Genomic Assembly Menggunakan Galaxy

Rully Meidyta
Program Studi Magister Matematika
Universitas Indonesia

Abstract

Genomic assembly adalah aspek penting dalam ilmu genomika yang berfokus pada konstruksi urutan genom lengkap dari data sekuensing DNA. Dalam tutorial ini, menggunakan Galaxy, platform analisis berbasis web yang kuat, untuk melakukan genome assembly dengan data bacaan dari bakteri Staphylococcus aureus yang dibayangkan. Data tersebut dihasilkan melalui sekuensing DNA whole genome shotgun dengan instrumen Illumina.

Proses assembly melibatkan langkah-langkah persiapan data, pemilihan alat assembly, pengaturan parameter, perakitan, evaluasi hasil, dan visualisasi. Dan juga menjelaskan cara membaca grafik kualitas data seperti ”Per Base Sequence Quality” dan ”Per Sequence Quality Scores,” serta grafik ”Sequence Duplicate Levels” untuk memahami kualitas data dan tingkat duplikasi. Galaxy menyediakan alat analisis yang ramah pengguna, memfasilitasi penelitian genomika dengan antarmuka yang intuitif.

I. Pendahuluan

Genome assembly adalah sebuah disiplin penting dalam ilmu genomika yang berfokus pada penggabungan jutaan fragmen DNA pendek menjadi urutan genom lengkap suatu organisme. Genom, atau kumpulan seluruh informasi genetik yang dimiliki oleh suatu organisme, memegang kunci untuk pemahaman tentang struktur, fungsi, dan evolusi kehidupan. Genome assembly menjadi fondasi bagi berbagai penelitian biologi, termasuk pemahaman tentang penyakit, keanekaragaman hayati, serta evolusi spesies.

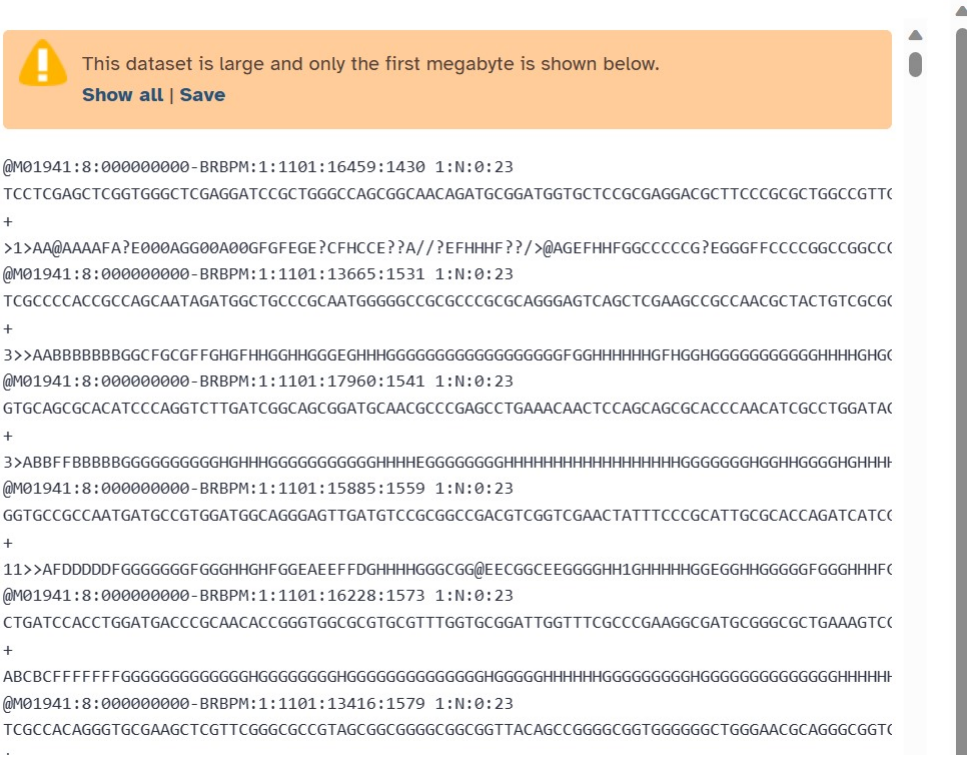
Proses genome assembly melibatkan sejumlah langkah yang rumit, dan teknologi serta perangkat lunak yang digunakan terus berkembang. Pada dasarnya, tujuan utamanya adalah mengonstruksi urutan DNA yang sesuai dengan urutan sebenarnya dalam genom suatu organisme. Namun, ini adalah tantangan yang rumit karena genom biasanya terdiri dari berjuta-juta pasangan basa DNA yang harus digabungkan dari bacaan-bacaan pendek yang dihasilkan dari sekuensing DNA.Galaxy adalah platform berbasis web yang kuat dan mudah digunakan untuk melakukan analisis data ilmiah. Dalam proses analisis, tindakan-tindakan dijalankan dengan menggunakan alat-alat yang tersedia di Galaxy, yang secara grafis mengartikan perintah-perintah yang biasanya diberikan melalui baris perintah perangkat lunak ke dalam antarmuka web yang ramah pengguna. Galaxy memiliki antarmuka web grafis yang intuitif dan menyediakan sejumlah besar alat analisis, serta sumber daya pelatihan berkualitas tinggi yang dikembangkan dan dipelihara oleh komunitas pengguna. Hal ini memungkinkan pemula dan ahli dalam analisis data untuk bekerja dengan cepat dan berinteraksi dengan data mereka secara efisien. Selama setiap langkah dalam analisis, Galaxy juga mencatat metadata yang mencakup informasi seperti identifikasi alat dan versi, input data, serta parameter yang digunakan. Hal ini membantu dalam menjaga reproduksibilitas hasil analisis.

Galaxy juga memungkinkan pengguna untuk dengan mudah berbagi alur kerja analisis dan data mereka dengan orang lain. Ini merupakan perangkat lunak sumber terbuka, yang berarti dapat diunduh dan diinstal secara lokal, atau digunakan melalui lebih dari 120 server publik yang tersedia.

II. Data

Untuk tutorial ini, kita memiliki sekumpulan bacaan (reads) dari bakteri Staphylococcus aureus yang dibayangkan dengan genom miniatur (197.394 pasangan basa). Set data bacaan dari strain mutan ini disekuensing dengan metode whole genome shotgun, menggunakan instrumen sekuensing DNA Illumina. Dari data bacaan ini, kita ingin membangun kembali bakteri Staphylococcus aureus yang dibayangkan kita melalui perakitan de novo dari sekumpulan bacaan pendek menggunakan perangkat perakit Velvet.

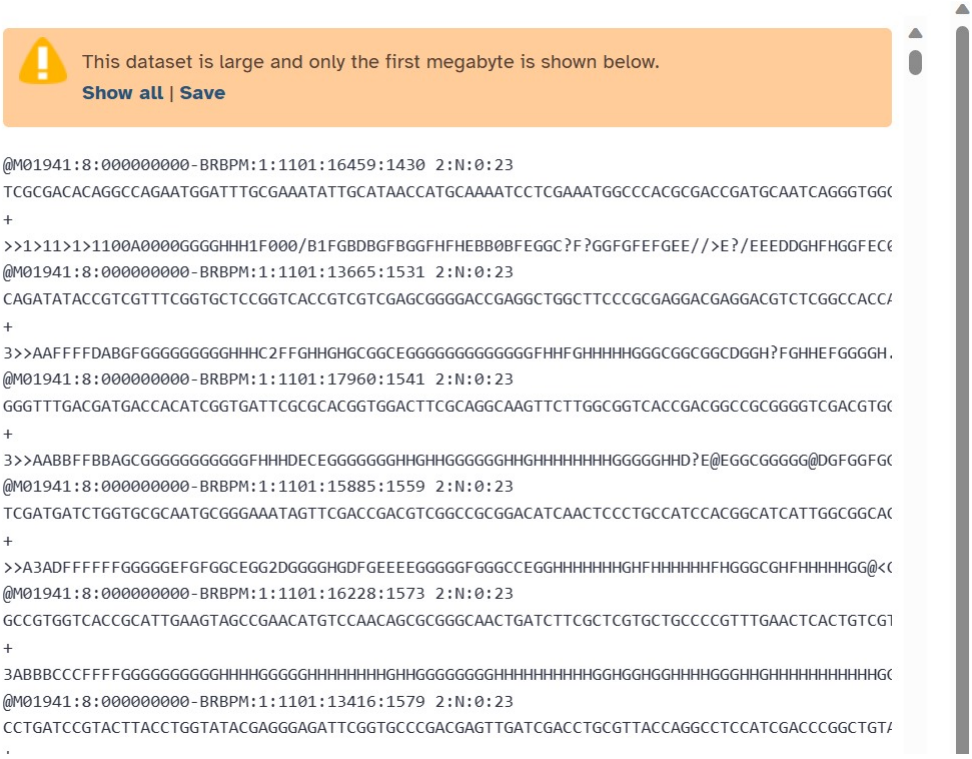
A. R1



This dataset is large and only the first megabyte is shown below.
Show all | **Save**

```
@M01941:8:000000000-BRBPM:1:1101:16459:1430 1:N:0:23
TCCTCGAGCTCGGTGGGCTCGAGGATCCGCTGGGCCAGCGGCAACAGATGCGGATGGTGCTCCGCGAGGACGCTTCCCGCGCTGGCCGTTC
+
>1>AA@AAAAFA?E000AGG00A00GFGFEGE?CFHCCE??A//?EFHHHF??/>@AGEFHHFGGCCCCCG?EGGGFFCCCCGGCCGGCCC
@M01941:8:000000000-BRBPM:1:1101:13665:1531 1:N:0:23
TCGCCCCACCGCCAGCAATAGATGGCTGCCCGCAATGGGGGCCGCGCCCGCGCAGGGAGTCAGCTCGAAGCCGCCAACGCTACTGTCGCGC
+
3>>AABBBBBBBGGCFGCGFFGHGFHHGGHHGGGEGHHHGGGGGGGGGGGGGGGGGFGGHHHHHHGFHGGHGGGGGGGGGGGHHHHGHGC
@M01941:8:000000000-BRBPM:1:1101:17960:1541 1:N:0:23
GTGCAGCGCACATCCCAGGTCTTGATCGGCAGCGGATGCAACGCCCGAGCCTGAAACAACTCCAGCAGCGCACCCAACATCGCCTGGATAC
+
3>ABBFFBBBBBGGGGGGGGGGHGHHHGGGGGGGGGGGHHHHEGGGGGGGHHHHHHHHHHHHHHHHHHGGGGGGGHGGHHGGGGHGHHHH
@M01941:8:000000000-BRBPM:1:1101:15885:1559 1:N:0:23
GGTGCCGCCAATGATGCCGTGGATGGCAGGGAGTTGATGTCCGCGGCCGACGTCGGTCGAACTATTTCCCGCATTGCGCACCAGATCATCC
+
11>>AFDDDDDFGGGGGGGFGGGHHGHFGGEAEEFFDGHHHHGGGCGG@EECGGCEEGGGGHH1GHHHHHGGEGGHHGGGGGFGGGHHHFC
@M01941:8:000000000-BRBPM:1:1101:16228:1573 1:N:0:23
CTGATCCACCTGGATGACCCGCAACACCGGGTGGCGCGTGCGTTTGGTGCGGATTGGTTTCGCCCGAAGGCGATGCGGGCGCTGAAAGTCC
+
ABCBCFFFFFFFGGGGGGGGGGGGGHGGGGGGGGHGGGGGGGGGGGGGGHGGGGGHHHHHHGGGGGGGGGHGGGGGGGGGGGGGHHHHHH
@M01941:8:000000000-BRBPM:1:1101:13416:1579 1:N:0:23
TCGCCACAGGGTGCGAAGCTCGTTCGGGCGCCGTAGCGGCGGGGCGGCGGTTACAGCCGGGGCGGTGGGGGGCTGGGAACGCAGGGCGGTC
.
```

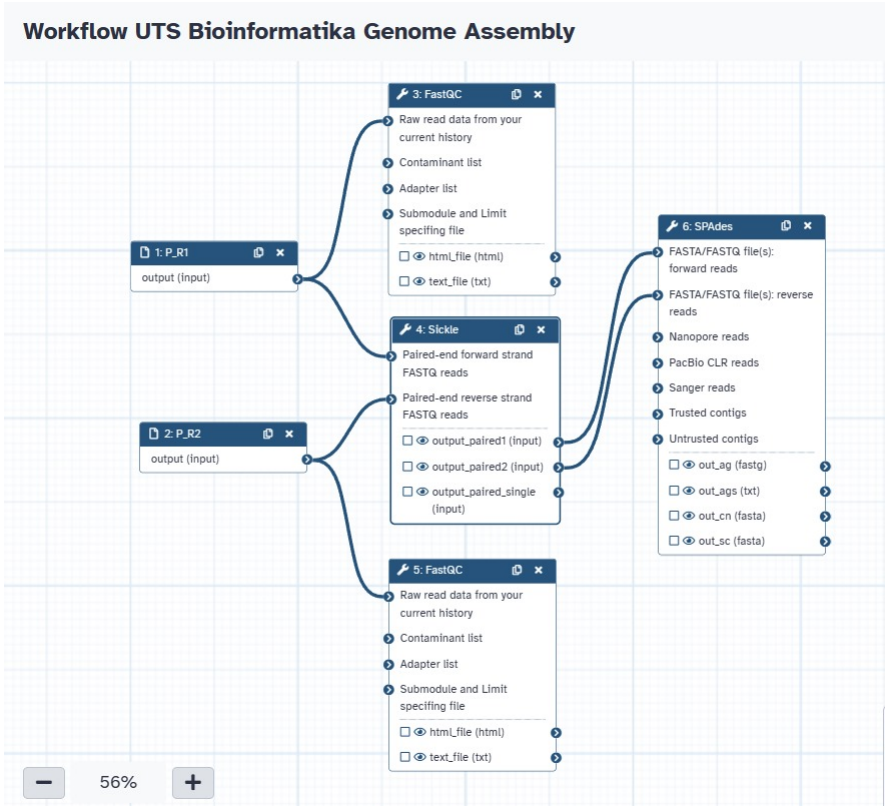
B. R2



This dataset is large and only the first megabyte is shown below.
Show all | **Save**

```
@M01941:8:000000000-BRBPM:1:1101:16459:1430 2:N:0:23
TCGCGACACAGGCCAGAATGGATTTGCGAAATATTGCATAACCATGCAAAATCCTCGAAATGGCCCACGCGACCGATGCAATCAGGGTGGC
+
>>1>11>1>1100A0000GGGGHHH1F000/B1FGBDBGFBGGFHFHEBB0BFEGGC?F?GGFGFEFGEE//>E?/EEEDDGHFHGGFECC
@M01941:8:000000000-BRBPM:1:1101:13665:1531 2:N:0:23
CAGATATACCGTCGTTTCGGTGCTCCGGTCACCGTCGTCGAGCGGGGACCGAGGCTGGCTTCCCGCGAGGACGAGGACGTCTCGGCCACCA
+
3>>AAFFFFDABGFGGGGGGGGGHHHC2FFGHHGHGCGGCEGGGGGGGGGGGGGGFHHFGHHHHHGGGCGGCGGCDGGH?FGHHEFGGGGH.
@M01941:8:000000000-BRBPM:1:1101:17960:1541 2:N:0:23
GGGTTTGACGATGACCACATCGGTGATTCGCGCACGGTGGACTTCGCAGGCAAGTTCTTGGCGGTCACCGACGGCCGCGGGGTCGACGTGC
+
3>>AABBFFBBAGCGGGGGGGGGGGFHHHDECEGGGGGGGGHHGHHGGGGGGHHGHHHHHHHHHGGGGGHHD?E@EGGCGGGGG@DGFGGFGC
@M01941:8:000000000-BRBPM:1:1101:15885:1559 2:N:0:23
TCGATGATCTGGTGCGCAATGCGGGAAATAGTTCGACCGACGTCGGCCGCGGACATCAACTCCCTGCCATCCACGGCATCATTGGCGGCAC
+
>>A3ADFFFFFFGGGGGEFGFGGCEGG2DGGGGHGDFGEEEEGGGGGFGGGCCEGGHHHHHHHHGHFHHHHHHFHGGGCGHFHHHHHGG@<(
@M01941:8:000000000-BRBPM:1:1101:16228:1573 2:N:0:23
GCCGTGGTCACCGCATTGAAGTAGCCGAACATGTCCAACAGCGCGGGCAACTGATCTTCGCTCGTGCTGCCCCGTTTGAACTCACTGTCGT
+
3ABBBCCCFFFFGGGGGGGGGGHHHHGGGGGHHHHHHHHGHHGGGGGGGGHHHHHHHHHHHGGHGGHGGHHHHGGGHHGHHHHHHHHHGC
@M01941:8:000000000-BRBPM:1:1101:13416:1579 2:N:0:23
CCTGATCCGTACTTACCTGGTATACGAGGGAGATTCGGTGCCCGACGAGTTGATCGACCTGCGTTACCAGGCCTCCATCGACCCGGCTGTA
.
```

III. WORKFLOW



IV. EVALUASI

Basic Statistics

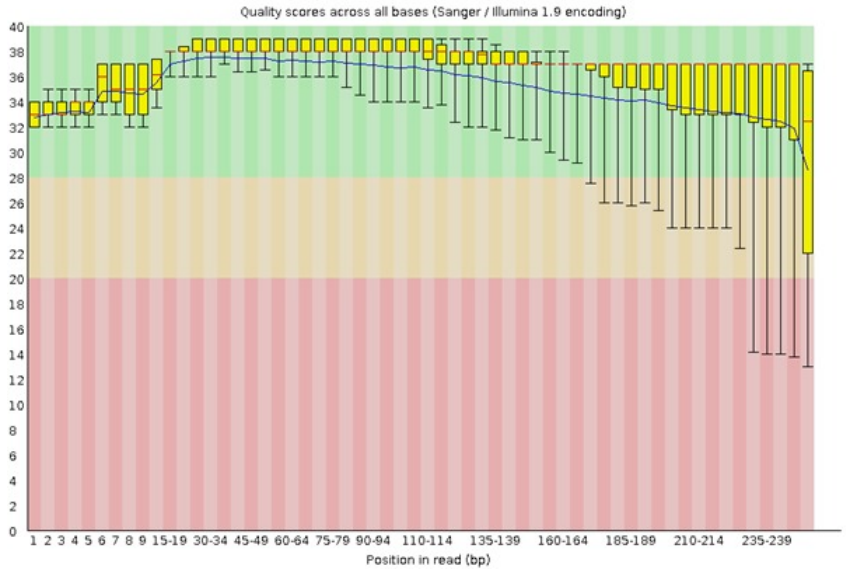
Measure	Value
Filename	P_R1.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	273071
Total Bases	61.7 Mbp
Sequences flagged as poor quality	0
Sequence length	35-251
%GC	64

A. Per Base Squence Quality

Pada grafik, akan melihat sumbu-X (horizontal) yang mewakili posisi dalam bacaan sekuensing (dalam urutan), dan sumbu-Y (vertikal) yang mewakili skor kualitas (biasanya dalam skala Phred). Grafik akan menampilkan serangkaian garis atau kurva yang menggambarkan skor kualitas (Phred) pada setiap posisi dalam bacaan. Garis-garis ini mewakili berbagai bacaan dalam dataset Anda.

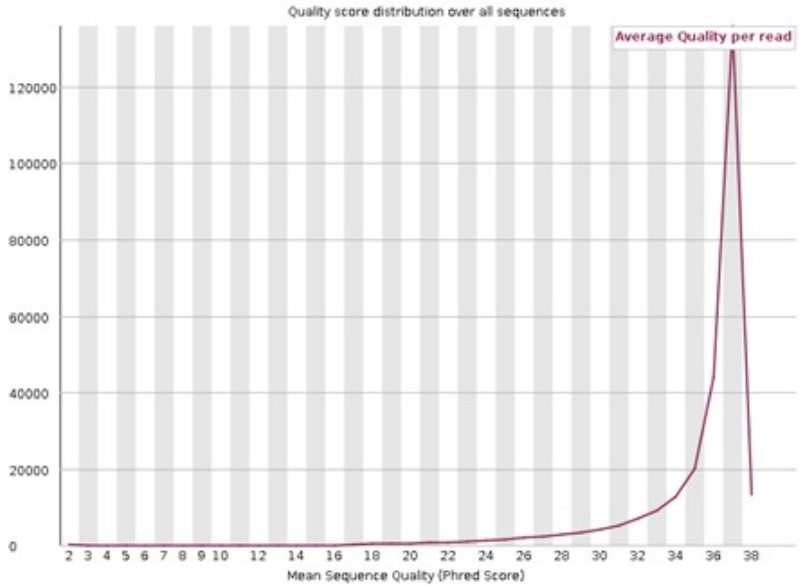
Pada umumnya, ingin melihat agar garis-garis tersebut tetap tinggi dan seragam pada semua posisi. Ini menunjukkan kualitas yang baik dan konsisten dari data sekuensing. Beberapa grafik FASTQ Sequence Quality menggunakan warna untuk menandai

kualitas. Jika ada area dengan warna merah, kuning, atau hijau, perhatikan apa yang warna tersebut wakili. Biasanya, merah menunjukkan kualitas rendah, sedangkan hijau menunjukkan kualitas tinggi. Perhatikan jika ada penurunan kualitas yang signifikan pada posisi tertentu dalam bacaan. Penurunan tajam dapat menjadi indikasi kesalahan atau masalah dalam data.



B. Per Sequence Quality Scores

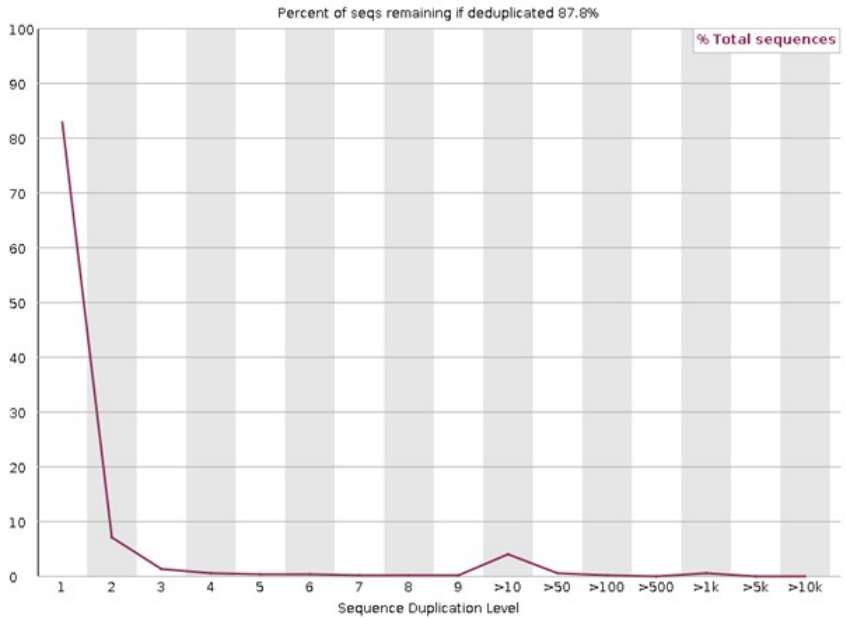
Setiap garis atau titik dalam grafik mewakili satu bacaan. Mereka tersebar di sepanjang sumbu-X sesuai dengan urutan bacaan dalam dataset. Garis atau titik tersebut pada sumbu-Y akan menunjukkan skor kualitas bacaan yang sesuai. Posisi tinggi pada sumbu-Y menandakan kualitas yang baik, sedangkan posisi rendah menandakan kualitas yang buruk. Perhatikan distribusi skor kualitas secara keseluruhan. Anda ingin melihat bahwa sebagian besar bacaan memiliki skor kualitas yang tinggi dan konsisten. Sejumlah besar bacaan dengan skor kualitas rendah atau dengan variabilitas yang tinggi dapat menjadi indikasi masalah dalam data. Perhatikan pola atau tren dalam grafik. Pola tertentu seperti perubahan mendadak dalam kualitas atau fluktuasi yang konsisten dapat memberikan wawasan tambahan.



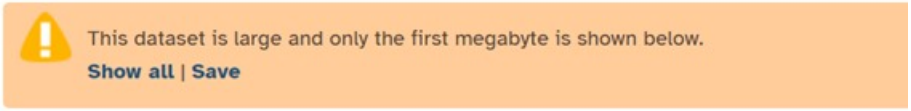
C. Squence Duplicate Level

Garis atau batang dalam grafik mewakili tingkat duplikasi yang berbeda (biasanya dalam persentase), dan tinggi garis atau batang tersebut pada sumbu-Y menandakan berapa banyak bacaan yang memiliki tingkat duplikasi tersebut. Perhatikan distribusi

tingkat duplikasi. Jika sebagian besar bacaan memiliki tingkat duplikasi rendah, ini menunjukkan bahwa dataset Anda memiliki variasi yang baik. Namun, jika ada banyak bacaan dengan tingkat duplikasi tinggi, ini bisa menjadi indikasi adanya duplikasi yang signifikan dalam data.



V. HASIL




```
>NODE_1_length_88090_cov_8.723054
GCCCCCGCCCTACCCGTCCCCCGACCCCTTTGAGCCGTTCGCCGTCGCCGCGGTGGAGCT
CGCCGACGAGGGGCTGATCATGCTAGGCAAAGTTGTCGAAGGCACGCTGGCCGCCGACTT
GAAGATCGGCATGGAGATGGAGCTGACGACCATGCCGCTGTTCACCGACGAGGACGGCGT
CAAACGCATCGTGTACGCCTGGAGGATCGCCTGATGAGTGCCCCCGAACCCCTTTACATC
CTTGGCGCCGGAATGCACCCGTGGGGCAAGTGGGGCAACGACTTCACCGAATACGGGGTC
GCTGCCGCCCGCGCCGCCCTGCGCGAAGCCGGTTTGGAATGGCGCCAGATCCAGCTGGTG
GCCGGCGCGGACACCATCCGCAACGGATACCCGGGCTTTGTGGCCGGCGCCACCTTCGCG
CAGAAGCTCGGCTGGAACGGCGTGCCGGTCAGTTCGAGCTACGCCGCGTGCGCCAGCGGC
TCCCAGGCGCTGCAAAGCGCGCGGGCCCAGATTCTGGCCGGATTCTGCGATGTTGCGTTG
GTGGTCGGCGCCGACACCACCCCGAAAGGGTTTTTCGCCCCGGTCGGCGGTGAGCGCAAA
AACGATCCCGATTGGCAGCGATTCCACCTGATCGGGGCTACCAACCCGGTCTACTTCGCC
TTGCTCGCGCGGCGCCGGATGGACCTGTACGGGGCCACCCTCGAGGATTTCGCCCAGGTG
AAAGTCAAGAACTCCCGGCACGGCCTGCAAAACCCCAACGCCCGCTACCGCAAGGAAGTC
TCGGTCGAGGACGTGCTGGCCAGCCAGGTAGTGGCTGAACCGTTGCGGTTGCTGGATATC
TGCGCCACCTCCGACGGCGCGGCGGCCTTGATCGTGGCCAGCGCAAAGTTCGCTCGCGAA
CACCTGGGCTCGCTCGACGGGGTGCCATCGGTGCGCGCGGTCAGCACCGTGACCCCGCGC
TACCCACAACATCTGCCCGAATTGCCGGACATCGCAACGGATTCCACGGCGGTAGTCCCC
GCACCGGAGCGAGTGTTCAAGGATCAGATCCTCGATGCGGCCTATACCGAAGCCGGCATA
GGCCCCAAGGACCTGAGCCTGGCTGAGGTGTACGACCTGTCCACCGCGCTAGAACTTGAC
TGGTACGAGCACCTGGGTCTGTGCCCGAAAGGTGAAGCCGAGGCGCTGCTGCGCAGCGGA
GCGACGACCATCGGCGGCAGGGTGCCGGTCAACCCGTCCGGTGGGCTGGCCTGCTTCGGC
GAGGCAATCCCCGCTCAGGCCATTGCCCAAGTCTGCGAGTTGAACTGGCAGCTGCGCGGT
CAGGCCACCGGCCGGCAAGTGGAGAACGCCAGGGTGGGCGTGACGGCGAACCAGGGCCTG
```

Gambar diatas merupakan hasil dari Genomic Assembly

VI. KESIMPULAN

Genomic assembly menggunakan Galaxy adalah pendekatan yang kuat dan mudah digunakan untuk menggabungkan data sekuensing DNA dari bakteri Staphylococcus aureus atau organisme lain. Dalam kasus konkret ini, data bacaan dari strain mutan Staphylococcus aureus telah disekuensing menggunakan metode whole genome shotgun dengan sekuensing DNA Illumina