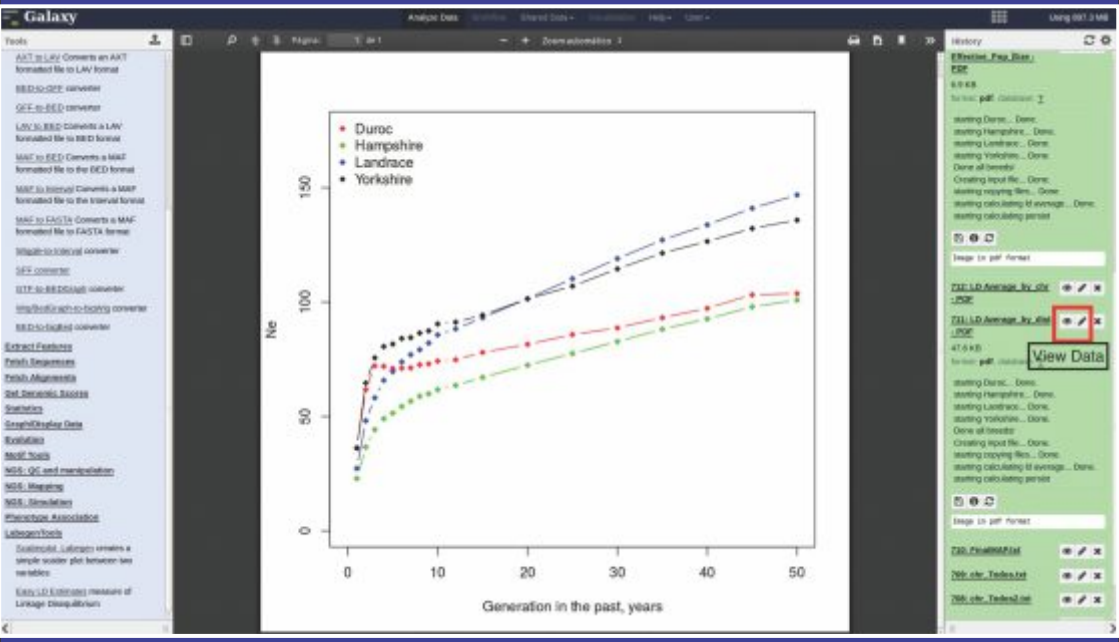
Easy LD Estimates for Galaxy

Manual do Usuário – Versão 1.0



Embrapa

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

Easy LD Estimates for Galaxy

Manual do Usuário – Versão 1.0

Willian Domingues Coelho
Leandro Lunardini Cardoso
Henry Gomes de Carvalho
Fernando Flores Cardoso

Embrapa
Brasília, DF
2018

Exemplares desta publicação podem ser adquiridos na:

Embrapa Pecuária Sul
BR 153, Km 632,9 Caixa postal 242
96401-970 - Bagé – RS
Fax: 55 53 3240-4650
www.embrapa.br/pecuaria-sul
www.embrapa.br/fale-conosco/sac

Supervisor editorial: Lisiane Bassols Brisolara
Revisor de texto: *Fernando Goss*
Normalização bibliográfica: *Graciela Olivella Oliveira*
Editoração eletrônica: *Washington Mogica - Murilo Gonçalves*
Imagem da capa: Reprodução *Embrapa Pecuária Sul*

1ª edição
Publicação digitalizada (2017)

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Pecuária Sul

Easy LD Estimates for Galaxy : manual do usuário – versão 1.0 / Willian Domingues Coelho ... [et al.]. — Brasília, DF : Embrapa, 2018.
PDF (22 p.) : il. color.

1. Bioinformática. 2. Plataforma web. 3. Análise de dados. I. Coelho, William Domingues. II. Embrapa Pecuária Sul.

CDD 570.285

Autores

Willian Domingues Coelho
Engenheiro de computação, Mestrando em Zootecnia na Universidade Federal de Pelotas,

Leandro Lunardini Cardoso
Zootecnista, Doutor em Zootecnia, Pós-doutorando da Embrapa Pecuária Sul,

Henry Gomes de Carvalho
Informata, Mestre em Ciência da Computação, Analista da Embrapa Pecuária Sul,

Fernando Flores Cardoso
Médico Veterinário, Pós-doutor *Michigan State University*, Pesquisador da Embrapa Pecuária Sul,

Apresentação

As publicações técnicas da Série Embrapa são importantes veículos de informação, destinados a produtores, técnicos, empresários do agronegócio, pesquisadores, estudantes e público em geral, interessados nas tecnologias desenvolvidas pela Empresa e seus colaboradores.

Tratam-se de publicações com distintas características, objetivos e público alvo, tais como: Boletim de Pesquisa e Desenvolvimento; Documentos; Circular Técnica; Comunicado Técnico; Sistemas de Produção; Livro e outros.

A Embrapa Pecuária Sul utiliza estes veículos para comunicar suas tecnologias produzidas, recomendações, práticas agrícolas e resultados de pesquisas e desenvolvimento, direcionando ao público interessado informações ligadas à produção de forrageiras e pastagens, bovinocultura de corte e de leite e ovinocultura dos Campos Sulbrasileiros.

É com satisfação que oferecemos mais esta obra, suplementando o trabalho vem sendo desenvolvido pela Embrapa Pecuária Sul, em Bagé, no sentido de desvendar o funcionamento dos genomas e promover o uso dessas informações no melhoramento genético em prol da melhoria da produtividade e sustentabilidade da pecuária sulina.

Esta publicação da Série Embrapa é um manual prático para utilização do sistema Easy LD Estimates, uma ferramenta de suporte às pesquisas com genômica em animais de produção. As informações contidas no documento visam orientar a utilização do sistema para calcular e visual de forma simplificada parâmetros estruturais dos genômas (desequilibrio de ligação e correlação de fase de ligação entre marcadores) e características populacionais para diferentes espécies e raças (tamanho efetivo e divergência populacional). Essas informações são essenciais para o delineamento de estudos genômicos e na aplicação prática da genômica no melhoramento animal. Esperamos que os leitores desfrutem deste Documento e sugerimos que, em caso de maior interesse no tema abordado ou necessidades de esclarecimentos, realizem o contato com nosso Serviço de Atendimento ao Cidadão (SAC)[1], ou pelo fone (53) 3240-4650. A Embrapa terá o máximo prazer em atendê-lo.

Fernando Flores Cardoso
Chefe Adjunto de PD&I

[1]Disponível em: <www.embrapa.br/faleconosco/sac/>.

Sumário

Introdução

Cálculo de desequilíbrio de ligação utilizando a ferramenta Easy LD Estimates para plataforma Galaxy – versão 1.0

Este documento tem por objetivo descrever a utilização e a funcionalidade da versão 1.0 do programa Easy LD Estimates para o ambiente Galaxy. Originalmente desenvolvido na Michigan State University em linguagem R por J. P. Steibel, no site da instituição estão disponibilizados os scripts, bem como o banco de dados[1]. A adaptação ao Galaxy foi desenvolvida no Laboratório de Bioinformática e Estatística Genômica da Embrapa Pecuária Sul.

O Galaxy Project é uma plataforma web, de código aberto, de mineração de dados, gerenciamento de dados e arquivamento eletrônico que visa tornar a bioinformática acessível a pesquisadores com nenhum conhecimento de programação de computador (GOECKS et al., 2010). Essa plataforma dispõe de diversas ferramentas pré-instaladas que realizam vários tipos de análise de dados[2]. Uma das principais características desta ferramenta é a capacidade de inserção de novos scripts, que podem rodar nas mais diversas linguagens de programação. A possibilidade de inserir novas ferramentas faz com que o Galaxy possa ser um mediador entre o usuário e as ferramentas executadas, pois elimina diversas etapas, tais como criação de arquivos de parâmetros, os quais podem ser construídos automaticamente pelo Galaxy a partir de formulários intuitivos. Também evita a sequência de linhas de comando executados via terminal.

Essa adaptação foi realizada com objetivo de facilitar a utilização dos scripts originais do LD Estimates de Steibel, que necessitam de diversas alterações em código fonte para que o processo de estimação se inicie, podendo dessa forma serem inseridos erros nos códigos fontes

[2]Disponível em: <https//msu.edu/~steibelj/JP_files/LD_estimate.html>.

ocasionando retrabalho e demanda de tempo para que o erro seja solucionado. Além disso, uma hierarquia de diretórios e arquivos de parâmetros precisa ser criada. A execução realizada via terminal pode ser um limitante para usuários não habituados com a interface de linha de comando ou a linguagem estatística R (R FOUNDATION FOR STATISTICAL COMPUTING, 2015).

Uma vez instalado na plataforma Galaxy, o Easy LD Estimates constrói a hierarquia de diretórios, cria o arquivo de parâmetros e realiza as alterações necessárias nos scripts de forma totalmente automatizada. Além disso, todos os scripts serão executados na ordem sequencial correta, de modo que as únicas ações tomadas pelo usuário são a de preencher um formulário no início do processo e a de coletar os resultados ao final da análise. Pode-se também integrar outras ferramentas do Galaxy antes ou depois da execução do Easy LD Estimates. Tal integração é feita através do recurso *Workflow* do Galaxy. O *Workflow* integra as ferramentas enfileirando os processos e sem a necessidade de aguardar um resultado para dar início ao processo seguinte.

O Easy LD Estimates utiliza a estimativa do r2 (considerando o quadro da correlação entre os genótipos de cada par de loci de um mesmo cromossoma para n indivíduos) como a medida de desequilíbrio de ligação (Linkage Disequilibrium - LD) calculando a média de r2 para todos os pares de combinações de SNPs, a média de r2 para os SNPs adjacentes, a correlação de fase dos marcadores em distâncias pré-determinadas, o tamanho efetivo da população e a divergência populacional. Dessa forma, a implementação do Easy LD Estimates na plataforma Galaxy permite uma fácil execução e visualização dos dados e resultados por meio de um browser de navegação da web, o que torna o trabalho do pesquisador mais fácil em relação ao uso da interface de linha de comando tradicional.

Dados de entrada e formatação dos dados

São necessários os arquivos de mapa dos marcadores que devem ser do tipo SNP (polimorfismos de base única) e o arquivo de genótipos já com a fase de ligação pré-determinada (haplótipos) como segue:

Mapa dos SNP

Deve conter três colunas: 1) nome do SNP; 2) cromossomo; 3) posição em pares de base. No mapa deve existir todos os SNPs constantes no arquivo de haplótipos. Pode-se acrescentar dados adicionais tais como SNPs não usados ou colunas adicionais, que serão ignoradas. O arquivo de mapa não deve conter cabeçalho.

Tabela 1. Formato do arquivo de mapa (não deve conter cabeçalho).

Nome do SNP	Cromossomo	Posição em pares de base
SNP1	1	240047
...	...	...
SNPn	n	n

Arquivo de haplótipos

O arquivo de haplótipos deve estar formatado conforme os arquivos de saída do software BEAGLE (BROWNING, 2011) em um arquivo único. A primeira coluna deve conter um identificador "I" (conforme descrito no manual oficial, a segunda coluna deve conter o nome do SNP e todas as colunas seguintes os haplótipos (alelo transmitido pelo pai e pela mãe de cada indivíduo). Os haplótipos podem ser codificados como A/T/C/G, A/B ou qualquer codificação numérica, com somente marcadores bialélicos sendo aceitos.

Tabela 2. Formato do arquivo de entrada dos haplótipos.

I	SNP	ID1_1	ID1_2	IDn_2
M	SNP1	A	T	C
M	...	...	...	...
M	SNPn	C	G	A

Como rodar o programa

Acesso ao Galaxy

O acesso a plataforma Galaxy pode ser feito na página do Laboratório Multiusuário de Bioinformática (LMB) da Embrapa Informática Agropecuária (CNPTIA). Conforme as políticas de acesso estão previstas as formas de acesso para laboratórios parceiros que necessitem apenas da infraestrutura computacional, e para os parceiros que necessitem da colaboração para planejar e executar a avaliação de bioinformática de seus projetos e podem ser acessadas no site institucional. A interface do Galaxy é baseada na plataforma web e pode ser acessada por quaisquer navegadores independentemente do sistema operacional.

Upload dos arquivos

O upload dos arquivos é feito via interface drag-and-drop (arrastar e largar) como demonstrado nas figuras 1 e 2. Para iniciar o processo de upload é necessário utilizar a opção Start da caixa de seleção de arquivos e então Close ao final do processo de upload. Os arquivos estarão prontos para uso quando todos os status de upload estiverem em 100% como demonstrado na Figura 3.

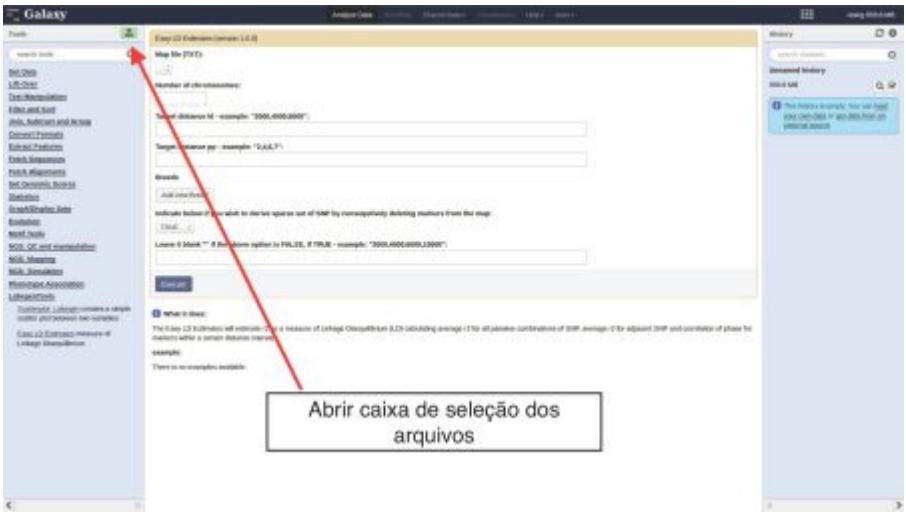


Figura 1. Seleção de arquivos para Upload

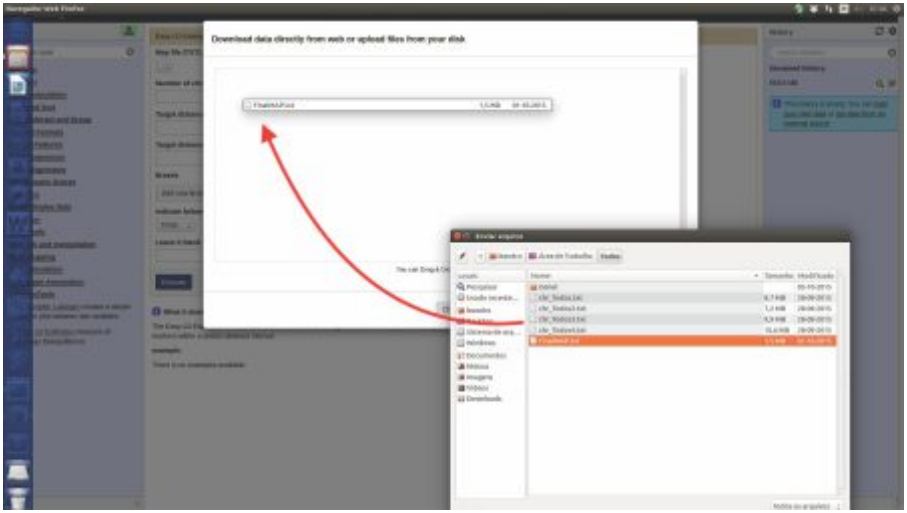


Figura 2. Upload dos arquivos no Galaxy - Easy LD Estimates.

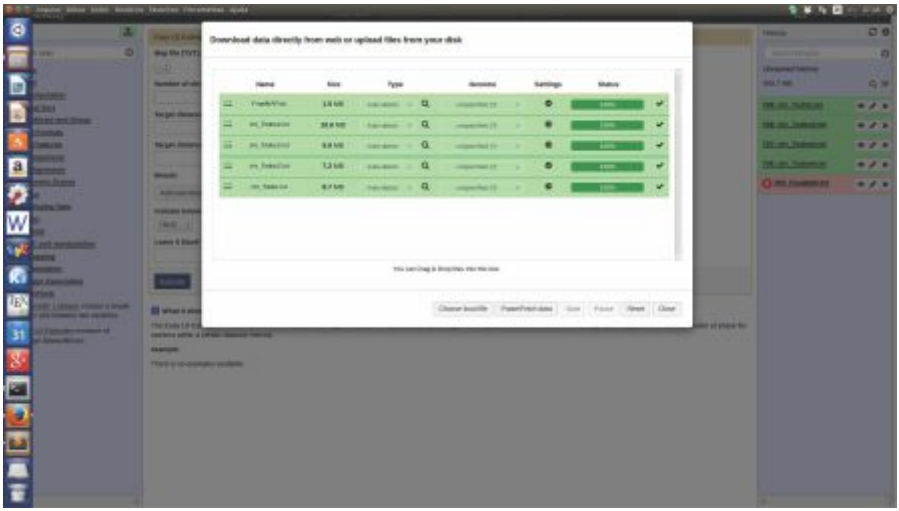


Figura 3. Arquivos prontos para serem enviados

Preenchimento do formulário e selecionar arquivos

Para a execução da ferramenta, deve-se preencher o formulário de parâmetros. O preenchimento do formulário requer atenção, pois alguns dos parâmetros são vetores e devem ser digitados seguindo um padrão específico. No formulário devemos selecionar os arquivos de entrada (Mapa de SNPs e Haplótipos) e informar os parâmetros nos campos Target distance ld, Target distance pp, Breed name (nome das raças na análise) e Sparse markers (TRUE ou FALSE) exemplificado na Figura 4. Os valores são determinados conforme a necessidade do usuário.

Parâmetros: o parâmetro *Target distance ld* (Figura 4) estima o desequilíbrio de ligação (r^2) médio nos intervalos em pares de base (pb) entre as distâncias especificadas pelo usuário. Por *default* o *script* estima automaticamente o r^2 médio e apresenta graficamente a cada 100 kb de 0 a 10 Mpb. O vetor numérico *Target distance pp* (Figura 4) contém os intervalos de distância no qual irá ser calculada a persistência de fase entre as populações em análise. Na análise de marcadores esparsos a opção TRUE (Figura 5) irá proceder a análise de desequilíbrio de ligação entre SNP adjacentes de painéis esparsos, neste

caso deve-se determinar a densidade dos marcadores dentro do painel. Como exemplo o vetor numérico esparso (Figura 5) é um conjunto de 2, 4 e 10 no qual irá resultar em três painéis de marcadores incluindo um a cada dois, quatro e dez marcadores, respectivamente. No caso da opção FALSE ser a elegida o programa não irá executar a análise de painéis esparsos.

O Easy LD Estimates for Galaxy pode ser utilizado para as medidas de desequilíbrio de ligação para diversas espécies. Após a indicação dos nomes das raças e dos arquivos de haplótipos relativos a cada raça assim como a definição dos parâmetros da análise, a execução da análise se dará pela seleção da opção *Execute* (Figura 5).

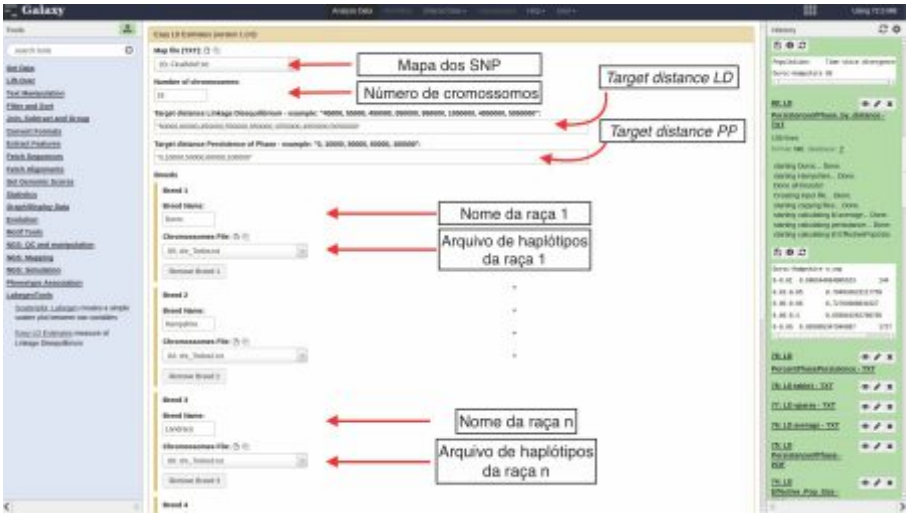


Figura 4. Interface do Easy LD Estimates, entrada dos arquivos e definição dos parâmetros da análise.

Figura 5. Interface do Easy LD Estimates, entrada dos arquivos e definição dos parâmetros da análise.

Arquivos de saída

O Easy LD Estimates entrega ao usuário diversos arquivos de saída após uma análise, alguns arquivos em formato PDF (contendo os gráficos) e outros em formato TXT. Esses arquivos em TXT já ficam salvos na base do Galaxy e podem ser utilizados em qualquer outra ferramenta no Galaxy que tenha arquivos neste formato como parâmetros ou entrada. Na Figura 6 podemos observar que existe uma coluna à direita onde é apresentado um histórico de ações executadas pelo usuário. Essas ações incluem upload de arquivos, execução de ferramentas, erros de execução e os arquivos de saída gerados. Cada arquivo de saída, seja um PDF ou TXT, será uma nova linha nesse histórico. Os arquivos são entregues um de cada vez para que estejam disponíveis na base do Galaxy e possam ser utilizados como parâmetros para outras ferramentas, da mesma forma como se tivessem sido enviados pelo usuário. A Figura 6 apresenta a saída gerada pela execução da Ferramenta Easy LD Estimates.

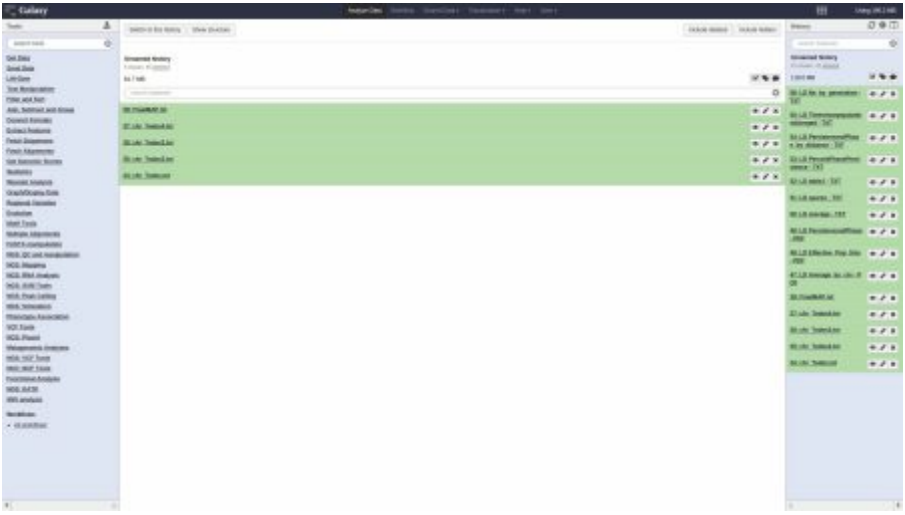


Figura 6. Coluna *History.*

Na coluna *History* basta clicar na opção *View Data* em qualquer uma das linhas de saída que o usuário deseja visualizar e a informação será apresentada na tela central do Galaxy como mostra a Figura 7. Além disso, existem outras opções como *download*, destacada na Figura 8, que permite baixar o arquivo de saída e a opção *Run this job again*.

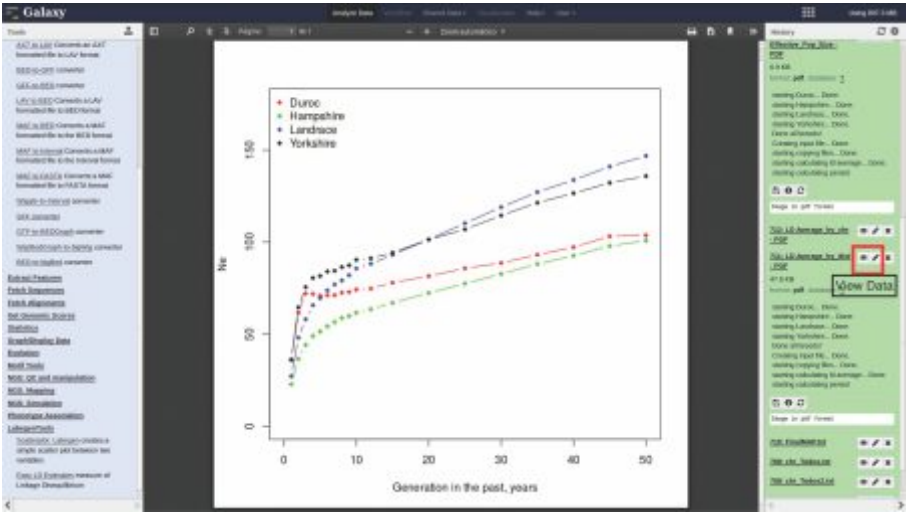


Figura 7. View Data.

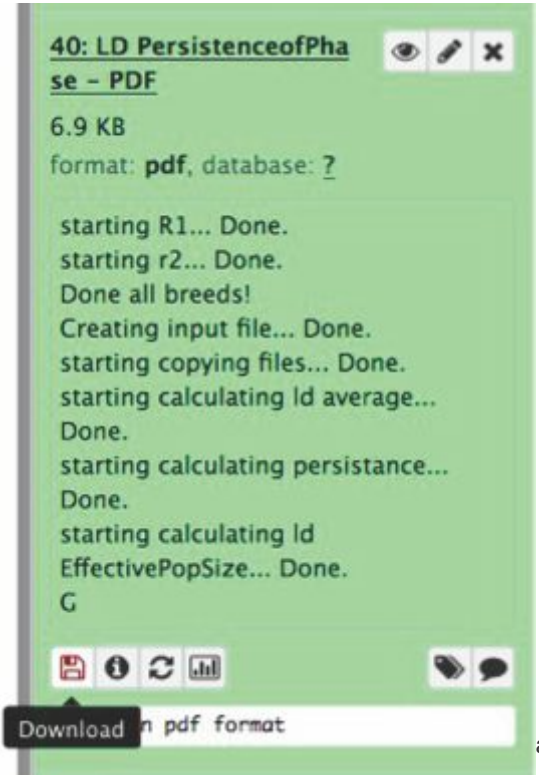


Figura 8. Download de arquivos de saída.

A opção *Run this job again*, retorna a página de execução do Easy LD Estimates devidamente preenchida com os parâmetros da análise anterior, permitindo ao usuário realizar as alterações que julgar necessário e rodar a análise novamente. Um exemplo de clique na opção *Run this job again* é apresentada na Figura 9.

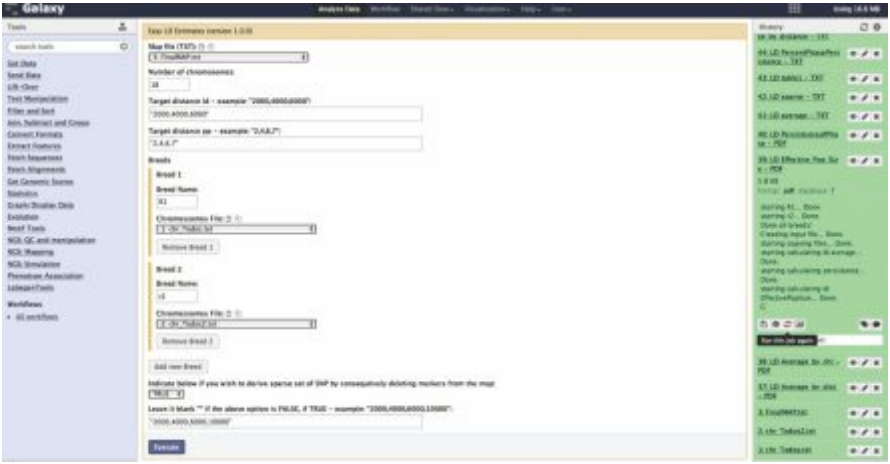


Figura 9. *Run this job again*

Considerações finais

As vantagens obtidas pela adaptação da ferramenta LD Estimates à Plataforma Galaxy são diversas. Contudo, cabe destacar a facilidade para executar análises, por construir a hierarquia de diretórios e realizar as modificações nos arquivos fonte de forma automática, sendo que o usuário não precisa ter conhecimentos em programação e nem saber utilizar uma interface de linha de comando. Mesmo usuários mais avançados podem tirar vantagem desse novo método de execução, que reduz o tempo e o esforço necessário para que uma nova análise seja iniciada. Ademais, o *software* garante que erros de sintaxe não serão inseridos nos arquivos fonte, garantindo que o *software* executará até o final de forma correta.

Outro destaque é a possibilidade de enfileirar processos do Galaxy para permitir que arquivos de saída de outras análises já finalizadas, ou ainda em execução, sejam usados como parâmetro de entrada de novas análises de outras ferramentas disponíveis na plataforma. Além disso, o pesquisador poderá acompanhar seus experimentos de qualquer lugar, desde que a plataforma Galaxy esteja instalada em um servidor com acesso liberado para conexões de qualquer origem (Acesso ao Galaxy), o que poderá ser feito de qualquer dispositivo com acesso à Internet.

Referências

BROWNING, B. L. **BEAGLE 3.3.2**. Washington, University of Washington, Department of Medicine, Division of Medical Genetics, 2011. 30 p. Disponível em: <https://faculty.washington.edu/browning/beagle/beagle_3.3.2_31Oct11.pdf>. Acesso em: 20 nov. 2015.

GOECKS, J.; NEKRUTENKO, A.; TAYLOR, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. **Genome Biology**, v. 11, n. 8, Aug. 2010. R86.

R FOUNDATION FOR STATISTICAL COMPUTING. **The R Project for Statistical Computing**. Vienna, [2015]. Disponível em: <http://www.R-project.org/>. Acesso em: 25 nov. 2015.

Literatura recomendada

BADKE, Y. M.; BATES, R. O.; ERNST, C. W.; SCHWAB, C.; STEIBEL, J. P. Estimation of linkage disequilibrium in four US pig breeds. **BMC Genomics**, v. 13, n. 24, Jan. 2012.

BLANKENBERG, D.; VON KUSTER, G.; CORAOR, N.; ANANDA, G.; LAZARUS, R.; MANGAN, M.; NEKRUTENKO, A.; TAYLOR, J. Galaxy: a web-based genome analysis tool for experimentalists. In: CURRENT protocols in molecular biology. New York: Wiley, 2010. Cap. 19.10.

BROWNING, B. L.; BROWNING, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. **American Journal of Human Genetics**, v. 84, n. 2, p. 210-223, Jan. 2009.

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

Embrapa

Pecuária Sul

CGPE 12950

Embrapa

MINISTÉRIO DA
AGRICULTURA, PECUÁRIA
E ABASTECIMENTO

BRASIL
GOVERNO FEDERAL