Caderno Computacional 2/6: Associação Biológica

Gepoliano Chaves, Ph. D.

Janeiro, 2021

Contents

Para plotagem PPT ou PDF, incluir os seguintes comandos:

output: powerpoint_presentation

ou

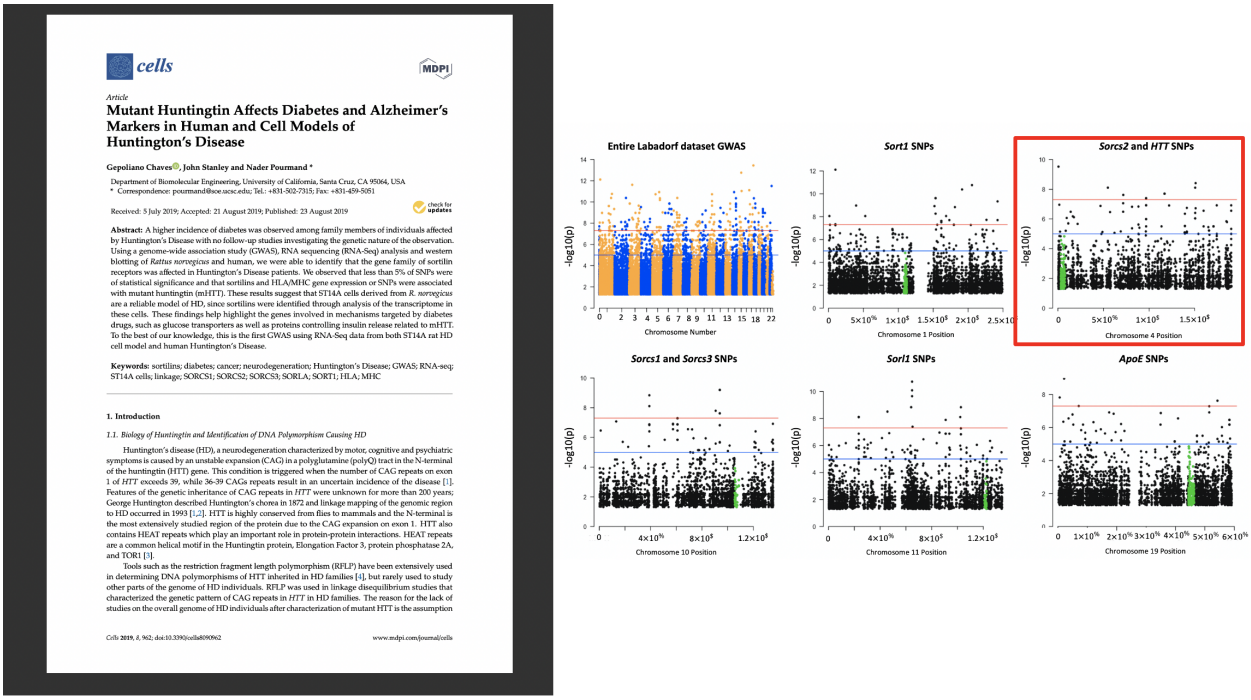
output: pdf_document: toc: yes toc_depth: '5' html_document: df_print: paged toc: yes number_sections: no toc_depth: 5 toc_float: yes, above

Carregar Livrarias

```
library(knitr)
```

1) Introdução

No Caderno 1, iniciamos a ambientação de estudantes com o universo da programação computacional através da instalação do RStudio, definição do uso do formato R Markdown para interação entre linguagem humana e linguagem de programação computacional e uso do primeiro comando em bash: head. A progressão do uso de programação computacional poderá ser interpretada como exponecial neste e nos próximos cadernos pois em si mesmos, estes cadernos não trazem ferramentas para extrapolação, entendimento e proficiêcia dos comandos utilizados. Neste Caderno 2, o objetivo é comunicar o conceito de Associação Biológica de forma relativamente simples e talvez intuitiva, sem preocupação de profundidade exagerada em conceitos científicos que embora de grande importância, não possuimos tempo hábil para lidar com sua demonstração. No presente Caderno, vamos aprender como foi plotada parte da Figura 3 presente no artigo científico ilustrado na Figura 1.



cells

MDPI

Article

Mutant Huntingtin Affects Diabetes and Alzheimer's Markers in Human and Cell Models of Huntington's Disease

Gepoliano Chaves, John Stanley and Nader Pourmand *

Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95064, USA
* Correspondence: pourmand@soe.ucsc.edu; Tel.: +831-502-7315; Fax: +831-459-5051

Received: 5 July 2019; Accepted: 21 August 2019; Published: 23 August 2019

Abstract: A higher incidence of diabetes was observed among family members of individuals affected by Huntington's Disease with no follow-up studies investigating the genetic nature of the observation. Using a genome-wide association study (GWAS), RNA sequencing (RNA-Seq) analysis and western blotting of *Rattus norvegicus* and human, we were able to identify that the gene family of sortilin receptors was affected in Huntington's Disease patients. We observed that less than 5% of SNPs were of statistical significance and that sortilins and HLA/MHC gene expression or SNPs were associated with mutant huntingtin (mHTT). These results suggest that ST14A cells derived from *R. norvegicus* are a reliable model of HD, since sortilins were identified through analysis of the transcriptome in these cells. These findings help highlight the genes involved in mechanisms targeted by diabetes drugs, such as glucose transporters as well as proteins controlling insulin release related to mHTT. To the best of our knowledge, this is the first GWAS using RNA-Seq data from both ST14A rat HD cell model and human Huntington's Disease.

Keywords: sortilins; diabetes; cancer; neurodegeneration; Huntington's Disease; GWAS; RNA-seq; ST14A cells; linkage; SORCS1; SORCS2; SORCS3; SORLA; SORT1; HLA; MHC

1. Introduction

1.1. Biology of Huntingtin and Identification of DNA Polymorphism Causing HD

Huntington's disease (HD), a neurodegeneration characterized by motor, cognitive and psychiatric symptoms is caused by an unstable expansion (CAG) in a polyglutamine (polyQ) tract in the N-terminal of the huntingtin (HTT) gene. This condition is triggered when the number of CAG repeats on exon 1 of *HTT* exceeds 39, while 36-39 CAGs repeats result in an uncertain incidence of the disease [1]. Features of the genetic inheritance of CAG repeats in *HTT* were unknown for more than 200 years; George Huntington described Huntington's chorea in 1872 and linkage mapping of the genomic region to HD occurred in 1993 [1,2]. HTT is highly conserved from flies to mammals and the N-terminal is the most extensively studied region of the protein due to the CAG expansion on exon 1. HTT also contains HEAT repeats which play an important role in protein-protein interactions. HEAT repeats are a common helical motif in the Huntingtin protein, Elongation Factor 3, protein phosphatase 2A, and TOR1 [3].

Tools such as the restriction fragment length polymorphism (RFLP) have been extensively used in determining DNA polymorphisms of HTT inherited in HD families [4], but rarely used to study other parts of the genome of HD individuals. RFLP was used in linkage disequilibrium studies that characterized the genetic pattern of CAG repeats in *HTT* in HD families. The reason for the lack of studies on the overall genome of HD individuals after characterization of mutant HTT is the assumption

Figure 1: Artigo científico ilustrando a associação biológica entre polimorfismos de DNA nos genes das proteínas Sortilinas e Doença de Huntington.

Em mais detalhes, o artigo científico original, de Chaves, Stanley e Pourmand (2019), trata da descoberta de marcadores genéticos comuns à Doença de Alzheimer e diabetes, em uma terceira patologia: a Doença de Huntington. A Doença de Huntington é uma doença monogenética que afeta o Sistema Nervoso Central com consequências psicológicas, metabólicas e motoras para o paciente. Por ser uma doença monogenética, isto é, causada por um único gene, a Doença de Huntington talvez seja uma das doenças genéticas mais compreendidas em termos de Associação Biológica entre fenótipo e sequência genômica. Na Doença de Huntington, ocorre acúmulo da proteína mutante, chamada huntingtina, representada por HTT, no citoplasma de células do paciente. A huntingtina mutante, chamada mHTT, interage com inúmeras proteínas do citoplasma celular e entre suas funções está o desempenho no transporte vesicular dentro da célula. De acordo com a literatura científica, a proteína mHTT liga-se a proteínas do Complexo Motor Molecular, como dineína, dinactina e quinesina (Figura 2).

Através da interação entre a huntingtina, o Complexo Motor Molecular e o citoesqueleto, o papel em condições normais e patológicas da huntingtina é desempenhado. A ligação de HTT e mHTT a elementos do cito-esqueleto celular permite então, que vesiculas contendo proteinas e outros mensageiros celulares, sejam transportadas através do citoesqueleto celular. Esta função da proteína HTT, normal e mutada, pode então

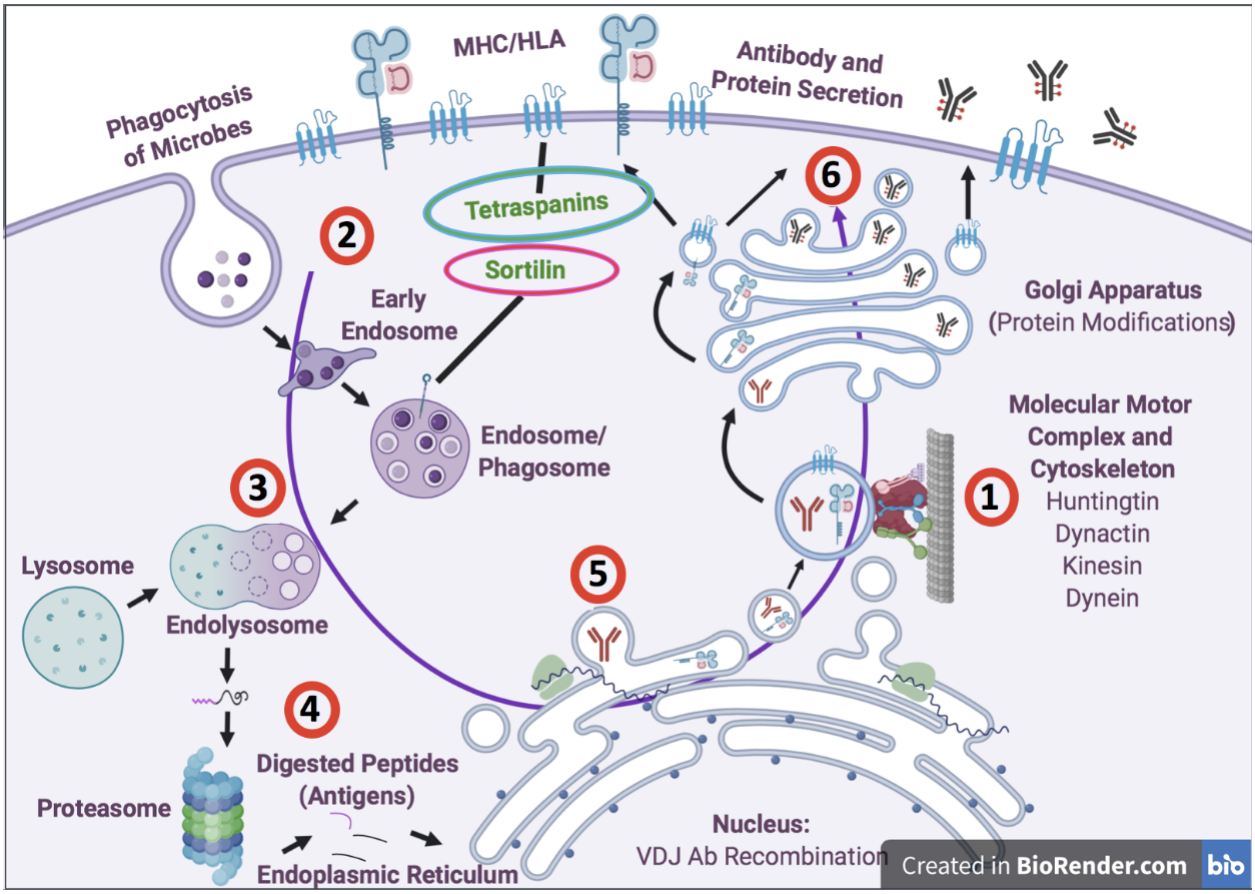


Figure 2: Mecanismo biológico da proteína HTT normal e mutante no tráfico intracelular via citoesqueleto. 1) Huntingtina tem como moléculas de interação, proteinas do Complexo Motor Molecular: dineína, quinesina e dinactina. 2) A interação da huntingtina com estas proteínas auxilia no transporte vesicular intracelular, inclusive na fagocitose de micro-organismos, utilizando o sistema de túbulos do citoesqueleto. Vesiculas contendo os mais variados tipos de proteínas, desde peptídeos apresentados ao sistema imunológico, processados nos endosomos (3) ou proteasomo (4) são transportadas usando o sistema de membranas intracelulares do retículo endoplasmático (5) e do complexo de Golgi (6), alcançando a membrana e o meio extra-celular.

ser usada para explicar os fundamentos biológicos para uma maior incidência de diabetes em indivíduos portadores da Doença de Huntington. Nestes indivíduos, o comprometimento da maquinaria de transporte intra-celular compromete, por consequência, a secreção de insulina nas células produtoras deste hormônio, as células do pâcreas, assim como de células receptoras do hormônio, como as células dos músculos e as células armazenadoras de gordura (também chamadas adipócitos). Em diabetes do tipo 1, as células do pâncreas são deficientes na secreção de insulina. Já na diabetes do tipo 2, o principal mecanismo biologico é a resistência à insulina. Neste caso, embora o pâcreas produza insulina, células musculares e adipócitas não conseguem absorver glicose, devido à resistêcia à insulina. Na resistência à insulina, a maquinaria de transporte intracelular responsavel pela internalização de vesiculas contendo receptores de glicose 4 (GLUT4), os quais realizam a captação celular de glicose, está comprometida (Figura 2).

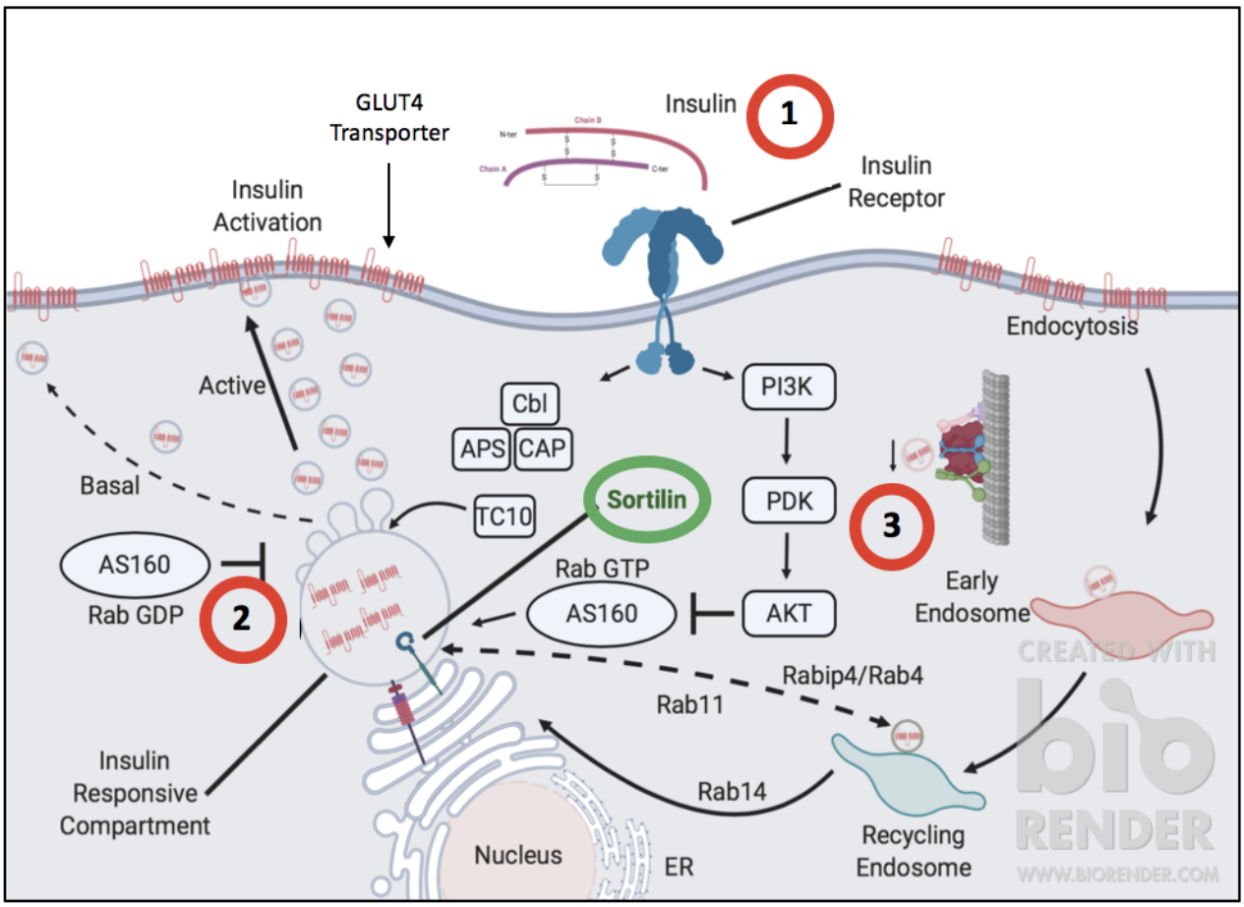


Figure 3: Mecanismo biológico de ação das proteínas sortilinas na captação de glicose, utilizando-se a ligação da insulina ao receptor de insulina e translocação do receptor de glicose 4 (GLUT4) para a membrana celular. Notar que a ligação da insulina a seu receptor (1) ocasiona translocação de receptores GLUT4 para a membrana (2) com utilização da maquinaria celular acoplada à huntingtina mutante e ao citoesqueleto (3).

Em teoria, indivíduos portadores da Doença de Huntington, por terem comprometimento de sua maquinaria de transporte intracelular, apresentam deficiêcia tanto nos mecanismos de diabetes tipo 1, quanto diabetes tipo 2. Assim, o uso de drogas para tratamento de diabetes, poderia ser também eficaz no tratamento da Doença de Huntington. No presente caderno, vamos demonstrar a associação biológica entre a Doença de Huntington e variantes genéticas identificadas em marcadores geneticos de diabetes, as proteínas sortilinas, em humanos. A Associação Biológica entre a Doença de Huntington e sortilinas foi possível graças a observação de que, em modelo celular da doença com células de *Rattus norvegicus*, a sortilina SORCS1 está aumentada nas células mutantes comparadas a células normais (Chaves, Stanley e Pourmand, 2019) (Figura 3).

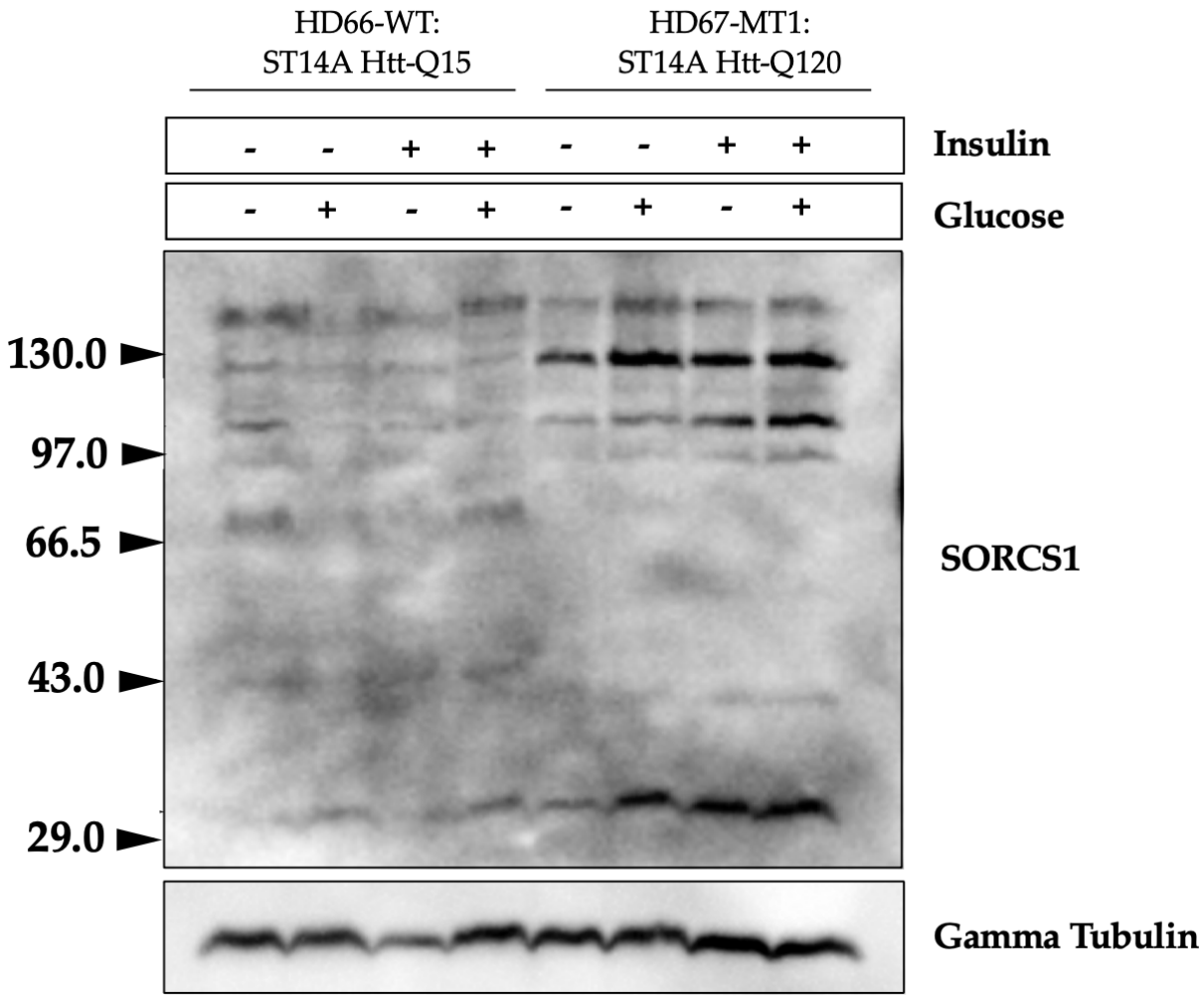


Figure 4: Aumento da expressão da sortilina SORCS1 devido à presença da huntingtina mutante mHTT em células de rato, possivelmente afetando a secreção de insulina e a translocação de GLUT4 em modelos de Doença de Huntington. Figura adaptada de Chaves, Stanley e Pourmand (2019).

Finalmente, usamos o pacote QQplot para visualização da Associação Biológica de mutação (ou mutações) de Sortilinas com o fenótipo representado pela Doença de Huntington, por meio de um gráfico conhecido como Manhattan plot.

2) Arquivo VCF gerado pela pipeline de identificação de variantes genéticas

Ao longo deste curso, utilizaremos pipelines, que são processamentos computacionais de dados, afim de produzir determinada análise. As áreas das Ciências de Dados e Tecnologias da Informação manejam amostras isoladas para análise na indústria da informação. O grande volume de dados produzidos e analisados por esta indústria faz necessária uma a construção de uma grande rede de conectividade entre os processos computacionais, para que a informação seja analisada e digerida da forma mais eficiente possível. Ao conjunto de passos que dão conectividade à análise, dá-se o nome de pipeline (Figura 4).

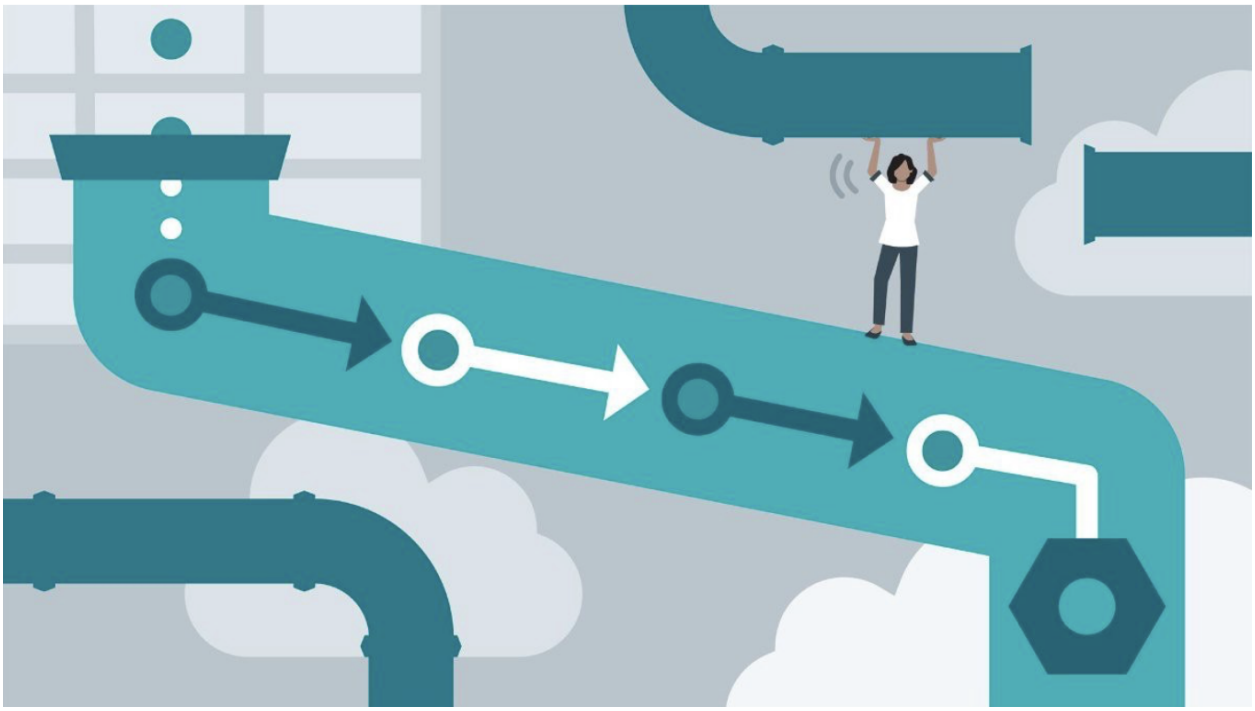


Figure 5: Pipelines são passos computacionais executados para produzir determinada análise. Neste curso, variantes genéticas são identificadas por meio de códigos computacionais executados ordenadamente, afim de obter variantes associadas a determinado fenótipo. Pipeline, em inglês, significa encanamento. Assim, o conceito encerra a ideia de que a informação deve seguir um fluxo lógico, muitas vezes matemático.

A Figura 4 foi extraída do seguinte website:

https://medium.com/@aayushdevgan/building-a-data-pipeline-framework-part-1-1653ab53dcba

Nossa pipeline de identificação de variantes genéticas sempre produzirá arquivos VCF (Variant Call Format). Esses arquivos contêm as posições das variantes genéticas identificadas em cada amostra. Desta forma, cada amostra, que pode ser um indivíduo sequenciado, uma população de células ou um swab nasal contendo material genético de SARS-CoV-2, terá um arquivo VCF correspondente. As posições das variantes são as coordenadas genômicas ao longo do cromossomo ou sequência genética analisada. No caso de humanos, há múltiplos cromossomos, ao passo que em SARS-CoV-2, a sequência genômica é um único arquivo.

2.1) Compressão e indexamento de arquivos VCF

Como a cada amostra corresponde um arquivo VCF, em nossa pipeline, precisamos mesclar todos os arquivos VCF correspondentes a todos os indivíduos em análise no nosso estudo. A palavra em inglês correspondente a mesclar é *merge*. Nos códigos executados abaixo, a palavra *merge* traz a ideia de que o arquivo que leva este nome, é produto de um comando de mesclagem. Para o comando de mesclagem que executaremos, é necessário os passos de compressão e indexamento, ilustrados abaixo.

```
for x in $(cat ~/Desktop/Gepoliano/UFSB/Arquivos/Labadorf_vcfs/files_list.txt); \
do /usr/local/bin/bgzip $x; done

for x in $(cat ~/Desktop/Gepoliano/UFSB/Arquivos/Labadorf_vcfs/files_list.txt) \
do /usr/local/bin/tabix $x".gz"; done
```

2.2) Mesclagem de arquivos VCF

É necessário mesclar o arquivo VCF proveniente de cada amostra, o qual foi comprimido e indexado no código anterior, formando um arquivo final, contendo todas as amostras do estudo. Como as amostras usadas aqui são provenientes do estudo publicado por Labadorf, o arquivo VCF final, contendo as mutações de todas as amostras, foi chamado de Labadorf_merged.vcf, indicando a ideia de que este arquivo é proveniente de um processo de mesclagem entre todas as amostras. O comando para mesclagem dos arquivos VCF, utilizando o software bcftools, está indicado abaixo.

```
cd ~/Desktop/Gepoliano/UFSB/Arquivos/Labadorf_vcfs/

/Users/gepolianochaves/anaconda3/bin/bcftools merge --missing-to-ref --force-samples \
SRR1747143_filtered_snps_final.vcf.gz \
SRR1747144_filtered_snps_final.vcf.gz \
SRR1747145_filtered_snps_final.vcf.gz \
SRR1747146_filtered_snps_final.vcf.gz \
SRR1747147_filtered_snps_final.vcf.gz \
SRR1747148_filtered_snps_final.vcf.gz \
SRR1747149_filtered_snps_final.vcf.gz \
SRR1747150_filtered_snps_final.vcf.gz \
SRR1747151_filtered_snps_final.vcf.gz \
SRR1747152_filtered_snps_final.vcf.gz \
SRR1747153_filtered_snps_final.vcf.gz \
SRR1747154_filtered_snps_final.vcf.gz \
SRR1747155_filtered_snps_final.vcf.gz \
SRR1747156_filtered_snps_final.vcf.gz \
SRR1747157_filtered_snps_final.vcf.gz \
SRR1747158_filtered_snps_final.vcf.gz \
SRR1747159_filtered_snps_final.vcf.gz \
SRR1747160_filtered_snps_final.vcf.gz \
SRR1747161_filtered_snps_final.vcf.gz \
SRR1747162_filtered_snps_final.vcf.gz \
SRR1747163_filtered_snps_final.vcf.gz \
SRR1747164_filtered_snps_final.vcf.gz \
SRR1747165_filtered_snps_final.vcf.gz \
SRR1747166_filtered_snps_final.vcf.gz \
SRR1747167_filtered_snps_final.vcf.gz \
SRR1747168_filtered_snps_final.vcf.gz \
SRR1747169_filtered_snps_final.vcf.gz \
```

```
SRR1747170_filtered_snps_final.vcf.gz \
SRR1747171_filtered_snps_final.vcf.gz \
SRR1747172_filtered_snps_final.vcf.gz \
SRR1747173_filtered_snps_final.vcf.gz \
SRR1747174_filtered_snps_final.vcf.gz \
SRR1747175_filtered_snps_final.vcf.gz \
SRR1747176_filtered_snps_final.vcf.gz \
SRR1747177_filtered_snps_final.vcf.gz \
SRR1747178_filtered_snps_final.vcf.gz \
SRR1747179_filtered_snps_final.vcf.gz \
SRR1747180_filtered_snps_final.vcf.gz \
SRR1747181_filtered_snps_final.vcf.gz \
SRR1747182_filtered_snps_final.vcf.gz \
SRR1747183_filtered_snps_final.vcf.gz \
SRR1747184_filtered_snps_final.vcf.gz \
SRR1747185_filtered_snps_final.vcf.gz \
SRR1747186_filtered_snps_final.vcf.gz \
SRR1747187_filtered_snps_final.vcf.gz \
SRR1747188_filtered_snps_final.vcf.gz \
SRR1747189_filtered_snps_final.vcf.gz \
SRR1747190_filtered_snps_final.vcf.gz \
SRR1747191_filtered_snps_final.vcf.gz \
SRR1747192_filtered_snps_final.vcf.gz \
SRR1747193_filtered_snps_final.vcf.gz \
SRR1747194_filtered_snps_final.vcf.gz \
SRR1747195_filtered_snps_final.vcf.gz \
SRR1747196_filtered_snps_final.vcf.gz \
SRR1747197_filtered_snps_final.vcf.gz \
SRR1747198_filtered_snps_final.vcf.gz \
SRR1747199_filtered_snps_final.vcf.gz \
SRR1747200_filtered_snps_final.vcf.gz \
SRR1747201_filtered_snps_final.vcf.gz \
SRR1747202_filtered_snps_final.vcf.gz \
SRR1747203_filtered_snps_final.vcf.gz \
SRR1747204_filtered_snps_final.vcf.gz \
SRR1747205_filtered_snps_final.vcf.gz \
SRR1747206_filtered_snps_final.vcf.gz \
SRR1747207_filtered_snps_final.vcf.gz \
SRR1747208_filtered_snps_final.vcf.gz \
SRR1747209_filtered_snps_final.vcf.gz \
SRR1747210_filtered_snps_final.vcf.gz \
SRR1747211_filtered_snps_final.vcf.gz > \
~/Desktop/Gepoliano/UFSB/Arquivos/Labadorf_vcfs/Labadorf_merged_test.vcf
```

2.3) Extração dos arquivos MAP e PED

Após a mesclagem dos arquivos VCF, é necessário extrair-se informação do arquivo mesclado de forma a obter um formato compatível com o processamento subsequente pelos softwares de interesse. Para o passo da Associação Biológica, utilizaremos o software PLINK, usado em estudos de Linkage Disequilibrium, ou Ligação Genética, em português. Arquivos MAP e PED foram largamente utilizados em estudos de Linkage, uma técnica de pesquisa genética, usada na Doença de Huntington para estudar a segregação do gene afetado pela doença nas famílias acometidas por esta condição. Arquivos MAP e PED são compatíveis com o software

PLINK, o qual será usado para calcular a Associação Biológica entre as SNPs e o fenótipo (Doença de Huntington). Para extrair arquivos MAP e PED do arquivo VCF, usamos o arquivo Labadorf_merged.vcf como input para vcftools. O comando abaixo produz os arquivos MAP e PED desejados.

```
## Parece que não consegui, inicialmente correr vcftools, pelo fato de o
## bash_profile não poder ser acessado daqui.
## Usando todo o path para vcftools, eu consegui:

/Users/gepolianochaves/anaconda3/bin/vcftools --vcf \
~/Desktop/Gepoliano/UFSB/Arquivos/Labadorf_merged.vcf \
--out ~/Desktop/Gepoliano/UFSB/Arquivos/Labadorf_merged --plink

## Este comando não funcionou, inicialmente:
# vcftools --vcf Labadorf_merged.vcf --out formatted_plink_output --plink
```

O arquivo PED construído nos comandos acima, necessita ainda ser modificado, na sexta coluna, de forma a refletir indivíduos afetados (1) e não afetados (2). A modificação pode ser efetuada editando-se o arquivo com Excel, porém esta estratégina não foi efetiva visto que o Excel corrompeu o arquivo ao salvá-lo. A modificação pôde então ser efetuada através do comando vim, na linha de comando. Para evitar confusões, o arquivo PED derivado do VCF Labadorf_merged.vcf foi salvo como Labadorf_merged_estavel.ped, para uso no futuro.

2.4) Visualização dos arquivos

Em bash, podemos usar o comando head, para visualizar os arquivos MAP e PED. Abaixo, o comando head não está funcionando para visualizar o arquivo PED. Então, devo precisar executar este comando em mimha linha de comando, quando for mostrar aos alunos.

```
head ~/Desktop/Gepoliano/UFSB/Arquivos/Labadorf_merged.map
#head ~/Desktop/Gepoliano/UFSB/Arquivos/Labadorf_merged.ped
```

Como mencionado, arquivos MAP e PED foram usados em estudos de Linkage, ou Ligação Genética. Linkage, Ligação Genética e Associação Biológica são conceitos relacionados. A palavre MAP, tem origem no inglês, significando mapa, indicando que este arquivo pode ser usado para localização das SNPs. PED vem de pedigree, e contém informações sobre as amostras e os genótipos de cada amostra, nas posições indicadas em MAP.

2.5) Associação Biológica pelo teste de Fisher

A Associação Biológica é detectada relacionando-se estatisticamente, as SNPs ao fenótipo de interesse. Neste caderno, o fenótipo é a Doença de Huntington. Usamos o programa PLINK para detecção da associação. O link abaixo traz um tutorial para a utilização de PLINK, o qual consultei.

https://www.staff.ncl.ac.uk/heather.cordell/msc2010casecon.html

O seguinte comando permite a construção do objeto plink.assoc.fisher, o qual contém os p-valores da Associação Biológica de cada SNP ao fenótipo:

```
cd ~/Desktop/Gepoliano/UFSB/Arquivos/

/Users/gepolianochaves/anaconda3/bin/plink \
--noweb \
--ped ~/Desktop/Gepoliano/UFSB/Arquivos/Labadorf_merged.ped \
--map ~/Desktop/Gepoliano/UFSB/Arquivos/Labadorf_merged.map \
```

```
--allow-no-sex \
--assoc fisher
```

Executar o comando acima, permitiu a construção do arquivo plink.assoc.fisher, o qual contém os p-valores da Associação Biológica. Vamos agora executar um comando que conta quantas das SNPs detectadas são significantivas, isto é, possuem um p-valor menor que 5% (0.05):

```
cat ~/Desktop/Gepoliano/UFSB/Arquivos/plink.assoc.fisher | \
awk 'header = $0; $8 < 0.05' | wc -l

head ~/Desktop/Gepoliano/UFSB/Arquivos/plink.assoc.fisher
```

```
##  2773261
##  CHR                SNP         BP   A1      F_A      F_U   A2           P           OR
##    0      GL000241.1:73         73    T      0.1        0    C    0.006319           NA
##    0      GL000208.1:77         77    C     0.05        0    T     0.08251           NA
##    0      GL000208.1:84         84    A     0.05        0    C     0.08251           NA
##    0      GL000208.1:98         98    A     0.05  0.02041    C      0.5794        2.526
##    0     GL000195.1:122        122    T        0  0.03061    A      0.5563            0
##    0     JH636052.4:135        135    A        0  0.02041    G           1            0
##    0     JH636052.4:140        140    A        0  0.04082    T       0.323            0
##    0     GL000220.1:140        140    G      0.1  0.02041    A     0.05851        5.333
##    0     GL000220.1:145        145    G      0.1  0.02041    A     0.05851        5.333
```

Agora, queremos preservar o cabeçalho ao filtrar o arquivo plink.assoc.fisher. O comando foi encontrado no seguinte link:

https://unix.stackexchange.com/questions/356080/get-all-rows-having-a-column-value-greater-than-a-threshold

A linha excedente resultante do comando abaixo, na contagem de linhas no arquivo, é o cabeçalho:

```
awk -F" " 'NR==1{print;next}$8<0.05' \
~/Desktop/Gepoliano/UFSB/Arquivos/plink.assoc.fisher | wc -l
```

```
##    82649
```

3) Arquivo Resultante da Associação Biológica feita por PLINK

Os objetos plink.assoc.fisher e final_manhattan_005_filtered.txt contêm informações equivalentes e foram construídos usando o programa PLINK, que calcula a associação estatística ou matemática, do fenótipo (Doença de Huntington) com o genótipo (mutações ou variações genéticas), como demonstrado com plink.assoc.fisher na sessão 2.5 acima. O p-valor, encontrado na coluna 8 do arquivo produzido por PLINK, informa uma estimativa de probabilidade de que a associação seja ao acaso. Quanto menor o p-valor, menos provável será que a associação biológica seja ao acaso, e menos provável será que de fato, não haja relação entre nosso achado (as SNPs) e o fenótipo. Em geral, é comum definir-se que uma associação não seja ao acaso, para p-valores menores que 5%, ou 0.05. Estudos Amplos de Associação do Genoma, em inglês *Genome Wide Association Studies* (GWAS), requerem p-valores mais estringentes que 5%, porém a fundamantação teórica para esta necessidade foge ao escopo do presente curso. Aqui, seguiremos a tradicional noção de que um p-valor de até 5% seja suficiente para considerar a Assiciação Biológica estatisticamente significante. Quando a Associação Biológica não é ao acaso, dizemos que a SNP está significativamente associada ao fenótipo, ou, simplesmente, que a SNP é significante. Assim, no objeto final_manhattan_005_filtered.txt, extraído do arquivo resultante da análise de PLINK, todas as SNPs têm p-valor menor que 0.05.

Visualize a constituição do objeto final_manhattan_005_filtered.txt, com o comando head em bash:

```
head ~/Desktop/Gepoliano/UFSB/Arquivos/final_manhattan_005_filtered.txt
```

```
## CHR SNP BP P
## 0 GL000241.1:73 73 0.006319
## 0 GL000195.1:1810 1810 0.006319
## 0 GL000195.1:1857 1857 0.005316
## 0 GL000220.1:3153 3153 0.02174
## 0 GL000224.1:3213 3213 0.006319
## 0 GL000220.1:4280 4280 0.001352
## 0 GL000220.1:4368 4368 0.02255
## 0 GL000220.1:5700 5700 0.03565
## 0 GL000219.1:7675 7675 0.006319
```

Note que o arquivo final_manhattan_005_filtered.txt possui as colunas CHR, SNP, BP e P. Estas colunas são, respectivamente: Cromossomo (CHR), Identificador da SNP (SNP), Posição da SNP, em pares de bases (BP), e p-valor (P) da associação com o traço fenotípico de interesse. Estas quatro colunas estavam presentes no arquivo plink.assoc.fisher, o qual obtivemos do processamento por PLINK na sessão 2.5. Sobre a localização das coordenadas genômicas, é muito importante observar que o genoma de qualquer organismo pode ser entendido como uma sequencia de nucleotídeos. Se cada nucleotídeo for contado, podemos então ter uma sequência que vai do primeiro ao último par de bases que forma um cromossomo. Assim, a coluna BP, no codigo anterior, informa a posição do par de bases na sequência genômica do cromossomo ou arquivo em questão. Se usarmos o comando grep e as iniciais correspondentes ao cromossomo 4, assim como o comando head, poderemos isolar partes do arquivo final_manhattan_005_filtered.txt correspondentes a localizações de variantes associadas à Doença de Huntington, no cromossomo 4:

```
grep chr4 ~/Desktop/Gepoliano/UFSB/Arquivos/final_manhattan_005_filtered.txt | head
```

```
## 4 chr4:128096 128096 0.01335
## 4 chr4:516586 516586 0.007626
## 4 chr4:523979 523979 0.002496
## 4 chr4:527217 527217 0.02174
## 4 chr4:566177 566177 0.0008988
## 4 chr4:578679 578679 0.01621
## 4 chr4:578790 578790 0.01037
## 4 chr4:579307 579307 0.03346
## 4 chr4:580259 580259 0.01904
## 4 chr4:585318 585318 0.03176
```

4) Plot de Manhattan (azul e laranja)

4.1) Plotar todos os cromossomos e realçar variantes no cromossomo 6

Visto termos noção de que a pipeline de identificação de variantes genéticas produz o arquivo VCF, contendo todas as variantes genéticas identificadas em um indivíduo, mostramos como mesclar os arquivos VCF de todos os indivíduos de nosso estudo, na sessão 2.2. Após termos mesclado os arquivos VCF de cada paciente, usamos o programa vcftools, na sessão 2.3, para extrair os formatos MAP e PED do arquivo VCF. Sendo os arquivos MAP e PED compatíveis com os requerimentos do programa PLINK para calcular a Associação Biológica entre as variantes genéticas e nosso fenótipo em estudo, usando o teste de Fisher, construímos então o objeto plink.assoc.fisher na sessão 2.5. Como mencionado anteriormente, o arquivo final_manhattan_005_filtered.txt possui colunas que foram isoladas do arquivo plink.assoc.fisher. Este isolamento foi necessário, pois o pacote Qqman, no R, requer apenas as colunas CHR, SNP, BP e P, definidas acima, para a visualização do plot de Manhattan. O plot de Manhattan mostra a coordenada das variantes genéticas no eixo X, e o p-valor da associação com o fenótipo, no eixo Y. A Figura 3A no artigo científico de

Chaves, Stanley e Pourmand, foi produzida com código similar ao encontrado abaixo, plotando todas as cerce de 80.000 SNPs associadas à Doença de Huntington anotadas no arquivo final_manhattan_005_filtered.txt. Podemos então proceder ao plot de Manhattan:

```
library("qqman")

GWAS_TABLE<-read.table(
  "~/Desktop/Gepoliano/UFSB/Arquivos/final_manhattan_005_filtered.txt",
  sep = " ",
  header = T)

GWAS_TABLE_Ommit<-na.omit(GWAS_TABLE)

## SNPs do cromossomo 6 anserem realçadas em verde:
highlight_these<-c("chr6:29910581","chr6:30038116","chr6:31244366","chr6:31244384",
                   "chr6:31244410","chr6:31244422","chr6:31244423","chr6:31244988",
                   "chr6:31245014","chr6:31245060","chr6:31245089","chr6:31245092",
                   "chr6:31245095","chr6:31322996","chr6:32487145","chr6:32489953",
                   "chr6:32545604","chr6:32545659","chr6:32547251","chr6:32605318",
                   "chr6:32608589","chr6:32610749","chr6:32623713","chr6:32624407",
                   "chr6:32624465","chr6:32625022","chr6:32625283","chr6:32625442",
                   "chr6:32627038","chr6:32627039","chr6:32627044","chr6:32627958",
                   "chr6:32628606","chr6:32631816","chr6:32632770","chr6:32632777")

manhattan(GWAS_TABLE_Ommit,
          highlight = highlight_these,
          col = c("blue4", "orange3"),
          genomewideline = -log10(0.05))
```

4.2) Plotar apenas SNPs do cromossomo 1

```
library(qqman)
GWAS_TABLE<-read.table("~/Desktop/Gepoliano/UFSB/Arquivos/Chr1.txt",
                       sep = " ",
                       header = T)

GWAS_TABLE_Ommit <- na.omit(GWAS_TABLE)

SNP_HIGHLIGHT<-c("rs464218", "rs2228604",
                 "rs12711447", "rs12711448",
                 "rs370088", "rs629301",
                 "rs12141363", "rs897171")

manhattan(GWAS_TABLE_Ommit,
          highlight = SNP_HIGHLIGHT,
          annotateTop = T,
          annotatePval = 0.20,
          genomewideline = -log10(0.10))
```

5) Filtrar cromossomo 4 para localizar genes *HTT* e *SORCS2*

5.1) Visualização de genes em navegador de genomas

Uma vez que conhecemos o arquivo que contém as variantes que desejamos plotar, podemos usar o comando sed, acompanhado as opções abaixo, para isolar as variantes presentes no cromossomo 4. No cromossomo 4, localizam-se os genes da huntingtina e de uma das proteínas sorlininas, a sortilina 2. O comando para o código abaixo foi encontrado com a seguinte busca:

https://stackoverflow.com/questions/9969414/always-include-first-line-in-grep

```
sed '1p;/chr4/!d' ~/Desktop/Gepoliano/UFSB/Arquivos/final_manhattan_005_filtered.txt > \
~/Desktop/Gepoliano/UFSB/Arquivos/chr4.txt
```

Os genes da HTT e da Sortilina 2 encontram-se relativamente próximos em termos de suas coordenadas genômicas, como pode ser visto na seguinte figura do navegador de genomas Ensembl:

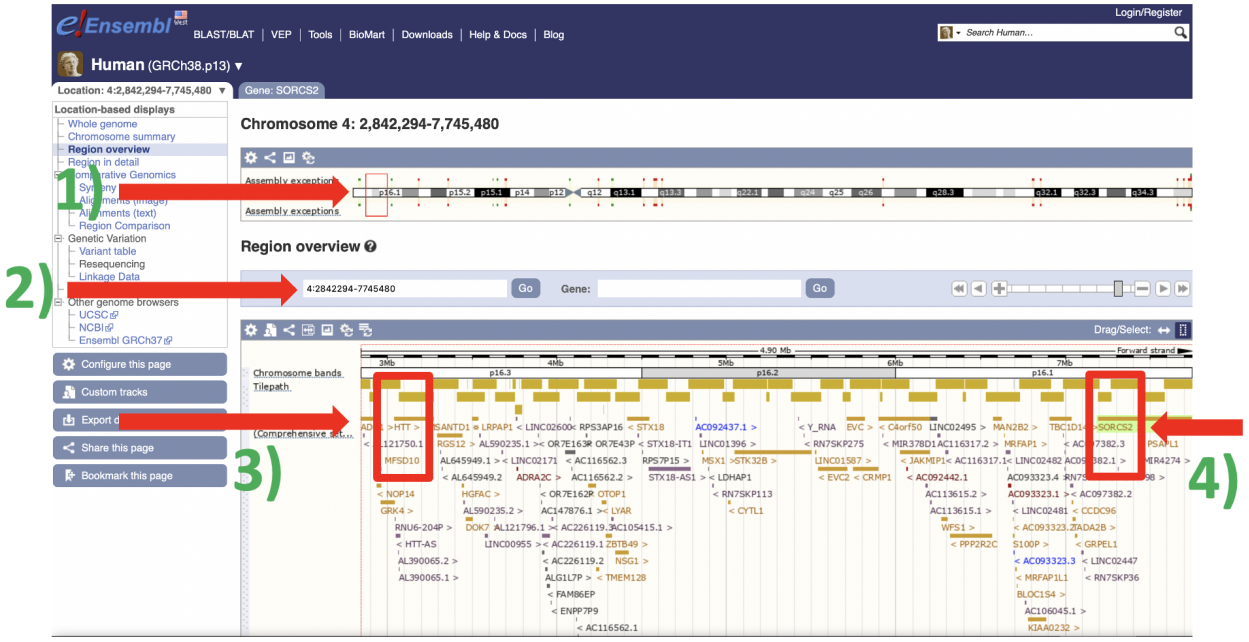


Figure 6: Localização relativa dos genes HTT e SORCS2 no genoma humano, visualizada no Browser de genomas Ensembl. 1) Visualizão global do cromossomo; 2) Posição ou coordenadas gênicas; 3) HTT; 4) SORCS2.

5.2) Visualização de cromossomo 4

5.2.1) Visualização sem realçamento de SNPs

Para plotarmos as SNPs do cromossomo 4 que acabamos de criar, basta repetir um comando parecido com o executado na sessão 4.2:

```
GWAS_TABLE<-read.table("~/Desktop/Gepoliano/UFSB/Arquivos/chr4.txt", sep = " ", header = T)

GWAS_TABLE_Ommit <- na.omit(GWAS_TABLE)
```

```
manhattan(GWAS_TABLE_Ommit, annotateTop = T, annotatePval = 0.20, genomewideline = -log10(0.10))
```

Na visualização das localizações genômicas no navegador de genomas Ensembl, vemos que cada gene possui um par de coordenadas. Por exemplo, as coordenadas para HTT são: 4:3041422-3243960 e para SORCS2, Chr4:7192538-7742836. Seria interessante então, se pudéssemos realçar a posição das variantes entre estes genes em nosso Manhattan plot. Para tanto, usaremos agora, o comando awk, para isolar 300 posições de variantes entre os genes. Ao isolarmos estas posições, usamos o comando sed, para incluirmos os caracteres "," entre as SNPs, visto ser este o padrão a ser oferecido à função que plota o plot Manhattan:

```
awk '$3 > 3041422' ~/Desktop/Gepoliano/UFSB/Arquivos/chr4.txt | \
awk '$3 < 7742836' | \
awk '{print $2}' | \
sed -n -e 'H;${x;s/\n/","/g;s/^,//;p;}'
```

5.2.2) Visualização com realçamento de SNPs entre *HTT* e *SORCS2*

Copiamos então, as variantes provenientes do código acima, para realçarmos no Manhattan plot:

```
library(qqman)

GWAS_TABLE<-read.table("~/Desktop/Gepoliano/UFSB/Arquivos/chr4.txt",
                       sep = " ", header = T)

GWAS_TABLE_Ommit <- na.omit(GWAS_TABLE)

SNP_HIGHLIGHT <- c("chr4:3043512","chr4:3043513","chr4:3048207","chr4:3224216",
                   "chr4:3231772","chr4:3233844","chr4:3235081","chr4:3235084",
                   "chr4:3236881","chr4:3236883","chr4:3241845","chr4:3243804",
                   "chr4:3263138","chr4:3265130","chr4:3265710","chr4:3314646",
                   "chr4:3380088","chr4:3409359","chr4:3411110","chr4:3415336",
                   "chr4:3415378","chr4:3438643","chr4:3446091","chr4:3449886",
                   "chr4:3473066","chr4:3476809","chr4:3480439","chr4:3487151",
                   "chr4:3496058","chr4:3496110","chr4:3506933","chr4:3508752",
                   "chr4:3510957","chr4:3512690","chr4:3517746","chr4:3518190",
                   "chr4:3529671","chr4:3532327","chr4:3533066","chr4:3746133",
                   "chr4:3747842","chr4:3748134","chr4:3765305","chr4:3765336",
                   "chr4:3944253","chr4:3944752","chr4:3944888","chr4:3946166",
                   "chr4:3946175","chr4:3969218","chr4:4051294","chr4:4076788",
                   "chr4:4103104","chr4:4103105","chr4:4109198","chr4:4109210",
                   "chr4:4240627","chr4:4242705","chr4:4243668","chr4:4245210",
                   "chr4:4245510","chr4:4245513","chr4:4245591","chr4:4245926",
                   "chr4:4245929","chr4:4246109","chr4:4246433","chr4:4246453",
                   "chr4:4246457","chr4:4246497","chr4:4249414","chr4:4249415",
                   "chr4:4249484","chr4:4271623","chr4:4275306","chr4:4304749",
                   "chr4:4318931","chr4:4318970","chr4:4319564","chr4:4319728",
                   "chr4:4319750","chr4:4322078","chr4:4709657","chr4:4732282",
                   "chr4:4789635","chr4:4822960","chr4:4824890","chr4:4825092",
                   "chr4:4825180","chr4:4865316","chr4:4865321","chr4:5018702",
                   "chr4:5812778","chr4:5814082","chr4:5833660","chr4:5833899",
                   "chr4:5835541","chr4:5851205","chr4:5862752","chr4:5862938",
                   "chr4:5862943","chr4:5901873","chr4:5905499","chr4:5906287",
                   "chr4:6018891","chr4:6019046","chr4:6020190","chr4:6020367",
                   "chr4:6025638","chr4:6025656","chr4:6025766","chr4:6026058",
```

```
                  "chr4:6083488","chr4:6204935","chr4:6235553","chr4:6237142",
                  "chr4:6238466","chr4:6239906","chr4:6240929","chr4:6245022",
                  "chr4:6245618","chr4:6245732","chr4:6245915","chr4:6246075",
                  "chr4:6246373","chr4:6246959","chr4:6290594","chr4:6292020",
                  "chr4:6294095","chr4:6298375","chr4:6316092","chr4:6321396",
                  "chr4:6324647","chr4:6324785","chr4:6327669","chr4:6328354",
                  "chr4:6328507","chr4:6333130","chr4:6333559","chr4:6333669",
                  "chr4:6335966","chr4:6435341","chr4:6435486","chr4:6435926",
                  "chr4:6437191","chr4:6437197","chr4:6457121","chr4:6457131",
                  "chr4:6457132","chr4:6568390","chr4:6570032","chr4:6570768",
                  "chr4:6596360","chr4:6613252","chr4:6613462","chr4:6620991",
                  "chr4:6624771","chr4:6626154","chr4:6641969","chr4:6642090",
                  "chr4:6644466","chr4:6644467","chr4:6644468","chr4:6647889",
                  "chr4:6648300","chr4:6662665","chr4:6663319","chr4:6663715",
                  "chr4:6674554","chr4:6678553","chr4:6678599","chr4:6690535",
                  "chr4:6698664","chr4:6698667","chr4:6698706","chr4:6720572",
                  "chr4:6911679","chr4:6985889","chr4:6987394","chr4:7002344",
                  "chr4:7004495","chr4:7004506","chr4:7005196","chr4:7005199",
                  "chr4:7024077","chr4:7024398","chr4:7029430","chr4:7031064",
                  "chr4:7044357","chr4:7044380","chr4:7048842","chr4:7052115",
                  "chr4:7055253","chr4:7064243","chr4:7067765","chr4:7073187",
                  "chr4:7074027","chr4:7677967","chr4:7701947","chr4:7702795",
                  "chr4:7703505","chr4:7703807","chr4:7704795","chr4:7704818",
                  "chr4:7709703","chr4:7712150","chr4:7714490","chr4:7733843",
                  "chr4:7735162","chr4:7735164","chr4:7736103","chr4:7736112")

manhattan(GWAS_TABLE_Ommit,
          highlight = SNP_HIGHLIGHT,
          annotateTop = T,
          annotatePval = 0.20,
          genomewideline = -log10(0.0001))
```

O plot Manhattan mostra a existência de muitas variantes entre as regiões dos genes *HTT* e *SORCS2*, significamente associadas à Doença de Huntington. O fato de a sortilina estar localizada tão próxima ao locus identificado por análises de Linkage Disequilibrium, ilustra ainda mais fortemente o envolvimento de sortilinas, e portanto, proteínas envolvidas na captação de glicose da corrente sanguínea, nos fenótipos relativos ao diabetes nesta doença, corroborando os dados de expressão protéica nas células de *Rattus norvegicus* e validando a observação do modelo celular, em dados humanos, quanto à Associação Biológica entre sortilinas e Doença de Huntington, tanto no modelo celular de *R. norvegicus*, quanto nos dados de transcriptômica em humanos. Deve-se notar também, que a mais significante na associação com a doença também apareceu nesta região, suportando ainda mais a noção da associação com a Doença de Huntington, e que há muito mais SNPs nesta região, sendo a visualização limitada a cerca de 300 SNPs pela ferramenta de visualização.