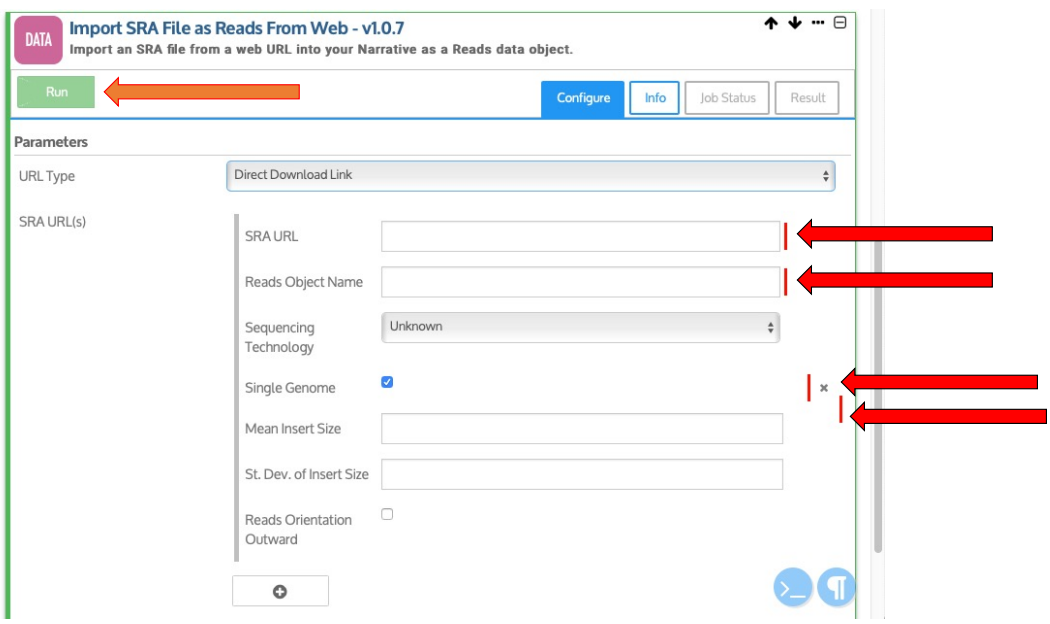

Figura 11. Importação dos dados do genoma para o KBase. Devemos inserir as informações apontadas pelas setas vermelhas. Após inserir as informações devemos rodar clicando no botão verde (Run) indicado pela seta laranja.

Tabela 1. Resumo dos dados carregados no KBase.

Name	Cpseudotuberculosis
Number of Reads	13,942,840
Type	Paired End
Platform	Illumina
Single Genome	Yes
Insert Size Mean	273.0
Insert Size Std Dev	61.0
Outward Read Orientation	No

IMS/CAT-UFBA - Rua Rio de Contas, 58 – Quadra 17 – Lote 58 – Bairro Candeias
Vitória da Conquista – BA - CEP 45.029-094/ Fone: (77) 3429 2709. E-mail: leandromartins@ufba.br

Controle de qualidade usando o FASTQC e Trimmomatic

Vamos selecionar no menu a esquerda a parte de *Read Processing* e usar o APP Assess Read Quality with FastQC - v0.11.5 (figura 12).

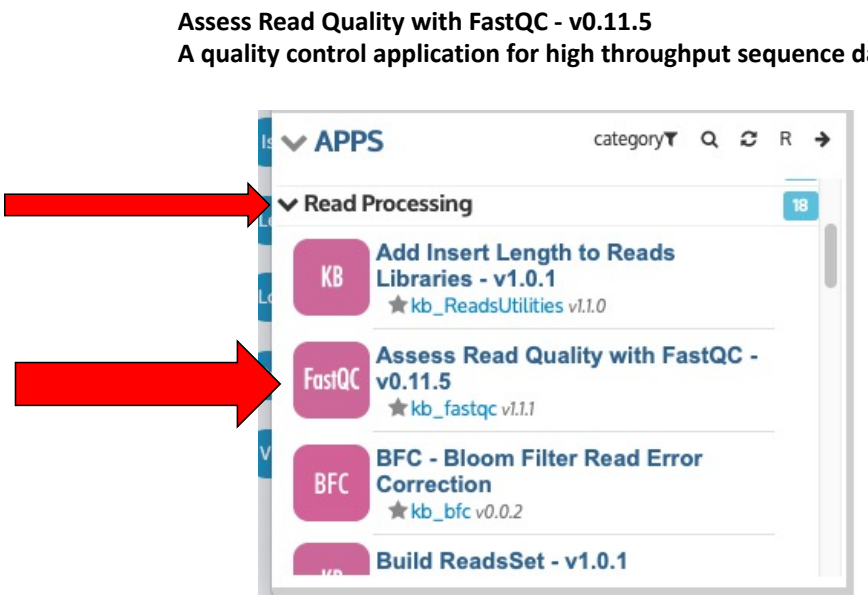

Figura 12. Seleção do Controle de qualidade dos reads usando o menu esquerdo do KBase, Read Processing, FASTQC (indicado pela seta vermelha).

Input Objects
Read Library/RNA-seq Sample Set: (Aqui selecionar os dados que foram carregados do genoma).

Após selecionar o genoma carregado, faça a análise da qualidade dos reads clicando no botão Run. O FASTQC irá retornar um relatório que mostra a quantidade de reads a qualidade dessas leituras.

Resultados do FASTQC

Basic Statistics

Measure	Value
Filename	Cpseudotuberculosis_75267_2_1.fwd.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	6971420
Sequences flagged as poor quality	0
Sequence length	40-301
%GC	48

Phred

Sanger Encoding = 33

Figura 13. Relatório estatístico após a análise do FASTQC.

O programa FASTQC faz uma análise da qualidade de cada base em cada *read* para fazer uma média de qualidade. As bases com qualidade PHRED inferior

a 20 são consideradas de baixa qualidade e não devem ser usadas na montagem do genoma.

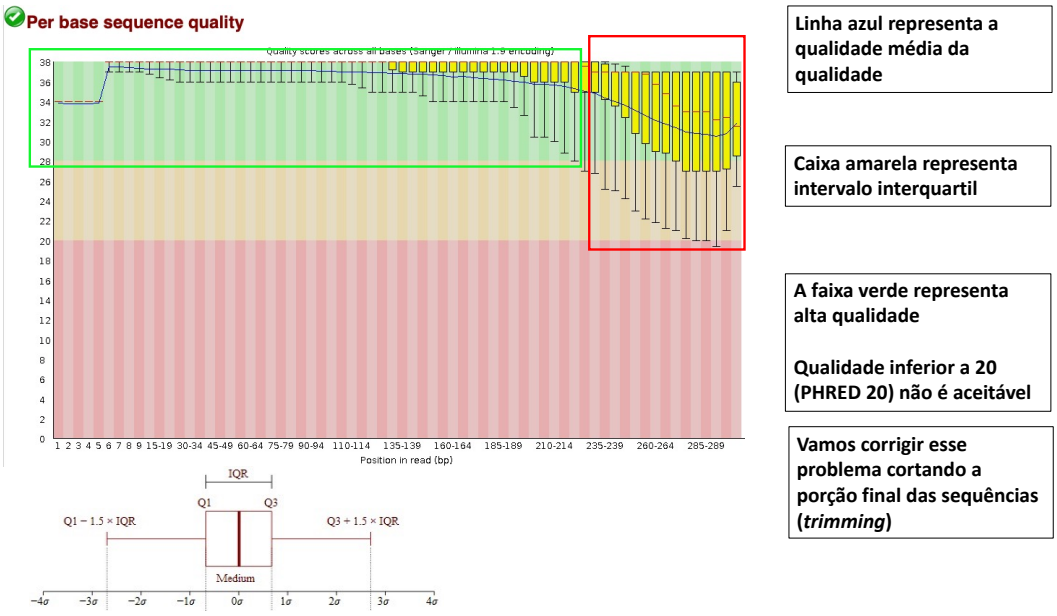

Figura 14. Relatório gráfico de qualidade dos reads após a análise do programa FASTQC. As leituras de baixa qualidade com PHRED inferior a 20 não podem ser usadas na montagem de genomas.

Vamos usar o programa Trimmomatic para fazer o corte das extremidades das leituras que apresentam bases de baixa qualidade e fazer a eliminação de leituras que não tenham muita qualidade. Dessa forma vamos evitar erros de pareamento de *reads* provocados por sequenciamento incorreto.

Removendo os reads de baixa qualidade

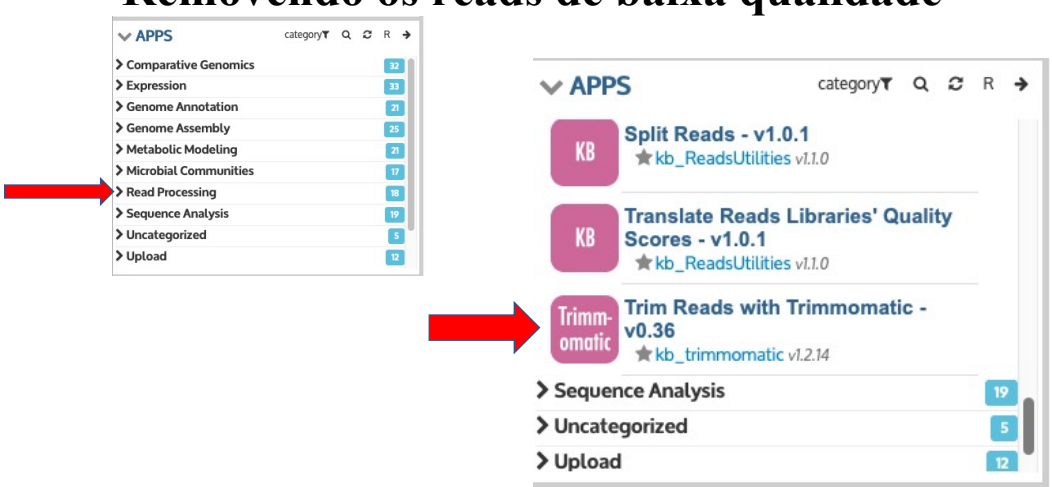

Figura 15. Selecionar no menu Read Processing o APP Trimmomatic (Trim Reads with Trimmomatic - v0.36) indicado pela seta vermelha.

Resultados do Trimmomatic

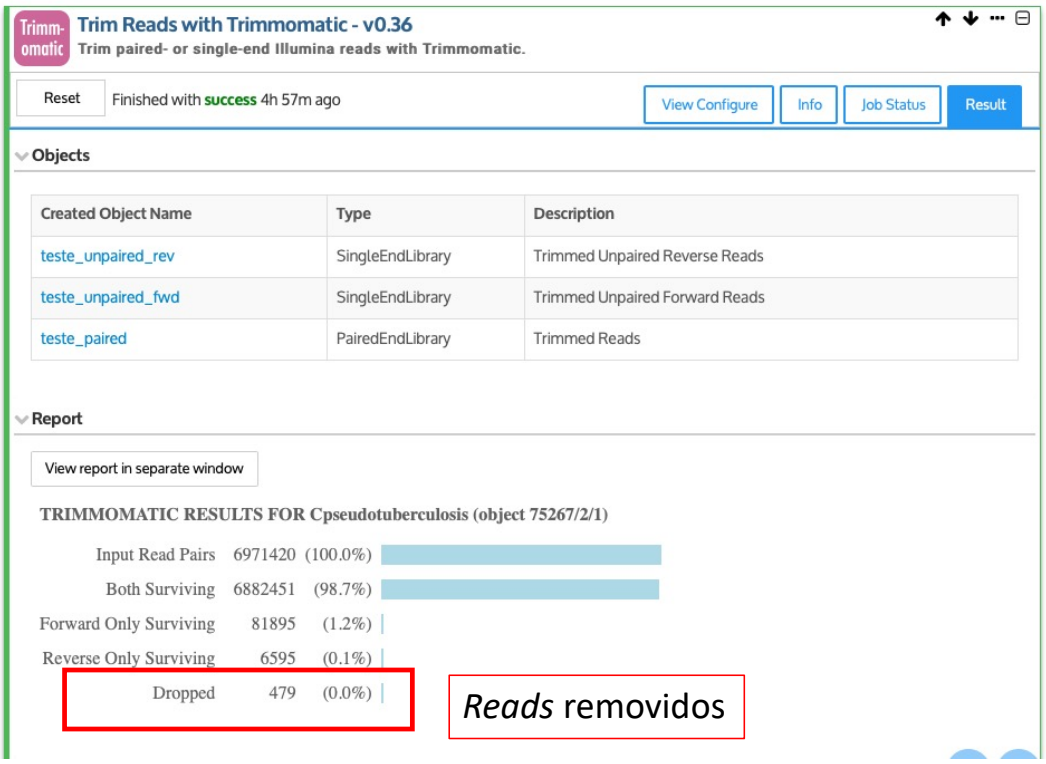

Figura 16. Resultado do Trimmomatic mostrando que foram eliminados 479 reads. Foram eliminados poucos reads, mas esses reads de baixa qualidade poderiam provocar erros na montagem do genoma.

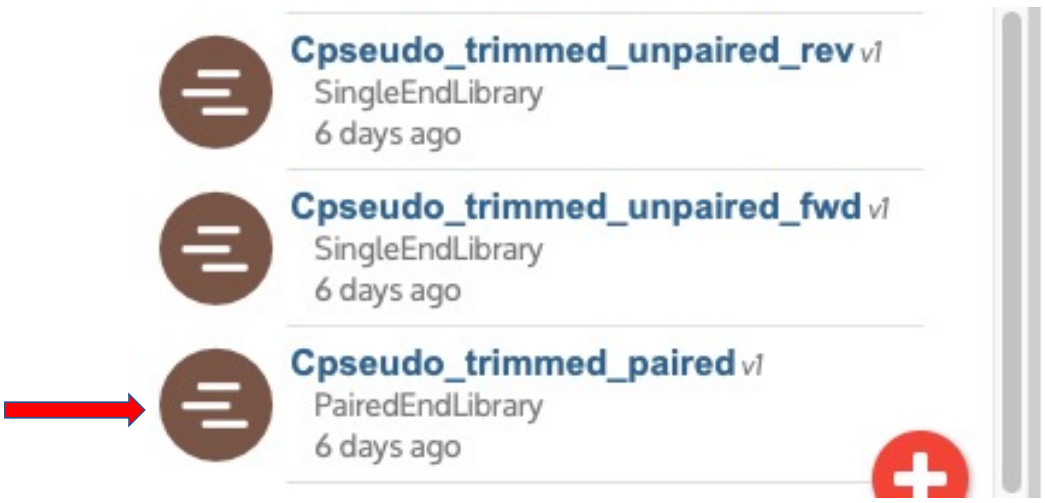

Figura 17. O Dashboard vai agora conter as leituras separadas. Vamos usar as leituras que mantiveram pareadas (Cpseudo_trimmed_paired) após o filtro do programa Trimmomatic.

Montagem do genoma com SPAdes

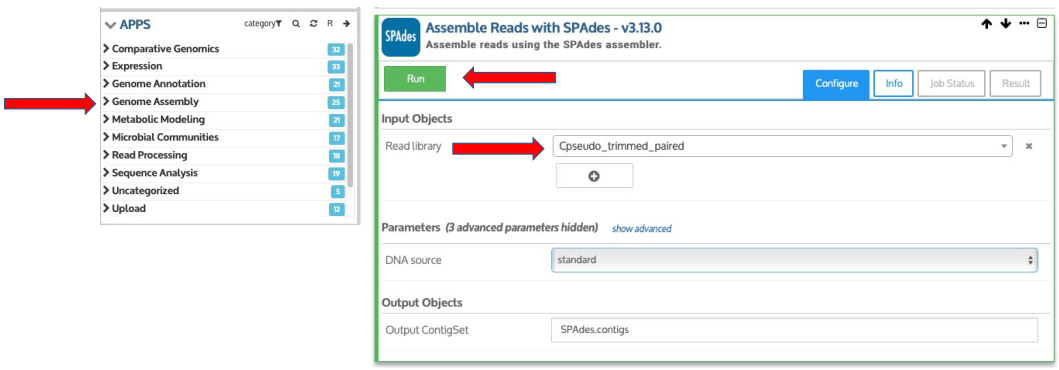

Figura 18. Vamos selecionar a montagem do genoma usando o programa SPAdes. Menu APPS, Genome Assembly, Assemble Reads with SPAdes - v3.13.0. Na opção *Read library* selecione "Escolher o resultado do Trimmomatic Paired". Depois clicar no botão Run.

Após a montagem do genoma foram construídos 248 contigs. Esses contigs são os *reads* que apresentaram sobreposição e montaram uma sequência maior. Não foi possível fechar o genoma de *C. pseudotuberculosis* somente com os *reads*. Isso é normal. Algumas regiões do genoma são mais difíceis de serem sequenciadas. O fechamento completo do genoma resulta em um *contig* que representa o genoma único da bactéria.

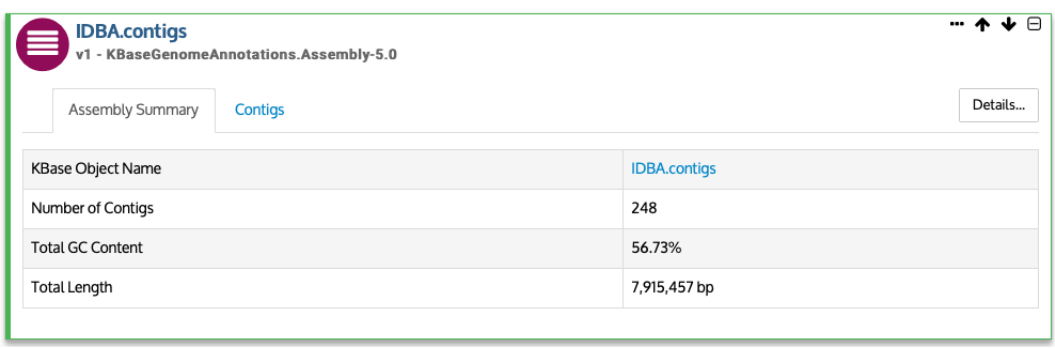

Figura 19. Resultado da montagem do genoma usando o programa SPAdes.

IMS/CAT-UFBA - Rua Rio de Contas, 58 – Quadra 17 – Lote 58 – Bairro Candeias
Vitória da Conquista – BA - CEP 45.029-094/ Fone: (77) 3429 2709. E-mail: leandromartins@ufba.br