UTS Bioinformatika Lanjut

Sayyid Abdullah

November 4, 2023

Abstract

Pengolahan data adalah elemen kunci dalam penelitian ilmiah dan analisis bioinformatika. Galaxy adalah platform perangkat lunak sumber terbuka yang memungkinkan para peneliti untuk membangun, menjalankan, dan membagikan alur kerja pengolahan data secara efisien. Dalam penelitian ini, kami menjelaskan bagaimana kami mereproduksi dua pipeline pengolahan data yang relevan dalam konteks analisis genetika menggunakan Galaxy.

Pertama, kami mereproduksi alur kerja untuk membuat pohon pilogenetik. Kami mengikuti langkah-langkah dalam Galaxy untuk lihat kekerabatan suatu sequens dengan menggunakan pohon pilogenetik. Kami memeriksa konfigurasi alur kerja dan parameter untuk memastikan reproduksi yang akurat.

Kedua, kami mereproduksi alur kerja untuk menemukan Exons dengan SNPs (Single-Nucleotide Polymorphism) terbesar. Kami mengikuti langkah-langkah yang terdokumentasi dengan baik dalam Galaxy untuk membuat alur kerja baru dan memasukkan data Exons dan SNPs ke dalam alur kerja. Hasil dari alur kerja ini adalah Exons dengan SNPs terbanyak.

Dalam kedua kasus, kami berhasil mereproduksi alur kerja dengan menggunakan Galaxy.

1 Pendahuluan

Spesies yang memiliki genom kompleks seperti manusia, seringkali memiliki sejumlah besar informasi genetik yang perlu diidentifikasi, diurutkan, dan dianalisis untuk memahami aspek-aspek biologisnya. Salah satu elemen penting dalam pemahaman genom adalah identifikasi dan karakterisasi ekson, yang merupakan bagian penting dari ekspresi gen. Ekson adalah bagian dari sekuens genom yang mengandung informasi yang di-transkripsi dan diterjemahkan menjadi produk protein. Penemuan ekson-ekson yang signifikan, terutama ekson terbesar dalam genom, merupakan tantangan penting dalam analisis genom. Ekson terbesar sering kali mengandung informasi genetik yang berperan dalam pengaturan ekspresi gen atau fungsi biologis yang penting. Oleh karena itu, identifikasi ekson terbesar memiliki potensi untuk mengungkap rahasia biologis yang penting.

Tidak hanya ekson saja yang bisa mengungkap rahasia biologis suatu spesies, kita bisa menggunakan sequens DNAnya untuk melihat kekerabatannya. Hal ini bisa melihat rahasia biologis dari suatu spesies dengan melihat biologis dari kerabatnya atau nenek monyangnya.

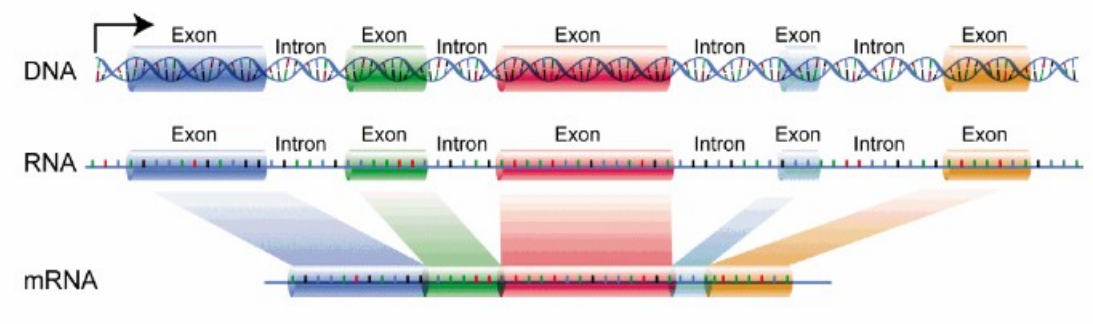

Figure 1: An original piece of DNA containing introns and exons has the introns cut out before the exons are joined together to form the mRNA

Dalam penelitian ini, kami akan menjelaskan pendekatan kami untuk mengembangkan sebuah pipeline analisis yang bertujuan untuk mencari dan mengidentifikasi ekson terbesar dalam genom spesies tertentu. Kami akan menggunakan metode bioinformatika dan komputasi yang canggih untuk mengelola, menganalisis, dan memproses data genom secara efisien. Penemuan ekson terbesar akan memberikan wawasan penting dalam pemahaman genom spesies tersebut dan dapat memiliki implikasi dalam berbagai aspek, termasuk pemahaman penyakit genetik, evolusi genom, dan pengembangan terapi genetik. Kemudian kami juga akan membuat pipeline untuk membuat pohon pilogenetik.

2 Metode

Untuk mengimplementasikan maksud dan tujuan pada pendahuluan. Terdapat langkah langkah sebagai berikut:

1. Mempersiapkan dataset.
2. Mempelajari dan memahami software atau tools yang akan digunakan. Dalam hal ini menggunakan Galaxy.
3. Membuat History baru.
4. Memasukkan datasetnya.
5. Membuat chanel yang berisi perintah dengan menggunakan tools yang sudah disediakan oleh Galaxy. Chanel disini disesuaikan dengan tujuannya.
6. Membuat workflow dari chanel yang sudah dibuat.

3 Hasil dan Pembahasan

Pada bagian ini akan dibahas mengenai cara pembuatan pipeline dengan menggunakan Galaxy. Pipeline yang akan dibuat adalah pipeline untuk menemukan Exons dengan SNPs terbesar dan pipeline untuk membuat pohon pilogenetik dari beberapa sequens.

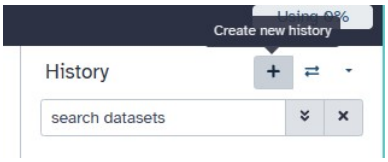

Figure 2: Membuat History Baru

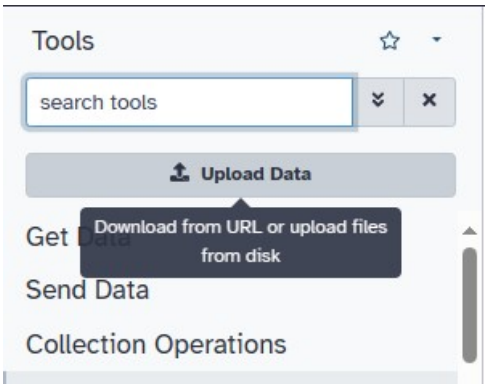

Figure 3: Tools upload data

3.1 Pipeline untuk menemukan Exons dengan SNPs terbesar

Langkah pertama adalah dengan mengunjungi link Galaxy. Selanjutnya adalah membuat history baru seperti pada Gambar 2. Nama history tersebut dapat diubah sesuai dengan kemauan user. Setelah membuat history baru, langkah selanjutnya adalah memasukkan data Ekson dan SNPs kedalam history yang sudah dibuat. Caranya adalah dengan menggunakan tools Upload data seperti pada Gambar 3. Tampilan selanjutnya adala seperti pada Gambar 4. Pada bagian ini user bisa memilih akan mengupload dataset darimana. Setelah dataset Exons dan SNPs terupload, langkah selanjutnya adalah perpotongan dari kedua dataset sersebut. Tools yang digunakan adalah bedtools Intersect intervals. Sesuaikan parameter seperti pada Gambar 5. Langkah selanjutnya adalah menghitung jumlah SNPs per Exons dengan menggunakan tools Datamash. Parameternya seperti pada Gambar 6. Langkah selanjutnya adalah mengurutkan hasil dari hitungan jumlah SNPs tadi dengan menggunakan tools Sort. Caranya seperti pada Gambar 7. Output dari proses tersebut adalah seperti pada Gambar 8 Output tersebut telah menghasilkan Exons dengan SNPS terbanyak. Dari hasil tersebut diperoleh bahwa lokasi exons dengan SNPs terbesar terdapat pada cromosom 22.

Analysis pada Galaxy bisa menghasilkan banyak history. User bisa mengontrol history tersebut seperti pada Gambar 9. Setelah dirasa pipeline berhasil, langkah selanjutnya adalah mengkonversinya menjadi workflow dengan mengekstak melalui 9. Hasil dari workflow tersebut ada pada menu workflow. Tampilan workflow seperti pada Gambar 10.

3.2 Pipeline untuk membuat pohon pilogenetik

Langkah pertama pada bagian ini adalah mengupload dataset sequens yang akan dibuat pohon pilogenetiknya. caranya sama seperti pada Gambar 3. Karena dataset yang kita upload memiliki nama yang berimbuhan .fastq.vcf, hal yang perlu dilakukan adalah den-

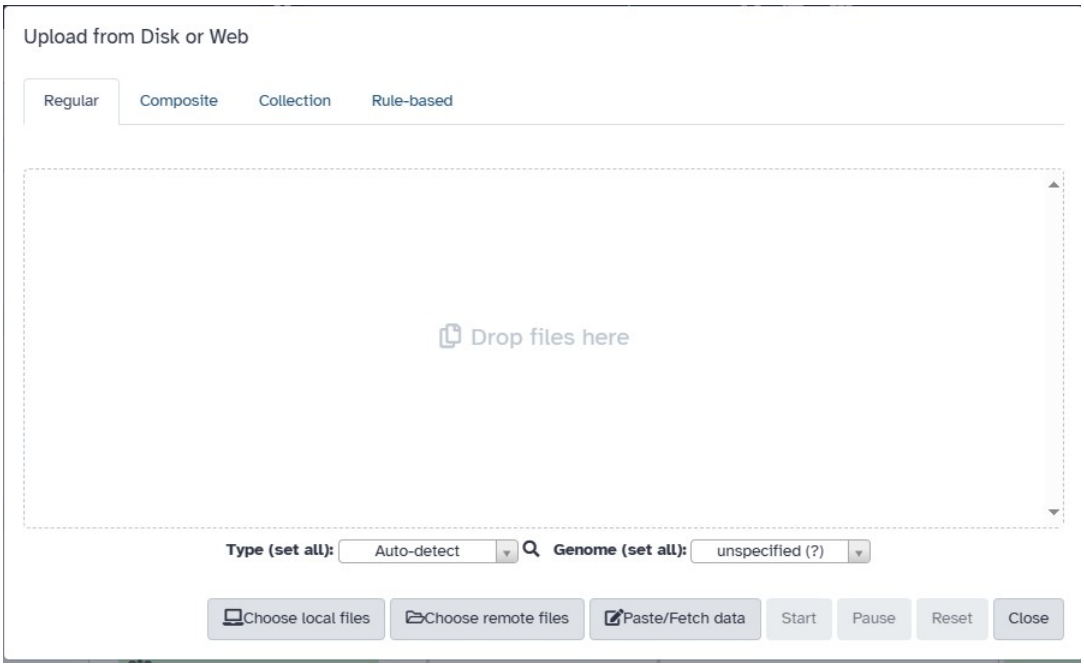

Figure 4: Upload data Exons dan SNPs

gan menghilangkan nama tersebut dengan menggunakan tools Replace Text. Parameter yang diisi seperti pada Gambar 11. Setelah data siap, selanjutnya adalah mencoba membuat pohon pilogenetiknya dengan menggunakan tools Phyogenetic reconstruction with RAxML. parameter yang perlu diedit seperti pada Gambar 12. Setelah itu kita bisa melihat hasilnya dengan membuka vizualisasi dari Best-scoring ML Tree dan hasilnya seperti pada Gambar 14 dan Gambar **??**. selanjutnya adalah mengkonversinya menjadi workflow dengan mengekstak seperti 9. Tampilan workflow seperti pada Gambar 15.

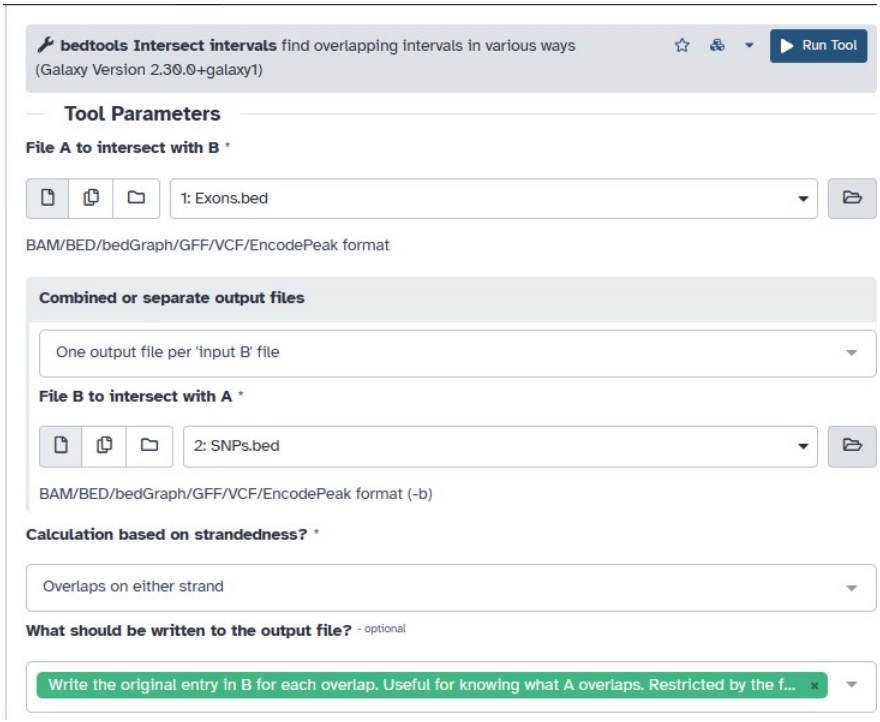

Figure 5: Parameter Intersection

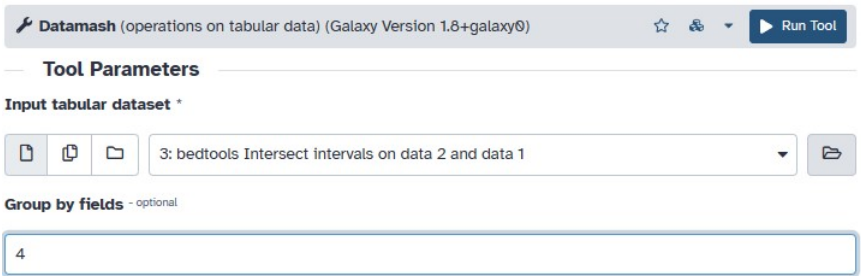

Figure 6: Parameter Datamash

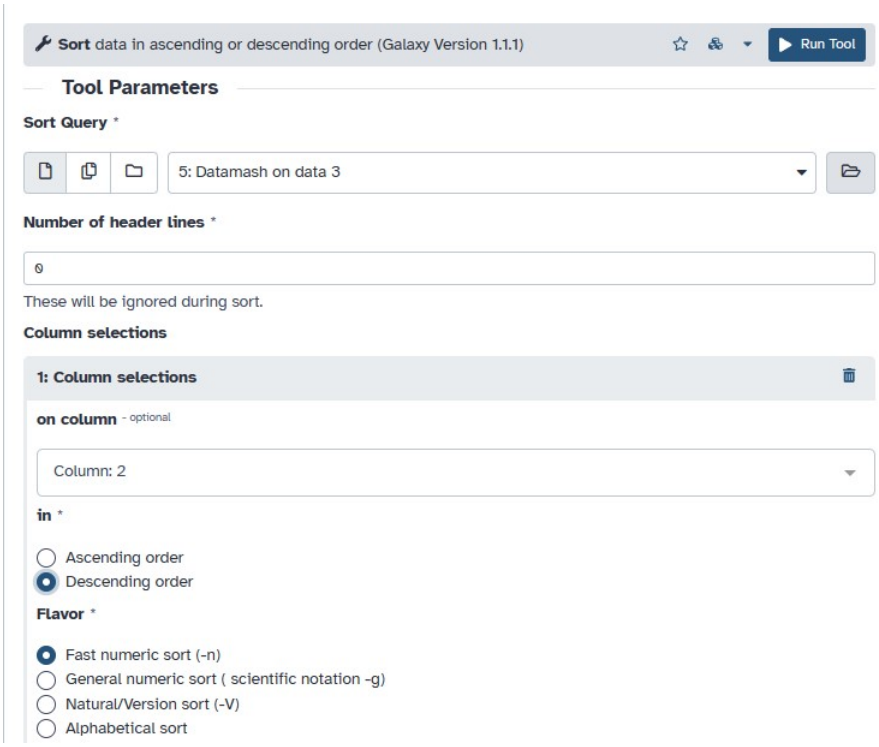

Figure 7: Sorting

Chrom	Start
ENST00000253255.7_cds_0_0_chr22_46256561_r	27
ENST00000648057.3_cds_0_0_chr22_50546244_f	26
ENST00000327423.11_cds_5_0_chr22_31712083_r	20
ENST00000302097.3_cds_0_0_chr22_22514002_r	14
ENST00000352371.5_cds_0_0_chr22_36191075_r	13
ENST00000616056.4_cds_0_0_chr22_36191075_r	13
ENST00000332987.5_cds_0_0_chr22_36191075_r	13
ENST00000262738.7_cds_34_0_chr22_46533627_r	13
ENST00000216268.6_cds_1_0_chr22_49883663_f	13
ENST00000644935.1_cds_6_0_chr22_37723185_f	11
ENST00000441493.7_cds_6_0_chr22_17817311_r	10

Figure 8: Output Sorting

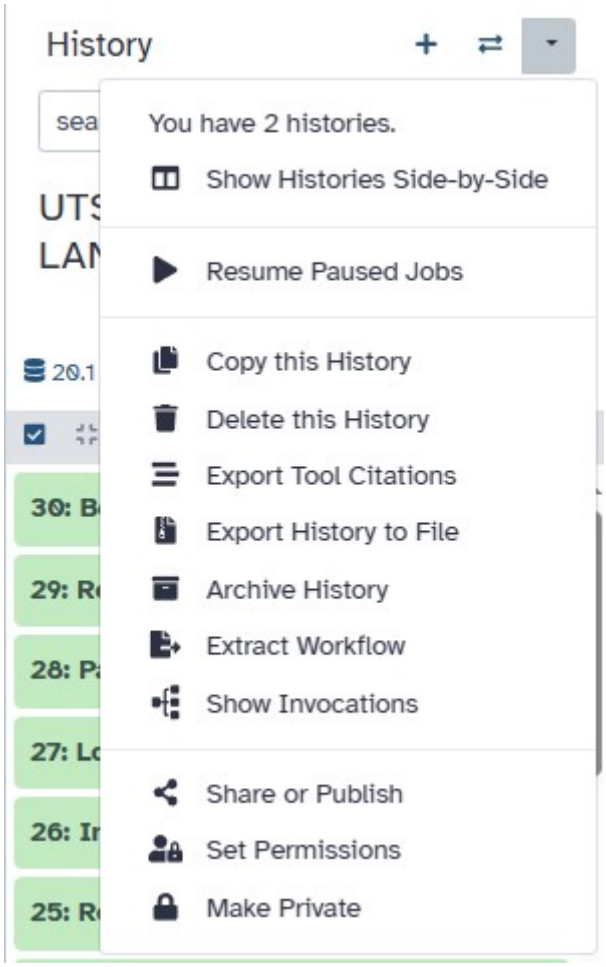

Figure 9: Control History

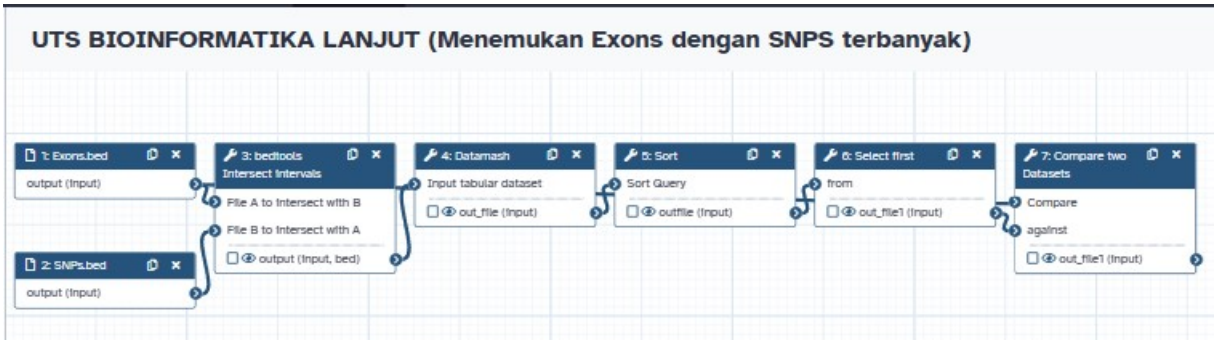

Figure 10: Workflow Menemukan Exons dengan SNPs terbanyak

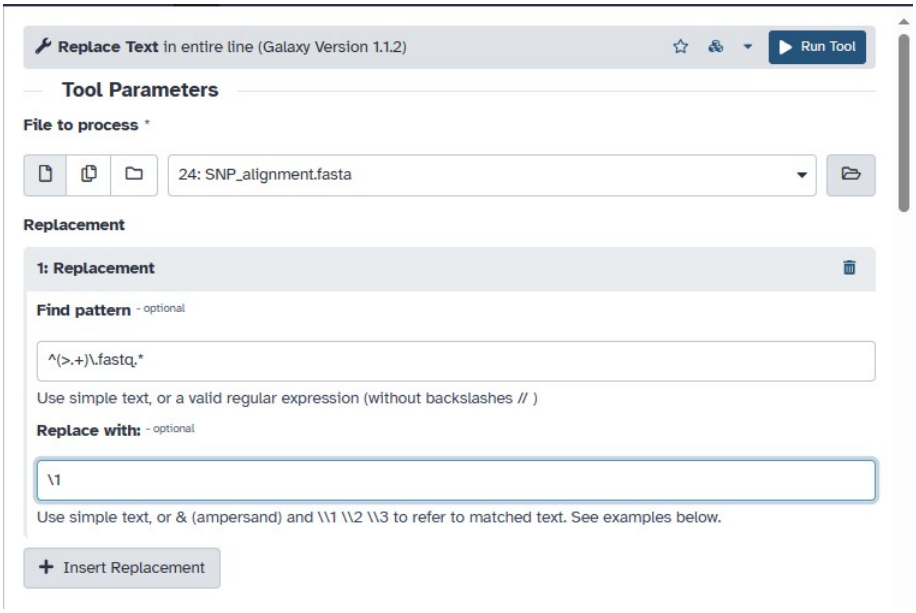

Figure 11: Replace text

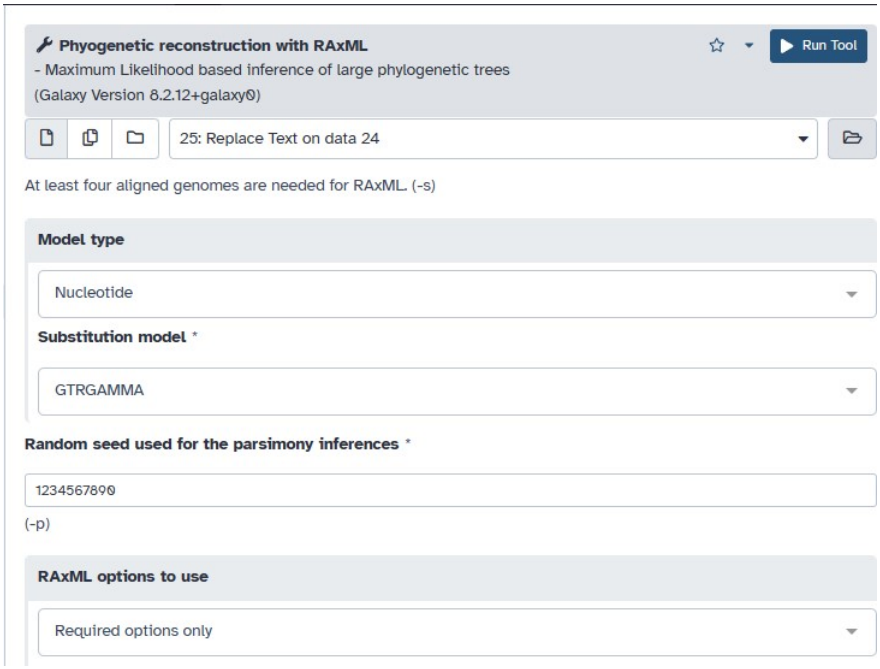

Figure 12: Phyogenetics parameter

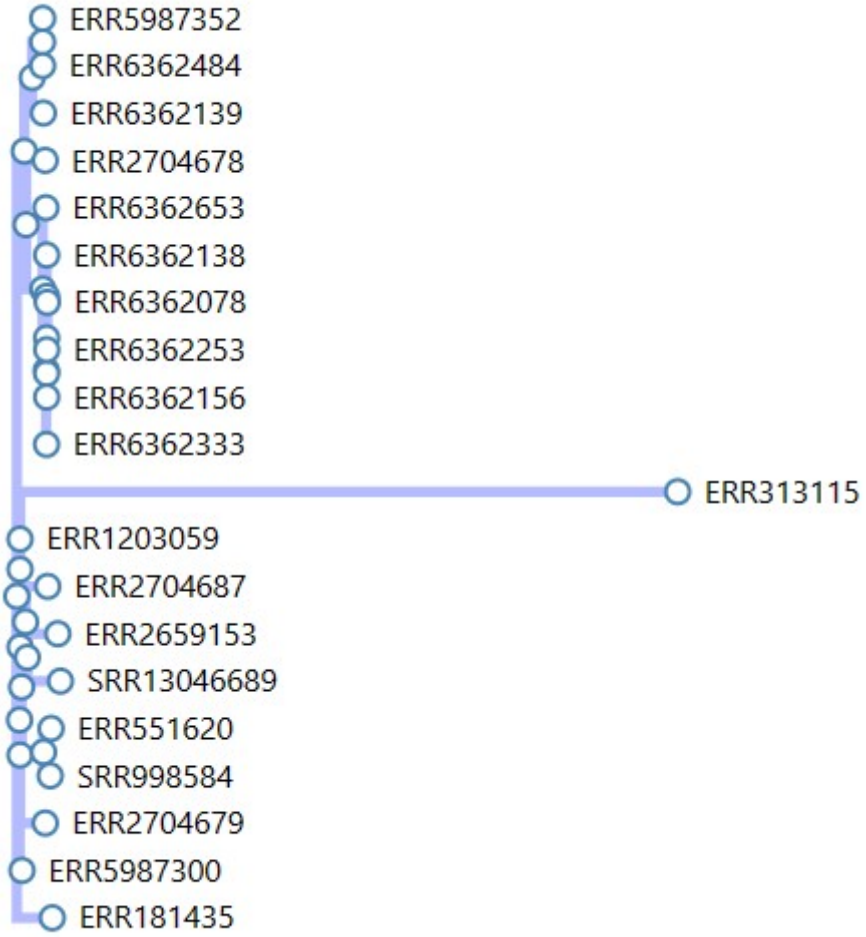

Figure 13: Pohon Pilogenetik

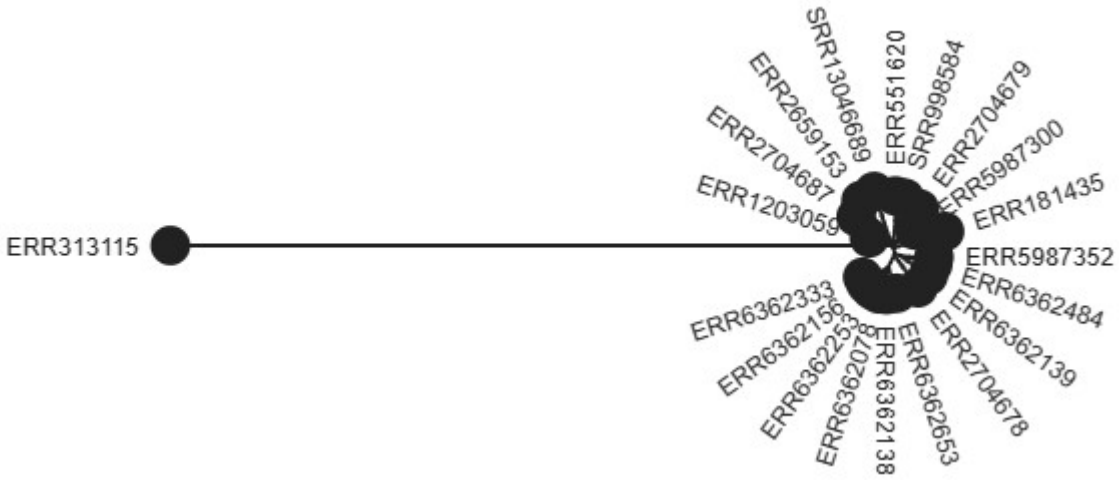

Figure 14: Pohon Pilogenetik

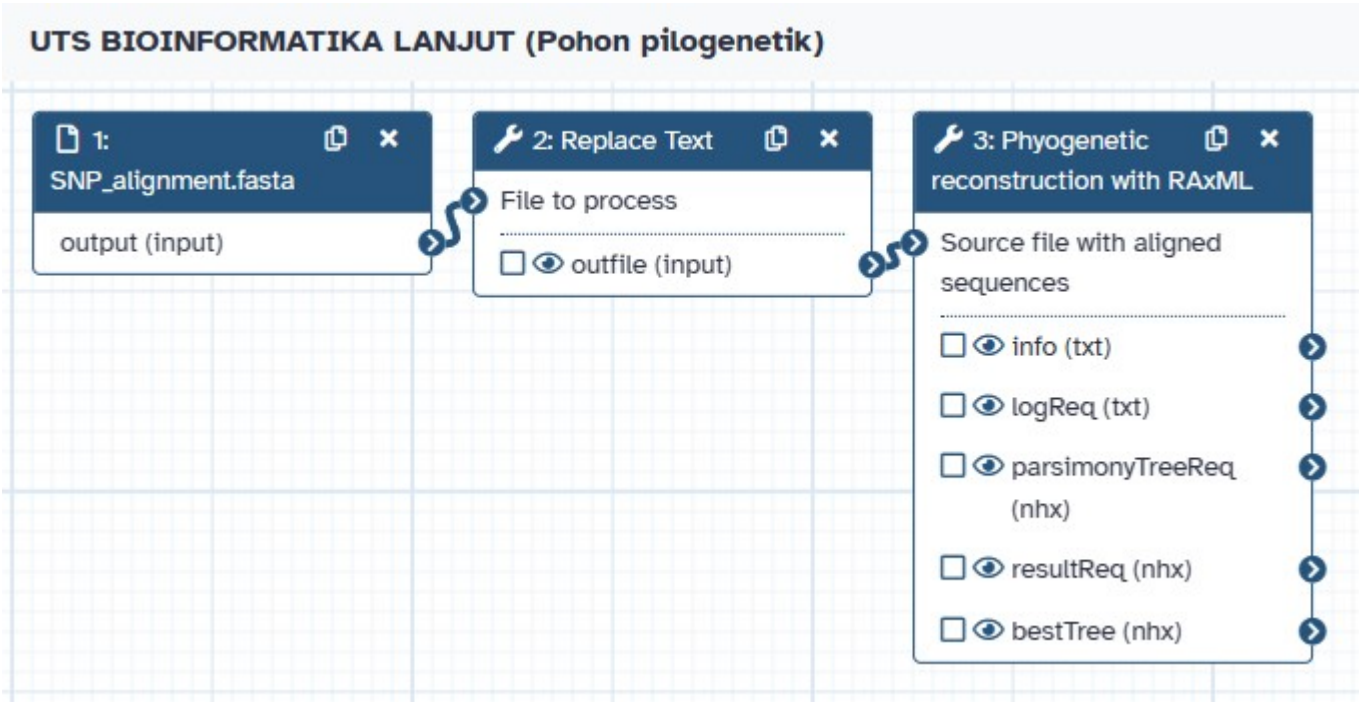

Figure 15: Workflow Pembuatan Pohon Pilogenetik