



# UNIVERSITÀ DI PISA

Dipartimento di Filologia, Letteratura e Linguistica

Corso di Laurea Magistrale in Informatica Umanistica

**REALIZZAZIONE DI UN SITO WEB PER LA RACCOLTA E L'ESPLORAZIONE DI DATI EPIGENOMICI  
SULLA SINDROME DI RETT**

Relatore:

Paolo Milazzo

Correlatore:

Ugo Borello

Candidato:

Marta Scalisi

Anno Accademico 2020/2021

# Indice

<b>Introduzione</b>	<b>5</b>
<b>1 La sindrome di Rett</b>	<b>7</b>
1.1 Nozioni di Genetica e Biologia Molecolare . . . . .	7
1.2 Cos'è la Sindrome di Rett? . . . . .	12
1.2.1 Altri geni: CDKL5 e FOXP1 . . . . .	13
1.3 Sintomatologia . . . . .	15
1.4 Diagnosi . . . . .	17
1.5 Prognosi e trattamento . . . . .	18
1.6 Importanza degli studi legati alla sindrome di Rett . . . . .	19
1.7 Studi del Dipartimento di Biologia dell'Università di Pisa . . . . .	20
<b>2 Genome Browsers</b>	<b>23</b>
2.1 Introduzione ai genome browsers . . . . .	23
2.2 Formati delle tracce di annotazione . . . . .	25
2.2.1 HiC . . . . .	25
2.2.2 BigWig . . . . .	25
2.2.3 Bed e Bedgraph . . . . .	26
2.3 UCSC, Ensembl, GIVE e WashU . . . . .	30

---

2.3.1	Ensambl . . . . .	30
2.3.2	UCSC . . . . .	32
2.3.3	Setup dei Genome Browser . . . . .	34
2.3.4	GIVE . . . . .	35
2.3.5	WashU . . . . .	39
2.3.6	GIVE e WashU: un breve confronto . . . . .	44
<b>3</b>	<b>Strumenti utilizzati per lo sviluppo del sito web</b>	<b>48</b>
3.1	Metodologie e strumenti utilizzati . . . . .	48
3.2	Design . . . . .	49
3.2.1	Figma . . . . .	50
3.2.2	Nicepage . . . . .	51
3.3	Implementazione . . . . .	51
3.3.1	HTML . . . . .	51
3.3.2	CSS . . . . .	52
3.3.3	Javascript . . . . .	53
3.3.4	JQuery . . . . .	54
3.3.5	Bootstrap . . . . .	54
3.4	Considerazione delle scelte . . . . .	55
<b>4</b>	<b>Contributo</b>	<b>57</b>
4.1	Organizzazione del lavoro . . . . .	57
4.2	Studio di Biologia Molecolare e Genetica . . . . .	58
4.2.1	Caricamento delle tracce . . . . .	59
4.3	Elaborazione dei wireframe e mockup . . . . .	62

4.3.1	Homepage . . . . .	64
4.3.2	Syndrome . . . . .	68
4.3.3	Database . . . . .	69
4.3.4	Genome Browser . . . . .	70
4.3.5	About us . . . . .	72
4.4	Implementazione pagina "Database" . . . . .	74
4.5	Implementazione pagina "Genome Browser" . . . . .	76
	<b>Prospettive Future</b>	<b>80</b>
	<b>Conclusioni</b>	<b>81</b>
	<b>Bibliografia</b>	<b>82</b>

*Alla mia famiglia*

---

# Introduzione

La sindrome di Rett è una malattia genetica rara che colpisce a livello neurologico prevalentemente soggetti di sesso femminile. Questa patologia congenita causa gravi deficit cognitivi poiché colpisce il sistema nervoso centrale e si manifesta solitamente dopo i 6-18 mesi di vita con: perdita della motricità, delle capacità manuali, dell'interesse all'interazione sociale.

Nella maggior parte dei casi la sindrome di Rett è causata da mutazioni del gene MECP2, localizzato sul cromosoma X. Tuttavia, sono anche stati osservati casi specifici in cui si verificavano delle mutazioni nei geni CDKL5 o FOXP1. Inizialmente la sindrome di Rett viene diagnosticata attraverso osservazione clinica, ma si ha una diagnosi definitiva solo quando viene individuato un difetto in uno dei geni citati in precedenza. Esistono comunque dei casi molto rari in cui in questi geni non viene individuata nessuna anomalia e ciò suggerisce che esistono altri geni mutati che possono causare dei sintomi analoghi a quelli della RTT.

L'obiettivo del lavoro di questa tesi è quello di sviluppare un sito web per la raccolta e l'esplorazione di dati epigenomici sulla sindrome di Rett. Il lavoro è stato strutturato in due fasi principali. La prima fase ha riguardato l'analisi e lo studio dei genome browser, cioè delle interfacce Web utilizzate per visualizzare la mappa fisica di un cromosoma. In particolar modo, è stata fatta un'analisi dei browser GIVE e WashU.

Sono state inoltre processate diverse tracce elaborate dai ricercatori del Dipartimento di Biologia dell'Università di Pisa, in seguito inserite all'interno dei due genome browser per poterne osservare la visualizzazione. Ciò ha permesso di comprendere al meglio analogie e differenze tra funzionalità dei rispettivi genome browser. Nella seconda fase invece si è proseguito con l'elaborazione dei *wireframe* e *mockup* per definire i contenuti di ogni pagina che, infine, sono stati implementati utilizzando gli strumenti HTML, CSS e Javascript. In questo modo è stato ottenuto un primo prototipo perfettamente funzionante del sito.

Questa tesi è stata svolta nell'ambito delle attività del progetto "Metodi Informatici Integrati per la Biomedica", Progetti di Ricerca di Ateneo (PRA) dell'Università di Pisa<sup>1</sup>.

---

<sup>1</sup>Project number: PRA\_2020-2021\_26

---

# La sindrome di Rett

## 1.1 Nozioni di Genetica e Biologia Molecolare

L'obiettivo di questo progetto è quello di sviluppare un sito web che permetta di raccogliere e esplorare dati epigenomici sulla sindrome di Rett. Al fine di inquadrare al meglio l'area di lavoro è stato necessario comprendere questa rara malattia genetica e le sue cause, focalizzandosi per prima cosa su alcuni concetti fondamentali di genetica. La genetica è una branca della biologia che studia i geni, l'ereditarietà e la variabilità genetica negli organismi viventi. Nello specifico, la genetica studia il genoma cioè l'insieme delle informazioni genetiche che caratterizzano ogni organismo vivente [13]. Nelle cellule che compongono l'organismo umano questa informazione è racchiusa in 46 cromosomi che si trovano nel nucleo, per la precisione, in 23 coppie di cromosomi

identici. A determinare il sesso sono solo due di questi 46 cromosomi. Di fatto, le femmine hanno due cromosomi X mentre i maschi un cromosoma X e uno Y. Ogni cromosoma a sua volta contiene parecchi geni, ossia porzioni di DNA che contengono le informazioni necessarie e sufficienti per sintetizzare una specifica proteina [14].

Il DNA è costituito da una sequenza lineare di quattro molecole differenti dette nucleotidi o basi azotate: adenina (A), timina (T), guanina (G) e citosina (C) e ha una struttura simile a una scala a chiocciola chiamata "a doppia elica" (Fig. 1.2). I nucleotidi presentano una struttura generale che comprende tre elementi molecolari: un gruppo fosfato, uno zucchero a 5 atomi di carbonio e una base azotata e, poiché è proprio quest'ultima che differenzia queste quattro subunità, da essa prendono il nome (Fig. 1.1) [1].

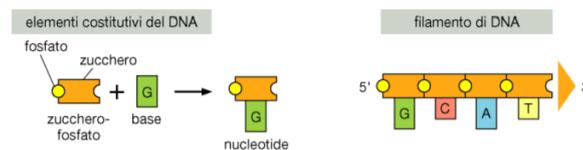


Figura 1.1: Basi chimiche che compongono il DNA

Attraverso gli zuccheri e i fosfati, i nucleotidi si uniscono con un legame detto covalente<sup>2</sup> tra loro in una catena polinucleotidica di DNA e tutte le informazioni necessarie per il corretto funzionamento della cellula sono racchiuse nel preciso ordine con cui queste basi si distribuiscono lungo il DNA.

Una molecola di DNA è costituita da due catene polinucleotidiche tra loro antiparallele e complementari. Sono dette antiparallele perché ogni filamento ha una sua polarità chimica o, in altre parole, ogni filamento ha una sua polarità chimica. Di

<sup>2</sup>In chimica, un legame covalente è un legame chimico in cui due atomi mettono in comune delle coppie di elettroni.

fatto, i nucleotidi formano un'ossatura in cui zucchero e fosfato si alternano tra loro. Ciò rende facilmente distinguibile l'estremità di ogni filamento poiché una ha una rientranza e l'altra una sporgenza (Fig. 1.1) [1].

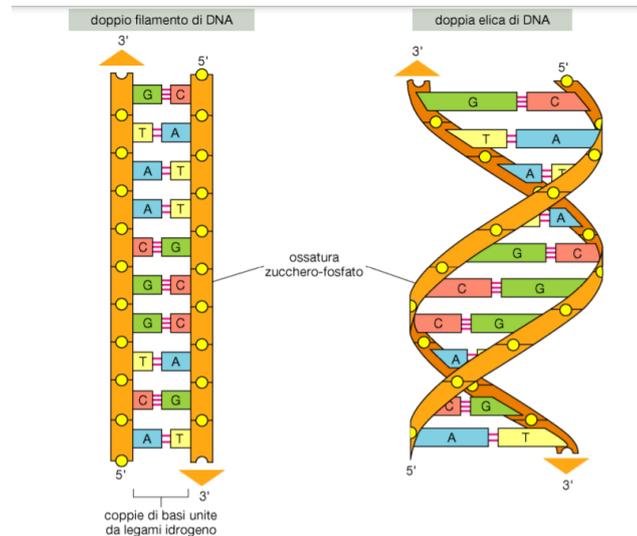


Figura 1.2: DNA: Doppio filamento e doppia elica

Le due catene che compongono la molecola del DNA sono tenute insieme da legami di idrogeno tra le basi dei diversi filamenti che si appaiano tra loro in modo complementare. Infatti, A si lega sempre con T formando due legami di idrogeno mentre G si lega sempre con C formando tre legami di idrogeno (Fig. 1.2).

Le sequenze di DNA vengono chiamate geni e ognuno di questi contiene le informazioni necessarie per sintetizzare una particolare proteina (o una serie di proteine) con funzioni specifiche all'interno del nostro organismo. Le proteine sono molecole complesse che nascono dall'unione di molecole più semplici, chiamate amminoacidi. La specifica sequenza dei diversi amminoacidi, che nell'uomo sono 20, che caratterizza la proteina, determina la sua forma e la sua funzione. Se nel DNA si verifica l'alterazione di una base necessaria per la produzione di una proteina, cambia la sequenza

di amminoacidi codificata e dunque la proteina prodotta può risultare più o meno alterata. Ad esempio, "il gene MECP2 è formato da circa 1500 basi che codificano informazioni per una proteina di circa 500 amminoacidi, chiamata MeCP2, che è fondamentale per le funzioni cerebrali e la regolazione di altri geni. Se una delle basi del DNA del gene MECP2 subisce una modifica, anche la sequenza della proteina MeCP2 cambia e questi cambiamenti possono alterarne o distruggerne le funzioni con gravi conseguenze" [25].

In particolare, durante la replicazione del DNA, possono verificarsi dei cambiamenti nelle sequenze di nucleotidi. La replicazione avviene durante la DNA polimerasi, un processo biologico che ha un ruolo centrale dei processi della vita perché permette di duplicare le nostre informazioni genetiche.

La DNA polimerasi avviene prima della divisione cellulare e permette la duplicazione del DNA nella cellula madre che ne passerà quindi una copia alle cellule figlie [12]. Le DNA polimerasi sono enzimi molto accurati, poiché creano ogni volta una copia esatta del DNA appaiando in modo preciso le basi che compongono i filamenti, creando una perfetta corrispondenza tra citosina a guanina e tra adenina e timina, e commettendo solo un errore ogni miliardo di basi. Inoltre, Dopo aver copiato ciascuna base, il processo della DNA polimerasi effettua un ulteriore controllo e corregge l'appaiamento qualora questo sia sbagliato (Fig. 1.3).

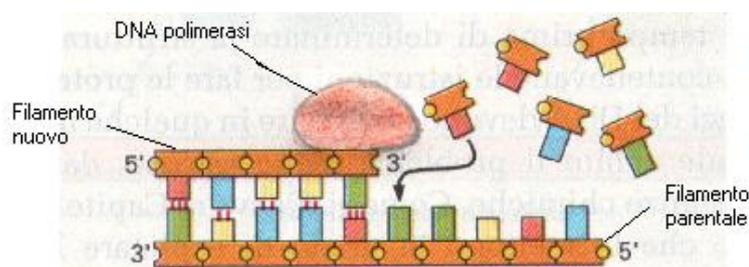


Figura 1.3: Processo della DNA polimerasi

Quando si verificano dei cambiamenti nelle sequenze di nucleotidi che compongono il DNA si ha una mutazione genetica che può generare un comportamento anomalo del gene coinvolto.

Una mutazione è un cambiamento nella successione o nel numero delle basi di un gene. In generale, si può avere una mutazione in qualsiasi cellula dell'organismo ma solo nel momento in cui questa si verifica nei gameti entra a far parte del patrimonio ereditario e, di conseguenza, verrà trasmessa da una generazione all'altra. Le mutazioni possono coinvolgere singoli nucleotidi o tratti di cromosoma. Nel primo caso parliamo di mutazioni genetiche, nel secondo di mutazioni cromosomiche [27].

Il DNA contiene inoltre le informazioni per la sintesi delle proteine. Di conseguenza, se la sequenza di DNA è mutata, queste informazioni risulteranno errate.

Della conversione le informazioni genetiche del DNA in proteine si occupa una macromolecola biologica chiamata l'RNA (acido ribonucleico). L'RNA si differenzia dal DNA in primo luogo per la struttura. Infatti mentre il primo è formato da due catene (o filamenti) di polinucleotidi, l'RNA è composto da una singola catena polinucleotidica. Inoltre, mentre lo zucchero del DNA è deossiribosio, nell'RNA si trova il ribosio, il quale può legarsi a quattro basi azotate: adenina, guanina e citosina, comuni al DNA, e all'uracile (U).

L'RNA agisce come una sorta di dizionario dell'informazione genetica che traduce il linguaggio codificato del DNA. Quindi i segmenti nucleotidici o geni vengono tradotti in amminoacidi, le unità strutturali primarie delle proteine [31].

Attraverso il processo dell'RNA polimerasi, che identifica una famiglia di enzimi complessi, costituiti da sub-unità proteiche, viene sintetizzato un filamento di RNA complementare a uno dei due filamenti di DNA che compongono la doppia elica. La RNA polimerasi si lega al DNA solo presso particolari sequenze, dette promotori, per iniziare la trascrizione di un gene.

Esistono diverse tipologie di RNA polimerasi. Nei casi specifici in cui vengono trascritti geni codificanti proteine, sono prodotti i cosiddetti RNA messaggeri (mRNA) necessari per la sintesi proteica.

Quindi nel caso in cui un gene sia mutato, in alcuni casi una proteina necessaria non viene prodotta, in altri casi viene prodotta nel momento e nel posto sbagliato o ancora la proteina sintetizzata ha caratteristiche diverse da quelle che ci si aspetta. Tutto questo, può alterare processi fondamentali della cellula poiché le proteine assumono ruoli estremamente importanti all'interno dell'organismo [32] [21].

## 1.2 Cos'è la Sindrome di Rett?

La sindrome di Rett (RTT) fu descritta per la prima volta nel 1966 da un neurologo di origine austriaca, Andreas Rett, da cui prende il nome [33].

È un disturbo dello sviluppo neurologico che si verifica quasi esclusivamente nelle femmine e che colpisce dopo un periodo di crescita normale di circa sei mesi [36].

Di fatto, inizialmente le bambine affette da questa sindrome nascono e crescono in

maniera apparentemente normale e solo in seguito si verifica in modo improvviso una fase di regressione che in breve tempo porta alla perdita delle abilità precedentemente acquisite: il cammino, il linguaggio, l'uso delle mani. Interessa 1 ogni 10 000-15 000 nascite femminili in tutto il mondo. La maggior parte dei casi registrati sono mutazioni casuali spontanee, mentre solo meno dell' 1% dei casi sono ereditari o trasmessi da una generazione alla successiva.

La sindrome Rett è generalmente causata da una mutazione nel gene metil CpG binding protein 2 (MECP2), localizzato sul cromosoma X, precisamente nella zona Xq28. In particolare, le bambine affette hanno un gene MECP2 difettoso ed un gene MECP2 normale. Questo gene è coinvolto nella produzione di una proteina chiamata proteina 2 (MeCP2) metil-citosina, necessaria per lo sviluppo del cervello. Questa proteina agisce come un interruttore biochimico che può sia aumentare l'espressione genica sia dire agli altri geni quando spegnere e fermare la produzione propria di proteine uniche [36].

Nella sindrome di Rett, dunque, la mutazione del gene MECP2 causa "forme strutturalmente anomale o inadeguate quantità della proteina e possono causare un'espressione anomala del gene in altri geni" [36].

Diversi studi hanno inoltre rivelato che questo gene è la causa di circa il 95% dei casi di RTT classica e di circa il 40-60% delle forme atipiche della stessa [15].

### 1.2.1 Altri geni: CDKL5 e FOXP1

Tuttavia, la sindrome di Rett non è sempre provocata da una mutazione MECP2, ma può essere provocata da delezioni parziali del gene, mutazioni in altre parti del gene MECP2, mutazioni in altri geni come CDKL5 e FOXP1 che influenzano lo sviluppo

del cervello nella sindrome di Rett atipica.

Negli ultimi 10 anni, mutazioni di questi ultimi geni si sono rivelate la causa di alcuni casi atipici di RTT. Ciò è stato riscontrato in circa il 10% dei casi in cui i soggetti non soddisfacevano i criteri che definiscono la forma classica della malattia [15].

Nello specifico, il gene CDKL5, sempre localizzato sul cromosoma X, viene generalmente associato alla variante di Hanefeld della sindrome di Rett o, più generalmente, ad encefalopatia epilettica precoce. Nel corso degli anni, a suo carico, sono state identificate più di 150 mutazioni. Questo gene, risulta essere anche associato ad altre gravi patologie, ad esempio l'encefalopatia epilettica infantile - che colpisce anche individui di sesso maschile - caratterizzata oltre che da crisi epilettiche nei primi mesi di vita, da forte ipotonia (perdita di tono di muscolo), ritardo nello sviluppo, compromissione motoria e tratti che ricordano la sindrome di Rett. CDKL5 contiene il codice necessario per la sintesi di una proteina appartenente alla famiglia delle "chinasi" fondamentali per regolare o modificare l'attività di proteine preesistenti nelle cellule. In particolare, le chinasi vengono sfruttate dai neuroni per rispondere in modo repentino agli stimoli che ricevono dall'esterno. Al momento, sono note poche proteine bersaglio dell'attività di CDKL5 (ancora oggetto di ricerca) ma è interessante notare che una di queste è MECP2. Di fatto, è stata avanzata l'ipotesi che gli aspetti clinici simili dei disordini associati a queste due proteine possano essere attribuite a delle loro funzioni comuni. Inoltre, diversi studi svolti finora hanno dimostrato che la mancanza di CDKL5 causa dei difetti nella maturazione dei neuroni e nella formazione della rete neuronale.

Più recentemente è stato scoperto un terzo gene: FOXP1. Questo gene è responsabile della forma più severa della sindrome: la variante congenita ad esordio precoce.

In questa variante le bambine presentano le stesse caratteristiche cliniche della Rett classica ma i sintomi cominciano ad insorgere già nei primi mesi di vita. I soggetti mostrano sintomi quali ipotonia, irritabilità, mancanza di reazione, ai quali si aggiungono in seguito grave microcefalia, ritardo nello sviluppo, crisi epilettiche, bruxismo (una condizione in cui si digrignano i denti) e stereotipie delle mani (ripetizione costante di alcuni movimenti delle mani) [34].

Il gene FOXP1 non si trova nel cromosoma X ma sul cromosoma 14. La proteina codificata da questo gene, come nel caso di MECP2 è una proteina che controlla l'espressione di altri geni e la sua presenza nel cervello sembra in grado di regolare il numero di neuroni generati durante lo sviluppo, spiegando così la comparsa di microcefalia nei bambini con mutazioni di FOXP1 [2]. La diagnosi, inizialmente fatta mediante osservazione clinica, è definitiva solo quando viene individuato un difetto in uno dei geni sopra menzionati. Tuttavia, esistono casi molto rari in cui non si riesce ad individuare alcuna anomalia in questi geni e ciò suggerisce che altri geni che possano causare sintomi analoghi debbano essere ancora individuati.

### 1.3 Sintomatologia

La prima manifestazione della malattia, il suo corso e la gravità dei sintomi variano in base al soggetto.

Questa sindrome causa di gravi deficit nell'acquisizione del linguaggio e della coordinazione motoria che rendono i soggetti che ne sono affetti completamente dipendenti dagli altri per tutta la vita.

Ciò che si verifica è un rallentamento nello sviluppo, che provoca "una regressione del-

le abilità con la perdita di utilizzo della mano con strizzamento e abitudini di lavaggio della mano compulsivi, rallentamento della crescita di testa e cervello, convulsioni, difficoltà a camminare e disabilità intellettiva" [36].

Gli stadi utilizzati per descrivere la sindrome di Rett sono quattro:

- **Stadio 1 o esordio precoce.** Solitamente si verifica un lieve rallentamento dello sviluppo nei primi 6-18 mesi di vita del bambino. Tra i sintomi si ha un minor contatto oculare, minor interesse nei giocattoli o ritardi in alcune attività motorie come sedersi o gattonare
- **Stadio 2 o regressione dello sviluppo o rapida fase distruttiva.** Inizia solitamente tra il primo e il quarto anno di vita, in modo rapido o graduale, e causa abilità manuali mirate e della lingua parlata. "Durante questa fase, cominciano i movimenti caratteristici della mano come strizzare, battere le mani, lavarle, toccare e portare più volte le mani alla bocca. I movimenti scompaiono durante il sonno. Possono verificarsi irregolarità di respirazione, come episodi di apnea e iperventilazione. La deambulazione può essere instabile, e avviare movimenti motori può essere difficile. Alcune ragazze possono anche avere sintomi simili a quelli dei disturbi dello spettro autistico come la compromissione dell'interazione sociale e della comunicazione" [36].
- **Stadio 3 o stadio pseudostazionario.** Si verifica tra i 2 e i 10 anni e può durare a lungo. In questa fase, sono molto comuni sintomi come convulsioni, deficit motori e aprassia. Mentre si verifica una diminuzione di sintomi simili all'autismo.
- **Stadio 4 o fase di deterioramento tardo motorio.** Caratteristiche comuni di questa fase comprendono scoliosi, diminuzione della mobilità, debolezza muscolare,

spasticità o rigidità.

Inoltre, i bambini possono sviluppare scoliosi e anomalie cardiache, come un intervallo QT prolungato. Inoltre, spesso si ha un rallentamento nella crescita e difficoltà a mantenere il peso.

## 1.4 Diagnosi

La diagnosi viene effettuata attraverso:

1. Valutazione clinica
2. Test genetici

Nello specifico, la sindrome di Rett viene diagnosticata clinicamente attraverso l'osservazione del bambino nelle prime fasi della crescita e il monitoraggio del suo stato fisico e neurologico. I test genetici per la mutazione MECP2 sul cromosoma X vengono utilizzati per integrare la diagnosi clinica.

Sono state fornite delle *guidelines* da parte del Neurological Disorders and Stroke (NINDS) utilizzate per confermare la diagnosi clinica della sindrome di Rett. Queste linee guida suddividono i criteri diagnostici clinici in tre diverse tipologie: principali, di supporto e di esclusione.

I principali criteri diagnostici comprendono la perdita di tutte o di parte delle abilità della mano; movimenti ripetitivi delle mani come, ad esempio, strizzare o spremere, battere o sfregare; la perdita di tutto o parte del linguaggio parlato; e anomalie nella deambulazione, come l'andamento in punta di piedi o una camminata instabile.

I criteri diagnostici di supporto possono presentarsi in alcuni individui ma non sono

sufficienti per effettuare una diagnosi. Questi sintomi possono essere più o meno gravi nel bambino e potrebbero svilupparsi con l'età, quindi non essere ancora presenti nei bambini molto piccoli. I criteri di supporto includono: scoliosi, digrignamento dei denti, sonno irregolare, anomalie del tono muscolare, mani e piedi troppo piccoli in relazione all'altezza, ridere o urlare in modo inappropriato, ridotta percezione del dolore. Ad un bambino che presenta dei criteri di supporto ma nessuno di quelli essenziali non può essere diagnosticata la Sindrome di Rett.

Esistono inoltre una serie di criteri che permettono ai medici di escludere la diagnosi della sindrome di Rett, si parla dei cosiddetti criteri di esclusione. Questi comprendono la presenza di altre patologie che causano sintomi simili, quali: lesioni cerebrali secondarie a traumi, malattie neuro metaboliche, infezioni gravi che causano problemi neurologici; e sviluppo psicomotorio fortemente anormale già presente nei primi 6 mesi di vita [36].

## 1.5 Prognosi e trattamento

Al momento non esiste una cura per questa malattia. Il trattamento di questo disturbo è sia sintomatico sia di supporto. Nel primo caso, quindi, ci si concentra sulla gestione dei sintomi, mentre nel secondo è richiesto un approccio multidisciplinare.

È necessario, inoltre, un attento monitoraggio nei soggetti che presentano scoliosi o anomalie cardiache. In casi eccezionali, queste ultime, insieme ad anomalie autonome, possono portare i bambini affetti dalla sindrome di Rett ad una morte improvvisa, ma solitamente i bambini sopravvivono in buone condizioni anche in età adulta.

In generale, si cerca di aiutare i bambini a svolgere in autonomia le normali attività

quotidiane come vestirsi o mangiare, utilizzando un tipo di terapia occupazionale. Mentre la terapia fisica e l'idroterapia cercano di prolungare la mobilità. Alcuni bambini possono aver bisogno di attrezzature speciali come tutori per la scoliosi, stecche per correggere i movimenti delle mani e programmi nutrizionali per cercare di controllare il peso.

In alcuni casi possono essere richiesti servizi accademici, sociali, professionali e di supporto speciale [36].

## 1.6 Importanza degli studi legati alla sindrome di Rett

Comprendere le cause di questo disordine è necessario sia per sviluppare nuove terapie, in modo da gestire al meglio sintomi specifici, sia per fornire dei migliori metodi di diagnosi.

Nel 1999, la scoperta del gene principale della Sindrome di Rett (MECP2) ha posto le basi per ulteriori studi genetici e ha consentito la sperimentazione su modelli animali come topi transgenici, che presentano una carenza di MECP2. Questi topi presentano anomalie neurologiche che possono essere invertite attraverso l'attivazione, in un secondo momento della vita, del gene MECP2.

Attualmente sono diverse le ricerche legate sulla Sindrome di Rett. Si è ormai consapevole che la presenza della proteina MeCP2 non correttamente funzionante disturba la funzione delle cellule cerebrali mature, ma non si conoscono ancora i meccanismi esatti con cui ciò accade.

Secondo alcuni ricercatori, tipi di mutazioni specifiche del gene MECP2 influiscono

sulla gravità dei sintomi legati alla sindrome di Rett e, di conseguenza, sono ancora in corso studi che cercando di comprendere meglio le caratteristiche di ogni mutazione. I ricercatori stanno anche cercando di trovare altri geni che potrebbero essere coinvolti nella sindrome di Rett, ma molto è ancora sconosciuto.

La ricerca legata alla RTT è quindi fondamentale sia per trovare una terapia che si riveli efficace sia per avere un quadro completo di questa malattia ancora, in parte, sconosciuta [36].

## 1.7 Studi del Dipartimento di Biologia dell'Università di Pisa

Gli studi dei ricercatori del gruppo guidato dal professor Borello relativi alla Sindrome di Rett sono stati effettuati a partire da alcuni lavori [24] [42] [10] in cui sono stati ricavati dati basati su un modello murino, quindi in funzione di esperimenti fatti su topo. Il gene su cui sono stati condotti gli studi è MECP2.

Questo gene, responsabile della maggior parte dei casi di sindrome di Rett, codifica per un rimodellatore della cromatina ed è espresso in tutte le cellule del corpo ma sembra essere particolarmente espresso a livello del sistema nervoso.

Il suo ruolo è fondamentale per il normale sviluppo del cervello. Di fatto, è un gene necessario per programmare la sintesi di una particolare proteina denominata proteina di legame metil-CpG che regola l'espressione e l'attività di molti altri geni. Di conseguenza, quando si verifica una mutazione si ha una carenza o totale mancanza della proteina da esso codificata e ciò si ripercuote sul funzionamento di tutti gli altri geni controllati da questa proteina, geni che tutt'ora non sono ancora stati completamente

identificati. La principale conseguenza di queste alterazioni genetiche “a cascata” è l’inadeguato sviluppo del cervello durante la prima parte dell’infanzia.

Lo scopo degli studi condotti è stato quello di osservare e scoprire altri geni che si trovassero sotto il controllo di MECP2. Gli esperimenti sono stati condotti su topi dalle 6 alle 11 settimane, periodo che sembra essere il migliore per studiare il ruolo di questo gene.

L’analisi è stata effettuata utilizzando una tecnica sperimentale che permette di analizzare le interazioni tra DNA e proteine. Questa tecnica è denominata *ChiP-Sequencing* o *ChiP-Seq* (chromatin immunoprecipitation sequencing).

La tecnica *ChiP-Seq*, usata per identificare i siti di legame tra le proteine e il DNA, combina l’immunoprecipitazione della cromatina (o ChiP) e la Next Generation Sequencing, permette di sequenziare moltissimi frammenti in parallelo. Di fatto, la *ChiP-Seq* è diventata uno dei metodi principali per effettuare ricerche epigenetiche, quindi per studiare l’interazione proteina-DNA su scala genomica.

La ChiP, tecnica sfruttata all’interno della *ChiP-Sequencing*, permette di localizzare i punti di legame tra DNA e proteine e, di conseguenza, capire quali sono le sequenze di DNA legate ad una particolare proteina di cui si vuole studiare il comportamento. Poiché il legame tra DNA e proteine nel nucleo cellulare ha una funzione di regolazione dell’espressione genica, lo studio dell’interazione tra DNA e proteina potrebbe favorire la scoperta di passaggi cruciali nella patogenesi di alcune malattie genetiche [9].

Quando si utilizza questa tecnica “le cellule vengono trattate con un agente che provoca la formazione di legami covalenti-crociati fra il DNA e le eventuali proteine ad esso legate. Successivamente la cromatina estratta dalle cellule viene frammentata, il

prodotto così ottenuto viene esposto all'anticorpo diretto contro la proteina. I frammenti di DNA recuperabili dall'immunoprecipitazione verranno purificati e si potrà determinarne la sequenza, con l'idea che essa sia correlata alla proteina. Oltre che ad evidenziare dove determinate proteine legano il DNA, è possibile individuare nuovi siti di legame e di conseguenza nuove interazioni" [19].

In questo caso specifico è stato utilizzato un anticorpo diretto sulla proteina MECP2 per capire in che punti del genoma questa si localizzava.

In altri casi invece, considerando una condizione di *wild type* mutato per MECP2, sono stati utilizzati anticorpi diretti contro particolari modificazioni istoniche<sup>3</sup>, con lo scopo di osservare come, in in diverse condizioni, MECP2 potesse andare ad influenzare la distribuzione delle modificazioni istoniche lungo il genoma. Il gruppo di lavoro del professor Borello ha quindi analizzato i risultati ottenuti da questi studi per renderli confrontabili e gestibili in un'unica piattaforma (ovvero il genome browser).

---

<sup>3</sup>Gli istoni sono proteine che condensano ed imballano il DNA nei cromosomi. La modifica degli istoni è un processo che si verifica dopo la traduzione del DNA, detta anche sintesi proteica.

---

# Genome Browsers

## 2.1 Introduzione ai genome browsers

In bioinformatica, un genome browser è un'interfaccia grafica che permette di visualizzare informazioni provenienti da un database biologico per dati genomici [40]. Questo permette ai ricercatori di visualizzare ed esplorare interi genomi che presentano dati annotati tra cui, ad esempio: struttura genica, proteine, espressione o regolazione.

Ad oggi, i genome browsers possono essere classificati in due tipologie:

- **Generali:** che comprendono diverse specie
- **Specifici:** che comprendono una sola specie

Per ogni specie si ha un genoma di riferimento comune in cui i ricercatori hanno la possibilità di inserire, in regioni specifiche, dati che solitamente provengono non da una sola fonte ma da più fonti diverse.

Le annotazioni sono organizzate secondo una coordinata genomica uniforme: il cromosoma come asse X e vari tipi di dati visualizzati lungo l'asse Y, ciò permette agli utenti di confrontare facilmente i dati di annotazione.

Esistono un gran numero di browser genomici. Molti di questi sono gratuiti e presentano al loro interno un database accessibile online. Dei più noti fanno parte UCSC, Ensembl [35] [30].

Questi browser permettono di fornire una visione d'insieme dei dati provenienti da database genomici e una valutazione comparativa di diverse sequenze genetiche tra più specie. Inoltre, la visualizzazione dei dati in più modi diversi permette di rendere più semplice la valutazione e l'interpretazione di questi dati complessi, poiché le informazioni vengono osservate in un preciso contesto genomico e non in maniera isolata [41]. È importante sottolineare che è proprio la visualizzazione in formato grafico dei dati a differenziare i genome browser dai normali database biologici.

Per la visualizzazione dei dati analizzati dal gruppo di ricercatori guidati dal professor Borello, sono stati considerati i Genome Browsers Ensembl [17] , UCSC [22], WashU e GIVE e, per ognuno, sono state analizzate le diverse caratteristiche: dalla visualizzazione dei dati all'usabilità.

## 2.2 Formati delle tracce di annotazione

I genome browser permettono di visualizzare diverse tracce di annotazione. Ogni riga nel file definisce una caratteristica particolare di visualizzazione per la traccia o, in alternativa, definisce un dato specifico all'interno stessa. I file di annotazione contengono tre tipi di linee iniziali: linee del browser, le linee di traccia e quelle di dati. Le righe vuote e quelle che iniziano con il simbolo "#" vengono automaticamente ignorate. Di seguito verranno analizzati le tipologie di formato maggiormente utilizzati.

### 2.2.1 HiC

Il formato *HiC* è un formato binario indicizzato progettato dall'Aiden Lab del Baylor College of Medicine che fornisce un rapido accesso casuale alle mappe di calore a matrice di contatti. I file *HiC* vengono utilizzati per visualizzare nei browser i dati relativi alla conformazione della cromatina [16]. In breve, la cromatina è un complesso di proteine e acidi nucleici. I suoi componenti principali sono il DNA e gli istoni, particolari proteine presenti nel nucleo cellulare. Gli istoni si legano al DNA permettendo alla molecola di compattarsi in diversi gradi di avvolgimento. Dunque, all'interno del nucleo cellulare le molecole di DNA si associano a molecole chiamate istoni, formando la cromatina.

### 2.2.2 BigWig

Il formato *BigWig* viene utilizzato per visualizzare dati densi e continui nei genome browsers. Sono creati partendo da file di tipi *Wiggle* (*Wig*) e poi convertiti attraverso il programma "wigToBigWig".

I file *BigWig*, come i file *HiC*, sono in un formato binario indicizzato. Il vantaggio principale di questi formati è che vengono trasferite nel genome browser solo le parti del file necessarie per la visualizzazione di una particolare regione. Di conseguenza, rispetto ai file *Wiggle*, le tracce *BigWig* riescono a gestire in modo più efficiente set di dati di grandi dimensioni poiché hanno delle prestazioni di visualizzazione notevolmente più veloci.

I dati di *Wiggle* devono essere continui e costituiti da elementi che presentano stesse dimensioni. Se i dati sono sparsi, solitamente si usa il formato *BedGraph* (vd 2.2.3) e non il *Wiggle*. Se il formato *Bedgraph* è troppo grande può essere convertito in *BigWig* usando il programma "*bedGraphToBigWig*" [4].

### 2.2.3 Bed e Bedgraph

Come già accennato in precedenza i file rielaborati dal gruppo di ricercatori guidati dal professor Borello sono basati su studi effettuati sul modello murino e sono sia in formato *Bed* sia in formato *Bedgraph*.

I file *Bed* (Browser Extensible Data) sono file di testo utilizzati per memorizzare regioni genomiche come coordinate e associarvi delle annotazioni.

Questo formato è stato sviluppato durante il progetto Human Genome Project e, ad oggi, è ampiamente utilizzato in bioinformatica. Uno dei più importanti vantaggi del *Bed* è che questo permette la manipolazione delle coordinate anziché delle sequenze nucleotidiche e ciò ottimizza la potenza e il tempo di calcolo quando si confrontano tutti o parte dei genomi. Inoltre, la sua semplicità permette di manipolare e leggere o analizzare coordinate o annotazioni utilizzando diversi linguaggi di programmazione come Python, Ruby or Perl.

Questi file presentano tre campi obbligatori e nove campi facoltativi aggiuntivi (Fig 2.1).

I primi tre campi obbligatori sono:

- **Chrom**: nome del cromosoma
- **ChromStart**: sito di inizio nel cromosoma
- **ChromEnd**: sito di fine nel cromosoma

Mentre i nove campi facoltativi sono riportati di seguito:

- **Name**: Definisce il nome di quella linea specifica del file *Bed*
- **Score**: Un valore che può variare tra 0 e 1000 che determina la tonalità di grigio con cui la feature sarà mostrata.
- **Strand**: definisce la direzione che viene rappresentata con "+", "-" o "." quando non se ne ha una.
- **thickStart**: La posizione di inizio in cui una specifica *feature* è disegnata in modo più fitto, come ad esempio la visualizzazione del codone all'inizio del gene.
- **thickEnd**: La posizione di fine in cui la *feature* è disegnata in modo più fitto.
- **ItemRgb** : è un valore RGB della forma R,G,B (es. 255,0,0). Se l'attributo track line itemRgb è impostato su "On", questo valore RGB determina il colore di visualizzazione dei dati contenuti in una specifica linea *Bed*.
- **BlockCount**: numero di esoni presenti nella linea *Bed*.

- **BlockSizes**: un elenco delimitato da virgole che presenta la stessa dimensione dei blocchi. Il numero di elementi in questo elenco dovrebbe corrispondere al `blockCount`.
- **BlockStarts**: Un elenco delimitato da virgole. Ogni posizione del `blockStart` dovrebbe essere calcolata rispetto a `chromStart`. Il numero di elementi in questo elenco dovrebbe corrispondere a `blockCount`.

```
chr7 127471196 127472363 Pos1 0 + 127471196 127472363 255,0,0
chr7 127472363 127473530 Pos2 0 + 127472363 127473530 255,0,0
chr7 127473530 127474697 Pos3 0 + 127473530 127474697 255,0,0
chr7 127474697 127475864 Pos4 0 + 127474697 127475864 255,0,0
chr7 127475864 127477031 Neg1 0 - 127475864 127477031 0,0,255
```

Figura 2.1: Formato *Bed* con campi facoltativi e obbligatori

I *Bedgraph* sono invece formati di tipo *Bed* che contengono al loro interno informazioni di maggiore complessità.

Rispetto al *Bed* che, come si è detto, lavora non con valori discreti i *Bedgraph* lavorano su valori continui e, di conseguenza si hanno più intervalli registrati. Rispetto al primo, si ha un ulteriore campo obbligatorio che è lo score, cioè una metrica che viene associata a ciascun intervallo continuo (Fig. 2.2) [3].

1. **Chrom** – nome del cromosoma
2. **ChromStart**: sito di inizio nel cromosoma
3. **ChromEnd**: sito di fine nel cromosoma

4. **Score:** valore continuo

```
chr1 59302000 59302300 -1.0
chr1 59302300 59302600 -0.75
chr1 59302600 59302900 -0.50
chr1 59302900 59303200 -0.25
chr1 59303200 59303500 0.0
```

Figura 2.2: Formato *Bedgraph*

Di conseguenza, in base al programma che si sta usando per fare *Peak Calling* (ovvero generare i file) questa metrica può essere un p-value associato all'intervallo o un arricchimento associato a quell'intervallo o il numero di conteggi associati a quell'intervallo. Questo può dipendere da ciò che si desidera vedere in quella colonna o da ciò che il programma associa di default a quel valore.

Il *Peak Calling* è un metodo computazionale che viene utilizzato dopo un esperimento di sequenziamento *ChIP* per identificare tutte quelle aree del genoma che sono state arricchite con letture allineate che permettono di confrontare sequenze di DNA o proteine.

In questo caso, la visualizzazione dei dati raccolti all'interno dei genome browser è importante perché permette di visualizzare i picchi, cioè tutti quei siti di legame della proteina che si sta studiando a livello di quella determinata sequenza di DNA. Permette quindi di vedere quali sequenze si hanno in quella specifica posizione e da lì si ricavano informazioni sulla funzione della proteina stessa.

## 2.3 UCSC, Ensembl, GIVE e WashU

Al fine di poter effettuare una scelta tra quale genome browser fosse più adeguato agli scopi del progetto sono stati presi in considerazione i seguenti:

- Ensembl<sup>4</sup>
- UCSC<sup>5</sup>
- GIVE<sup>6</sup>
- WashU<sup>7</sup>

In particolare, poiché la scelta tra WashU e GIVE (ovvero i due genome browser meglio conosciuti all'interno del gruppo di ricerca) ha occupato gran parte del lavoro ci si focalizzerà in modo particolare su questi ultimi.

### 2.3.1 Ensembl

Ensembl nasce da un progetto congiunto tra l'EBI (European Bioinformatics Institute) e l'istituto Wellcome Trust Sanger. Questo browser fornisce agli utenti un portale per sequenziare dati, annotazioni/predizioni genetiche e altri tipi di dati che è possibile trovare nei vari database Ensembl. Inoltre, supporta diverse specie e ne rende disponibili i dati attraverso download diretti di database e API. Ensembl è scritto per diverse tipologie di utenti quali ad esempio, ricercatori sia di biologia molecolare tradizionale che di estrazione clinica o bioinformatici [5].

---

<sup>4</sup><https://www.ensembl.org/index.html>

<sup>5</sup><https://genome.ucsc.edu/>

<sup>6</sup>[https://zhong-lab-ucsd.github.io/GIVE\\_homepage/](https://zhong-lab-ucsd.github.io/GIVE_homepage/)

<sup>7</sup><https://epigenomegateway.wustl.edu/>

Per quanto riguarda l'interfaccia grafica, Ensembl (Fig. 2.3) fornisce una stessa interfaccia per ogni organismo. Il corpo centrale dell'interfaccia contiene due pannelli: nel pannello di sinistra troviamo un elenco che permette di visualizzare la posizione dell'intero genoma su diversi livelli; quello principale invece è organizzato in tre sezioni dall'alto verso il basso, che forniscono agli utenti diverse scale di analisi del genoma [40]. Quest'ultimo ha inoltre alcune caratteristiche particolari progettate per consentire agli utenti di accedere in modo più semplice alle annotazioni.

Facendo *click* sull'elemento grafico nel pannello principale è possibile visualizzare le informazioni dettagliate di una specifica unità di annotazione. Spostando il mouse sul nome della traccia verranno visualizzati, alla sua sinistra, diverse icone. Spostando invece il mouse sull'icona delle informazioni verrà mostrata la descrizione dettagliata della traccia.

Inoltre, lo stile di visualizzazione di ogni traccia può essere modificato spostandosi sull'apposita icona di configurazione. Gli utenti hanno la possibilità di riordinare le tracce liberamente in modo da avere una visualizzazione personalizzata.

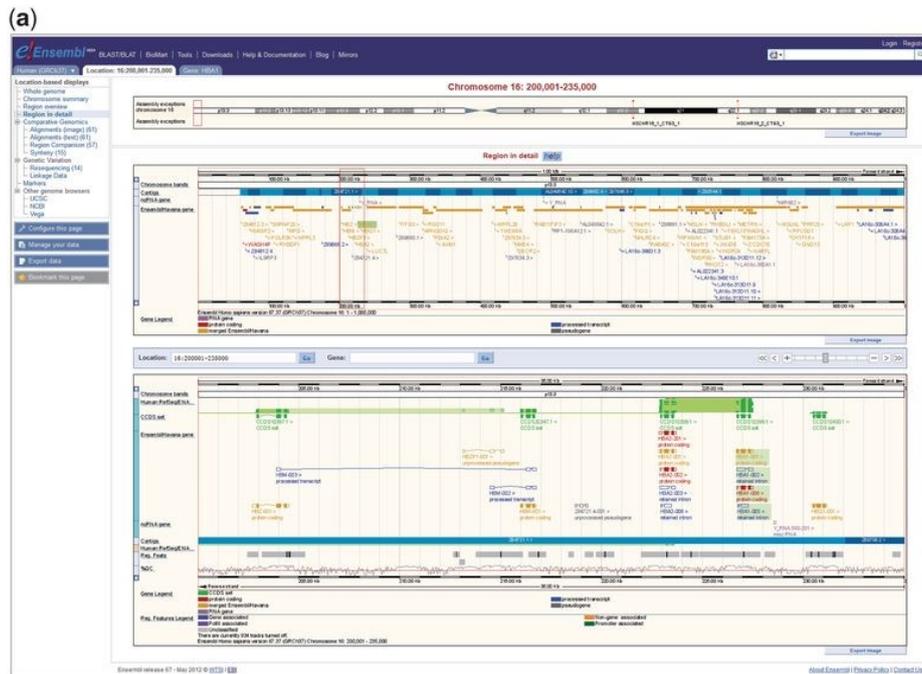


Figura 2.3: Interfaccia grafica di Ensembl

### 2.3.2 UCSC

UCSC è un genome browser basato su web sviluppato dell'Università della California Santa Cruz. Questo browser permette di visualizzare su qualsiasi scala una porzione di genoma, accompagnata da una serie di "tracce" di annotazione allineate. Inoltre, tutte le informazioni rilevanti per una regione vengono presentate in un'unica finestra in modo da poter facilitare sia l'analisi sia l'interpretazione biologica.

UCSC permette di visualizzare, scaricare e manipolare tutte le tabelle del database in cui sono contenute le tracce del Genome Browser usando un'altra applicazione web chiamata UCSC Table Browser. Gli utenti hanno la possibilità di caricare dati come tracce di annotazioni personalizzate in entrambi i browser per scopi di ricerca o didattici [20].

È possibile accedere al software e ai dati del browser Genome su Internet dal sito Web dell'UCSC Genome Bioinformatics Group. UCSC (Fig. 2.4) ha alcune caratteristiche comuni con Ensembl, incluso il layout generale per la visualizzazione grafica delle coordinate cromosomiche. Tuttavia, la *user experience* in UCSC è molto diversa da quella di Ensembl.

A differenza del sistema Ensembl, la barra delle colonne è collegata alla descrizione della traccia e alla pagina di configurazione e le tracce possono essere riordinate semplicemente cliccando e trascinando il nome della traccia.

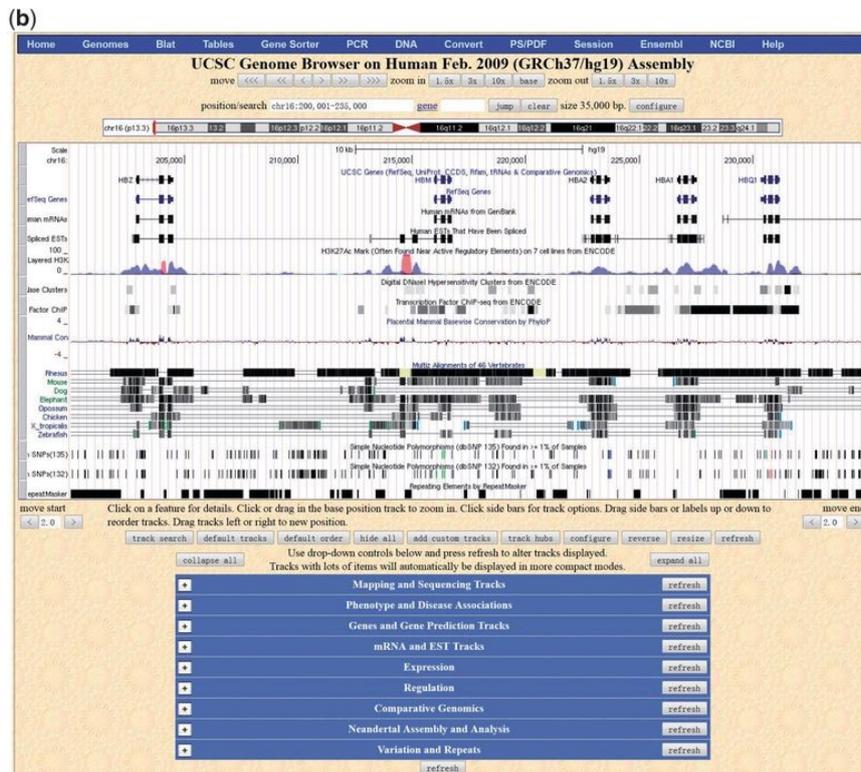


Figura 2.4: Interfaccia grafica di UCSC

### 2.3.3 Setup dei Genome Browser

Prima di procedere con l'analisi di GIVE e WashU è necessario introdurre Docker<sup>8</sup>, poiché questo strumento ha permesso di installare in locale entrambi i browser semplificando tutto il processo di configurazione delle piattaforme.

Docker è una piattaforma software che permette di creare, testare e distribuire applicazioni con la massima rapidità. Raccoglie il software in unità standardizzate chiamate container che offrono tutto il necessario per la loro corretta esecuzione, incluse librerie, strumenti di sistema, codice e *runtime*.

Ciò accade perché spesso, quando si crea un'applicazione, può succedere che questa funzioni bene nella propria macchina, ma non nell'ambiente di produzione. Le ragioni possono essere diverse:

- Dipendenze
- Librerie e versioni
- Framework
- OS Level features
- Microservizi

Tutti strumenti, che lo sviluppatore ha nella propria macchina ma non in ambiente di produzione.

I container rendono quindi più semplice la creazione, il rilascio e l'esecuzione di applicazioni. Docker permette di racchiudere un'applicazione e le sue dipendenze in un

---

<sup>8</sup><https://www.docker.com/>

contenitore virtuale eseguibile su ogni server di Linux. Inoltre, ogni container viene eseguito come un processo isolato nello spazio utente e la sua architettura "a livelli" permette di occupare meno spazio rispetto ad una classica macchina virtuale.

In poche parole, con Docker, è possibile distribuire e ricalibrare le risorse per un'applicazione in qualsiasi ambiente, tenendo sempre sotto controllo il codice eseguito.

### 2.3.4 GIVE

GIVE, acronimo di Genomic Interaction Visualization Engine, è una libreria open source, sviluppata in HTML e JavaScript con cui è possibile creare genome browser personalizzati da poter inserire all'interno della propria pagina web.

#### 2.3.4.1 Caratteristiche di GIVE

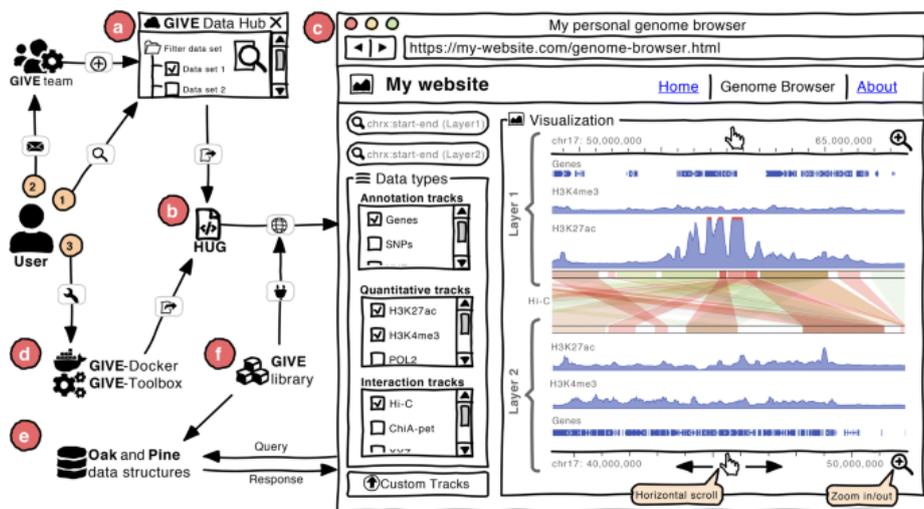


Figura 2.5: Caratteristiche di GIVE

Le caratteristiche principali caratteristiche di GIVE (Fig. 2.5) sono [23]:

- 
- (a) Il Data Hub GIVE, motore che permette la visualizzazione dell'interazione genomica, è una central repository gestita direttamente dal team GIVE, che consente all'utente di cercare i dati di interesse ed esportarli come HTML incapsulabile.
  - (b) HUG, un'interfaccia grafica su Data Hub che genera automaticamente il codice HTML per dati selezionati, che possono in seguito essere incorporati in qualsiasi sito web.
  - (c) Esempio dell'interfaccia utente di GIVE dopo aver incorporato il codice generato da HUG in un sito web. Nell'esempio vengono mostrate molte delle opzioni disponibili, inclusa la visualizzazione dei dati in più layer, le caselle che permettono la selezione dei dati da visualizzare e le opzioni di scorrimento e zoom.
  - (d) GIVE-Toolbox e GIVE-Docker che permettono di configurare rapidamente una versione locale del server GIVE.
  - (e) Oak e Pine, strutture dati ad albero che trasferiscono i dati dal server GIVE al browser web in risposta a una query richiesta dall'utente
  - (f) La libreria GIVE portatile utilizza componenti web che permette il trasferimento e la visualizzazione dei dati e incapsula gli algoritmi delle strutture dati Oak e Pine.

### 2.3.4.2 Allestire un genome browser

Per poter aggiungere e gestire i propri dati, l'utente deve scaricare ed lanciare GIVE-Docker. Questo, può essere eseguito su tutti i sistemi operativi mainstream senza dover configurare il sistema in modo specifico. Quando viene eseguito GIVE-Docker configura automaticamente un web server e in database. All'interno dell'eseguibile è inoltre presente GIVE-Toolbox, con cui è possibile automatizzare le operazioni sul database in una riga di comando su Linux.

Dopo aver scaricato ed eseguito GIVE-Docker, è necessario inizializzare il genoma di riferimento utilizzando i comandi forniti da GIVE-Toolbox, aggiungere le annotazioni genetiche e caricare i propri dati [38]. In seguito, eseguendo un comando specifico sarà possibile visualizzare la struttura ad albero con le tracce create nel genoma di riferimento (Fig. 2.6).

```
root@ac0ae5738edb:/# bash list_tracks.sh -u root -p Admin2015 -r hg38
|
|__ hg38
|   |
|   |__ genes
|   |   |__ --- knownGene
|   |
|   |__ genomic_interactions
|   |   |__ --- exampleInteractions
|   |
|   |__ RNA_seq
|   |   |__ --- exampleBW
|   |
|   |__ TAD
|   |   |__ --- exampleBed
|
Summary:
  There are 4 track groups in reference genomes "hg38":
  genes,genomic_interactions,RNA_seq,TAD
```

Figura 2.6: Struttura ad albero delle tracce genomiche caricate

Il sito web possiede alcune funzionalità interattive implementate attraverso JavaScript. Ciò permette all'utente di scegliere il sottoinsieme di tracce da visualizzare, di modificare le coordinate del genoma trascinando le tracce con il mouse a destra

o a sinistra o, ancora, ingrandire e rimpicciolire il genoma scorrendo la rotellina del mouse tenendo il puntatore nell'area delle coordinate del genoma.

Per visualizzare il browser sarà in seguito necessaria l'aggiunta di alcuni tag HTML.

### 2.3.4.3 Embedding

GIVE, come accennato precedentemente fornisce un modo semplice per configurare un genome browser inserendo dei tag HTML specifici che permettono di aggiungere ad una pagina web personale la visualizzazione interattiva di più tipi di tracce di dati: annotazioni genomiche in formato *Bed*, dati quantitativi in formato Wig/BigWig e interazioni genomiche in formato matrice di interazione *HiC*.

Questa piattaforma include inoltre un'interfaccia grafica chiamata GIVE HUG (HTML Universal Generator). Quest'interfaccia genera in modo interattivo del codice HTML che permette di visualizzare il set di dati scelto dall'utente. Il codice quindi, dopo essere stato inserito all'interno del proprio sito web permette di visualizzare al suo interno i dati interattivi.

Uno sviluppatore web può importare l'intera libreria GIVE in una pagina web inserendo le due righe seguenti (Fig. 2.7).

```
<script src="https://www.givengine.org/bower_components/webcomponentsjs/webcomponents-loader.js"></script>  
<link rel="import" href="https://www.givengine.org/components/chart-controller/chart-controller.html">
```

Figura 2.7: File da importare per l'*embedding* di GIVE

Mentre per visualizzare i dati genomici, lo sviluppatore web può utilizzare il tag *chart-controller* o il tag *chart-area* (Fig. 2.8).

Questo tag l'attributo "num-of-sub" permette di definire il numero di volte in cui

si desidera visualizzare le tracce inserite. Ad esempio, se si desidera confrontare due regioni diverse di uno stesso dato sarà necessario inserire il valore "2".

```
<!-- embed the browser in the web page -->
<chart-controller title-text="A 2-minute starter of building a genome browser with GIVE" ref="hg19" num-of-subs="2"
  coordinate='["chr18:19140000-19450000", "chr18:19140000-19450000"]' group-id-list=["genes", "Chi-C_promoter"]'>
</chart-controller>
```

Figura 2.8: Codice per la creazione del genome browser nella pagina web

### 2.3.5 WashU

WashU è un sistema che permette di visualizzare grandi set di dati genomici sequenziati su tutto o parte del genoma su specifici set di geni.

Gli utenti possono esaminare centinaia di set di dati su diversi livelli che vanno dall'intero genoma fino ai singoli nucleotidi, permettendo di avere sia una visione d'insieme sia una più nel dettaglio. Inoltre, ogni traccia può essere ordinata, organizzata, trascinata e rilasciata individualmente o in combinazione con i propri metadati e i dati sono visualizzabili in base a diverse caratteristiche genomiche selezionate, come ad esempio promotori o qualsiasi set di geni o coordinate genomiche.

WashU rende inoltre possibile fare delle analisi statistiche standard e visualizzare i risultati graficamente sul browser. Queste funzionalità aiutano l'utente a ottenere rapidamente informazioni dai dati su scala genomica che possono essere utilizzati per testare o persino generare ipotesi.

2.3.5.1 Caratteristiche principali di WashU



Figura 2.9: Caratteristiche di WashU

Le caratteristiche principali di WashU (Fig. 2.9) sono [28]:

- (a) Elementi modulari del codebase. Il nome di ogni componente è indicato nella colonna sinistra.
- (b) Funzioni per ogni voce del menu.
- (c) Interfaccia di gestione dei metadati e la mappa dei colori dei metadati. Ogni metadato è rappresentato con un valore univoco, ad esempio il rosa significa IMR900. Questa colonna è inoltre personalizzabile e fornisce un modo per selezionare tracce adiacenti che condividono gli stessi valori di metadati.
- (d) Visualizzazione di diverse tracce di dati relative al gene HOXA.

- (e) quattro strumenti attivabili che modificano il comportamento del mouse, tra cui trascinamento, zoom e riordino delle tracce; una serie di pulsanti di panoramica e zoom; e gli strumenti "Annulla", "Ripristina" e "Cronologia" (da sinistra verso destra).

Questa nuova architettura suddivide l'interfaccia utente in parti modulari che gestiscono le attività in modo indipendente (Fig 2.10). Anche dal punto di vista dello sviluppatore si hanno numerosi vantaggi, ad esempio la creazione di nuove funzionalità richiede solo la modifica di una sola parte del codice anziché di molte .

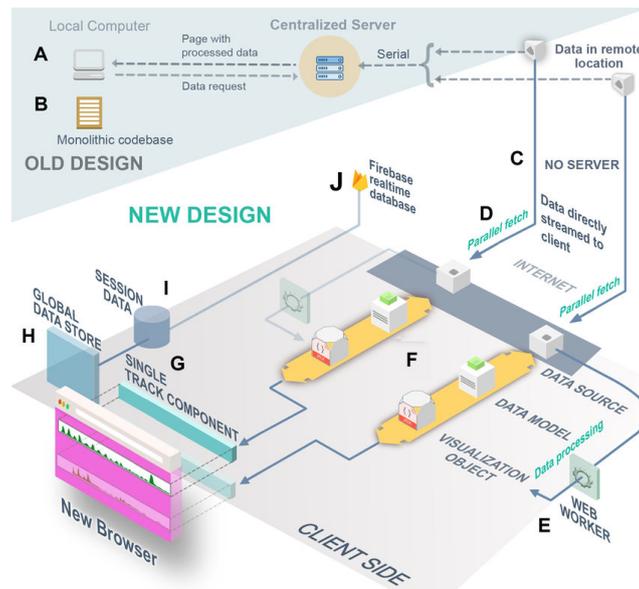


Figura 2.10: Architettura software della nuova versione di WashU

In futuro, questa architettura potrebbe facilmente incorporare un sistema di estensione in grado di caricare nuove funzionalità in modo dinamico, proprio come il sistema di estensione di Google Chrome. Le previsioni sono che i bioinformatici sviluppino e condividano estensioni e questo permetterebbe di arrivare ad un livello di personaliz-

zazione senza precedenti.

Gli sviluppatori non sono gli unici a vedere i benefici. Come risultato della riscrittura del codice, gli utenti hanno riscontrato una serie di miglioramenti delle prestazioni e della stabilità. Nella versione precedente, se una traccia aveva un bug o non riusciva a recuperare i dati, poteva causare il fallimento anche delle altre tracce, con conseguente perdita di lavoro. Nella nuova versione, i problemi relativi ad una specifica traccia non influiranno in alcun modo su altre tracce e gli utenti potranno tentare un ricaricamento senza perdere lavoro [28].

### 2.3.5.2 Tracce locali e data hubs

WashU possiede una funzionalità che permette ai ricercatori di visualizzare i propri file senza necessariamente configurare un Web server. Infatti, è possibile leggere dati da file locali o attraverso URL.

L'utente può quindi caricare singoli file o intere cartelle raggruppate come hub di dati. Questa funzione è fondamentale per i ricercatori che desiderano garantire la sicurezza dei dati, in particolare per i dati clinici in cui è necessario proteggere la privacy del paziente. Alcuni file, come *BigWig* e *HiC*, che hanno un formato binario, possono essere usati direttamente sul browser. Mentre altri invece, devono prima essere elaborati.

Ad esempio, i formati *Bedgraph* e *Bed*, di cui ci siamo occupati principalmente, devono prima essere compressi in "bgzip" e indicizzati in seguito con "tabix".

Ad esempio, un file "track.bedgraph" dovrà prima essere ordinato con un generico comando di Linux e, in seguito, sarà compresso utilizzando "bgzip" e indicizzato con "tabix" (Fig. 2.11).

```
# Using Linux sort
sort -k1,1 -k2,2n track.bedgraph > track.bedgraph.sorted
# Using bedSort
bedSort track.bedgraph track.bedgraph.sorted
# Using sort-bed
sort-bed track.bedgraph > track.bedgraph.sorted

bgzip track.bedgraph.sorted
tabix -p bed track.bedgraph.sorted.gz
```

Figura 2.11: Comandi generici *sort* Linux (sinistra) e "bgzip" e "tabix" (destra)

I file ottenuti saranno quindi "track.bedgraph.sorted.gz" e "track.bedgraph.sorted.gz.tbi", entrambi necessari per visualizzare correttamente il primo file nell'interfaccia grafica di WashU.

### 2.3.5.3 Embedding

È possibile incorporare il browser in qualsiasi file HTML in due modi differenti.

In un caso, sarà necessario importare alcuni file specifici in modo da poter sfruttarne le specifiche funzionalità. Dopodiché bisognerà creare un file JSON al cui interno verrà indicato il genoma di riferimento e le tracce che si desiderano visualizzare.

All'interno del file JSON è possibile indicare il tipo di file, il nome, l'URL, le opzioni e i metadati. È necessario ricordare che, nel caso in cui si volessero visualizzare dei file *Bed* o *Bedgraph* questi dovranno sempre essere sempre processati in "tabix" per ottenere i file "gz" e "tbi" corrispondenti.

Inoltre, per poterlo visualizzare correttamente basterà solo indicare l'URL del file "gz", mentre il "tbi" ad esso associato dovrà semplicemente trovarsi all'interno dello stesso percorso. Le tracce vengono infatti elaborate utilizzando uno script che ricerca all'interno del percorso in cui si trova il "gz" il "tbi" corrispondente.

Infine, basterà utilizzare la funzione *renderBrowserInElement(contents, container)*

che al suo interno prende due parametri: il *div* in cui si desidera visualizzare l'elemento e il file JSON creato. Chiamando la funzione il browser verrà immediatamente visualizzato all'interno della propria pagina.

Un'alternativa a questo, è la creazione, all'interno della pagina HTML, di un tag di un elemento *iframe* a cui bisognerà aggiungere alla proprietà *src* il percorso della pagina WashU specificando il genoma di riferimento che si desidera visualizzare. Se insieme al browser si desidera mostrare delle tracce di default basterà specificare il percorso del file JSON, contenente le tracce, dopo l'indirizzo della pagina di WashU(). È necessario che il file sia messo online e, per ragioni di sicurezza, non saranno visualizzati JSON salvati localmente.

### 2.3.6 GIVE e WashU: un breve confronto

L'attenta analisi delle principali funzionalità di entrambi i genome browser ha messo in luce alcune somiglianze e differenze. GIVE e WashU si sono rivelati molto simili su alcuni aspetti fondamentali:

- effettuare la ricerca utilizzando sia il nome del gene sia le sue coordinate.
- aggiunta di tracce sia locale sia attraverso URL.
- ordine, rimozione e confronto delle tracce visualizzate nell'interfaccia grafica.

Tuttavia, differiscono per altre caratteristiche.

#### 2.3.6.1 GIVE: punti di forza e limiti

Una funzionalità interessante che troviamo in GIVE è quella presente all'interno della sua interfaccia grafica. Come già accennato precedentemente infatti, questa presenta

due layers che permettono di visualizzare, confrontare e analizzare diverse regioni dello stesso genoma. Dunque, è possibile aggiungere al browser due coordinate genomiche in parallelo, com'è possibile vedere nella parte superiore del pannello di visualizzazione (Fig. 2.12).

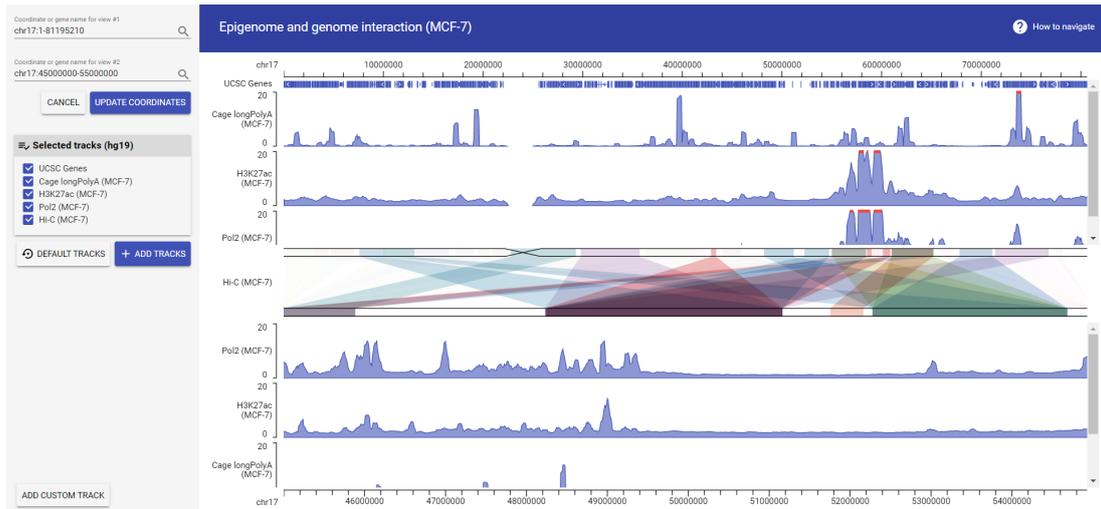


Figura 2.12: Interfaccia grafica di GIVE

GIVE supporta al momento, solamente tre tipo di tracce:

- *BedTrack*
- *BigWig Track*
- *Interaction Track*

Inoltre, i dati all'interno dell'interfaccia grafica possono essere visualizzati sempre nello stesso ordine di grandezza.

### 2.3.6.2 WashU: punti di forza e limiti

WashU rispetto a GIVE supporta più tipologie di formato. I dati presenti nell'interfaccia grafica, tuttavia, vengono mostrati sempre in un solo layer e, di conseguenza è possibile confrontare diverse tracce solo all'interno della stessa area del genoma. Un'ulteriore funzione interessante di WashU è la pendenza di una barra che permette la visualizzazione delle tracce sotto più prospettive diverse (confrontare Fig. 2.13 e Fig. 2.14).

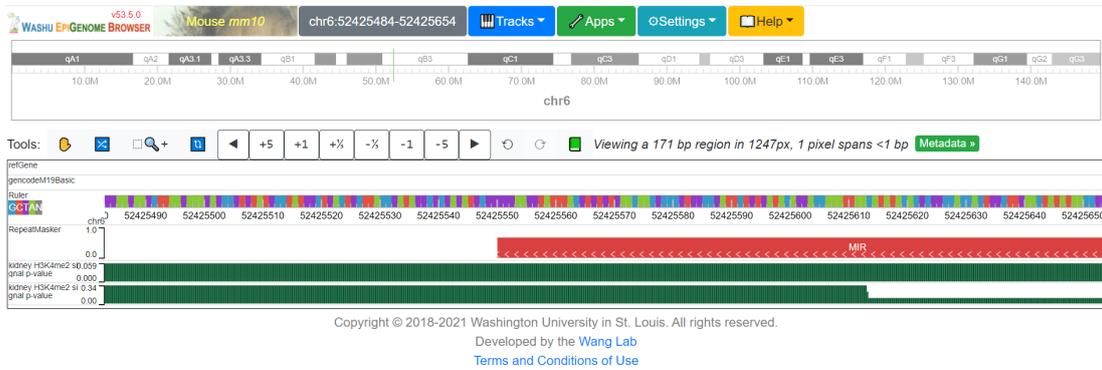


Figura 2.13: Visualizzazione di una regione di 171 bp in 1247px

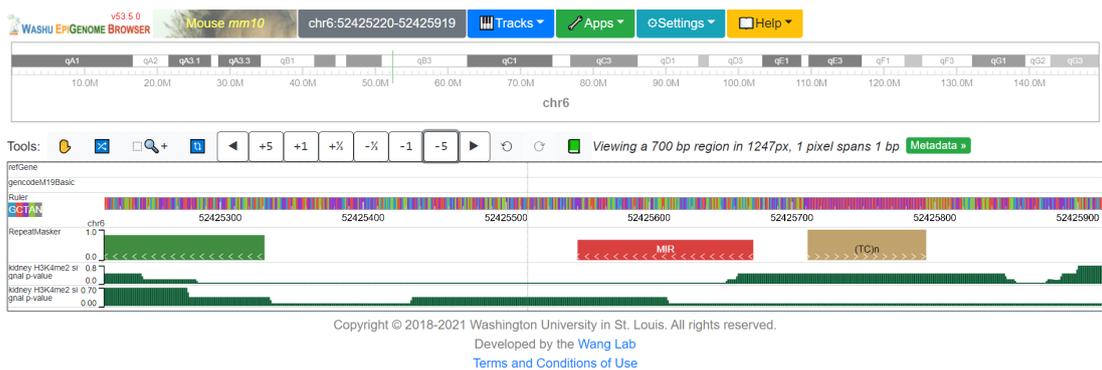


Figura 2.14: Visualizzazione di una regione di 700 bp in 1247px

### 2.3.6.3 Valutazione

Dopo aver effettuato un'attenta analisi i ricercatori del Dipartimento di Biologia hanno optato per l'esclusione di Ensembl e UCSC, mentre scegliere tra GIVE e WashU è stato molto più arduo. Di fatto, sono state effettuate, per entrambi i genome browser, diversi tipi di analisi prima di poter arrivare ad una scelta definitiva. Ad esempio, per ognuno sono stati creati più *embedding* all'interno di diverse pagine HTML e sono state caricate molte tracce di più formati diversi.

Infine, dopo aver analizzato scrupolosamente punti di forza e debolezza di entrambi i genome browser la scelta è ricaduta su WashU per diverse motivazioni. Prima di tutto, WashU permette il caricamento di più formati di dati. Inoltre, questo genome browser consente all'utente di visualizzare le diverse tracce a più livelli di dettaglio. Infine, l'*embedding* di WashU attraverso l'*iframe* garantisce una maggiore scalabilità. Ad esempio, in futuro permetterà di aggiungere, in modo molto semplice, un ulteriore genome browser per visualizzare dati su un diverso genoma.

---

# Strumenti utilizzati per lo sviluppo del sito web

## 3.1 Metodologie e strumenti utilizzati

Il lavoro relativo allo sviluppo del sito web è stato organizzato in due fasi principali:

1. Design
2. Implementazione

Nella prima fase, per ogni pagina sono stati organizzati i contenuti utilizzando inizialmente carta e penna. Sono state quindi create diverse bozze in cui ad ogni elemento veniva assegnata una posizione specifica all'interno della pagina in cui sarebbe stato

contenuto. In seguito, attraverso l'utilizzo di Figma<sup>9</sup> e Nicepage<sup>10</sup> sono stati realizzati prima *wireframe* e *mockup*. In breve, un *wireframe* è una rappresentazione a bassa fedeltà di un design mentre un *mockup* è una rappresentazione molto vicina a come verranno visualizzati tutti gli aspetti finali. In particolare, Figma è stato utilizzato nella prima fase iniziale per definire la grafica della navbar e del titolo della pagina, mentre Nicepage per il corpo.

Per la fase di programmazione invece sono stati utilizzati diversi strumenti, quali: HTML, Javascript e CSS. Di fatto, l'utilizzo di HTML e Javascript è stata una scelta dovuta principalmente all'*embedding* dei genome browser effettuato, per l'appunto, utilizzando questi strumenti. Infine, si è scelto di utilizzare JQuery per sviluppare alcune funzionalità del sito in modo più semplice e scrivere un codice più pulito e Bootstrap per avere un maggiore controllo sul responsive delle pagine.

## 3.2 Design

Per la prima parte di design è stato scelto Figma, una piattaforma finalizzata allo sviluppo di siti web che permette l'utilizzo di strumenti responsive.

Su Figma è stata creata una prima bozza della struttura del sito ponendo l'attenzione, in modo particolare, del logo e della navbar.

Per la progettazione effettiva delle singole pagine e dei loro contenuti, invece, è stato utilizzato Nicepage. In particolare, Nicepage ha consentito il passaggio dalla fase di progettazione a quella di implementazione effettiva senza soluzione di continuità.

---

<sup>9</sup><https://www.figma.com/files/recent?fuid=977202153177403752>

<sup>10</sup><https://nicepage.com/>

### 3.2.1 Figma

Figma è un tool per la progettazione di interfacce pensato per realizzare il design di siti web e applicazioni.

Uno dei vantaggi legato a questo software è che può essere utilizzato online, quindi non è necessaria l'installazione di nessun programma. Ciò permette all'utente di poter lavorare semplicemente attraverso il browser. Questa modalità permette inoltre il salvataggio automatico del proprio lavoro che, può comunque essere salvato, in qualsiasi altro momento, anche in locale in formato FIG o, nel caso delle immagini, in formato PNG, JPG, SVG e PDF. È comunque possibile l'utilizzato offline di Figma attraverso due diverse applicazioni installabili rispettivamente su macOS e Windows [37].

Figma permette inoltre di lavorare al design della propria pagina utilizzando diversi *frame*, in modo da ottenere un perfetto responsive design.

In breve, quando si parla di *textitframe* si fa riferimento alla singola pagina su cui lavora. In altri programmi di grafica vengono anche chiamati Artboards o tavole da disegno [37]. Ciò permette di creare un design per il proprio sito che si adatti perfettamente ad ogni tipo di dispositivo su cui verrà visualizzato.

Questo tool permette l'interazione tra i *textitframe* attraverso l'uso di animazioni. Questa funzione è estremamente utile quando si vuole realizzare un prototipo del sito che sia visualizzabile in modo interattivo, ad esempio comprendere come verrà visualizzato un *dropdown* della propria pagina. Sono presenti diversi tipi di interazioni, quali *click*, *drag*, *hover*; diversi tipi di azioni come *navigate to*, *go back*, *scroll to*, *open overlay*; e infine diverse tipologie di animazioni come *istant*, *dissolve*, *smart animate* o *move in*.

Una volta definito il design della propria pagina, una funzione di Figma estremamente utile all'utente è *inspect*, poiché questa permette di visualizzare le proprietà di uno specifico elemento grafico in codice CSS, XML (per Android) o Swift (per iOS).

### 3.2.2 Nicepage

Nicepage è un'applicazione di Automated Web Design per sistemi Mac e Windows. Contiene al suo interno una serie di template che possono essere utilizzati come base per creare la propria pagina web. È possibile, attraverso le diverse funzionalità dell'applicazione personalizzare il design della pagina web che si vuole creare e renderlo responsive, grazie a una particolare funzionalità di Nicepage che permette di organizzare e visualizzare i contenuti su diversi *device*.

## 3.3 Implementazione

A seguito della fase di progettazione e, in particolare, grazie alla funzionalità di Nicepage di poter esportare il design delle singole pagine in HTML, è stato possibile passare alla fase di implementazione delle effettive funzionalità del sito utilizzando i seguenti linguaggi e strumenti.

### 3.3.1 HTML

HTML, acronimo di HyperText Markup Language, è un linguaggio markup la cui sintassi è stabilita dal World Wide Web Consortium (W3C), che permette di definire il modo in cui disporre gli elementi all'interno di una pagina [29].

La prima versione del linguaggio è stata definita nel 1991 da Tim Barners Lee e

nel corso del tempo si è evoluta fino ad arrivare allo standard HTML5, attualmente utilizzato. Questo linguaggio utilizza un insieme definito di tag o marcatori, inclusi tra parentesi angolari, utilizzati per descrivere gli elementi di una pagina web. L'inserimento dei tag all'interno del documento HTML fornisce al browser le istruzioni necessarie per strutturare il documento che verrà visualizzato sullo schermo. Ovviamente, il marcatore è nascosto e il browser mostra solo il contenuto dell'elemento, dunque solo quello che c'è tra il tag[18].

### 3.3.2 CSS

L'acronimo CSS sta per Cascading Style Sheets e designa un linguaggio di stile utilizzato per definire la formattazione di documenti HTML, XHTML e XML. A partire dal 1996, W3C emana una serie di direttive al cui interno sono contenute le regole per comporre un CSS [11].

I file CSS forniscono al browser le istruzioni necessarie per presentare il documento all'utente, definendo, ad esempio, il font, i colori, le immagini, il layout o lo stile dei diversi elementi presenti sulla pagina.

La storia del CSS procede di pari passo con la storia dell'HTML, poiché questi due linguaggi sono complementari. Di fatto, un linguaggio HTML dovrebbe essere visto semplicemente come un linguaggio strutturale, alieno da qualunque scopo attinente alla presentazione di un documento. Per arricchire l'aspetto visuale e la presentazione di una pagina, si utilizza invece il CSS. In questo modo è possibile separare il contenuto dalla presentazione [26].

La prima specifica ufficiale di CSS risale al dicembre del 1996, mentre nel maggio 1998 è stata la volta della seconda versione: CSS 2. Quest'ultimo non fa altro che

aggiungere al primo alcune nuove proprietà, valori di proprietà e definizioni per stili non canonici come quelli rivolti alla stampa o alla definizione di contenuti audio [26].

### 3.3.3 Javascript

Javascript fu ideato nel 1995 da Netscape e venne rilasciato per la prima volta con la prima versione del browser Netscape Navigator, inizialmente sotto il nome di LiveScript e, in un secondo momento, con l'attuale nome, creando non poca confusione con Java che veniva lanciato proprio quell'anno.

È un linguaggio di programmazione orientato agli oggetti e agli eventi per la creazione, in siti web e applicazioni web. Inizialmente utilizzato solo nella programmazione Web lato client fu esteso, in seguito, anche al lato server.

I programmi che sfruttano questo linguaggio sono chiamati script e sono eseguiti in automatico al caricamento della pagina. Gli script vengono scritti ed eseguiti come testo semplice e, proprio per questo, non è necessaria alcuna fase di preparazione o compilazione. Sotto questo aspetto, JavaScript è estremamente differente da Java.

In quest'ultimo linguaggio invece la sintassi va passata ad un compilatore che produce un file binario in cui la sintassi scompare. Ad esempio, i programmi di Windows sono dei file compilati riconoscibili dal formato "EXE". Al loro interno, non c'è traccia della sintassi originaria o, in altre parole, del codice sorgente.

Javascript invece è non è compilato dunque, all'interno di una pagina HTML è possibile visualizzare in qualsiasi momento il codice Javascript.

Inoltre, dire che è un linguaggio di scripting sottintende anche il fatto che si tratta di un linguaggio interpretato: ovvero l'assenza del compilatore impone che sia lo stesso browser a «tradurre» i comandi in operazioni mediante un apposito «motore di

scripting» che legge le parti di codice JavaScript.

### 3.3.4 JQuery

Negli ultimi anni sono nati diversi framework di Javascript nati con lo scopo di rendere più veloce e semplice lo sviluppo di interfacce utente (UI). Di questi framework fa parte JQuery una libreria Javascript nata nel 2006 per opera di John Resign.

Rispetto a Javascript, JQuery risulta essere molto più semplice e flessibile, diventando in poco tempo uno dei framework più utilizzati e diffusi tra i webmaster di tutto il mondo.

Tra i punti di forza di jQuery è importante citare la sua elevata compatibilità con la maggior parte dei browser in circolazione e la sua capacità di lavorare in simultanea con altre librerie senza creare conflitti.

Le funzionalità integrate nella libreria permettono di manipolare i CSS, selezionare elementi del DOM, personalizzare gli eventi generati dall'utente, creare chiamate AJAX, senza tralasciare i numerosi effetti grafici.

Infine, un altro aspetto importante di JQuery è la possibilità di utilizzare i numerosi plugin disponibili e scaricabili gratuitamente plugins [7].

### 3.3.5 Bootstrap

Bootstrap nasce nel 2010 per opera degli sviluppatori Mark Otto e Jacob Thornton. Viene pensato inizialmente come un progetto interno a Twitter, ma diventa in seguito indipendente e utilizzabile dagli sviluppatori di tutto il mondo come base per la realizzazione di interfacce web. Al momento è considerato framework più efficiente per rendere i siti web responsive.

La sua definizione esatta è "HTML, CSS, and JS toolkit from Twitter", ovvero una raccolta di strumenti grafici, stilistici e di impaginazione che permettono di avere a disposizione una gran quantità di funzionalità e di stili modificabili e adattabili a seconda delle proprie esigenze [8].

Il suo utilizzo risulta molto versatile, in quanto è compatibile con tutte le ultime versioni dei principali browser. Dalla versione 2.0 supporta anche il responsive web design. Ciò vuol dire che è in grado di adattarsi dinamicamente a seconda della grandezza e delle caratteristiche del dispositivo utilizzato, sia esso desktop, tablet o telefono cellulare. Questa raccolta di strumentazione di tipo grafico si pone, perciò, come una libreria multi dispositivo e multiplatforma [6].

### 3.4 Considerazione delle scelte

Nel campo del design, si è scelto Figma perché si voleva utilizzare un tool specifico per la creazione di *wireframe*, *mockup* e prototipi finalizzati alla creazione di siti web. Inoltre, Figma mette a disposizione degli studenti, in modo totalmente gratuito, tutte le funzionalità del piano professionale. Questo aspetto ha favorito la scelta di Figma rispetto ad altri tool, come ad esempio Adobe XD o Sketch). Di fatto, nonostante quest'ultimo sia gratuito molte delle sue funzionalità sono a pagamento.

I primi *wireframe* del sito sono stati tutti realizzati su Figma. Tuttavia, in un secondo momento si optato per utilizzare, in concomitanza, anche Nicepage (v.d 3.2.2). Questa scelta è stata presa in quanto si voleva partire da un template di base, ma allo stesso tempo che potesse essere riadattato in base a specifiche richieste. Nicepage permette, dopo aver definito e personalizzato il design della propria pagina web, di

poterlo riadattare e visualizzare su diversi dispositivi. Questa funzione si è rivelata particolarmente utile per definire un layout del sito che fosse anche responsive.

Insieme, Nicepage e Figma possiedono una funzione che permette di sportare direttamente l'HTML e CSS del sito creato. Tuttavia, su Figma il risultato si è rivelato non soddisfacente per i fini del lavoro di questa tesi e, di conseguenza, considerando che il *mockup* della navbar era stato disegnato proprio su Figma, la scelta è stata quella di ricrearla direttamente nel codice, implementando le sue complete funzionalità senza utilizzare alcun tool grafico. La navbar è stata creata infine utilizzando Bootstrap, principalmente per motivi di responsive design.

La scelta di HTML, CSS e Javascript è dovuta principalmente al fatto che questi sono gli strumenti più classici utilizzati nel web design, ma non solo. Come mostrato nel capitolo precedente (vd. 2.3.5.3 e 2.3.4.3), infatti, nei tutorial per l'*embedding* in una pagina web dei genome browser venivano utilizzati in entrambi i casi questi strumenti. Di conseguenza, questo fattore ha permesso di prendere una decisione definitiva.

Inoltre, tutte le funzionalità del sito sono state implementate non solo attraverso l'uso del linguaggio di programmazione Javascript, ma anche attraverso la sua libreria JQuery. Questo ha permesso di:

- semplificare lo scripting lato client di HTML.
- utilizzare un numero inferiore di righe per eseguire un'attività.
- manipolare più facilmente il contenuto della pagina.

---

# Contributo

## 4.1 Organizzazione del lavoro

Il contenuto di questa tesi si struttura in quattro fasi principali:

1. Studio e approfondimento di argomenti di natura biologica per affinare la comprensione dei genome browser e del contenuto dei dati da processare per il Dipartimento di Biologia dell'Università di Pisa.
2. Elaborazione di *wireframe* per le diverse sezioni.
3. Implementazione del sito.

Nella prima fase è stato necessario lo studio e l'approfondimento della biologia e, in modo particolare della genetica. Grazie a questo è stato possibile ottenere una

maggior consapevolezza sull'utilizzo dei genome browser e sull'importanza che questi hanno nella ricerca. La prima fase si è inoltre rivelata fondamentale anche per la comprensione ed elaborazione delle tracce analizzate dal gruppo di ricercatori guidati dal professor Borello.

Dopo aver ultimato la prima fase, la seconda e la terza si sono susseguite senza soluzione di continuità. Per esempio, la necessità di aggiungere una nuova funzionalità ad una pagina già implementata ha portato allo sviluppo di nuovi *wireframe* e, di conseguenza a nuove scelte implementative.

Il risultato finale ottenuto è stato un primo prototipo del sito completamente funzionante.

## 4.2 Studio di Biologia Molecolare e Genetica

Nella prima fase, lo studio e l'approfondimento di elementi di biologia molecolare e genetica ha permesso di comprendere meglio le finalità del progetto e il suo scopo: creare una piattaforma per la raccolta e l'esplorazione di dati epigenomici sulla Sindrome di Rett.

Grazie a questi studi è stata acquisita una maggior consapevolezza sul lavoro condotto dal Dipartimento di Biologia dell'Università di Pisa e quindi sulle tracce da loro elaborate, ma non solo. Come diretta conseguenza di questi studi, è stato possibile sviluppare una maggior comprensione dei genome browser, il loro funzionamento e loro importanza nella ricerca scientifica.

Di fatto, come già precedentemente accennato (vd. 2.1), questi browser permettono non solo di fornire una visione d'insieme dei dati ma anche di poter effettuare una va-

lutazione comparativa tra diverse sequenze genetiche. Questo poiché - se si considera che le informazioni vengono osservate nell'insieme del genoma e non isolatamente - la visualizzazione dei dati ne rende più semplice l'interpretazione.

### 4.2.1 Caricamento delle tracce

Prima di effettuare una scelta definitiva sul genome browser da inserire all'interno della pagina web, si è proceduto con il caricamento di alcune tracce elaborate dai colleghi di biologia dopo aver eseguito l'*embedding* sia di WashU sia di GIVE.

È importante sottolineare che i file devono essere processati e caricati in modo differente a seconda del browser e, di conseguenza, parte del lavoro svolto è stato quello di prendere familiarità con questi strumenti in modo da poter avere una corretta visualizzazione dei dati. Le tracce di annotazione su cui ci si è focalizzati sono state principalmente quattro: *BigWig*, *Bed*, *BedGraph* e *HiC*.

Su WashU, il caricamento dei file in formato *BigWig* e *HiC* è quello più immediato poiché le tracce di questi formati non richiedono nessun tipo di processamento per poter essere visualizzate. Per i file in formato *Bed* e *Bedgraph* invece è prima necessario la compressione in "bgzip" e l'indicizzazione in "tabix" (vd. 2.3.5.2).

Infine, per far visualizzare le tracce all'interno della pagina web è stato necessario creare un file JSON, seguendo una struttura specifica indicata nella documentazione di WashU, contenente al suo interno gli indirizzi di riferimento delle tracce.

Su GIVE invece, ogni traccia può essere caricata senza effettuare alcun tipo di modifica. È importante ricordare che GIVE non supporta i file di tipo *Bedgraph*, dunque per quest'ultimo formato non è stato possibile effettuare nessun tipo di test.

In questo caso, la visualizzazione delle tracce all'interno di una pagina web si è rive-

lato essere più complesso rispetto a quello di WashU. Infatti, per poter visualizzare le tracce nell'interfaccia grafica di GIVE è necessario per prima cosa che queste vengano inserite all'interno dell'immagine creata in Docker. Per questo motivo è stato prima necessario, utilizzando uno specifico comando sul terminale di Linux, creare una nuova directory che ospitasse il volume creato appositamente per Docker (Fig. 4.1).

```
mkdir ~/container-data
```

Figura 4.1: Comando per la creazione di una directory

Un volume Docker è semplicemente una directory che si trova nel file system host e non fa parte dell'UFS (Union File System) del container. Nonostante questo, però, è possibile salvare al loro interno i dati del container [39].

In seguito, è bastato combinare il comando base di Docker per la creazione dell'immagine con il comando specifico per l'aggiunta di un nuovo volume (Fig. 4.2).

```
docker run -d -it -p 40080:80 -p 40443:443 --name give -v ~/container-data:/data zhonglab/give
```

Figura 4.2: Comando Docker per la creazione dell'immagine e aggiunta del volume

Dopo aver creato il container quindi, questo includerà una directory "data" creata sul volume host in "container-data" in cui dovranno essere inserite le tracce che si intende caricare e dopodiché sarà possibile gestirle utilizzando GIVE-Toolbox (vd 2.3.4.2) e in seguito visualizzarle all'interno della propria pagina web.

Prima di poter visualizzare le tracce all'interno dei genome browser è stato però necessario un lavoro preliminare. Alcuni dei dati rielaborati dal gruppo di ricercatori guidati dal professor Borello presentavano una struttura un po' diversa da quella

mostrata precedentemente (Fig. 2.2) . Nella prima colonna, infatti, ad essere mostrato era solo il valore del cromosoma, mentre il riferimento "chr" veniva omissso (Fig. 4.3). Di conseguenza, i file non venivano visualizzati nell'interfaccia grafica di GIVE, mentre in WashU non era possibile l'esecuzione dei comandi Linux "tbi" e "tabix".

```
1 59302000 59302300 -1.0
1 59302300 59302600 -0.75
1 59302600 59302900 -0.50
1 59302900 59303200 -0.25
1 59303200 59303500 0.0
```

Figura 4.3: Formato *Bedgraph* non leggibile dai genome browser

Per cercare di risolvere questo problema, è stato creato un piccolo script in Python che, partendo dal file originale, andasse ad aggiungere su tutte le righe il prefisso mancante e che restituisse in output il file *Bed* o *Begraph* corretto (Fig. 4.4).

```
my_file = open('h3k36me3_treat_pileup.bdg')
all_the_lines = my_file.readlines()
x = "chr"
f = open("h3k36me3_treat_pileup.bedgraph", "w")
for i in all_the_lines:
    add = x + i
    f.write(add)
f.close()
```

Figura 4.4: Codice Python per la correzione dei file

In questo modo è stato possibile far leggere e visualizzare correttamente i dati ai diversi genome browser.

### 4.3 Elaborazione dei wireframe e mockup

Dopo aver approfondito alcuni argomenti di Genetica, come descritto nella sezione precedente, e dopo aver analizzato in maniera attenta e scrupolosa i diversi genome browser (sia per le loro funzionalità sia per la visualizzazione dei dati elaborati dal Dipartimento di Biologia dell'Università di Pisa) si è passati alla definizione dell'interfaccia utente e degli elementi di design. Questa fase è stata preliminare e propedeutica a quella realizzativa vera e propria.

Nello specifico, dopo aver identificato quali fossero gli utenti target (ricercatori esperti, non esperti e bioinformatici), è stato possibile proseguire, prima attraverso la definizione dei *wireframe* e, in seguito, attraverso la realizzazione dei *mockup*.

In questa fase, il processo di lavoro non è stato così lineare. Nel momento in cui i primi *wireframe* e *mockup* venivano elaborati, i contenuti da inserire all'interno della pagina erano ancora in fase di definizione e, di conseguenza, spesso i *mockup* creati venivano abbandonati per poi ripartire da nuovi *wireframe*.

Un *wireframe* è la prima bozza del sito e viene utilizzato per mostrare la disposizione degli elementi che saranno inseriti al suo interno. Dunque è fondamentale per definire una prima struttura degli elementi all'interno della pagina.

Principalmente lo scopo del *wireframe* è quello di:

- definire la struttura di base del sito.
- distribuire i contenuti all'interno della pagina secondo logiche di priorità.
- avere il consenso del cliente prima di procedere all'elaborazione grafica degli elementi

I *wireframe* possono essere creati attraverso programmi di elaborazione grafica, software specifici o semplicemente in modo "analogico", quindi utilizzando carta e penna. Nel caso specifico di questo progetto, si è preferito procedere utilizzando per primo quest'ultimo approccio per poi rielaborare il tutto in versione digitale.

L'elaborazione dei *wireframe* è stata seguita dalla creazione dei *mockup*. In questa fase, alla struttura base del sito, precedentemente definita vengono applicati elementi e stili grafici meglio definiti: colori, immagini, testi, font e così via. È quindi possibile definire un *mockup* come una bozza grafica del sito con un maggior grado di fedeltà, il cui obiettivo è quello di dare al cliente un'idea dell'aspetto finale del lavoro.

Per la preparazione dei *wireframe* si è seguito un processo di tipo "top down": è stata infatti individuata una prima struttura generale del sito per poi scendere nel dettaglio nei singoli elementi.

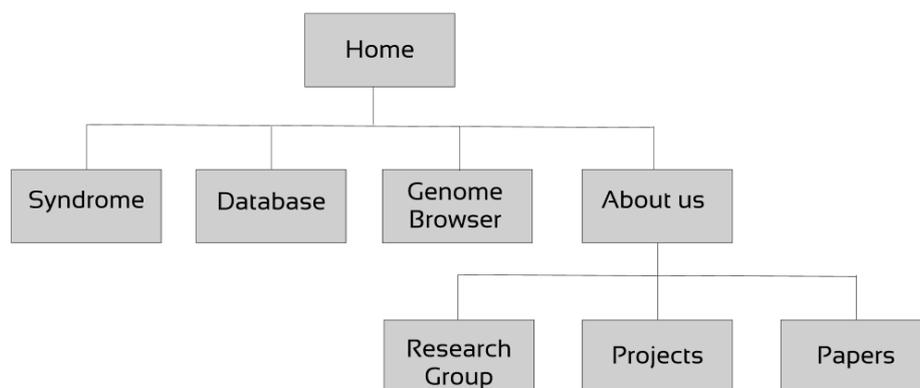


Figura 4.5: Sitemap

La struttura del sito prevede diverse pagine (Fig. 4.5):

- **Home:** la homepage del sito.

- **Syndrome:** pagina contenente informazioni relative alla Sindrome di Rett.
- **Database:** pagina contenente un database con tutti i file rielaborati dal Dipartimento di Biologia dell'Università di Pisa.
- **Genome Browser:** la pagina che contiene il genome browser WashU.
- **Research Group:** la pagina che contiene informazioni relative ai partecipanti del progetto.
- **Project:** la pagina che contiene informazioni sul progetto e ad altri ad esso correlati.
- **Papers:** la pagina che contiene la lista e i link che rimandano ai papers utilizzati.

### 4.3.1 Homepage

Un primo approccio utilizzato nella creazione dei *wireframe* è stato quello di creare un prototipo che rassomigliasse, nella disposizione dei contenuti, al sito all'interno del quale sarebbe poi stato inserito: BioMedical and Health Informatics Lab<sup>11</sup> (Fig. 4.6) che si occupa di sviluppare metodologie computazionali per la raccolta, l'elaborazione, l'analisi e l'interpretazione di dati biologici, medici e sanitari.

---

<sup>11</sup><https://biomedinfo.di.unipi.it/>

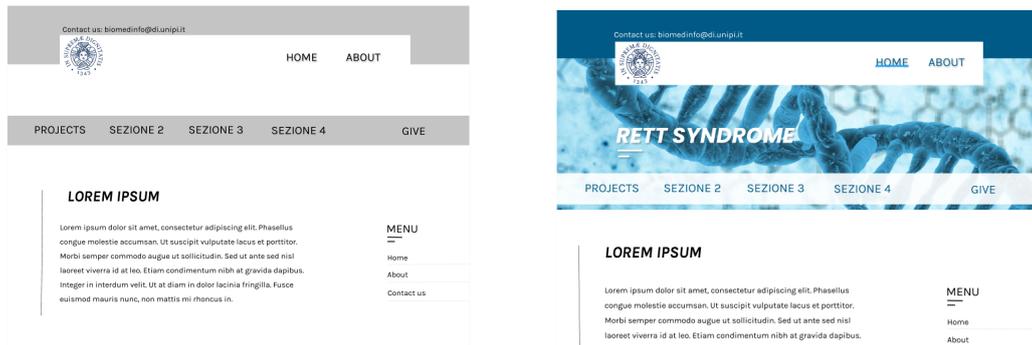


Figura 4.6: Primi *wireframe* e *mockup* elaborati

Di conseguenza, i primi *wireframe* e *mockup* sono stati elaborati avendo sempre come punto di riferimento il sito sopra citato.

Tuttavia, l'idea di creare una struttura ed elementi grafici troppo simili a quelli del sito di partenza si è rivelata essere insoddisfacente ed è stata velocemente abbandonata. Si è preferito invece avere un sito che richiamasse in qualche modo quello del laboratorio ma che avesse un aspetto ad esso indipendente.

Il rischio di creare una grafica troppo simile al sito BioMedical and Health Informatics Lab sarebbe stato quello di far apparire la pagina sulla sindrome di Rett come una pagina di quel sito, mentre in realtà ciò che si voleva creare era semplicemente un sotto ad essa collegato.

Di fatto, i *wireframe* a seguire sono stati progettati in modo completamente diverso rispetto al primo. È stato aggiunto un elemento grafico che fosse coerente con i contenuti presenti all'interno del sito, ovvero una rappresentazione del DNA. Per la navigazione tra le pagina, si è scelto invece di utilizzare una navbar che rendesse più semplice l'orientamento e la navigazione all'utente (Fig. 4.7).



Figura 4.7: *Wireframe* e *mockup* della seconda versione del sito

Il primo prototipo del sito è stato strutturato attraverso due riquadri principali. Il primo di colore bianco in cui erano contenuti a sinistra il logo dell'Università di Pisa e a destra un piccolo menu con tre voci: Home, Contact us e About us. Il riquadro blu invece ospitava il nome del sito arricchito con un'immagine del DNA e la navbar. Il corpo del sito è stato realizzato in modo estremamente semplice, specialmente nelle pagine in cui venivano inseriti contenuti testuali. L'idea era infatti quella di creare una pagina che fosse il più possibile minimale e che contenesse al suo interno semplicemente titolo e corpo separati tra loro da una semplice linea retta orizzontale.

Questa prima versione del sito realizzata senza utilizzare nessun tipo di template di base veniva gestita principalmente attraverso la programmazione ad eventi in Javascript. Di fatto, il corpo dell'HTML veniva ricreato dinamicamente in base alle interazioni dell'utente, mentre a rimanere fissi erano solo intestazione e il footer.

In questa fase, i contenuti del sito stavano sempre più perfezionandosi e ciò ha portato a rivedere anche alcuni aspetti della grafica già implementati. Ad esempio, il titolo del sito era troppo generico e di conseguenza si è pensato a qualcosa che potesse essere più specifico e che allo stesso tempo potesse rappresentare meglio i contenuti della

pagina.

Ripensare ad una grafica per il titolo ha portato ad una revisione pressoché totale della parte superiore del sito. Gli elementi principali sono stati mantenuti: stile del titolo, immagine del DNA e la linea orizzontale che interseca tutti questi elementi. Tuttavia, la loro disposizione è cambiata in modo da poter inserire tutto all'interno della navbar (Fig. 4.8). Inoltre, titolo e immagine sono stati combinati assieme come un unico elemento grafico.

Nel nuovo e ultimo prototipo del sito, la navbar contiene al suo interno tutti gli elementi che prima erano disposti su più riquadri e scompare il rettangolo bianco contenente il piccolo menu e il logo dell'Università di Pisa. La navbar diventa l'unico elemento di navigazione dei contenuti.

Anche il corpo del sito cambia radicalmente. È stato utilizzato come punto di partenza un template di base presente su Nicepage (vd), i cui elementi grafici sono stati riadattati in base alle specifiche esigenze della pagina.

L'idea della HomePage è stata quella di dare una visione d'insieme dei contenuti del sito, quindi mostrare all'utente una panoramica di alcuni dei contenuti principali del sito (per esempio, la pagina esplicativa sulla sindrome di Rett e il genome browser) e allo stesso tempo fornire una breve descrizione sull'obiettivo principale del sito, che viene poi approfondito nella sezione specifica.

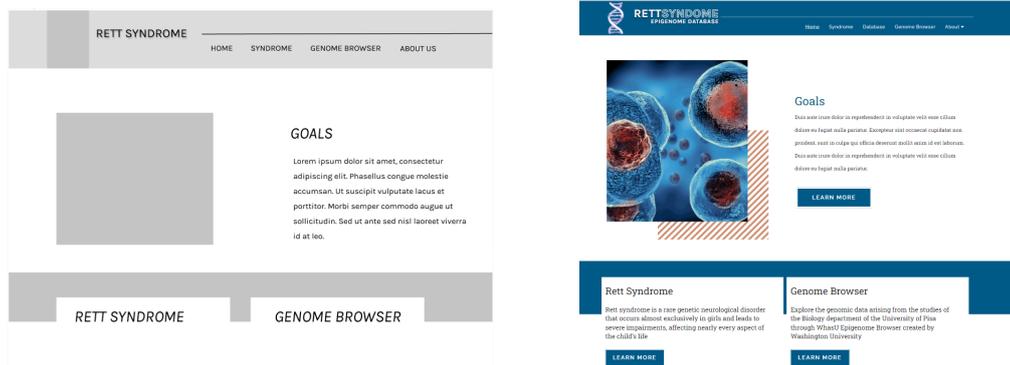


Figura 4.8: *Wireframe* e *mockup* della "Homepage"

Per quanto riguarda la scelta dei colori sono stati considerati dei colori che richiamassero quelli dell'Università di Pisa. Di conseguenza, i colori predominanti all'interno delle pagine sono sempre blu e bianco. In particolare, la tonalità di blu è  $\text{rgb}(0, 90, 135)$ .

### 4.3.2 Syndrome

La pagina esplicativa sulla sindrome di Rett ha una struttura molto semplice e permette all'utente di avere alcune informazioni fondamentali sulla malattia. In breve, viene quindi spiegato cos'è la malattia, com'è possibile diagnosticarla e i geni coinvolti attualmente conosciuti (Fig. 4.9).

L'obiettivo di questa pagina è quindi quello dare una visione d'insieme molto generale della malattia all'interno del sito, ma allo stesso tempo fare in modo che si abbiano tutti i riferimenti necessari per poter approfondire ognuno degli argomenti trattati. Sotto ogni sezione è possibile trovare un bottone "Read more" che rimanda l'utente a delle pagine specifiche di approfondimento.

Infine, in fondo alla pagina è stata aggiunta un'ultima sezione "Other Reference" in modo da poter aggiungere altri link utili di approfondimento su tematiche specifiche della malattia.

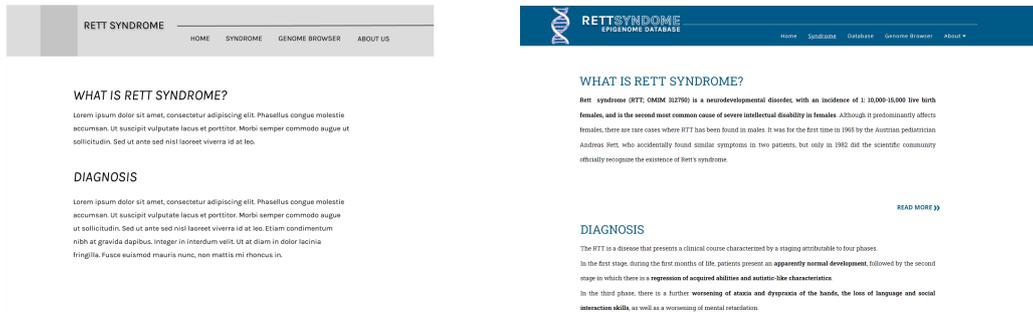


Figura 4.9: Wireframe e mockup della pagina "Syndrome"

### 4.3.3 Database

Il database, insieme al genome browser le due principali funzionalità che il sito fornisce ai ricercatori. La pagina del Database si compone di due elementi fondamentali: una tabella e una breve descrizione del campione inserito (Fig. 4.10). All'interno della tabella sono presenti dei dati che forniscono al ricercatore maggiori informazioni sulle tracce visualizzate nel genome browser. Inoltre, per ogni campione inserito all'interno della tabella vengono fornite diverse informazioni che è possibile approfondire attraverso i link forniti per ciascun GSM (v.d 4.4).

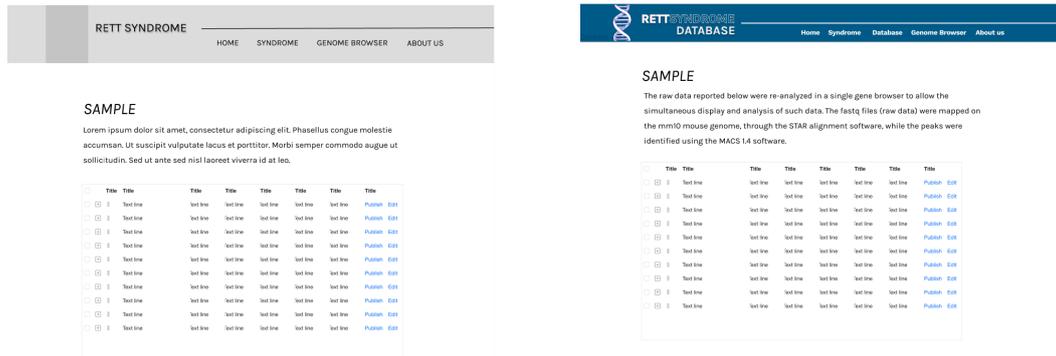


Figura 4.10: Wireframe e mockup della pagina "Database"

### 4.3.4 Genome Browser

La pagina del genome browser è quella che nel corso del tempo è stata riprogettata più volte (Fig.4.11 e Fig.4.12). Inizialmente, nella pagina era presente una lista di checkbox che permetteva all'utente di aggiungere e rimuovere dinamicamente le tracce presenti nell'interfaccia grafica. In seguito però, dopo aver approfondito le funzionalità di WashU, è stata considerata l'idea di rimuovere i checkbox sia perché esiste una funzione già implementata di WashU che permette il riordinamento delle tracce, sia perché si andava incontro ad altri tipi di problematiche che questa implementazione comportava (ad esempio, il genome browser veniva nuovamente ricaricato, v.d. 4.5). Un aspetto su cui si è stata posta particolare attenzione è l'utilizzo poco intuitivo di alcune funzioni di WashU. Inizialmente quindi, una prima soluzione è stata l'aggiunta di una nuova pagina al sito in cui erano disponibili alcuni tutorial consultabili liberamente dall'utente. Per ragioni di praticità, tuttavia, si è cercato di mantenere tutto all'interno della stessa pagina in modo che l'utente non fosse costretto ad uscire dalla pagine del genome browser per visualizzare i tutorial.

Un'altra problematica a cui si è andati incontro era scegliere la parte della pagina in cui posizionare i tutorial. Questa scelta è stata effettuata secondo il seguente ragionamento: si ponga, ad esempio, un utente esperto come un ricercatore di biologia, costui non avrà bisogno di alcun tutorial perché questi strumenti già sono ampiamente utilizzati nel campo. Tuttavia, si è voluto comunque considerare un pubblico più ampio. Dunque, un ricercatore che ha meno esperienza o poca dimestichezza con questi strumenti informatici potrebbe aver bisogno di una guida iniziale.

Dunque, inizialmente è stato inserito un bottone in alto a destra del genome browser che, se cliccato, eseguiva uno scroll-down della pagina in modo da portare l'utente nell'area specifica in cui il tutorial era stato inserito. In un secondo momento però è stata ridotta la grandezza dei bottoni: in questo modo è stato possibile collocarli nella parte alta della pagina senza che togliessero spazio al genome browser. Infine, cliccando su ogni bottone viene mostrato un tutorial specifico sotto forma di galleria che l'utente può scorrere (v.d 4.5).



Figura 4.11: Primi *wireframe* della pagina contenente il genome browser

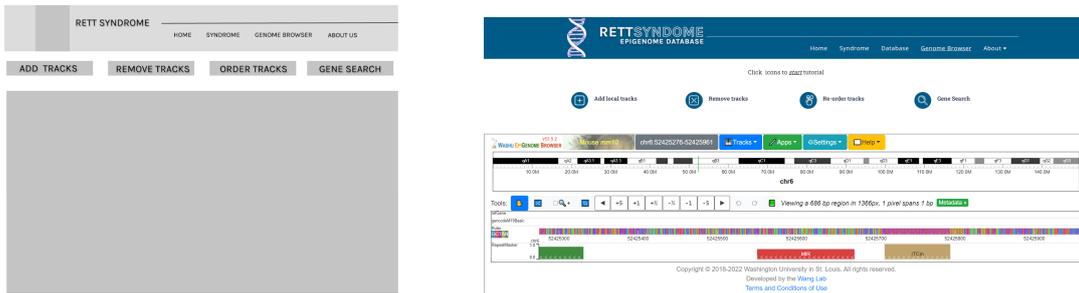


Figura 4.12: Wireframe e mockup finali della pagina del "Genome Browser"

### 4.3.5 About us

Attraverso la voce "About us" è possibile accedere a tre diverse pagine del sito: Research Group, Projects e Papers.

#### 4.3.5.1 Research Group

Nella pagina Reserach Group è possibile vedere chi ha collaborato al progetto (Dipartimento di Biologia dell'Università di Pisa e Dipartimento di Bioinformatica dell'Università di Pisa) e avere una breve descrizione del lavoro svolto (Fig. 4.13).

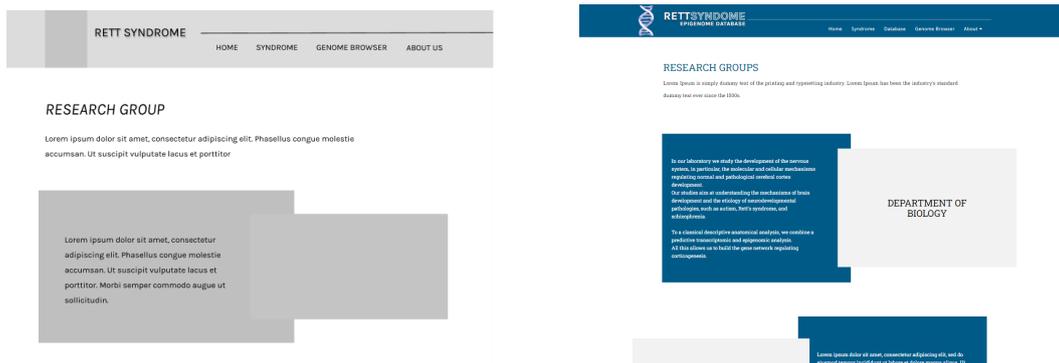


Figura 4.13: Wireframe e mockup della pagina "Research Group"

### 4.3.5.2 Projects

Nella pagina "Projects" viene descritto in modo più approfondito lo scopo del progetto e l'obiettivo principale del sito. Si è inoltre posta ulteriore attenzione sullo scopo del database presente all'interno della pagina (Fig. 4.14).



Figura 4.14: Wireframe e mockup della pagina "Project"

### 4.3.5.3 Papers

Infine, la pagina Papers contiene i riferimenti ai principali articoli utilizzati dal Dipartimento di Biologia dell'Università di Pisa per effettuare le analisi sui campioni (Fig. 4.15).



Figura 4.15: Wireframe e mockup della pagina "Papers"

## 4.4 Implementazione pagina "Database"

La tabella presente nella pagina del database contiene informazioni relative ai file GSM già accennati nel capitolo precedente (vd 4.3.3).

I file GSM rappresentano dei campioni e sono il risultato di un esperimento fatto in condizioni specifiche. Su ognuno di questi file sono state condotte le analisi effettuate dal Dipartimento di Biologia dell'Università di Pisa e, ognuno di essi, è esplorabile nel genome browser.

Questa pagina vuole fornire al ricercatore la possibilità di avere informazioni in più sui file visualizzati nel genome browser (Fig. 4.16). La tabella contiene, per ogni GSM, alcune informazioni fondamentali:

- **Marker:** indica una molecola che identifica la presenza di un determinato tessuto
- **PMID:** PubMed unique Identifier<sup>12</sup> associato a ciascun campione
- **Tissue/Site:** sito o tessuto da cui il campione ha origine.
- **Species:** specie di riferimento
- **GSM:** numero di identificazione del campione.

Nel caso in cui, per ciascun campione, il ricercatore volesse avere qualche informazione in più o volesse esplorarli più approfonditamente, è possibile accedere alla pagina specifica di ogni campione cliccando il link presente su ogni numero di identificazione

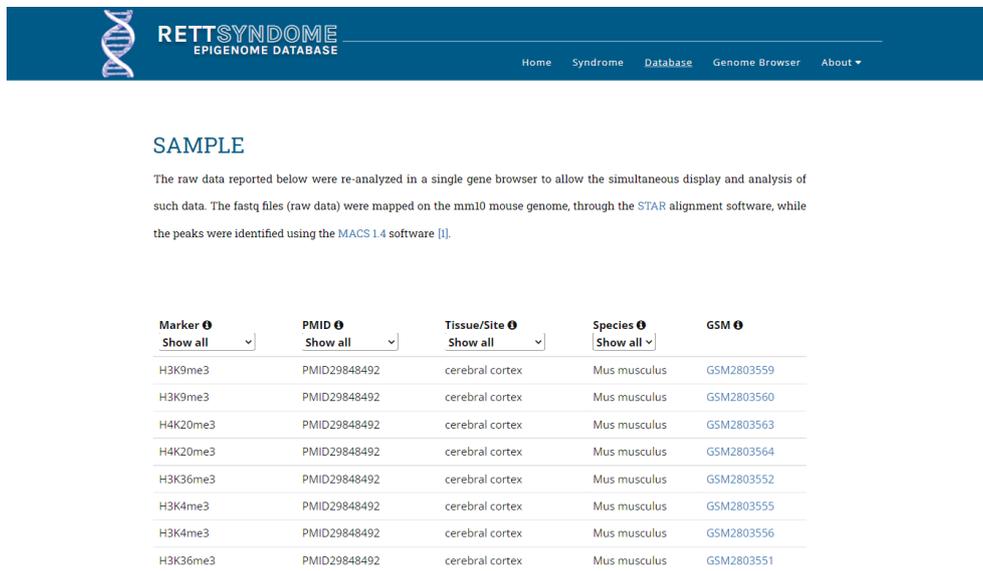
---

<sup>12</sup>Search Field Descriptions and Tags, su [ncbi.nlm.nih.gov](http://ncbi.nlm.nih.gov), National Center for Biotechnology Information. URL consultato il 14 agosto 2010.

di un file specifico.

Per rendere più semplice la ricerca degli elementi nella tabella, sono stati aggiunti dei filtri per cercare elementi specifici nelle colonne Marker, PMID, Tissue/Site e Species. Di fatto, la tabella al momento presenta solo una parte di dati, cioè quelli ottenuti da esperimenti fatti su topi. In futuro però in tabella saranno aggiunti anche quelli fatti su esseri umani e ciò comporterà un notevole incremento dei dati presenti. Di conseguenza, anche se al momento i filtri non sarebbero stati necessari, si è deciso comunque di implementarli in prospettiva futura.

La tabella è stata implementata utilizzando Bootstrap in modo da renderla responsive e visualizzabile su diversi dispositivi. Inoltre, in ogni colonna è stata inserita una piccola icona che contiene una breve descrizione dell'informazione contenuta nella colonna. L'informazione viene visualizzata attraverso l'evento *onHover*, dunque quando l'utente passa il mouse sopra l'icona.



The screenshot shows the website header with the logo and navigation menu. Below the header, there is a section titled "SAMPLE" with a paragraph of text explaining the data source. Below the text is a table with five columns: Marker, PMID, Tissue/Site, Species, and GSM. Each column has a dropdown menu with a "Show all" option. The table contains 8 rows of data.

Marker 	PMID 	Tissue/Site 	Species 	GSM 
H3K9me3	PMID29848492	cerebral cortex	Mus musculus	GSM2803559
H3K9me3	PMID29848492	cerebral cortex	Mus musculus	GSM2803560
H4K20me3	PMID29848492	cerebral cortex	Mus musculus	GSM2803563
H4K20me3	PMID29848492	cerebral cortex	Mus musculus	GSM2803564
H3K36me3	PMID29848492	cerebral cortex	Mus musculus	GSM2803552
H3K4me3	PMID29848492	cerebral cortex	Mus musculus	GSM2803555
H3K4me3	PMID29848492	cerebral cortex	Mus musculus	GSM2803556
H3K36me3	PMID29848492	cerebral cortex	Mus musculus	GSM2803551

Figura 4.16: Implementazione finale della pagina del database

## 4.5 Implementazione pagina "Genome Browser"

Al fine di inserire WashU all'interno della pagina del Genome Browser sono state considerate due scelte implementative:

1. Utilizzando uno script per ricrearlo all'interno della pagina.
2. Utilizzando un *iframe*.

Nel primo caso sono stati importati nella pagina HTML tutti gli script e fogli di stile necessari per effettuare l'*embedding* di WashU. Quindi è stata utilizzata la funzione specifica `renderBrowserInElement(contents, container)` che prende come primo parametro un file JSON contenente i riferimenti alle tracce che devono essere visualizzate dal browser e le informazioni relative al genoma di riferimento e come secondo parametro l'elemento contenitore di HTML all'interno del quale si vuole inserire il genome browser. In questo modo è stato possibile includere WashU all'interno della pagina HTML. Inoltre, per poter permettere all'utente di selezionare o rimuovere le tracce da visualizzare dentro il browser è stata implementata una lista di checkbox in cui venivano elencate le tracce elaborate dal Dipartimento di Biologia dell'Università di Pisa disponibili per la visualizzazione.

Ben presto però sono emersi diversi limiti che hanno portato ad una revisione e ristrutturazione della pagina. Di fatto, per l'aggiunta e la rimozione di tracce attraverso le checkbox era necessario ricostruire l'intero genome browser. Se, ad esempio, un ricercatore sta osservando una particolare regione del genoma e vuole, in quella regione specifica, aggiungere una traccia spuntando l'apposita checkbox, il ricaricamento dell'elemento HTML avrebbe portato alla perdita della regione del genoma

e, di conseguenza il ricercatore avrebbe dovuto inserirla nuovamente ogni volta che una traccia veniva aggiunta o rimossa. Un ulteriore limite di questo primo *embedding* era dato dal fatto che l'interfaccia grafica di WashU non presentava l'apposita sezione per permettere all'utente di poter caricare le sue tracce personali in locale. Di conseguenza, in questo modo, non avrebbe potuto confrontare le proprie tracce con quelle elaborate dal Dipartimento di Biologia dell'Università di Pisa, già presenti nel genome browser. Per questi motivi si è cercato di pensare ad una possibile soluzione alternativa.

Tutti i limiti discussi sono stati superati utilizzando all'interno della pagina HTML l'elemento con il tag *frame* contenente il genome browser (Fig. 4.17). Ciò ha comportato una riorganizzazione dei contenuti all'interno della pagina. Le checkbox sono state rimosse e all'apertura della pagina vengono mostrate nell'interfaccia grafica tutte le tracce elaborate dal Dipartimento di Biologia dell'Università di Pisa. L'utente ha comunque la possibilità di rimuoverle o ordinarle in un secondo momento attraverso una specifica funzione di WashU. Inoltre, la presenza della voce "Local Track" dà la possibilità di aggiungere delle proprie tracce al genome browser in modo da poter fare un confronto con quelle già presenti.

L'utente principale del sito è il ricercatore di biologia, quindi che ha già esperienza con i genome browser e, di conseguenza, sa già come utilizzarli ed esplorarli. Tuttavia, è stato preso in considerazione anche il ricercatore non esperto che, non avendo familiarità con questi strumenti potrebbe avere bisogno di una piccola guida che lo accompagni nelle diverse operazioni (v.d 4.3.4). Per questo motivo è stato inserito all'interno della pagina un piccolo tutorial che illustra all'utente alcune delle funzioni fondamentali di WashU:

- Aggiungere una traccia.
- Rimuovere una traccia.
- Ordinare le tracce.
- Eseguire la ricerca per gene o per regione.

Il tutorial è stato realizzato utilizzando una galleria di JQuery che, in seguito, è stata riadattata in base al risultato che si voleva ottenere nella pagina. Inizialmente, le diverse gallerie per i diversi tutorial erano già visibili all'interno della pagina: all'utente bastava quindi cliccare su quella desiderata e sfoglarla. In un secondo momento però questa soluzione è sembrata essere troppo "ingombrante" ed è sembrato più opportuno che i tutorial occupassero uno spazio minimo all'interno della pagina. Per questo motivo, prima del genome browser sono stati aggiunti dei bottoni che gestiscono l'apertura della galleria.

In breve, la galleria di base non viene mostrata all'utente poiché viene nascosta attraverso il CSS. Quando l'utente clicca su un bottone, si apre la galleria corrispondente. Ciò viene gestito attraverso l'evento "trigger" che attiva gli eventi associati all'elemento selezionato senza bisogno della relativa interazione da parte dell'utente.

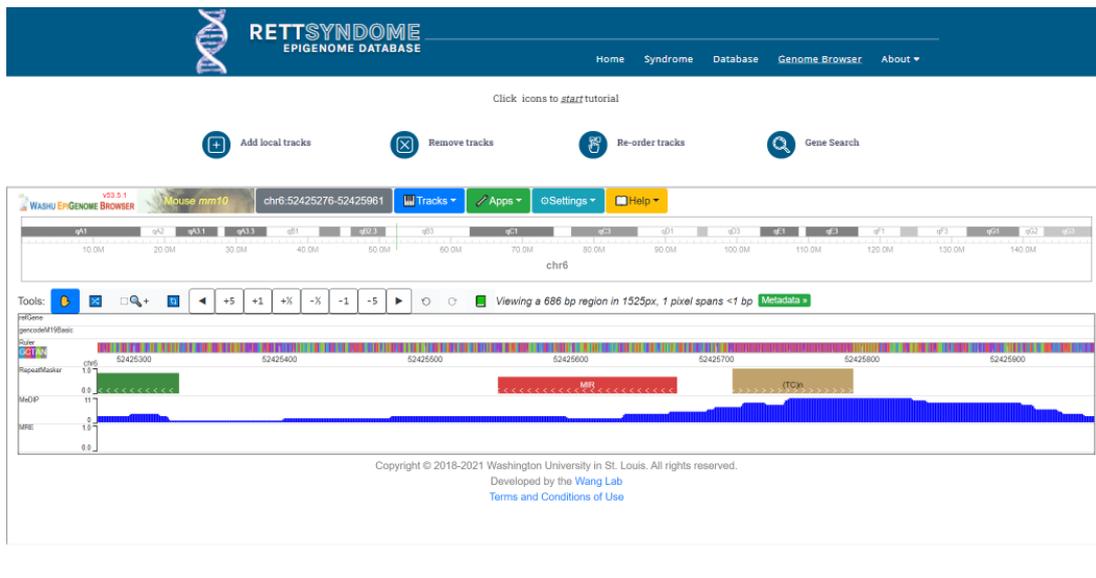


Figura 4.17: Implementazione finale della pagina del genome browser

---

## Prospettive future

Al momento, il sito contiene solo parte dei dati epigenomici, mentre molti altri saranno aggiunti in un secondo momento. Questo comporterà un'espansione della tabella nella pagina del database e l'aggiunta di una nuova pagina contenente un diverso *iframe* di WashU impostato su genoma umano hg19.

Per migliorare l'usabilità del sito sarà necessario apportare alcune modifiche alla navbar. Ad esempio, l'utente che scorre la pagina per poter accedere ad un'altra sezione, è costretto a dover tornare in cima per visualizzare la navbar. Per ragioni di usabilità quindi la navbar dovrà diventare di tipo *sticky*, quindi rimanere fissa in alto anche quando l'utente scorre la pagina. In questo caso, inoltre, le sue dimensioni andrebbero ridotte in modo che lo spazio occupato nella pagina sia minimo.

In ultima istanza, al momento i test sono stati effettuati sul gruppo di ricerca guidato dal professor Borello. Tuttavia, potrebbero essere necessari ulteriori test su una platea di utenti target più ampia in modo tale da poter effettuare, se necessario, ulteriori cambiamenti nella progettazione del sito.

---

# Conclusioni

Il lavoro svolto in questa tesi è stato suddiviso principalmente in tre parti:

- Analisi dei genome browser e elaborazione delle tracce
- Elaborazione di *wireframe* e *mockup*
- Implementazione del sito

Per la scelta del genome browser e l'elaborazione e il caricamento delle tracce è stata necessaria una fase preliminare di acquisizione di concetti e elementi fondamentali di Biologia e Genetica. Ciò ha permesso non solo di avere una visione più ampia del progetto e del suo fine ultimo ma anche di avere maggiore consapevolezza delle scelte effettuate in quest'ambito.

A seguito della scelta del genome browser da poter inserire all'interno della pagina e il caricamento delle prime tracce, sono stati definiti i contenuti della pagina. A sua volta, per ognuna delle pagine è stato definito un layout attraverso l'elaborazione di *wireframe* e, in seguito di *mockup*.

Infine, nell'ultima fase sono stati implementati gli elementi presenti nei *wireframe* e nei *mockup*, creando quindi un primo prototipo funzionante del sito. Questa prima versione sarà la base per gli sviluppi successivi a livello progettuale, valutativo e implementativo.

---

## Bibliografia

- [1] Bruce Alberts et al. *Biologia molecolare della cellula*. Zanichelli Bologna, 2004.
- [2] *Altri geni: CDKL5 e FOXP1*. URL: <https://prorett.org/altrigeni.php>.
- [3] *BedGraph Track Format*. URL: <https://genome.ucsc.edu/goldenPath/bedgraph.html>.
- [4] *BigWig Track Format*. URL: <https://genome.ucsc.edu/goldenPath/help/bigWig.html>.
- [5] Ewan Birney et al. “An overview of Ensembl”. In: *Genome research* 14.5 (2004), pp. 925–928.
- [6] *Bootstrap (framework)*. URL: <https://it.wikipedia.org/wiki/Bootstrap>.
- [7] Massimiliano Bossi. *Introduzione a jQuery*. URL: [https://www.mrw.it/jquery/introduzione-jquery%5C\\_10443.html](https://www.mrw.it/jquery/introduzione-jquery%5C_10443.html).
- [8] Simone Braconcini. *Bootstrap: cos'è e perché usarlo*. URL: [https://verytech.smartworld.it/bootstrap-cos-e-e-perche-usarlo-353740.html#steps%5C\\_1](https://verytech.smartworld.it/bootstrap-cos-e-e-perche-usarlo-353740.html#steps%5C_1).
- [9] *ChIP-Seq Service*. URL: <https://www.biosense.it/prodotto/chip-seq-service>.

- 
- [10] Adam W Clemens et al. “MeCP2 represses enhancers through chromosome topology-associated DNA methylation”. In: *Molecular cell* 77.2 (2020), pp. 279–293.
- [11] *CSS*. URL: <https://it.wikipedia.org/wiki/CSS>.
- [12] *DNA Polimerasi*. URL: <https://www.pianetachimica.it/>.
- [13] *Genetica*. URL: <https://it.wikipedia.org/wiki/Genetica>.
- [14] *Genoma, DNA e geni*. URL: <https://www.humanitas-care.it/news/genoma-dnageni>.
- [15] Wendy Anne Gold e John Christodoulou. “The utility of next-generation sequencing in gene discovery for mutation-negative patients with Rett syndrome”. In: *Frontiers in cellular neuroscience* 9 (2015), p. 266.
- [16] *HiC Track Format*. URL: <https://genome.ucsc.edu/goldenPath/help/hic.html>.
- [17] Tim Hubbard et al. “The Ensembl genome database project”. In: *Nucleic acids research* 30.1 (2002), pp. 38–41.
- [18] *HyperText Markup Language (HTML)*. URL: <https://www.informarsi.net/introduzione-html/>.
- [19] *Immunoprecipitazione della cromatina*. URL: [https://it.wikipedia.org/wiki/Immunoprecipitazione%5C\\_della%5C\\_cromatina](https://it.wikipedia.org/wiki/Immunoprecipitazione%5C_della%5C_cromatina).
- [20] Donna Karolchik, Angie S Hinrichs e W James Kent. “The UCSC genome browser”. In: *Current protocols in bioinformatics* 40.1 (2012), pp. 1–4.

- [21] Gerald Karp e James G Patton. \* *Biologia cellulare e molecolare: concetti ed esperimenti*. EdiSES, 2015.
- [22] W James Kent. “BLAT—the BLAST-like alignment tool”. In: *Genome research* 12.4 (2002), pp. 656–664.
- [23] Aziz Khan e Xuegong Zhang. “Making genome browsers portable and personal”. In: *Genome biology* 19.1 (2018), pp. 1–3.
- [24] Benyam Kinde et al. “DNA methylation in the gene body influences MeCP2-mediated gene repression”. In: *Proceedings of the National Academy of Sciences* 113.52 (2016), pp. 15114–15119.
- [25] *La genetica della sindrome di Rett*. URL: <https://prorett.org/genetica.php>.
- [26] Cesare Lamanna. *Introduzione*. URL: <https://www.html.it/pag/14209/introduzione1/>.
- [27] *Le mutazioni*. URL: <https://www.sapere.it/sapere/strumenti/studiafacile/biologia/L-eredit%E2%80%93delle-caratteristiche/Le-basi-molecolari-dell-eredit-/Approfondimenti/Le-mutazioni.html>.
- [28] Daofeng Li et al. “WashU epigenome browser update 2019”. In: *Nucleic acids research* 47.W1 (2019), W158–W165.
- [29] Andrea Marzilli. *Introduzione all’HTML, 22 Marzo 2016*. URL: <https://www.html.it/pag/16026/introduzione22/>.
- [30] Jonathan Pevsner. *Bioinformatics and functional genomics*. John Wiley & Sons, 2015.

- [31] *RNA: che cos'è, a che serve, differenza con il DNA*. URL: <https://www.unicusano.it/blog/universita/rna-che-cose>.
- [32] David Sadava et al. “La nuova biologia. blu”. In: *Genetica, DNA e* (2012).
- [33] *Sindrome di Rett*. URL: [https://it.wikipedia.org/wiki/Sindrome\\_di\\_Rett](https://it.wikipedia.org/wiki/Sindrome_di_Rett).
- [34] *Sindrome di Rett congenita: modelli cellulari e murini per lo studio del ruolo di FOXP1 nella neurogenesi*. URL: <https://www.telethon.it/cosa-facciamo/ricerca/progetti-finanziati/sindrome-di-rett-congenita-modelli-cellulari-e-murini-per-lo-studio-del-ruolo-di-foxp1-nella-neurogenesi/>.
- [35] Michael R Speicher. “Chromosomes”. In: *Vogel and Motulsky's Human Genetics*. Springer, 2010, pp. 55–138.
- [36] Stephen Brian Sulkes. *Sindrome di Rett*. URL: <https://www.msmanuals.com/it-it/professionale/pediatria/disturbi-dell-apprendimento-e-dello-sviluppo/sindrome-di-rett>.
- [37] Giulio Tiseo. *Figma: il miglior tool per realizzare il design di siti web e app*. URL: <https://creativecode.it/articoli/realizzare-design-siti-o-app-con-figma>.
- [38] *Use GIVE-Toolbox to manage data tracks*. URL: <https://github.com/Zhong-Lab-UCSD/Genomic-Interactive-Visualization-Engine/blob/master/tutorials/3-GIVE-Toolbox.md>.

- 
- [39] Jack Wallen. *Docker Basics: How to Share Data Between a Docker Container and Host*. URL: <https://thenewstack.io/docker-basics-how-to-share-data-between-a-docker-container-and-host/>.
- [40] Jun Wang et al. “A brief introduction to web-based genome browsers”. In: *Briefings in Bioinformatics* 14.2 (2013), pp. 131–143.
- [41] *What is a genome browser and why do we need one?* URL: <https://www.futurelearn.com/info/courses/bacterial-genomes-access-and-analysis/0/steps/47043>.
- [42] Ying-Tao Zhao et al. “Long genes linked to autism spectrum disorders harbor broad enhancer-like chromatin domains”. In: *Genome research* 28.7 (2018), pp. 933–942.