

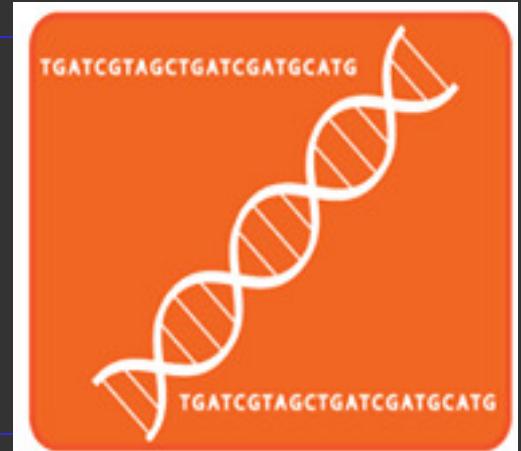
CURSO DE CURTA DURAÇÃO - 2017

BIOINFORMÁTICA

BioME – CENTRO MULTIUusuÁRIO DE BIOINFORMÁTICA - UFRN

NEXT GENERATION SEQUENCING

Análise de Dados de Sequenciadores de Segunda Geração



Prof. Dr. JORGE ESTEFANO SANTANA DE SOUZA

E-mail: jorge@imd.ufrn.br



Bioinformatics
Multidisciplinary
Environment



Centro
Multiusuário
de Bioinformática



Instituto de
Bioinformática e
Biotecnologia

Introdução à Bioinformática:

Habilidades, ferramentas iniciais e formatos

Agenda:

- 1 - Introdução de Bioinformática.**
- 2 - Estrutura mínima.**
- 3 - Principais bancos de dados disponíveis.**
- 4 - Linux.**
- 5 - MySQL.**
- 6 - Perl.**
- 7 - R.**

Passos para ser um bioinformata:

- Saber Estatística

- Conhecer os principais testes estatísticos
- Saber onde e quando aplicar estatística
- Saber usar a linguagem R

- Saber Computação

- Pensar como computador (Lógica, lógica de programação)
- Programar bem em 2 ou 3 linguagens (Perl, Python, PHP e R)
- Banco de dados (linguagem SQL) (MySQL e / ou Postgre)
- Linux (Shell script) (bash e linguagem shell script)
- Conhecer um pouco de redes e adm. de sistemas
- Conhecer um pouco de Hardware
- Gostar de tecnologia e estar atualizado

- Acompanhar a literatura técnica

- Nature / Science / Cell
- Gen. Research, PLoS, Gen. Biology
- Bioinformatics, BMC Bioinf., PLoS Comp. Bio.
- Revistas técnicas do foco do seu estudo

- Ser organizado

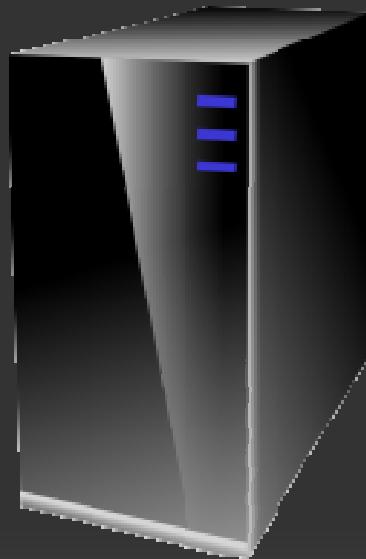
- Anotação eficientemente (caderno de lab, físico ou virtual)
- Ter uma toolbox (todos os seus programas organizados e anotados)

- Saber Biologia

- Biologia molecular
- Pensar como biólogo (foco, pequena escala)
- Conhecer o modelo / fenômeno do seu estudo

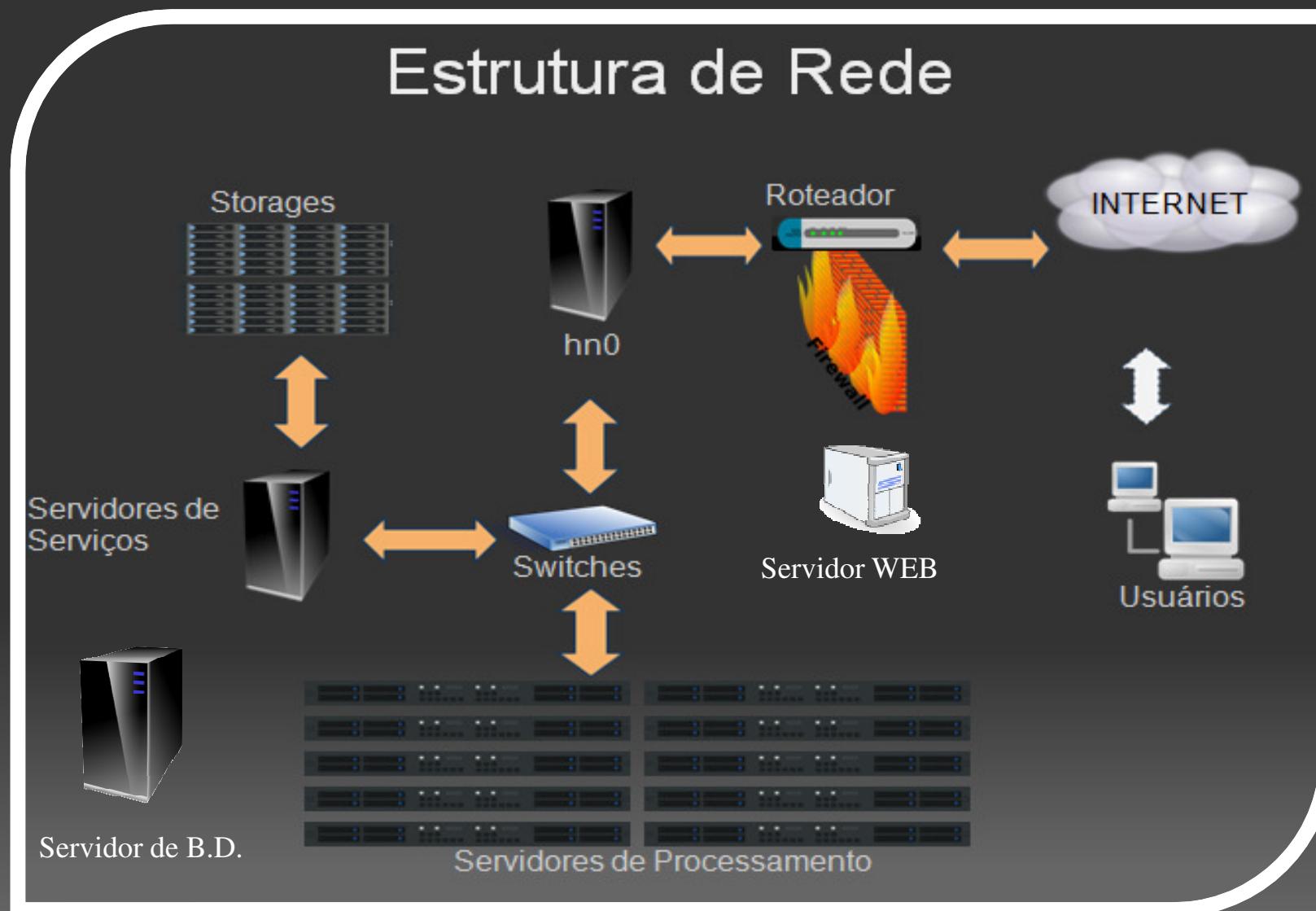
Estrutura mínima.

O Bioinformata errante.



- 8 cores de Processamento.
- $\geq 2,5$ GHz
- 32 Gb Ram.
- 3 TB HD.
- **Sistema Operacional**
 - Linux - Fedora
 - Linux - CentOS
- **Banco de dados**
 - Servidor (MySQL)

Estrutura mínima.



Agenda:

- 1 - Introdução de Bioinformática.
- 2 - Estrutura mínima.
- 3 - Principais bancos de dados disponíveis.**
- 4 - Linux.
- 5 - MySQL.
- 6 - Perl.
- 7 - R.
- 8 - Formatos e qualidade.

Repositórios de dados públicos

<http://www.ncbi.nlm.nih.gov/>

NCBI Resources How To Sign in to NCBI

All Databases Search

NCBI Home Resource List (A-Z) All Resources Chemicals & Bioassays Data & Software DNA & RNA Domains & Structures Genes & Expression Genetics & Medicine Genomes & Maps Homology Literature Proteins Sequence Analysis Taxonomy Training & Tutorials Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | Research | RSS Feeds

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

Education Resources

Central point of access for help of protein, and molecular modeling, teaching materials, news, outlets, and other educational resources.



II 1 2 3 4 5 6 7 8

Popular Resources

PubMed Bookshelf PubMed Central PubMed Health BLAST Nucleotide Genome SNP Gene Protein PubChem

NCBI Announcements

New version of Genome Workbench available 06 Sep 2012 An integrated, downloadable application for viewing and analyzing sequence data

NCBI's July Newsletter is on the Bookshelf 13 Aug 2012

Repositórios de dados públicos

<http://genome.ucsc.edu/>

UCSC Genome Bioinformatics

[Genomes](#) - [Blat](#) - [Tables](#) - [Gene Sorter](#) - [PCR](#) - [VisiGene](#) - [Session](#) - [FAQ](#) - [Help](#)

[Genome Browser](#)
[ENCODE](#)
[Neandertal](#)
[Blat](#)
[Table Browser](#)
[Gene Sorter](#)
[In Silico PCR](#)
[Genome Graphs](#)
[Galaxy](#)
[VisiGene](#)
[Utilities](#)
[Downloads](#)
[Release Log](#)

About the UCSC Genome Bioinformatics Site

Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to the [ENCODE](#) and [Neandertal](#) projects.

We encourage you to explore these sequences with our tools. The [Genome Browser](#) zooms and scrolls over chromosomes, showing the work of annotators worldwide. The [Gene Sorter](#) shows expression, homology and other information on groups of genes that can be related in many ways. [Blat](#) quickly maps your sequence to the genome. The [Table Browser](#) provides convenient access to the underlying database. [VisiGene](#) lets you browse through a large collection of *in situ* mouse and frog images to examine expression patterns. [Genome Graphs](#) allows you to upload and display genome-wide data sets.

The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the Center for Biomolecular Science and Engineering ([CBSE](#)) at the University of California Santa Cruz ([UCSC](#)). If you have feedback or questions concerning the tools or data on this website, feel free to contact us on our [public mailing list](#).

News  [News Archives ▶](#)

To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the [genome-announce](#) mailing list.

17 September 2012 – Retiring the Proteome Browser

We are announcing the imminent retirement of the Proteome Browser. We introduced the Proteome Browser in 2003 to provide access to protein-specific information independent of the genomic details presented in the Genome Browser. Since then we've incorporated more of this information in the UCSC Genes details page accessible from the Genome Browser, and in the columns of the Gene Sorter. Since very few people are still accessing the Proteome Browser, we've decided to retire it to focus our work on these other two tools. Access to the databases that supported the Proteome Browser (uniProt and proteome) will still be available through the Table Browser and our public MySQL server.

Repositórios de dados públicos

<http://www.1000genomes.org/>

The screenshot shows the homepage of the 1000 Genomes Project. At the top, there is a dark banner with the project's name "1000 Genomes" in large yellow letters and the subtitle "A Deep Catalog of Human Genetic Variation" in smaller white letters. To the right of the banner is a decorative background image of numerous colorful, glowing DNA helixes against a dark background. Below the banner is a navigation bar with links: Home, About, Data, Analysis, Participants, Contact, Browser, Wiki, and FTP search. To the right of the navigation bar is a search input field and a "Search" button. The main content area has a light gray background. On the left, there is a section titled "LATEST ANNOUNCEMENTS" with a sub-section for "WEDNESDAY OCTOBER 31, 2012" announcing "An integrated map of genetic variation from 1092 human genomes". It describes the publication in Nature and provides a link to the ftp site. Below this is another announcement for "MONDAY JANUARY 06, 2014" about cell lines and DNA samples available from the Coriell Cell Repository. On the right side, there is a "NAVIGATION" sidebar with a link to "Frequently Asked Questions". Below it is a "LINKS" sidebar with four items: "All Project Announcements" (megaphone icon), "Sample and Project Information" (DNA helix icon), "Media Archive" (newspaper icon), and "Download the 1000" (star icon).

1000 Genomes
A Deep Catalog of Human Genetic Variation

Home About Data Analysis Participants Contact Browser Wiki FTP search Search

LATEST ANNOUNCEMENTS

WEDNESDAY OCTOBER 31, 2012

An integrated map of genetic variation from 1092 human genomes

The Phase 1 publication, [An Integrated map of genetic variation from 1092 human genomes](#) is now available from [Nature](#) and can be downloaded directly from the [ftp site](#). The paper is distributed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported licence. Please share our paper appropriately.

All the data files associated with this paper can be found in our [phase1 analysis results directory](#).

Recent project announcements

MONDAY JANUARY 06, 2014

Cell lines and DNA samples and panels are available from the Coriell Cell Repository

All the samples from the 1000 genomes are available as lymphoblastoid cell lines (LCLs) and LCL derived DNA from the [Coriell Cell Repository](#) as part of the [NHGRI Catalog](#). In addition Standard Population DNA Panels for the 1000 Genomes and HapMap projects are available at \$1000 or less each.

NAVIGATION

- Frequently Asked Questions

LINKS

- All Project Announcements
- Sample and Project Information
- Media Archive
- Download the 1000

Repositórios de dados públicos

<http://hapmap.ncbi.nlm.nih.gov/>



International HapMap Project

[Home](#) | [About the Project](#) | [Data](#) | [Publications](#) | [Tutorial](#)

[中文](#) | [English](#) | [Français](#) | [日本語](#) | [Yoruba](#)

The International HapMap Project is a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals. See "[About the International HapMap Project](#)" for more information.

Project Information	News
About the Project HapMap Publications HapMap Tutorial HapMap Mailing List HapMap Project Participants	<ul style="list-style-type: none">• 2013-06-14: HapMap data conversion tool There are several inquiries for a conversion tool to convert HapMap data into the VCF format. Please take a look of The Genome Analysis Toolkit (by Broad Institute).• 2012-12-06: Downtime for hardware maintenance From December 15 - 16, Hapmap site will be taken offline for an internal hardware maintenance. Sorry for the inconvenience.• 2011-06-13: HapMap help desk announcement There was a problem with the HapMap help desk system. In the past several weeks, emails sent to hapmap-help@ncbi.nlm.nih.gov did not reach the help desk, and thus user requests were not addressed. Please resend your email request if you sent emails to the HapMap help desk in the past several weeks. Sorry for the inconvenience.• 2011-04-20: Hapmap help desk service interruption notice There will be no help desk support from 05/03/2011 to 05/23/2011. Sorry for the inconvenience.• 2011-02-02: Haploview issues with rel 28 data

Repositórios de dados públicos

<https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp>

The screenshot shows the homepage of The Cancer Genome Atlas (TCGA) Data Portal. At the top, there are logos for the National Cancer Institute and the National Human Genome Research Institute. The main header features the text "The Cancer Genome Atlas Data Portal" next to a globe icon, with the tagline "Understanding genomics to improve cancer care". Below the header is a navigation bar with links for Home, Query the Data, Download Data, Tools, About the Data, and Publication Guidelines. The "Home" link is currently selected. The main content area has a blue background and contains sections for "TCGA Data Portal Overview", "Available Cancer Types", and "Announcements". The "TCGA Data Portal Overview" section explains the portal's purpose and mentions the CGHub and dbGaP repositories. The "Available Cancer Types" table lists four types: Acute Myeloid Leukemia [LAML], Bladder Urothelial Carcinoma [BLCA], and Brain Lower Grade Glioma [LGG]. The "Announcements" section includes details about planned downtime on November 13th and completed maintenance on October 10th.

National Cancer Institute

National Human Genome Research Institute

The Cancer Genome Atlas Data Portal
Understanding genomics to improve cancer care

Home | Query the Data | Download Data | Tools | About the Data | Publication Guidelines | TCGA Home | Contact Us | For the Media

Home

TCGA Data Portal Overview

We provide 3 ways to download data: The Cancer Genome Atlas (TCGA) Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA. It contains clinical information, genomic characterization data, and high-throughput sequencing analysis of the tumor genomes.

The TCGA Data Portal does not host lower levels of sequence data. NCI's [Cancer Genomics Hub \(CGHub\)](#) is the new secure repository for storing, cataloging, and accessing sequence related data. New users must still apply for authorized access through NCBI's [Database of Genotypes and Phenotypes \(dbGaP\)](#).

[Query the Data](#)

[Download Data](#)

Search summarized data for genes, patients and pathways

Choose from three ways to download data

Available Cancer Types	# Patients with Samples	# Downloadable Tumor Samples	Date Last Updated (mm/dd/yy)
Acute Myeloid Leukemia [LAML]	202	200	11/05/12
Bladder Urothelial Carcinoma [BLCA]	153	126	11/05/12
Brain Lower Grade Glioma [LGG]	222	181	11/07/12

Announcements

11/9/2012 - DCC Planned Downtime

On Tuesday, November 13th, the DCC will perform a software release starting at 8AM and lasting for approximately two hours. The TCGA Data Portal will not be available during this time. A complete list of the issues addressed in this release can be found on the [TCGA Wiki](#).

If you have any questions or concerns about this release, contact tcga-dcc-binf-l@list.nih.gov.

10/10/2012 - DCC Maintenance Completed

The maintenance scheduled for this morning has been completed. If you notice any problems, contact the DCC at tcga-dcc-binf-l@list.nih.gov.

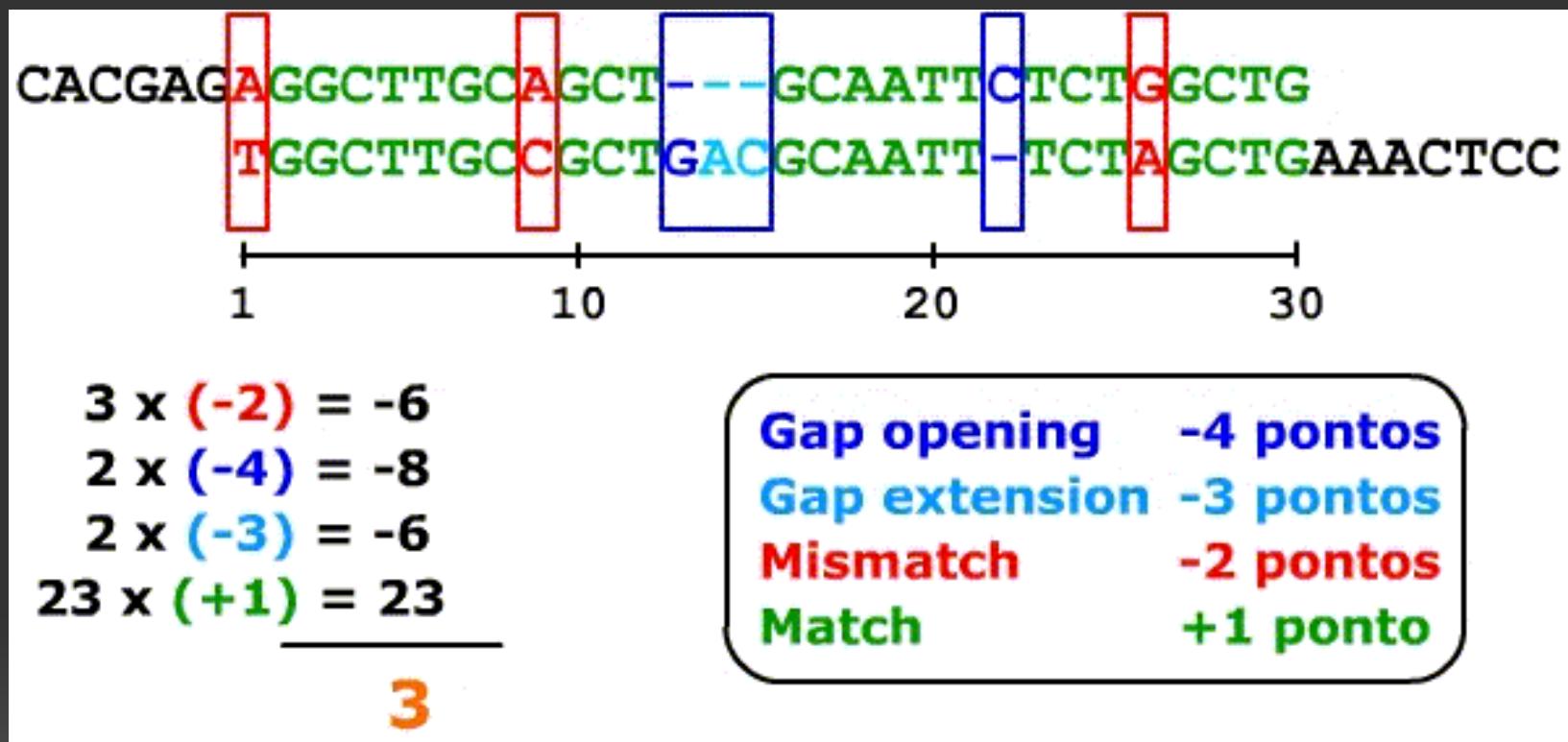
[See all announcements](#)

Repositórios de dados públicos

Principais métodos de consulta/busca aos bancos públicos:

- Web;
- FTP;
- APIs;
- Mysql;
- SSH;
- BioPerl, BioPython, etc;
- R;
- Busca por palavras chaves;
- Busca por padrões;
- Alinhamento.

Elementos de um alinhamento



Bioinformática na WEB.

Sequência de exemplo:

```
>SEQ-EXE
TTAAAAGAGTTAAGGACTCTGAAGATGTACCTATGGCCTAGTAG
GAAATAAAATGTGATTGCCTCTAGAACAGTAGACACAAAAACAGG
CTCAGGACTTAGCAAGAAGTTATGGAATTCTTTATTGAAACAT
CAGCAAAGACAAGACAGGGTGTGATGATACTTCTATACATTAGT
TCGAGAAATTGAAAACATAAAGAAAAGATGAGCAAAGATGGTAA
AAAGAAGAACAAAGAAGTCAAAGACAAAGTGTGAATTATGTAAAT
ACAATTGTACTTTCTTAAGGCATACTAGTACAAGTGGTAAT
TTTGTAATTACACTAAATTATTAGCATTGTTAGCATTACC
TAATTTCCTGCTCCATGCAGACTGTTAGCTTTACCTTAAA
TGCTTATTAAA
```

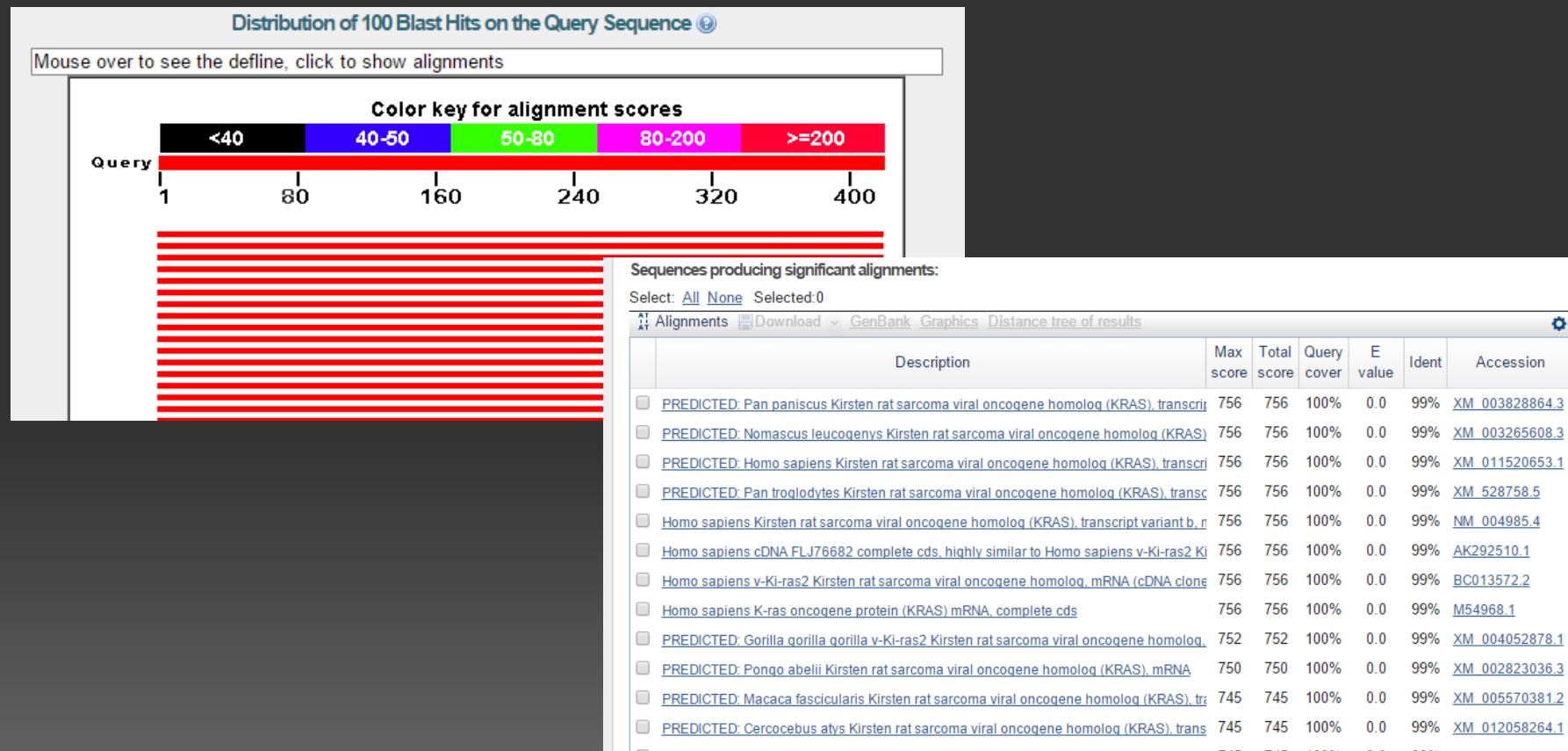
Bioinformática na WEB.

Identificar as espécies das quais as sequências são provenientes.

The screenshot shows the NCBI BLAST search interface. At the top, there's a navigation bar with links for Home, Recent Results, Saved Strategies, Help, My NCBI, Sign In, and Registered. Below the navigation bar, the title "Basic Local Alignment Search Tool" and "NCBI/BLAST/blastn suite" are displayed. The main section is titled "Standard Nucleotide BLAST". It features a "Enter Query Sequence" input field containing a nucleotide sequence (SEQ-EXE) starting with >SEQ-EXE. To the right of the sequence input is a "Query subrange" section with "From" and "To" fields. Below the sequence input, there are fields for "Or, upload file" (with a button to "Escolher arquivo" which says "Nenhum arquivo selecionado"), "Job Title" (set to "SEQ-EXE"), and a "Descriptive title" input field. There's also a checkbox for "Align two or more sequences". The "Choose Search Set" section includes a "Database" dropdown set to "Human genomic + transcript", an "Organism" dropdown set to "Nucleotide collection (nr/nt)", and various optional settings like "Exclude" and "Limit to". At the bottom, there's an "Entrez Query" input field, a YouTube link, and a "Create custom database" link.

Bioinformática na WEB.

Identificar as espécies das quais as sequências são provenientes.



Bioinformática na WEB.

strand (fita) positivo e negativo.

The screenshot shows the Human BLAT Search interface. At the top, there's a navigation bar with links for Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, and Help. Below that is a sub-navigation bar for the Human BLAT Search, with tabs for BLAT Search Genome, BLAT Search Assembly, and BLAT Search Results.

Search parameters are set to Genome: Human, Assembly: Feb. 2009 (GRCh37/hg19), Query type: BLAT's guess, Sort output: query, score, and Output type: hyperlink.

A sequence snippet is displayed:

```
>SEQ-EXE
TTAAAAGAGTTAAGGACTCTGAAGATGTACCTATGGTCTAGTAGGAAATAATGTGATTGCCTCTAGAACAGTAG
ACACAAACAGGGCTCAGGACTTAGCAAGAAGTTATGAAATTCTTTATTGAAACATCAGCAAAGACAAGACAGGGTG
TTGATGATACTTCTATACATTAGTCGAGAAATTGAAACATAAAGAAAAGATGAGCAAAGATGGTAAAAAGAAAGAA
CAAGAAGTCAAAGACAAAGTGTAAATTGTAAATACAATTGTACTTTCTTAAGGCATACTAGTACAAGTGGT
AATTTTGTACATTACACTAAATTATTAGCATTTGTTAGCATTACCTAATTTCCTGCTCCATGCAGACTGTT
AGCTTTACCTTAATGCTTATTTAAA
```

The main content area is titled "BLAT Search Results". It includes a link to "Go back to chr12:25362579-25378699 on the Genome Browser".

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	SEQ-EXE	412	1	418	418	99.6%	12	-	25378699	16121	
browser details	SEQ-EXE	345	1	418	418	91.9%	6	+	54635690	54636082	393
browser details	SEQ-EXE	27	202	238	418	83.4%	1	+	93633768	93633802	35
browser details	SEQ-EXE	25	201	228	418	84.7%	1	+	48828463	48828488	26
browser details	SEQ-EXE	21	200	220	418	100.0%	3	+	29133833	29133853	21
browser details	SEQ-EXE	20	281	300	418	100.0%	2	-	229809517	229809536	20
browser details	SEQ-EXE	20	45	64	418	100.0%	2	-	52604417	52604436	20

Bioinformática na WEB.

Determinar se são sequências de DNA ou RNA.

Alignment of SEQ-EXE

SEQ-EXE
Human chr12
block1
block2
together

00000001 taaaaagagttaaggactctgaagatgtacatatggccttaggaaat 00000050
 <<<<<< ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| <<<<<<
 25378699 taaaaagagttaaggactctgaagatgtacatatggccttaggaaat 25378650

00000051 aatgtgatttgcctctagaacagtagacacaaaacaggctcaggactt 00000100
 <<<<<< ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| <<<<<<
 25378649 aatgtgatttgcctctagaacagtagacacaaaacaggctcaggactt 25378600

00000101 agcaagaagtatggatttcctttattgaaacatcagcaaagacaagac 00000150
 <<<<<< ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| <<<<<<
 25378599 agcaagaagtatggatttcctttattgaaacatcagcaaagacaagac 25378550

00000151 ag 00000152
 <<<<<< ||| <<<<<<
 25378549 ag 25378548

00000153 ggttgtatgtat 00000164
 <<<<<< ||||| ||||| <<<<<<
 25362845 ggttgtatgtat 25362834

00000166 cttctatacattatgtcgagaaattcgaaaacataaaagaaaaatgagca 00000215
 <<<<<< ||||| ||||| ||||| ||||| ||||| ||||| ||||| <<<<<<
 25362831 cttctatacattatgtcgagaaattcgaaaacataaaagaaaaatgagca 25362782

The figure displays genomic tracks for chromosome 12. At the top, a scale bar shows positions from 13.31 to 25.38 Mb. Below the scale bar, a red vertical line marks the position of the KRAS gene. A blue arrow points to the KRAS gene location. The main panel shows sequence alignments for various blocks (block1, block2, together) across different genomic coordinates. Red arrows point to specific alignment lines: one to the first sequence block, one to the second sequence block, and one to the KRAS gene alignment. The bottom section shows a detailed view of the KRAS gene region, including gene models for KRAS and CASC1, and a BLAT search visualization.

Bioinformática na WEB.

Identificar alelos variantes (mutações) .

```

00000166 cttctatacattagttcgagaaattcgaaaaacataaagaaaagatgagca 00000215
<<<<< ||||| ||||| ||||| ||||| ||||| <<<<<
25362831 cttctatacattagttcgagaaattcgaaaaacataaagaaaagatgagca 25362782

00000216 aagatggtaaaaagaagaacaagaagtc aaagacaaaagtgttaattatg 00000265
<<<<< ||||| ||||| ||||| ||||| <<<<<
25362781 aagatggtaaaaagaagaaaaagaagtc aaagacaaaagtgttaattatg 25362732

00000266 taaatacaatttgcacttttcttaaggcatactagtacaagtggtaat 00000315
<<<<< ||||| ||||| ||||| ||||| <<<<<
25362731 taaatacaatttgcacttttcttaaggcatactagtacaagtggtaat 25362682

00000316 ttttgtagattacactaaattttagcatttttagcattacctaatt 00000365
<<<<< ||||| ||||| ||||| <<<<<
25362681 ttttgtagattacactaaattttagcatttttagcattacctaatt 25362632

```

Oncotator

Toggle Annotations: Genomic Protein Cancer Non-Cancer Specific Columns

Show 10 entries Search:

Gene	Variant Classification	Variant Type	Hgvs Genomic Change	Hgvs Protein Change
KRAS	3'UTR	DEL	12:g.25362832_25362833delGC	
KRAS	3'UTR	SNP	12:g.25362762T>G	

Showing 1 to 2 of 2 entries Previous 1

[Download MAF](#) [Download VCF](#) [Download JSON](#)

PROVEAN

J. Craig Venter INSTITUTE

Search

PROVEAN HUMAN GENOME VARIANTS

PROVEAN Tools PROVEAN Protein

INPUT	PROTEIN SEQUENCE CHANGE								PROVEAN PREDICTION				SIFT PREDICTION	
	PROTEIN_ID	LENGTH	STRAND	CODON_CHANGE	POS	RESIDUE_REF	RESIDUE_ALT	TYPE	SCORE	PREDICTION (cutoff=-2.5)	#SEQ	#CLUSTER	SCORE	PREDICTION (cutoff=0.05)
12,25362832,GC,,	ENSP00000308495	188	-1					Frameshift	NA	NA	366	30	NA	NA
	ENSP00000452512	75	-1					Frameshift	NA	NA	481	30	NA	NA
12,25362762,T,G	ENSP00000308495	188	-1	AAG AA[A/C]AAG	178	K	N	Single AA Change	-0.71	Neutral	366	30	0.250	Tolerated
	ENSP00000452512	75	-1	AAG AA[A/C]AAG	65	K	N	Single AA Change	-0.31	Neutral	481	30	0.000	Damaging

<https://digitalworldbiology.com/tutorial/blast-for-beginners>

The screenshot shows a website header with a colorful logo, followed by navigation links: Home, Products ▾, Services, Community ▾, Blog, and About ▾. The main title "BLAST for beginners" is centered above a "Color key for alignment scores" chart. The chart has a red background with a black border. It includes a legend at the top with five color-coded boxes: <40 (red), 40-50 (blue), 50-80 (green), 80-200 (purple), and >=200 (dark red). Below the legend, the word "Query" is written in blue, followed by a sequence of numbers: 1080956, 1081456, 1081956, 1082456, 1082956, and 1083456. A horizontal line with small colored dashes (red, blue, green, purple) corresponds to these numbers. To the right of the chart, there are social media sharing icons for Facebook, Twitter, LinkedIn, and Pinterest. Below the chart, a text block describes the tutorial's purpose: "BLAST for beginners introduces students to blastn, a commonly used tool for comparing nucleotide sequences (DNA and RNA). This popular tutorial shows how to do a blast search with a nucleotide sequence, highlights information in the search results, and shows how to interpret the E value and alignment scores." Further down, it says "The complete set of materials consists of:" followed by a numbered list: 1. The BLAST for beginners tutorial., 2. A set of unknown sequences., 3. A set of questions for students to answer. On the right side, there is a "Get Molecule World for iPhone!" section with an "App Store" download button and a "Sign up for our mailing list" link.

Color key for alignment scores

Query 1080956 1081456 1081956 1082456 1082956 1083456

<40 40-50 50-80 80-200 >=200

1080956 1081456 1081956 1082456 1082956 1083456

BLAST for beginners introduces students to blastn, a commonly used tool for comparing nucleotide sequences (DNA and RNA). This popular tutorial shows how to do a blast search with a nucleotide sequence, highlights information in the search results, and shows how to interpret the E value and alignment scores.

The complete set of materials consists of:

1. The BLAST for beginners tutorial.
2. A set of unknown sequences.
3. A set of questions for students to answer.

Get Molecule World for iPhone!

Download on the App Store

Sign up for our mailing list

Get tips and ideas!

A 3D molecular model composed of spheres in various colors (grey, blue, red) representing atoms. The model is shown from a slightly elevated angle, highlighting its three-dimensional structure.

Primeiros passos: Do usuário ao bioinformata

- Familiarizar-se com Linux.
- Aprender a usar bancos relacionais.
- Familiarizar-se com MatLab e R.
- Aprender a programar (PERL).

- Familiarizar-se com Linux.

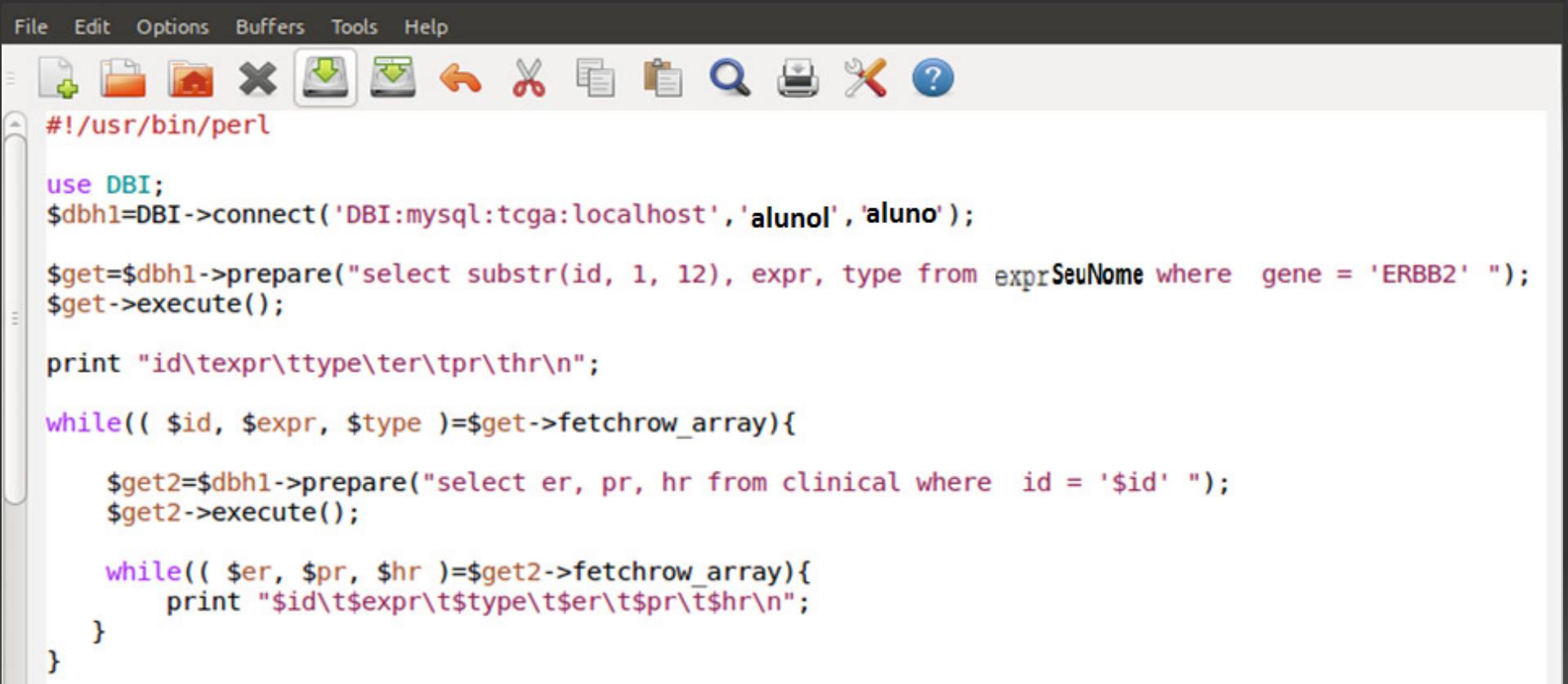
→ Tutorial Linux

- Aprender a usar bancos relacionais.

→ Tutorial Mysql.

PERL:

Fazer um perl para dar select no banco e fazer a união das tabelas.



The screenshot shows a Perl script in a code editor. The menu bar includes File, Edit, Options, Buffers, Tools, and Help. The toolbar contains icons for new file, open, save, cut, copy, paste, find, print, and help. The code itself is as follows:

```
#!/usr/bin/perl

use DBI;
$dbh1=DBI->connect('DBI:mysql:tcga:localhost', 'alunol', 'aluno');

$get=$dbh1->prepare("select substr(id, 1, 12), expr, type from exprSeuNome where gene = 'ERBB2' ");
$get->execute();

print "id\texpr\ttype\tter\tpr\thr\n";

while(( $id, $expr, $type )=$get->fetchrow_array){

    $get2=$dbh1->prepare("select er, pr, hr from clinical where id = '$id' ");
    $get2->execute();

    while(( $er, $pr, $hr )=$get2->fetchrow_array){
        print "$id\t$expr\t$type\t$er\t$pr\t$hr\n";
    }
}
```

R - <http://tryr.codeschool.com/>

The screenshot shows the landing page for the 'Try R' course. On the left, there's a vertical navigation menu with numbers 1 through 8. The main content area features a large, ornate logo of a stylized 'R' with the word 'TRY' integrated into its top. To the right of the logo, the text 'Try R is Sponsored By:' is followed by the 'O'REILLY®' logo. Below that, 'Created By:' is followed by the 'code school' logo. A paragraph of text explains what R is used for, mentioning statistics, data modeling, and graphics. Below the text are social sharing links for Twitter, Facebook, and LinkedIn. A 'Start the Course' button is visible, along with a progress bar consisting of colored segments. The bottom section contains a 'Table of Contents' and a 'Chapter Badges' section.

Table of Contents

1. R Syntax: A gentle introduction to R expressions, variables, and functions
2. Vectors: Grouping values into vectors, then doing arithmetic and graphs with them
3. Matrices: Creating and graphing two-dimensional data sets
4. Summary Statistics: Calculating and plotting some basic statistics: mean, median, and standard deviation
5. Factors: Creating and plotting categorized data
6. Data Frames: Organizing values into data frames, loading frames from files and merging them

Chapter Badges

A grid of circular badges, each containing a number from 1 to 7. The badges are arranged in three rows: the first row has three badges, the second row has three badges, and the third row has one badge in the center. The badges are red with white numbers. The eighth badge, located at the bottom right, is the 'TRY R' logo.

R - <http://tryr.codeschool.com/>

Expressions

1.1

Type anything at the prompt, and R will evaluate it and print the answer.

Let's try some simple math. Type the below command.

[Or, if you prefer, click on the command and it will be typed into the console for you!]

1 + 1

```
> 1 + 1  
[1] 2  
>
```

Type the string "Arr, matey!". (Don't forget the quotes!)

"Arr, matey!"

```
> "Arr, matey!"  
[1] "Arr, matey!"  
>
```

R - <http://tryr.codeschool.com/>

Logical Values

1.2

Some expressions return a "logical value": `TRUE` or `FALSE`. (Many programming languages refer to these as "boolean" values.) Let's try typing an expression that gives us a logical value:

```
3 < 4
```

```
> 3 < 4  
[1] TRUE  
>
```

And another logical value (note that you need a double-equals sign to check whether two values are equal - a single>equals sign won't work):

```
2 + 2 == 5
```

```
> 2 + 2 == 5  
[1] FALSE  
> █
```

CURSO DE CURTA DURAÇÃO - 2017

BIOINFORMÁTICA

BioME – CENTRO MULTIUΣUÁRIO DE BIOINFORMÁTICA - UFRN

Obrigado.

E-mail: jorge@imd.ufrn.br



Bioinformatics
Multidisciplinary
Environment

Centro
Multiusuário
de Bioinformática



Instituto de
Bioinformática e
Biotecnologia