Documentos

ISSN 1677-9274
Dezembro, 2016
149

Análise de dados de RNA-Seq utilizando o Galaxy

Embrapa

ISSN 1677-9274
Dezembro, 2016

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Informática Agropecuária
Ministério da Agricultura, Pecuária e Abastecimento

Documentos 149

Análise de dados de RNA-Seq utilizando o Galaxy

Adhemar Zerlotini Neto
Leandro Carrijo Cintra

Embrapa Informática Agropecuária
Campinas, SP
2016

Embrapa Informática Agropecuária
Av. André Tosello, 209 - Barão Geraldo
Caixa Postal 6041 - 13083-886 - Campinas, SP
Fone: (19) 3211-5700
www.embrapa.br/informatica-agropecuaria
SAC: www.embrapa.br/fale-conosco/sac/

1ª edição

publicação digitalizada 2016

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Informática Agropecuária

Zerlotini Neto, Adhemar.

Análise de dados de RNA-Seq utilizando o Galaxy / Adhemar Zerlotini Neto, Leandro Carrijo Cintra.- Campinas : Embrapa Informática Agropecuária, 2016.

36 p. : il.: cm. - (Documentos / Embrapa Informática Agropecuária, ISSN 1677-9274; 149).

1. Biologia molecular computacional. 2. Pipeline. 3. Processamento distribuído. 4. Genes. 5. Workflow. I. Zerlotini Neto, Adhemar. II. Cintra, Leandro Carrijo. III. Embrapa Informática Agropecuária. IV. Título. V. Série.

CDD 572.80285

Autores

Adhemar Zerlotini Neto
Cientista da computação, Doutor em Bioinformática
Pesquisador da Embrapa Informática Agropecuária, Campinas, SP

Leandro Carrijo Cintra
Cientista da Computação, Doutor em Bioinformática
Analista da Embrapa Informática Agropecuária, Campinas, SP

Apresentação

Os equipamentos de sequenciamento de nova geração nos possibilitam mensurar a quantidade de RNA transcrito e, consequentemente, identificar modulações na expressão dos genes correlacionadas com diferentes estágios de desenvolvimento ou condições experimentais dos mais diversos organismos vivos. Esta metodologia, o RNA-Seq, é hoje a técnica mais utilizada para identificação de expressão diferencial de genes, pois possibilita ainda a obtenção da sequência completa do RNA e a identificação de diferentes formas de *splicing*.

Esses equipamentos produzem milhões de sequências pequenas, variando entre 100pb e 250pb, e o processamento desses dados geralmente requer uma considerável infraestrutura computacional. A Bioinformática é a área do conhecimento que busca superar tais desafios, por meio da congregação de métodos da computação, biologia, matemática e estatística. O Linux é o sistema operacional adotado pela comunidade de Bioinformática e diversos softwares foram desenvolvidos para análise de dados de RNA-Seq neste sistema.

Cientistas de áreas como a Biologia enfrentam grandes dificuldades no processamento desses dados, uma vez não possuem treinamento formal na utilização do sistema operacional Linux, bem como em paralelização de processos em clusters de computadores.

Neste documento, serão apresentados métodos computacionais para facilitar o processo de análise de dados de RNA-Seq, por meio de ferramentas acessíveis via navegadores. Esta metodologia possibilita o processamento distribuído e o compartilhamento de grandes volumes de dados de RNA-Seq, com o objetivo de efetivamente identificarmos as diferenças de expressão de genes para elucidar mecanismos biológicos ligados à produtividade e a doenças.

Silvia Maria Fonseca Silveira Massruhá
Chefe-geral
Embrapa Informática Agropecuária

Sumário

Análise de dados de RNA-Seq utilizando o Galaxy

Adhemar Zerlotini Neto
Leandro Carrijo Cintra

1. Introdução

1.1. Análise de dados de RNA-Seq

A tecnologia de sequenciamento de RNA, RNA-Seq, possibilita a identificação de genes e transcritos diferencialmente expressos entre amostras biológicas. O volume e a complexidade deste tipo de dados exige programas eficientes e escaláveis. Neste artigo, serão apresentados programas de código fonte aberto que nos permitem analisar grandes conjuntos de dados de RNA-Seq para identificar novos genes, formas de splicing alternativo e comparar genes e transcritos de duas ou mais condições experimentais. Os programas utilizados foram desenvolvidos pelo grupo de pesquisa do Dr. Cole Trapnell, do Center for Computational Biology, na Johns Hopkins University. Este grupo tem publicado regularmente artigos científicos e textos online relativos às ferramentas para análise de dados de RNA-Seq (TRAPNELL et al., 2012).

A análise de dados de RNA-Seq que será apresentada consiste no mapeamento das sequências de RNA em um genoma de referência, sua quantificação e aplicação de testes estatísticos para determinação de expressão

diferencial entre as amostras. As ferramentas utilizadas para esta análise, estão descritas na figura Figura 1.

Como especificado no artigo (TRAPNELL et al., 2012), as ferramentas devem ser utilizadas em um terminal do Linux em modo texto. O Galaxy possibilita utilizá-las diretamente na interface gráfica por meio de um navegador web.

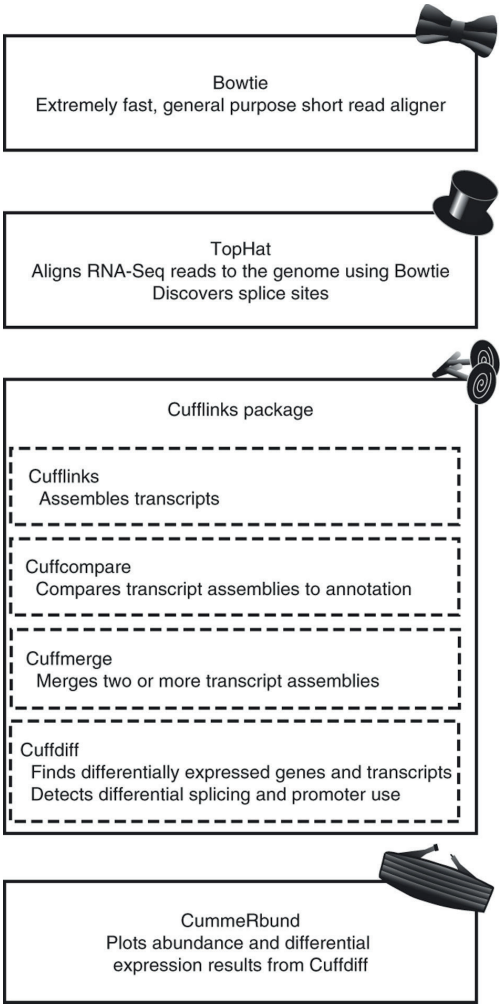

Figura 1. Ferramentas de análise de dados de RNA-Seq.

Fonte: Trapnell et al. (2012).

1.2. Galaxy

O Galaxy (GIARDINE et al., 2005) é uma plataforma web de código fonte aberto para pesquisas biomédicas que analisam grandes volumes de dados. Seja utilizando o servidor público (usegalaxy.org) ou instalando sua própria instância local (galaxyproject.org), você pode executar, reproduzir e compartilhar análises de dados.

O objetivo do Galaxy é tornar a análise de grandes volumes de dados mais acessível, transparente e reprodutível, por meio de um ambiente web, em que os usuários podem executar análises computacionais complexas e ter todos os detalhes de cada etapa da execução registrados para posterior inspeção, publicação ou reutilização.

1.3. Instância do Galaxy no LMB

A instância do Galaxy para análises de bioinformática, instalada no servidor do Laboratório Multiusuário de Bioinformática da Embrapa (LMB), pode ser acessada no seguinte endereço web: https://www.lmb.cnptia.embrapa.br/galaxy. O acesso é restrito por senha e a criação de um usuário de acesso deve ser solicitada pelo formulário disponível em: https://www.lmb.cnptia.embrapa.br/web/lmb/politicas-de-uso.

2. Caso de uso

A análise de dados de RNA-Seq que será demonstrada neste documento foi extraída de um exercício proposto por um membro da equipe de desenvolvimento do Galaxy em: https://usegalaxy.org/u/jeremy/p/galaxy-rna-seq-analysis-exercise. Este exercício introduz as ferramentas necessárias para a execução completa de uma análise de dados de RNA-Seq utilizando-se um genoma de referência e conceitos gerais da plataforma Galaxy.

2.1. Conjunto de dados

Os arquivos de entrada são amostras extraídas do projeto Illumina

BodyMap 2.0 (http://www.ensembl.info/blog/2011/05/24/human-bodymap-2--0-data-from-illumina/). São arquivos no formato fastq (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847217/) que contém pares de sequências de 50 nucleotídeos. Essas amostras contêm sequências localizadas em uma região de 500 nucleotídeos do cromossomo humano 19, provenientes de dois tecidos: cérebro (*brain*) e glândula suprarrenal (*adrenal*).

Existe ainda, neste conjunto de dados, um arquivo que contém características dos genes humanos contidos no cromossomo 19, que será utilizado em etapas posteriores da análise.

Os arquivos se chamam: adrenal_1.fastq, adrenal_2.fastq, brain_1.fastq, brain_2.fastq e 'iGenomes UCSC hg19, chr19 gene annotation'. Todos os 5 arquivos devem ser baixados do link do exercício de RNA-Seq para posterior processamento, utilizando-se o link representado por um disquete (Figura 2).

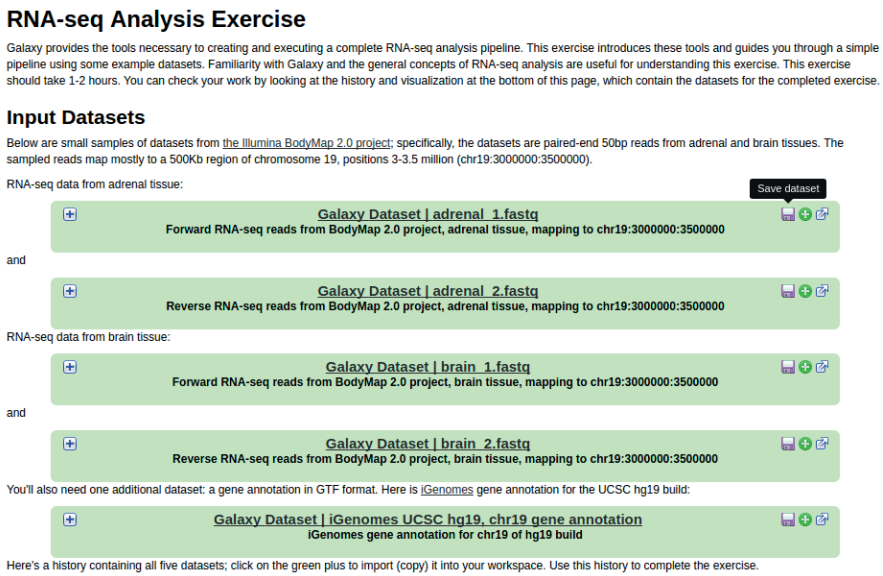

Figura 2. Fonte de dados e descrição da análise que será realizada.

2.2. Página inicial do Galaxy

A página inicial do Galaxy é composta de uma barra superior e três painéis, conforme a Figura 3. A barra superior contém um menu que possibilita acesso às análises (*Analyze Data*), *workflows*, dados compartilhados (*Shared Data*), visualizadores genômicos (*Visualization*), administração (*Admin*), ajuda (*Help*) e configurações de usuário (*User*).

O painel da esquerda, chamado Tools (Ferramentas), contém uma ferramenta de busca de programas de bioinformática (*search tools*) e links para os programas separados em categorias. O ícone de upload, no canto superior direito desse painel, possibilita carregamento de arquivos. Arquivos carregados no Galaxy serão exibidos no painel da direita.

O painel da direita, chamado *History* (Histórico), possui uma ferramenta de busca nos itens do histórico (*search datasets*) e uma lista de arquivos que foram carregados ou gerados por meio da execução de ferramentas. Ao clicar no texto *Unnamed History* é possível dar um nome para o histórico atual. Históricos adicionais podem ser criados a partir do link representado por uma engrenagem no canto superior direito desse painel, além de outras opções relativas aos históricos.

Por último, o painel central que é utilizado para visualização dos formulários de opções dos programas contidos na barra de ferramentas e do conteúdo e informações dos arquivos contidos na barra histórico.

Figura 3. Página inicial do Galaxy implementado no Laboratório Multiusuário de Bioinformática.

2.3. Carregamento de arquivos

No canto superior direito do painel de ferramentas (Figura 4) há um ícone de carregamento de arquivos (*upload*). Esta ferramenta também pode ser acessada por meio do link *Get Data* > *Upload File*.

Ao abrir a ferramenta de carregamento de arquivos será disponibilizada uma janela, na qual é possível arrastar arquivos de um gerenciador de arquivos diretamente para ela (Figura 5). Nesta janela, arquivos podem ser submetidos de três formas: arquivos do seu computador (*Choose local*

Figura 4. Link para carregamento de arquivos.

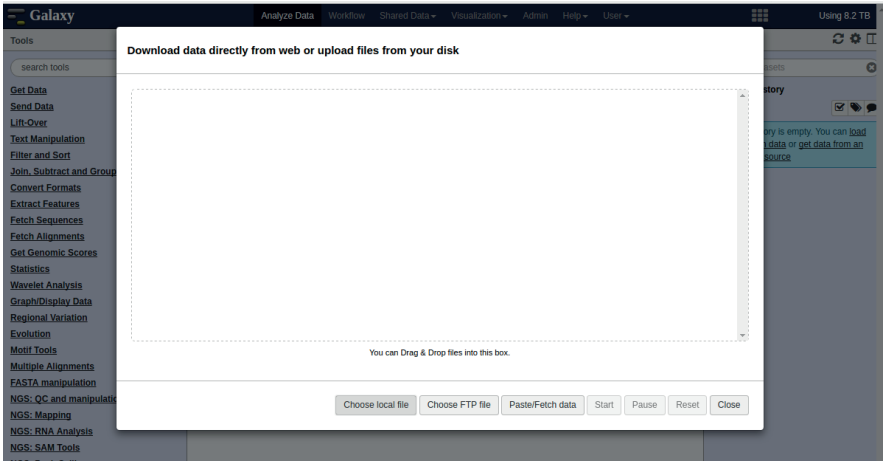

Figura 5. Arquivos podem ser carregados do próprio computador, de um link de FTP, de uma URL ou colado em uma caixa de textos.

file), arquivos enviados previamente por um servidor de FTP (*Choose FTP file*) e colar/digitar conteúdo ou URL de um arquivo (*Paste/Fetch data*).

A primeira opção é, geralmente, a mais utilizada, porém existe uma limitação de tamanho máximo de aproximadamente 2Gb. Quando arquivos excedem este limite, precisam ser enviados previamente para um servidor FTP disponibilizado pelo administrador local do Galaxy, ou seja, para enviar arquivos por este método é necessário contatar o administrador e obter instruções específicas de conexão no servidor FTP.

A opção de colar o conteúdo ou uma URL de um arquivo é particularmente interessante. Esta opção facilita a inserção de dados manualmente, como uma lista de identificadores ou um conjunto pequeno de sequências (*primers*, adaptadores, …). Ela possibilita ainda carregar arquivos diretamente de um endereço da web, facilitando baixar arquivos de sequência de sites como o NCBI, Kegg, Uniprot, e outros.

Os arquivos fastq e gtf baixados anteriormente deverão ser carregados utilizando o botão *Choose local file* ou arrastando os mesmos para a janela (Figura 6)

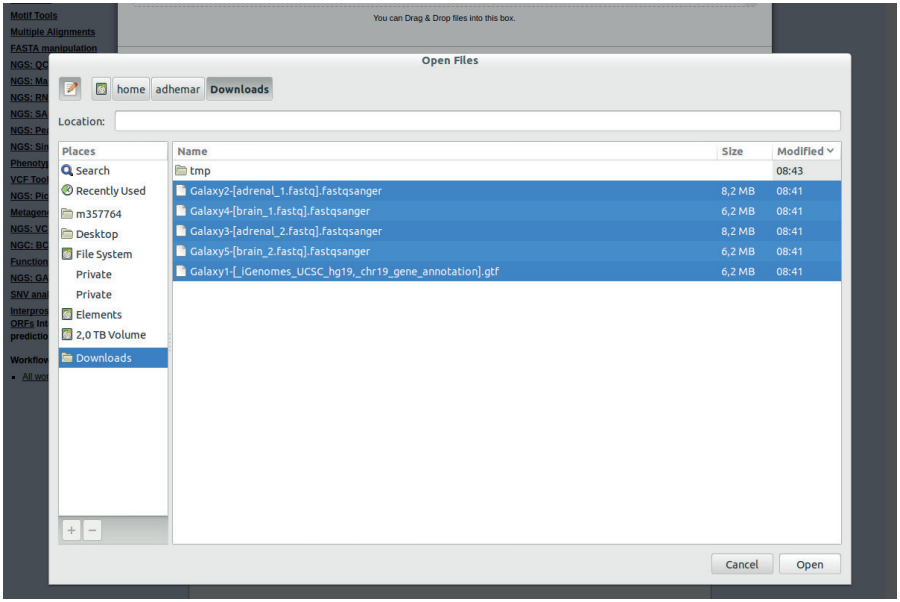

Figura 6. Para a análise de dados de RNA-Seq deste documento, serão necessários os arquivos selecionados (azul).

Antes de clicar no botão *Start* para início do carregamento dos arquivos é muito importante definir o seu formato utilizando a coluna *Type*. O sistema de carregamento de arquivos tentará detectar o formato automaticamente, porém existem diferentes tipos de arquivos fastq e este sistema irá selecionar o formato genérico fastq. As ferramentas de bioinformática disponíveis no Galaxy exigem que os arquivos de sequência estejam em um formato fastq específico conhecido como fastqsanger.

Arquivos no formato fastq (https://en.wikipedia.org/wiki/FASTQ_format) são constituídos de uma ou mais entradas compostas por 4 linhas, sendo elas: o identificador da sequência e uma descrição; a sequência de nucleotídeos; o sinal +; e uma sequência de caracteres que representa a qualidade de cada nucleotídeo. A principal variação existente entre arquivos fastq está na linha de qualidade da sequência, que pode ser computada subtraindo-se 33 do valor ASCII de cada caractere (tipo Sanger) ou 64 (tipo Solexa). Os arquivos produzidos por sequenciadores Illumina, como os utilizados neste documento, possuem valores de qualidade Sanger e para informarmos ao Galaxy dessa característica, é necessário selecionar a opção fastqsanger na coluna *type* de cada um dos arquivos de sequência conforme a figura Figura 7.

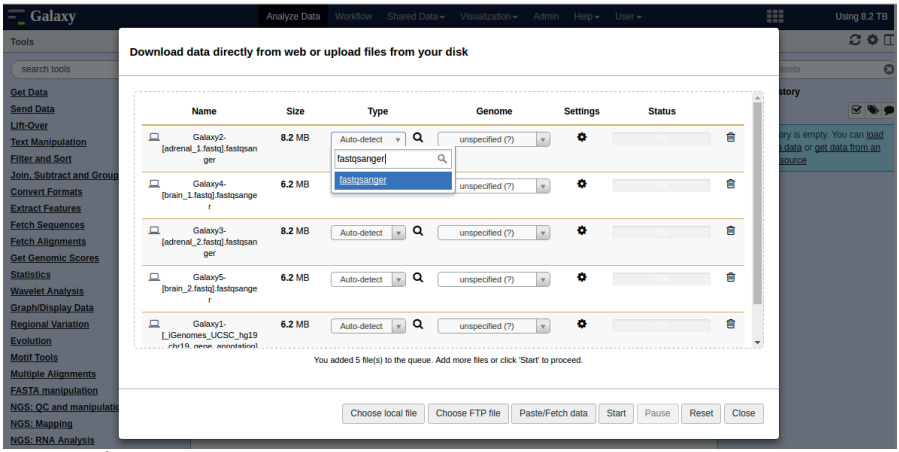

Figura 7. É imprescindível selecionar o tipo 'fastqsanger' no campo '*type*' nos arquivos com extensão .fastq uma vez que as ferramentas reconhecem apenas este formato.

Existem também variações nos formatos de arquivos de anotação, gtf (http://www.ensembl.org/info/website/upload/gff.html), e, por segurança,

devemos selecionar a opção gtf na coluna type do arquivo '*iGenomes UCSC hg19, chr19 gene annotation*' conforme Figura 8.

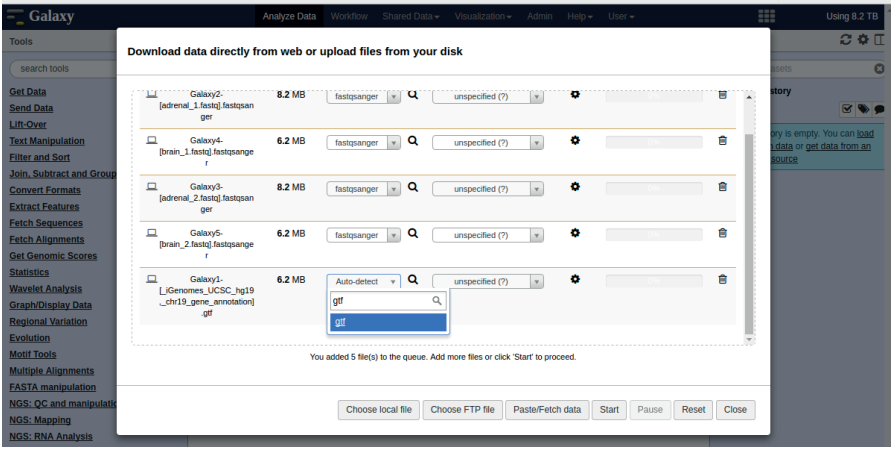

Figura 8. É imprescindível selecionar o tipo 'gtf' no campo '*type*' nos arquivos com extensão .gtf.

Ao clicar no botão *Start* os arquivos serão carregados no Galaxy e aparecerão entradas no histórico (Figura 9). Inicialmente elas aparecerão na cor cinza a qual indica que o comando para carga dos dados foi disparado. Logo que entrar em execução, a entrada mudará sua cor para amarelo. Ao fim da execução a cor mudará para verde em caso de sucesso, ou vermelho em caso de erro. Cada uma dessas entradas podem ser expan-

Figura 9. Os arquivos carregados aparecem no painel '*History*'. Os itens deste painel tem cores diferentes indicando os estados: espera (cinza), execução (amarelo), concluído (verde) e erro (vermelho).

didas com um clique para obtenção de informações adicionais como pré-visualização do conteúdo, detalhes da execução e parâmetros utilizados e relatório de erros.

2.4. Mapeamento no genoma de referência

A ferramenta utilizada para mapear sequências de RNA-Seq em um genoma de referência se chama Tophat (TRAPNELL et al., 2009). No painel de ferramentas (*Tools*) podemos localizar o Tophat usando a busca ou abrir a categoria *NGS: RNA Analysis*. Os parâmetros da ferramenta selecionada serão exibidos no painel central conforme a Figura 10.

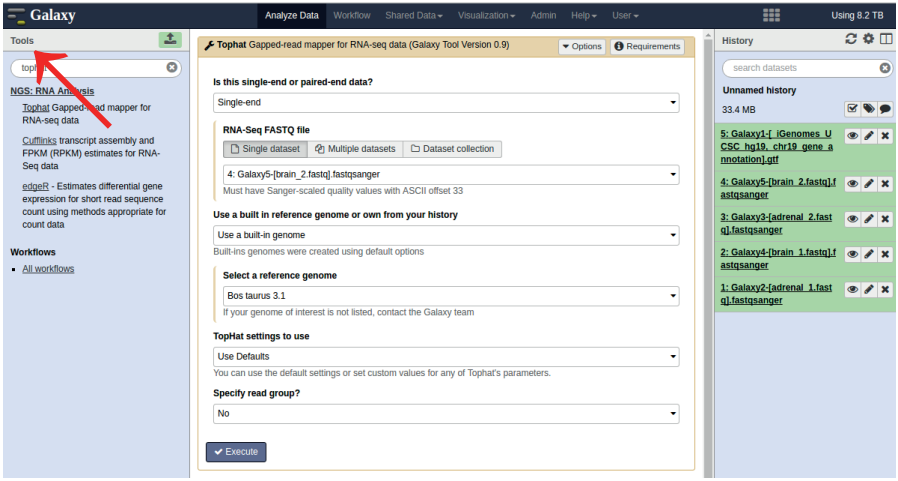

Figura 10. O painel '*Tools*' contém um campo chamado '*search tools*' que possibilita localizar ferramentas.

O primeiro parâmetro a ser configurado no Tophat é o tipo do conjunto de dados: *single-end* ou *paired-end*. Os arquivos carregados são *paired-end* onde adrenal_1.fastq e brain_1.fastq são sequências 5' (*forward*) e adrenal_2.fastq e brain_2.fastq são sequências 3' (*reverse*). Inicialmente iremos mapear a amostra *adrenal* indicando os arquivos adrenal_1.fastq em *RNA-Seq FASTQ file, forward reads* e adrenal_2.fastq em *RNA-Seq FASTQ file*, *reverse reads* conforme Figura 11. O texto original do exercício nos informa que a distância média entre os pares de sequências é 110 nucleotídeos e essa informação deve ser informada no campo *Mean Inner*

Distance between Mate Pairs.

Todas as outras opções serão mantidas como padrão exceto o genoma de referência que deverá ser selecionado no campo *Use a built in reference genome or own from your history*. Como não foi feito o carregamento de um genoma de referência, devemos selecionar a opção *Use a built in genome*. Será exibido um novo campo chamado *Select* a reference genome no qual deve ser selecionada a opção *Human (Homo sapiens)*: hg19. Para iniciar a execução desta ferramenta basta clicar no botão *Execute* ao final do painel central. Uma vez concluída a execução, serão gerados 5 novos arquivos no histórico. O Galaxy utiliza um formato específico para nomear arquivos de resultados de análise composto pelo nome do programa, os números dos conjuntos de dados e um título em caso de múltiplos resultados. Os arquivos gerados por meio da execução do Tophat terão nomes

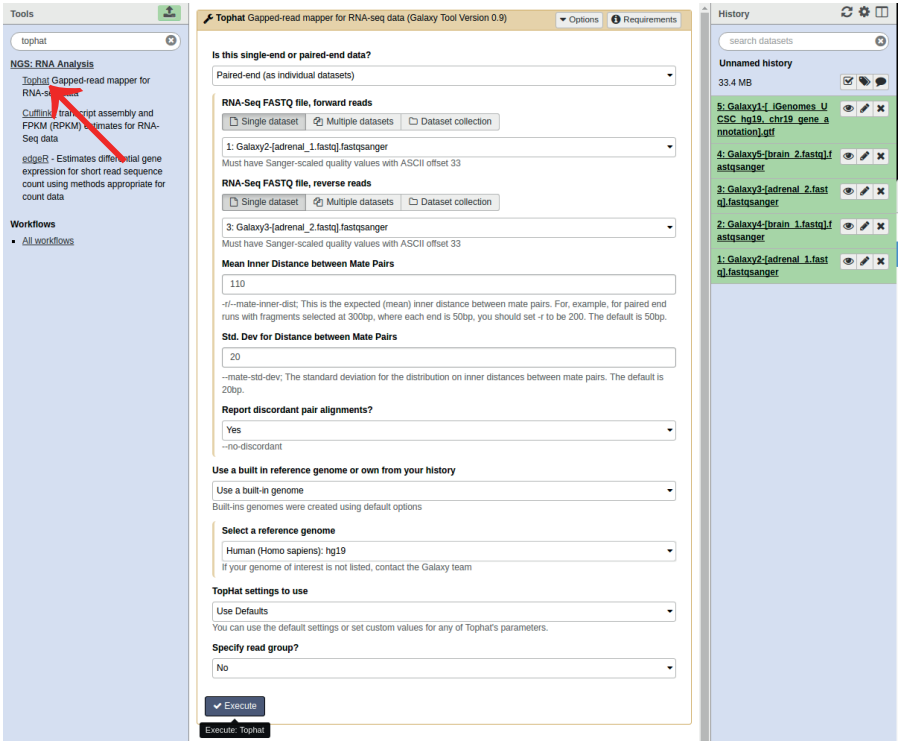

Figura 11. O Tophat é uma ferramenta para mapear sequências de RNA-Seq em genomas de referência.

iniciados por *Tophat on data 3 and data 1: título*, onde título pode ser *align_summary, insertions, deletions, splice_junctions e accepted_hits*.

Para visualizar o conteúdo destes resultados, basta clicar no ícone representado por um olho no canto superior direito de cada arquivo conforme Figura 12. O arquivo *align_summary*, por exemplo, é um relatório do número de sequências mapeadas e o Dentre os arquivos produzidos pelo Tophat, o mais importante se chama accepted_hits. Este é um arquivo no formato .bam (https://samtools.github.io/hts-specs/SAMv1.pdf) que contém todas as informações de alinhamento das sequências dos arquivos fastq no genoma de referência.

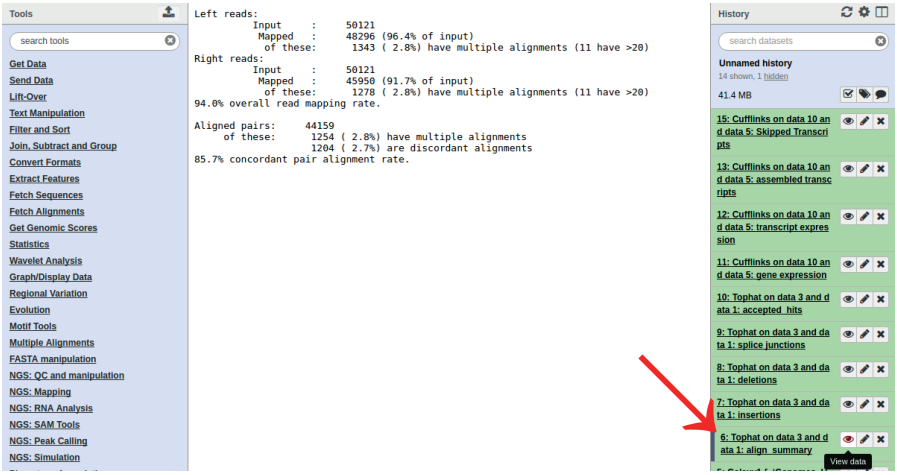

Figura 12. O arquivo '*align summary*', resultante do processamento do Tophat, apresenta um resumo do número de *reads* mapeadas.

2.5. Identificação de genes e transcritos

A ferramenta utilizada para efetuar a identificação de genes e transcritos por meio da evidência de alinhamento de sequências de RNA-Seq mapeadas se chama *Cufflinks* (TRAPNELL et al., 2010). No painel de ferramentas (*Tools*) podemos localizar o *Cufflinks* usando a busca ou abrir a categoria *NGS: RNA Analysis*. Os parâmetros da ferramenta selecionada serão exibidos no painel central conforme Figura 13.

O primeiro parâmetro a ser configurado no *Cufflinks*, *SAM or BAM file of*

aligned RNA-Seq reads, é o arquivo de entrada no formato SAM ou BAM. Como efetuamos o mapeamento da amostra adrenal utilizando o Tophat, temos apenas uma opção a escolher, *Tophat on data 3 and data 1: accepted hits*. Todas as outras opções deverão ser mantidas como padrão exceto o parâmetro *Reference Annotation*. Neste campo deve ser selecionado o arquivo de anotação '*iGenomes UCSC hg19, chr19 gene annotation*', para que o *Cufflinks* atribua os mesmos identificadores para os genes encontrados por evidência de sequências de RNA-Seq mapeadas no genoma.

Para iniciar a execução desta ferramenta basta clicar no botão *Execute* ao final do painel central. Uma vez concluída a execução, serão gerados 5 novos arquivos no histórico. Como mencionado anteriormente, o Galaxy

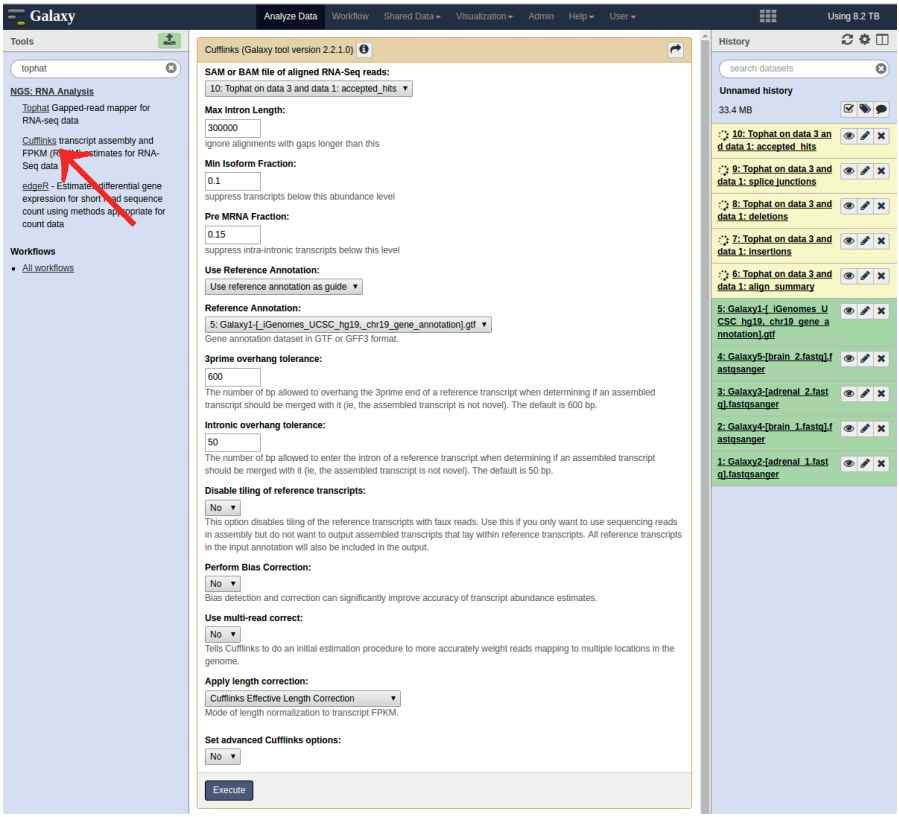

Figura 13. O *Cufflinks* é uma ferramenta para identificação de genes e transcritos por meio da localização das sequências de RNA-Seq no genoma de referência.

irá criar arquivos com nomes específicos, iniciados por *Cufflinks on data 20 and data 5: título*, onde título pode ser *Skipped Transcripts, assembled transcripts, transcript expression e gene expression*.

Para visualizar o conteúdo destes resultados, basta clicar no ícone representado por um olho no canto superior direito de cada arquivo conforme Figura 14. Dentre os arquivos produzidos, o mais importante se chama *assembled transcripts*. Este é um arquivo no formato .gtf que contém as informações dos transcritos encontrados.

Para visualizar o conteúdo destes resultados, basta clicar no ícone representado por um olho no canto superior direito de cada arquivo conforme Figura 14. Dentre os arquivos produzidos, o mais importante se chama assembled transcripts. Este é um arquivo no formato .gtf que contém as informações dos transcritos encontrados.

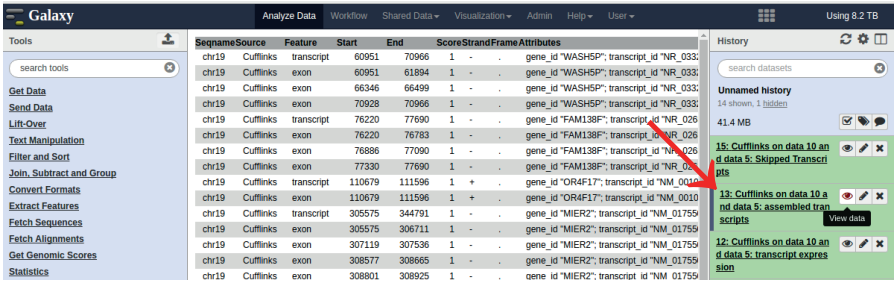

Figura 14. O arquivo '*assembled transcripts*', produzido pelo Cufflinks, contém a localização dos transcritos e seus exons.

2.6. Organização do histórico

Os nomes automaticamente gerados pelo Galaxy, apesar de indicarem com precisão a ferramenta e o conjunto de dados que foram utilizados, não contém o nome original da amostra. A medida que executarmos essas ferramentas para amostras adicionais, iremos nos deparar com um histórico poluído visualmente, que pode nos levar a cometer erros no agrupamento de amostras em etapas posteriores da análise.

Uma solução para este tipo de problema é renomear os arquivos de resultados. Para isso basta clicar no ícone representado por um lápis

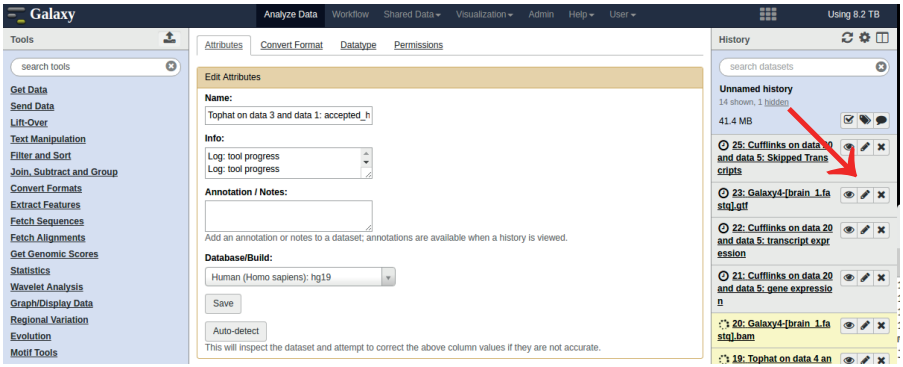

Figura 15. Cada item do painel '*History*' contém um ícone de um lápis que permite alterar informações ou formatos e incluir anotações. Um dos principais arquivos gerados pelo Tophat se chama '*tophat on data 3 and data 1: accepted hits*'. Recomenda-se renomear este arquivo para conter o nome da amostra (eg. adrenal.bam).

contido em cada entrada do painel histórico (*History*). Como informado anteriormente, o principal resultado do Tophat é o arquivo *accepted_hits* e do Cufflinks, *assembled transcripts*. Devemos editar os atributos destes arquivos para que os nomes contenham o nome da amostra e sua identificação seja imediata ao buscá-los no painel histórico (*History*). Desta forma, o arquivo do Tophat, *accepted_hits*, passaria a se chamar adrenal.bam (Figura 15) e o arquivo do Cufflinks, *assembled transcripts*, adrenal.gtf (Figura 16).

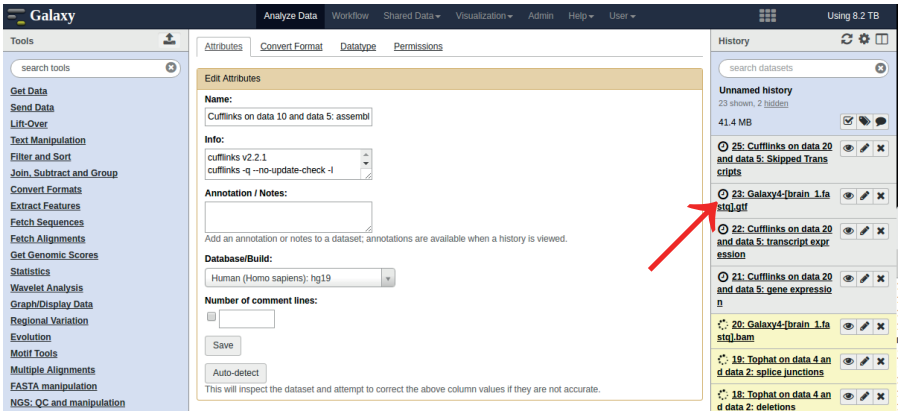

Figura 16. Um dos principais arquivos gerados pelo Cufflinks se chama '*cufflinks on data 10 and data 5: assembled transcripts*'. Recomenda-se renomear este arquivo para conter o nome da amostra (eg. adrenal.gtf).

2.7. *Workflows*

Em um experimento tradicional de análise de dados de RNA-Seq são analisadas várias bibliotecas. Cada uma dessas bibliotecas precisa ser mapeada no genoma de referência (sessão 2.4) e, posteriormente, utilizar os dados de mapeamento para identificação dos transcritos (sessão 2.5). O Galaxy possibilita a execução automatizada dessas etapas da análise para cada uma das bibliotecas, garantindo que sejam utilizados exatamente os mesmos parâmetros.

Os *workflows* podem ser acessados a partir do item '*Workflow*' na barra de menu superior, sendo possível gerenciá-los: criar novos, alterar, excluir ou importar. Para criar um *workflow*, o Galaxy oferece uma interface gráfica que possibilita inserir as diferentes ferramentas que compõem determinada análise, e conectá-las. Também é possível criar um novo *workflow* a partir de um histórico.

Para criar um novo *workflow* a partir de um histórico basta clicar na engrenagem no canto superior do painel histórico (*History*) e selecionar a opção '*Extract workflow*' conforme a Figura 17. No painel principal serão exibidos todos o itens do histórico, possibilitando eliminar arquivos de entrada ou programas antes de criar o *workflow* (Figura 18). Existe um campo chamado '*Workflow name*' que nos possibilita atribuir um novo nome ao *workflow* que será criado. Iremos colocar o nome *TophatAndCufflinks* para identificar o *workflow* que contém as análises executadas no nosso histórico (Tophat e Cufflinks). Para prosseguir, devemos clicar no botão '*Create workflow*'.

No painel central será exibida uma tela de resultado da criação do *workflow* (Figura 19) com duas opções 'edit' e 'run', para editar ou executar o *workflow* recém criado, respectivamente. Antes de executar, iremos editar o *workflow* para nos certificarmos das etapas e parâmetros desta análise. Para editar o *workflow* basta clicar no link '*edit*'.

Na tela de edição de *workflows* serão exibidas caixas representando arquivos de entrada ou ferramentas conforme o que havia sido executado no painel *History* (histórico) Figura 20 ilustra tal situação. As caixas identificadas por *Input dataset* representam os arquivos de entrada e são exibidas 5 caixas representando os 5 arquivos que foram carregados (sessão 2.3). As análises que efetuamos (sessões 2.4 e 2.5) processaram apenas 3 arqui-

Figura 17. No topo do painel '*History*' há um ícone de uma engrenagem que permite acessar configurações deste painel. Dentre as opções disponíveis, a '*Extract Workflow*' permite criar um passo a passo das análises executadas.

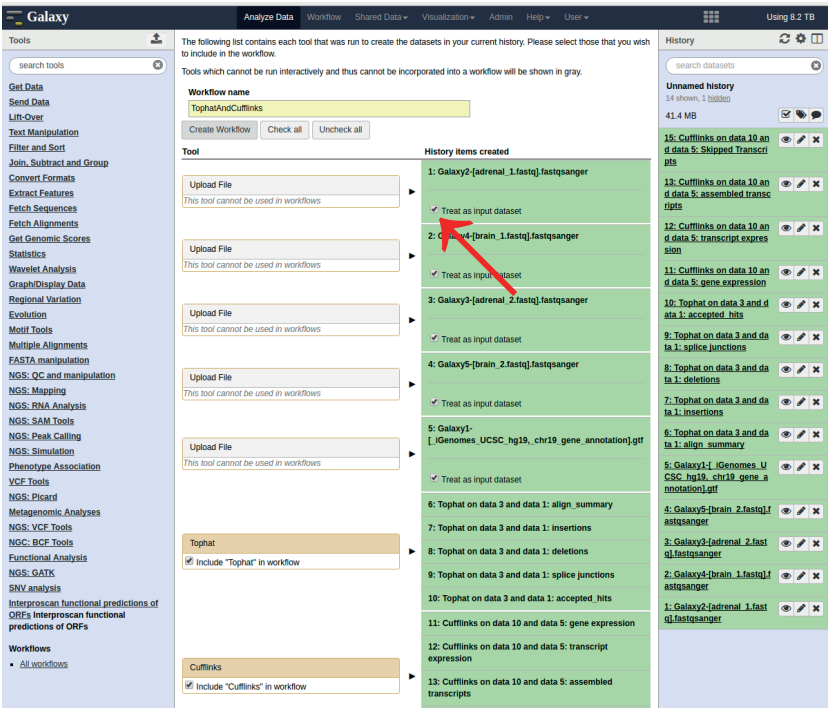

Figura 18. Ao criar um '*workflow*' a partir dos itens do painel '*History*' é possível selecionar itens.

vos (‘adrenal_1.fastq’, ‘adrenal_2.fastq’ e ‘*iGenomes UCSC hg19, chr19 gene annotation*’), desta forma, apenas 3 caixas estão conectadas nas caixas de ferramentas Tophat e Cufflinks. Em outras palavras, para efetuar as análises de mapeamento e identificação de transcritos precisaremos apenas de 3 caixas *Input dataset* e, portanto, podemos eliminar as caixas

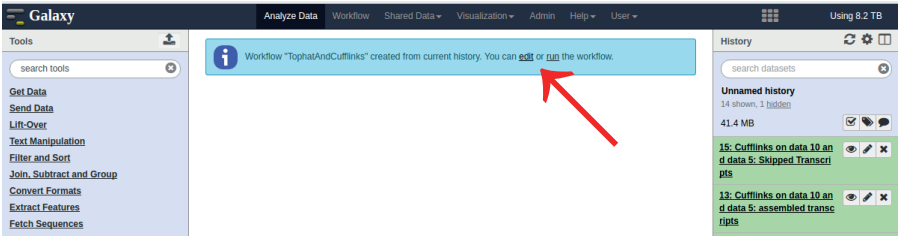

Figura 19. Após a criação de um ‘*workflow*’ por meio da ferramenta ‘*Extract Workflow*’ podemos editá-lo ou executá-lo.

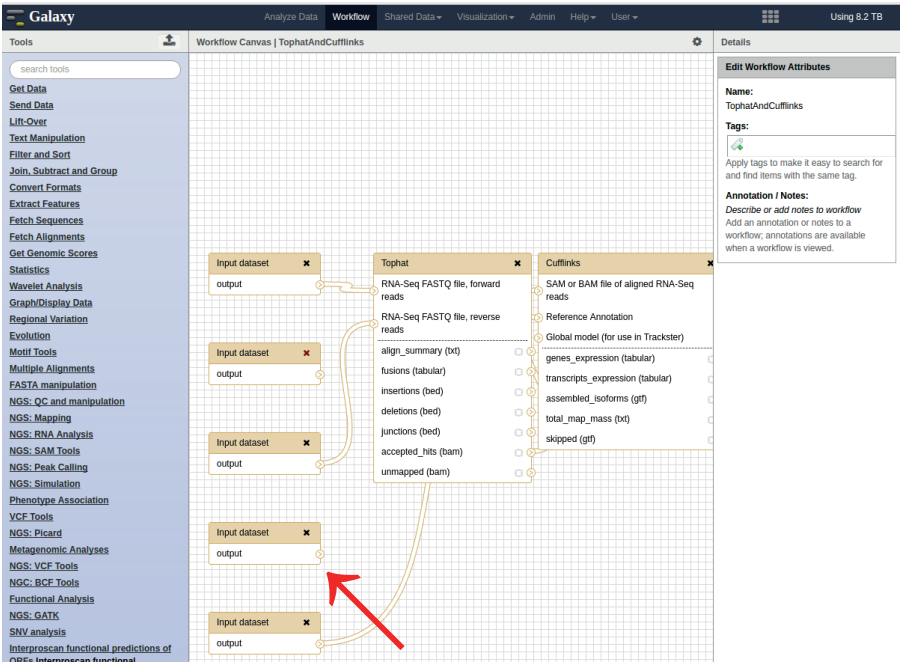

Figura 20. A ferramenta gráfica de edição de ‘*workflows*’ possibilita inserir ou excluir itens, bem como acessar seus parâmetros. Neste exemplo, gerado automaticamente por meio da função ‘*Extract Workflow*’, pode-se perceber que existem itens que não estão ligados e, portanto, podem ser excluídos.

que não estão ligadas a lugar algum. Dessa forma, teremos um *workflow* que representa a análise de mapeamento e identificação de transcritos que é constituída de 3 arquivos de entrada e as ferramentas Tophat e Cufflinks.

Ao clicar em qualquer caixa do *workflow* temos acesso aos parâmetros no painel do lado direito chamado '*Details*' (Detalhes) conforme a Figura 21. Se selecionarmos a ferramenta *Tophat*, por exemplo, podemos observar que os mesmos parâmetros utilizados anteriormente (sessão 2.4) já estão selecionados.

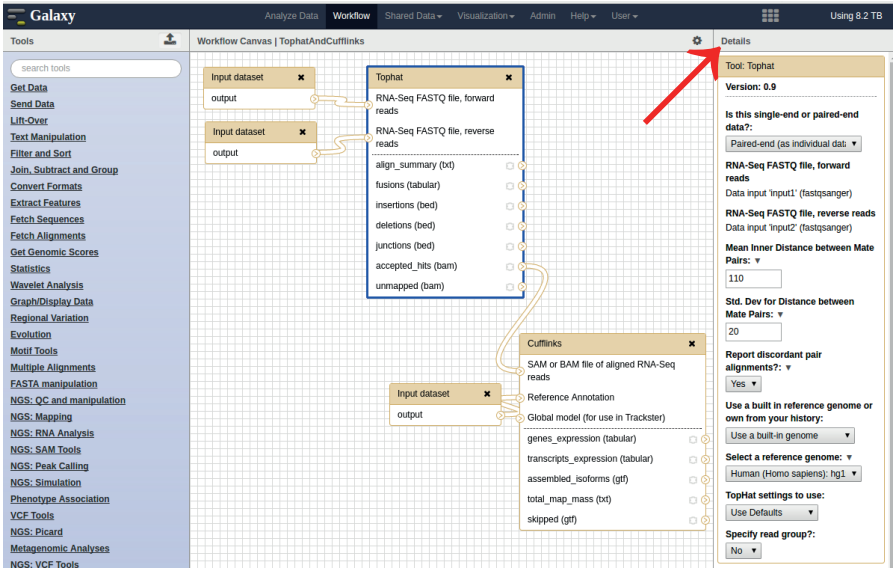

Figura 21. Ao selecionar o item Tophat, pode-se observar que todos os seus parâmetros são exibidos no painel '*Details'*.

No painel *Details* é possível também incluir uma etapa para renomear os arquivos de resultado, eliminando a necessidade de fazê-la manualmente como na sessão 2.6. Para isso, devemos rolar até a sessão *Edit Step Actions* e selecionar *Rename Dataset* (Figura 22). Logo abaixo, devemos selecionar o arquivo de resultado a ser renomeado, neste caso *accepted_hits*, e clicar no botão *Create*. Será exibida um novo campo chamado *Rename Dataset on accepted_hits* que nos permite inserir uma expressão que renomeia o arquivo utilizando o nome do arquivo de entrada. No Tophat, os arquivos de entrada se chamam *input1* e *input2*, conforme

exibido abaixo da caixa *New output name* em *Available inputs are*. Para utilizar esses nomes, devemos utilizar a seguinte notação: #{input1} ou #{input2}. O Galaxy permite ainda nos livrarmos da extensão do arquivo usando o parâmetro |basename. Desta forma, ao utilizar a notação #{input1|basename} em um arquivo chamado brain_1.fastq, obteríamos o resultado: brain_1.

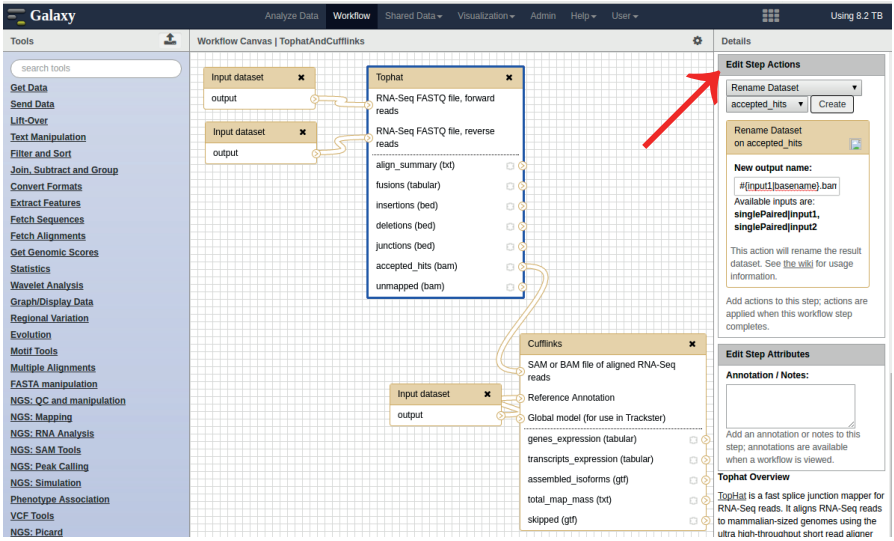

Figura 22. No painel '*Details*' é possível incluir um passo adicional para renomear o arquivo, de forma que ele contenha o nome da amostra.

Iremos renomear alguns arquivos de entrada da mesma forma que fizemos na sessão 2.6. O arquivo de resultados *accepted_hits* produzido pelo Tophat será renomeado para #{input1|basename}.bam ao incluirmos esta expressão dentro do campo *New output name*.

O mesmo deve ser feito para o arquivo de resultado *assembled_transcript*, utilizando a notação #{input|basename}.gtf, na caixa do Cufflinks (Figura 23).

Detalhes a respeito das possibilidades de utilização da ferramenta *Rename Dataset* podem ser obtidos em:

https://wiki.galaxyproject.org/Learn/AdvancedWorkflow/variablesEdit?action=show&redirect=Learn%2FAdvancedWorkflow%2FVariables.

Ao final, basta clicar no ícone representado por uma engrenagem no canto direito superior do painel central e clicar em *Save*. Uma boa prática é clicar em *Save* várias vezes ao longo da edição do *workflow*. Para executar este *workflow*, basta clicar na engrenagem e depois em *Run* (Figura 23).

Na tela de execução de *workflow* (Figura 24) serão exibidos todos os passos do *workflow*, sendo que os campos referentes a arquivos de entrada de dados, *Input dataset*, deverão ser utilizados para indicar os arquivos que serão processados. Como já executamos as amostras adrenal, devemos selecionar as amostras brain. No primeiro *Input Dataset* devemos selecionar brain_1.fastq, no segundo brain_2.fastq e no terceiro o arquivo de anotação '*iGenomes UCSC hg19, chr19 gene annotation*'. Para executar basta clicar no botão *Run workflow*.

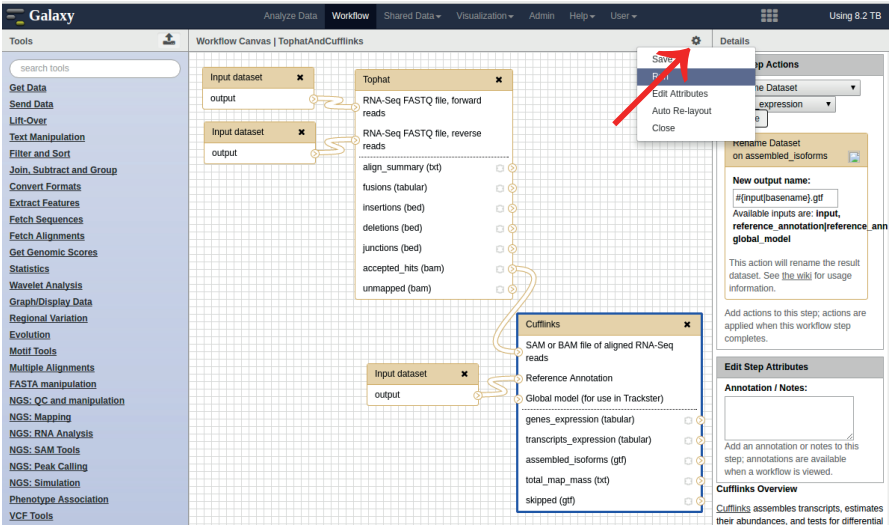

Figura 23. No ícone de engrenagem disponível no painel central é possível gravar e executar o '*workflow*'.

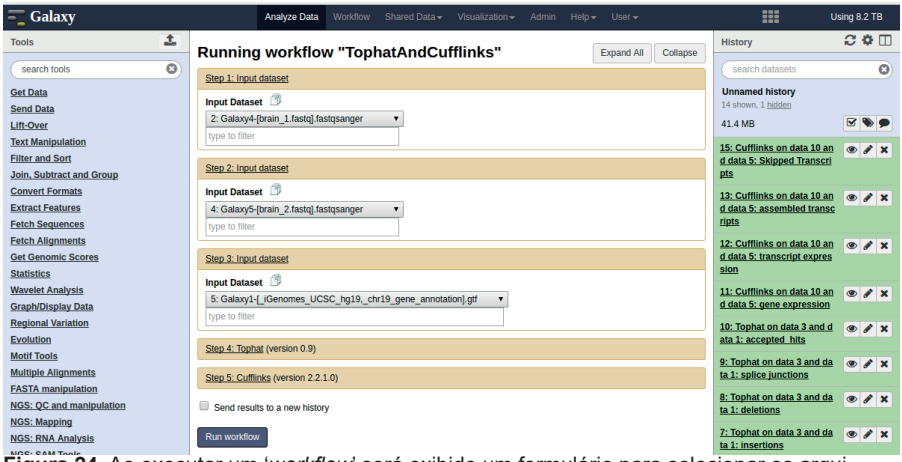

Figura 24. Ao executar um '*workflow*' será exibido um formulário para selecionar os arquivos de entrada. Parâmetros podem ser alterados ao clicar na barra de título de cada passo (bege).

2.8. Construção do transcriptoma de referência

Nesta etapa da análise iremos criar um transcriptoma de referência por meio da união dos transcritos encontrados nas amostras *adrenal* e *brain*. Para recapitular, o arquivo .gtf de transcritos da amostra *adrenal* foi obtido por meio da execução do Tophat (sessão 2.4) e do Cufflinks (sessão 2.5) e os da amostra *brain* foi obtido por meio da execução do Tophat e do Cufflinks dentro do *workflow* (sessão 2.7).

A ferramenta utilizada para unir os arquivos de transcriptoma se chama Cuffmerge (TRAPNELL et al., 2010). No painel de ferramentas (*Tools*) podemos localizar o Cuffmerge usando a busca ou abrir a categoria *NGS: RNA Analysis*. Os parâmetros da ferramenta selecionada serão exibidas no painel central conforme a Figura 25.

O primeiro parâmetro a ser configurado no Cuffmerge é o de seleção dos arquivos gerados pelo Cufflinks: *GTF file(s) produced by Cufflinks*. Deverão ser selecionados os arquivos de transcritos adrenal.gtf e brain.gtf (segurar o CTRL para selecionar múltiplos arquivos).

O Cuffmerge possibilita anotar os transcritos encontrados com informa-

ções de um transcriptoma de referência. Para isso, iremos selecionar Yes em *Use reference annotation* e selecionar o arquivo de anotação 'iGenomes UCSC hg19, chr19 gene annotation' no campo *Reference Annotation*. Todas as outras opções serão mantidas como padrão. Para executar o Cuffmerge basta clicar no botão *Execute*.

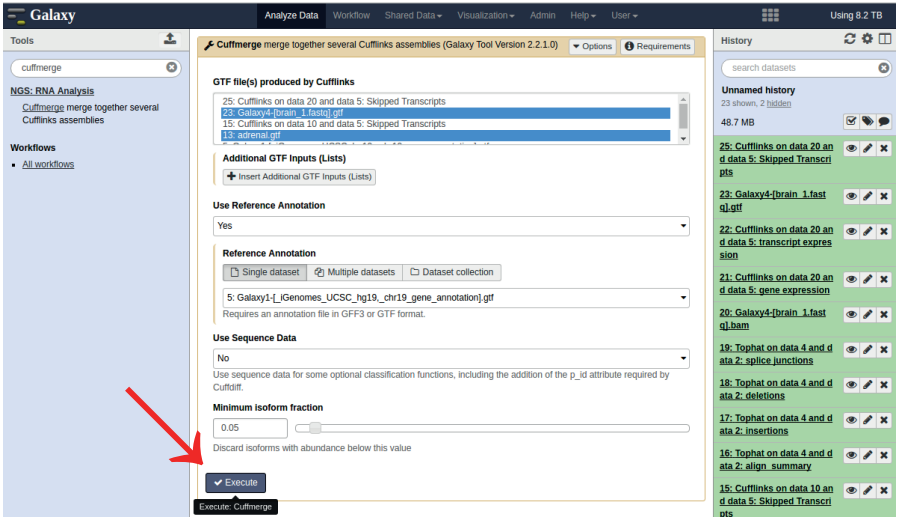

Figura 25. A ferramenta Cuffmerge possibilita gerar um arquivo de transcriptoma contendo transcritos encontrados em todas as amostras e também anotá-los por meio de um transcriptoma de referência.

2.9. Análise de expressão diferencial

Nesta etapa da análise iremos quantificar o número de *reads* de RNA-Seq que foram mapeadas nos transcritos que constituem nosso transcriptoma de referência (sessão 2.8) e efetuar o cálculo para identificação de expressão diferencial entre as amostras.

A ferramenta utilizada para fazer a análise de expressão diferencial se chama Cuffdiff (TRAPNELL et al., 2010). No painel de ferramentas (Tools) podemos localizar o Cuffdiff usando a busca ou abrir a categoria . Os parâmetros da ferramenta selecionada serão exibidas no painel central conforme a Figura 26.

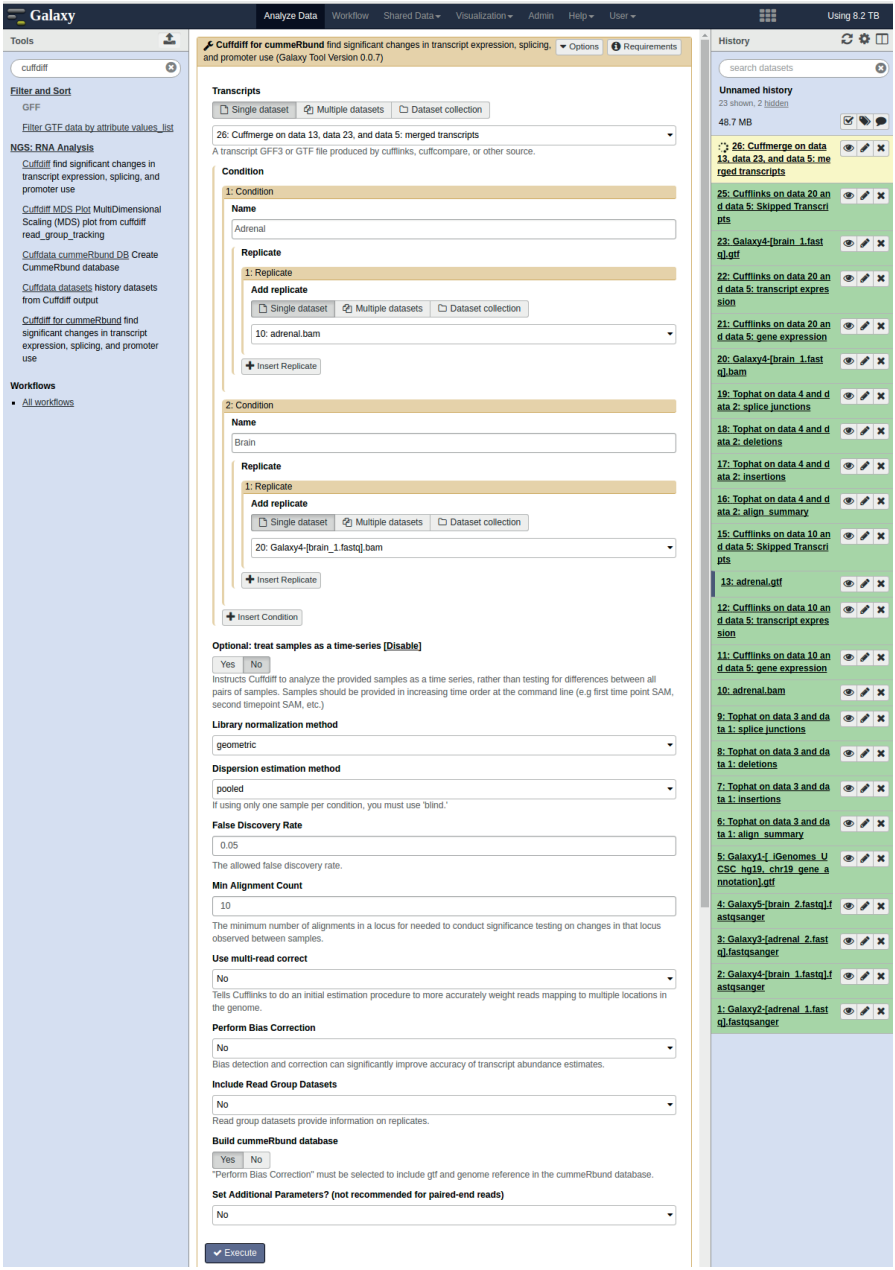

Figura 26. A ferramenta Cuffdiff permite a análise de expressão diferencial de amostras de RNA-Seq.

O primeiro parâmetro a ser configurado no Cuffdiff é o de seleção do transcriptoma de referência: *Transcripts*. Deverá ser selecionado o arquivo de transcritos gerado pelo Cuffmerge.

O segundo parâmetro, *Condition*, é destinado ao delineamento experimental, ou seja, quais são as condições experimentais e réplicas. Por padrão, o Cuffdiff apresenta uma tela com duas condições experimentais. Devemos preencher os campos *Name* com os valores: Adrenal e Brain. Na réplica da primeira condição devemos selecionar o arquivo adrenal.bam e na réplica da segunda condição o arquivo brain.bam.

Todas as outras opções serão mantidas como padrão. Para executar o Cuffdiff basta clicar no botão *Execute*.

Após a execução serão criados 15 arquivos de três tipos diferentes: *FPKM tracking, differential expression testing* e *read group tracking.* A especificação do formato dos arquivos, bem como a descrição do seu conteúdo está detalhada em: https://cole-trapnell-lab.github.io/cufflinks/cuffdiff.

Tomemos, por exemplo, os arquivos relacionados a genes. Os arquivos *FPKM tracking* contém os valores de FPKM de cada gene por amostra, ou seja, o número de fragmentos dividido por milhares bases dividido por milhões de reads. Os arquivos *read group tracking* contém a contagem de sequências mapeadas em cada gene por amostra. Por último, o arquivo *differential expression testing* contém o resultado o teste de expressão diferencial dos genes entre as condições experimentais. Neste documento iremos examinar apenas os arquivos *differential expression testing*.

Para visualizar o resultado do teste de expressão diferencial dos transcritos identificados devemos clicar no ícone representado por um olho no canto superior direito do arquivo *transcript differential expression testing* conforme Figura 27. No painel central será exibido o conteúdo do arquivo que possui 14 colunas. As colunas de 1 a 4 identificam o transcrito (test_id, gene_id, gene e locus). As colunas 5 e 6 identificam as condições experimentais comparadas (sample_1 e sample_2). A coluna 7 (status) irá apresentar OK caso o teste tenha sido bem sucedido. As colunas 8 e 9 (value_1 e value_2) contém o valor de FPKM em cada condição experimental. Por fim, as colunas 10 a 14 contém o resultado da análise de expressão diferencial, destacando-se log2(fold_change) (o valor do log2 da divisão de FPKMy por FPKMx, onde x é a amostra 1 e y a amostra 2), p (o valor p

do teste estatístico) e q (o valor p corrigido por FDR, correção em comparações de múltiplas hipóteses). A última coluna (*significant*) irá apresentar yes quando o valor q for menor que o FDR (0,05).

Para visualizar apenas os genes diferencialmente expressos, podemos utilizar a ferramenta 'Filter' para selecionar apenas as linhas que contém 'yes' na última coluna. No painel de ferramentas (*Tools*) podemos localizar a ferramenta 'Filter' usando a busca ou abrir a categoria *Filter and sort*. Os parâmetros da ferramenta selecionada serão exibidos no painel central conforme a Figura 28. O primeiro parâmetro a ser configurado é o de seleção do resultado do arquivo *transcript differential expression testing*.

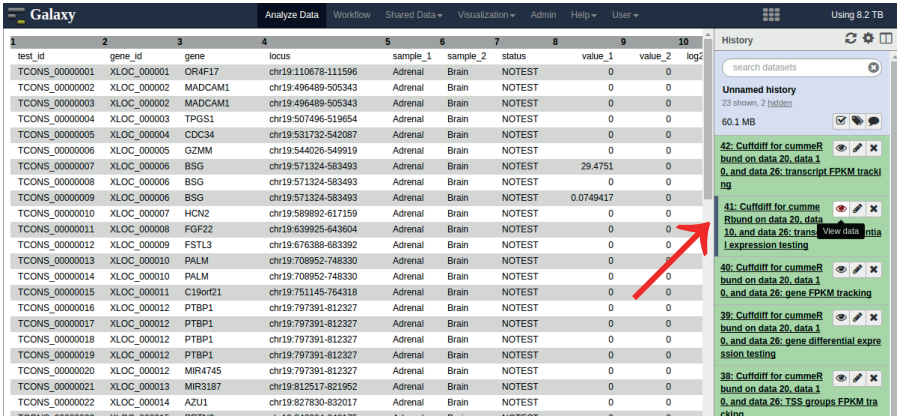

Figura 27. Arquivo de resultado do Cuffdiff que lista informações dos genes e o teste estatístico de expressão diferencial.

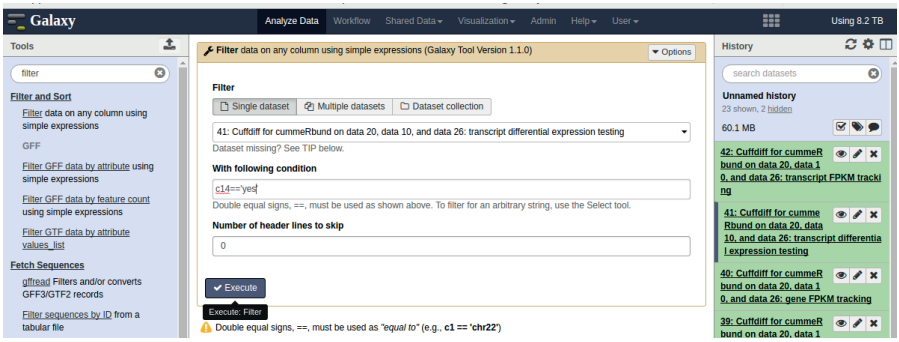

Figura 28. Utilizando-se a ferramenta '*Filter*' é possível selecionar apenas os genes diferencialmente expressos, ou seja, os que tem a palavra '*yes*' na coluna 14.

O segundo parâmetro, With following condition, deverá ser preenchido com a expressão: c14==’yes’, onde c14 é a especificação da última coluna do arquivo. Logo abaixo dos parâmetros existem duas sessões, *Syntax* e *Example*, que detalham como criar expressões de filtros. Para executar o Filter basta clicar no botão *Execute*.

O arquivo gerado pelo *Filter* contém as mesmas colunas do arquivo original, porém estão contidas apenas as linhas que passaram no filtro, conforme Figura 29.

Este filtro nos possibilitou identificar os transcritos diferencialmente expressos entre as amostras. Para identificar os genes diferencialmente expressos, devemos repetir os passos anteriores, selecionando o arquivo de genes na ferramenta *Filter*. Para visualizar estes arquivos, basta clicar o ícone de olho no canto superior direito de cada arquivo conforme Figuras 29 e 30.

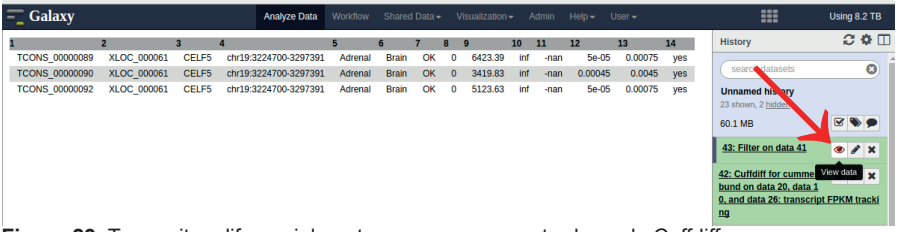

Figura 29. Transcritos diferencialmente expressos encontrados pelo Cuffdiff.

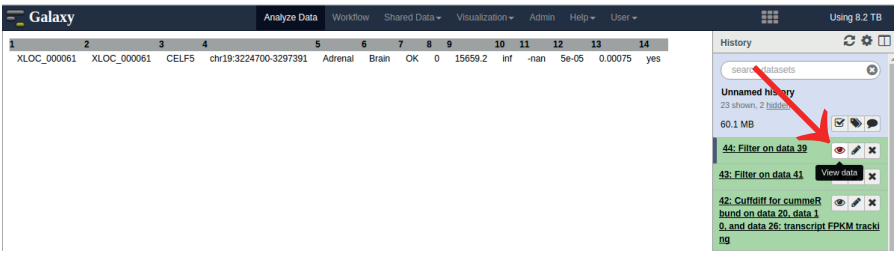

Figura 30. Gene diferencialmente expresso encontrado pelo Cuffdiff.

3. Conclusão

O conjunto de ferramentas apresentado nos possibilita a identificação de genes e transcritos diferencialmente expressos entre amostras de RNA-Seq utilizando-se um genoma de referência. Os autores destas ferramentas publicam artigos científicos e textos online com regularidade, apresentando novas funcionalidades e detalhes sobre cada programa e parâmetros.

A utilização da plataforma Galaxy possibilita atualizações de ferramentas regularmente, de forma transparente ao usuário. O Galaxy possibilita ainda maior facilidade de visualização de tabelas e gráficos, o que o tornam uma alternativa interessante, principalmente para pesquisadores que não têm experiência com a execução de comandos na interface texto.

A técnica de análise de dados de RNA-Seq está sendo aperfeiçoada continuamente e o pesquisador que for analisar tal tipo de dado deve estar atento às publicações recentes e constantes.

4. Referências

GIARDINE, B.; RIEMER, C.; HARDISON, R. C.; BURHANS, R.; ELNITSKI, L.; SHAH, P.; ZHANG, Y.; BLANKENBERG, D.; ALBERT, I.; TAYLOR, J.; MILLER, W.; KENT, WJ.; NEKRUTENKO, A. Galaxy: a platform for interactive large-scale genome analysis. **Genome Research**, v. 15, n. 10, p. 1451-1455, 2005.

TRAPNELL, C.; PACHTER, L.; SALZBERG, S. L. TopHat: discovering splice junctions with RNA-Seq. **Bioinformatics**, v. 25, n. 9, p. 1105-1111, 2009.

TRAPNELL, C.; ROBERTS, A.; GOFF, L.; PERTEA, G.; KIM, D.; KELLEY, D. R.; PIMENTEL, H.; SALZBERG, S. L.; RINN, J. L.; PACHTER, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. **Nature Protocols**, v. 7, n. 3, p. 562-578, 2012. DOI: 10.1038/nprot.2012.016

TRAPNELL, C.; WILLIAMS, B. A.; PERTEA, G.; MORTAZAVI, A.; KWAN, G.; BAREN, M. J. van; SALZBERG, S. L.; WOLD, B. J.; PACHTER, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. **Nature Biotechnology**, v. 28, n. 5, p. 511-515, 2010.

Embrapa

Informática Agropecuária

MINISTÉRIO DA
AGRICULTURA, PECUÁRIA
E ABASTECIMENTO

BRASIL
GOVERNO FEDERAL

CGPE 13454