

CORSO DI METODI MOLECOLARI E BIOINFORMATICA

LM Biologia Evoluzionistica

Università di Padova

Anno accademico 2022/23

Computational Genomics Lab.

compgen.bio.unipd.it



Prof. Stefania Bortoluzzi



Alessia Buratin, Post. Doc
(alessia.buratin@unipd.it)



Enrico Gaffo, Post. Doc
(enrico.gaffo@unipd.it)



Silvia Orsi, PhD student
(silvia.orsi@studenti.unipd.it)

Programma del corso

- 02/11 Genome browser (UCSC)
- 11/11 Comparazione di biosequenze ed
 Allineamenti multipli (Blast, Clustal Omega)
- 18/11 Struttura sistema Unix + Comandi Unix
- 02/12 Comandi Unix
- 07/12 Analisi di dati DNA-seq
- 14/12 Analisi di dati RNA-seq
- 21/12 Analisi di matrici d'espressione con R

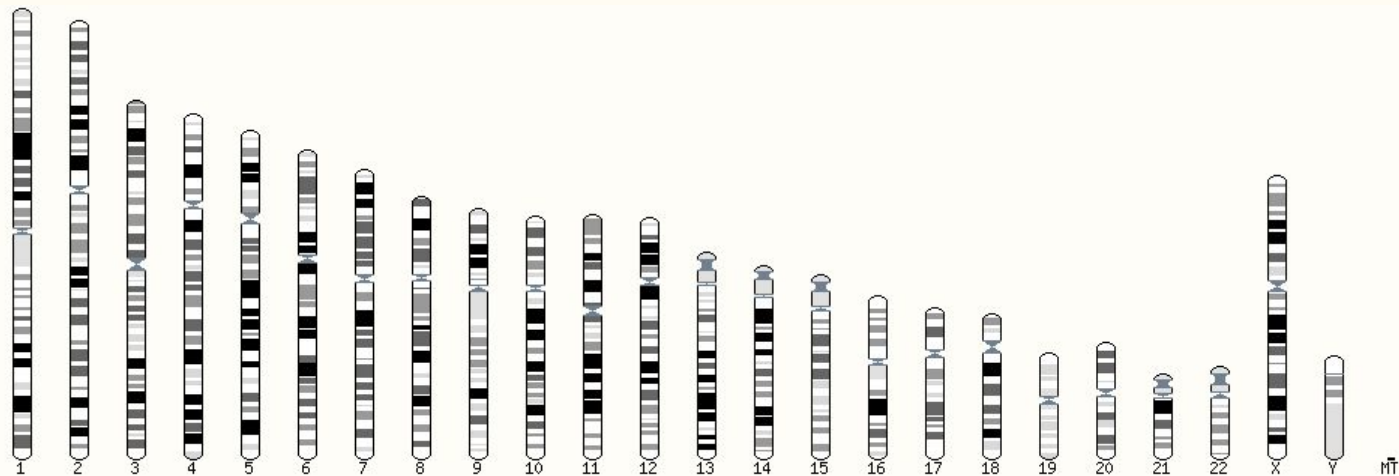
Programma del corso

- 02/11 Genome browser (UCSC)
- 11/11 Comparazione di biosequenze ed
Allineamenti multipli (Blast, Clustal Omega)
- 18/11 Struttura sistema Unix + Comandi Unix
- 02/12 Comandi Unix
- 07/12 Analisi di dati DNA-seq
- 14/12 Analisi di dati RNA-seq
- 21/12 Analisi di matrici d'espressione con R

Conosciamo tutti il nostro genoma... ma come possiamo muoverci da un cromosoma all'altro per identificare geni, esoni, SNP, promotori, ecc?

Teniamo presente che le dimensioni del genoma umano assemblato sono esattamente di **3,609,003,417 paia di basi**

Corrispondenti a **20,418** geni codificanti proteine, **22,107** geni non codificanti e **15,195** pseudogeni per oltre **200,000** trascritti!



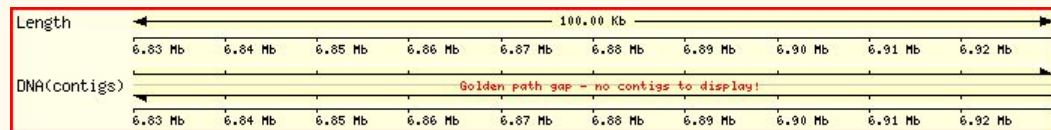
Come è possibile «rappresentare graficamente un genoma» di alcuni miliardi di paia di basi?



Jim Kent, uno dei più grandi bioinformatici della storia, ha definito un genoma come:

«Well, it has a lot of G, C, A and Ts»

Ed effettivamente è così ... senza un'**annotazione**, una sequenza genomica, di per sé, non ha significato ...



Nasce la necessità di creare i Genome Browser per:

- 1) Esplorare regioni cromosomiche
- 2) Esplorare regioni di regolazione fiancheggiando un gene
- 3) Effettuare ricerche di elementi (per parola chiave o similarità di sequenza) su scala dell'intero genoma
- 4) Comparare l'architettura del genoma in organismi differenti (comparative genomics)
- 5) ...

articles

Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium*

**A list of authors and their affiliations appears in the Supplementary Information*

The sequence of the human genome encodes the genetic instructions for human physiology, as well as rich information about human evolution. In 2001, the International Human Genome Sequencing Consortium reported a draft sequence of the euchromatic portion of the human genome. Since then, the international collaboration has worked to convert this draft into a genome sequence with high accuracy and nearly complete coverage. Here, we report the result of this finishing process. The current genome sequence (Build 35) contains 2.85 billion nucleotides interrupted by only 341 gaps. It covers ~99% of the euchromatic genome and is accurate to an error rate of ~1 event per 100,000 bases. Many of the remaining euchromatic gaps are associated with segmental duplications and will require focused work with new methods. The near-complete sequence, the first for a vertebrate, greatly improves the precision of biological analyses of the human genome including studies of gene number, birth and death. Notably, the human genome seems to encode only 20,000–25,000 protein-coding genes. The genome sequence reported here should serve as a firm foundation for biomedical research in the decades ahead.

ANNOTAZIONE

Per «annotazione» si intende il processo con cui vengono assegnate informazioni relative alla posizione di geni, trascritti, varianti ed elementi di regolazione su un genoma

Tutte le annotazioni sono basate su un **complesso mix di dati** di espressione genica, predizioni computazionali ed evidenze basate sulla similarità di sequenza con altri organismi (basate su BLAST)

L'annotazione può essere di tipo **strutturale** (assegnare le coordinate rispetto ad un cromosoma) o **funzionale** (che nome devo dare ad un determinato gene? Che ruolo biologico ha?)



CHE TIPO DI INFORMAZIONI SONO DISPONIBILI IN UN GENOME BROWSER?

ANNOTAZIONI «BASE» CON COORDINATE RISPETTO AD UNO DEI CROMOSOMI

- ✓ Geni (introni, esoni, 5' e 3' UTR)
- ✓ Trascritti (comprensivi di splicing alternativi, di solito indicati CDS ed UTR per quelli codificanti)
- ✓ Non-coding RNAs (rRNA, tRNA, lncRNAs, ecc.)
- ✓ Pseudogeni
- ✓ Link ad altre informazioni collegate (es. scheda della proteina codificata da un determinato mRNA)

CHE TIPO DI INFORMAZIONI SONO DISPONIBILI IN UN GENOME BROWSER?

ANNOTAZIONI «AVANZATE»

- ✓ Varianti genetiche (SNPs, STRs, indels, ecc.)
- ✓ Sequenze ripetitive (LINE, SINE, DNA transposons, ecc., spesso «mascherate» in un genoma, cioè mostrate come lunghe stretch di «N»)
- ✓ Dati di espressione (allineamenti con ESTs, o dati relativi a microarray oppure esperimenti di RNA-sequencing)
- ✓ Allineamenti con regioni genomiche omologhe in altre specie (strumenti per studi di genomica comparata)
- ✓ E molte altre annotazioni...

GENOME BROWSERS – ACCESSIBILITA'

1) ENSEMBL

<http://www.ensembl.org/index.html>

→ 2) UCSC genome Browser Gateway



<https://genome-euro.ucsc.edu/cgi-bin/hgGateway>

3) NCBI Genome Data Viewer

<https://www.ncbi.nlm.nih.gov/genome/gdv/>

4) Genome browser personalizzabili, di solito legati a specifici progetti genomici o centri di ricerca

UCSC GENOME BROWSER









Genome Browser Gateway

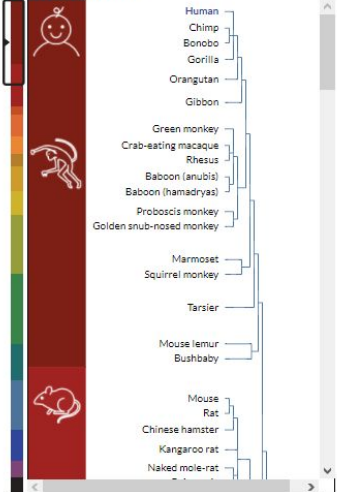
GenomesGenome BrowserToolsMirrorsDownloadsMy DataHelpAbout Us

Browse/Select Species

POPULAR SPECIES

HumanMouseRatFruitflyWormYeast

REPRESENTED SPECIES



- Human
- Chimp
- Bonobo
- Gorilla
- Orangutan
- Gibbon
- Green monkey
- Crab-eating macaque
- Rhesus
- Baboon (anubis)
- Baboon (hamadryas)
- Proboscis monkey
- Golden snub-nosed monkey
- Marmoset
- Squirrel monkey
- Tarsier
- Mouse lemur
- Bushbaby
- Mouse
- Rat
- Chinese hamster
- Kangaroo rat
- Naked mole-rat

Find Position

Human Assembly
Dec. 2013 (GRCh38/hg38)


Position/Search Term
Enter position, gene symbol or search terms
Current position: chr16:176,680-177,522

GO

Human Genome Browser - hg38 assembly

view sequences

UCSC Genome Browser assembly ID: hg38
Sequencing/Assembly provider ID: Genome Reference Consortium Human GRCh38 (GCA_000001405.15)
Assembly date: Dec. 2013
Accession ID: GCA_000001405.15
NCBI Genome ID: 31 (Homo sapiens (human))
NCBI Assembly ID: 883148 (GRCh38, GCA_000001405.15)
BioProject ID: 31257



Search the assembly:

- By position or search term: Use the "position or search term" box to find areas of the genome associated with many different attributes, such as a specific chromosomal coordinate range; mRNA, EST, or STS marker names; or keywords from the GenBank description of an mRNA. More information, including sample queries.
- By gene name: Type a gene name into the "search term" box, choose your gene from the drop-down list, then press "submit" to go directly to the assembly location associated with that gene. More information.
- By track type: Click the "track search" button to find Genome Browser tracks that match specific selection criteria. More information.

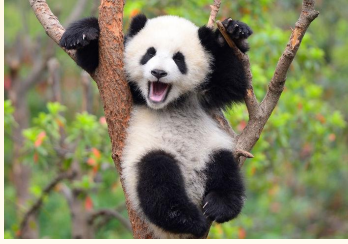
Download sequence and annotation data:

- Using rsync (recommended)
- Using FTP
- Using HTTP
- Data use conditions and restrictions
- Acknowledgments

Assembly Details

The GRCh38 assembly is the first major revision of the human genome released in more than four years. As with the previous GRCh37 assembly, the Genome Reference Consortium (GRC) is now the primary source for

l'UCSC Genome Browser nasce inizialmente per ospitare il genoma umano, ma oggi presenta ben 46 genomi



Interesse di tipo **evolutivo** (primati, ma non solo), di **biologia di base** (*Xenopus*)
economico (pensiamo ai pesci, al pollo o alla pecora) o di **conservazione** (panda)

COME SONO GESTITE TUTTE QUESTE INFORMAZIONI IN UN GENOME BROWSER?

UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr16:176,680-177,522 843 bp.

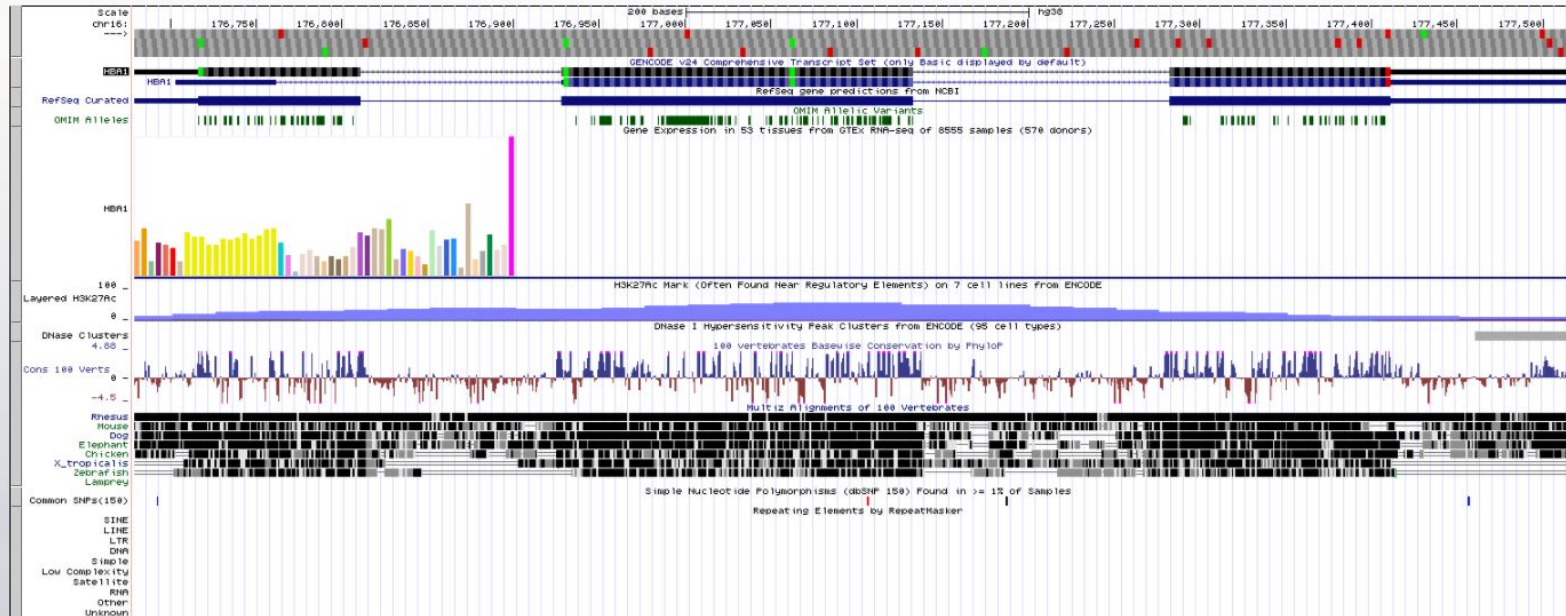
enter position, gene symbol, HGVS or search terms

go

ZOOM IN /OUT
BARRA DI RICERCA

chr16 (p13.3) 16p13.3 p13.2 16p12.3 p12.2 16p12.1 16p11.2 16p11.1 16p12.1 16p12.213 16q21 16q22.1 22.2 16q23.1 23.2 23.3 23.4 24.1

POSIZIONE SUL CROMOSOMA



mRNA (INTRONI/ESONI)
VARIANTI OMIM

TRACCIA
GENE EXPRESSION

TRACCIA CONSERVAZIONE
TRA SPECIE

TRACCIA ELEMENTI
RIPETUTI

