Le banche dati

Un pò di storia …
le banche dati

Per la gestione dei dati biologici è necessario disporre di accurati **sistemi di archiviazione** e strumenti adeguati per **accedere** alle informazioni.

Cosa sono?

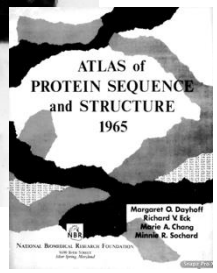
- **contenitori ordinati** di informazioni costruiti per inserire e conservare dati di tipo biologico, permettendone una facile consultazione o «query»
- Raccoglitori di informazioni e dati derivati dalla letteratura, cioè da analisi effettuate in laboratorio oppure risultati di analisi bioinformatiche.

Che caratteristiche hanno?

- Sono **accessibili liberamente** e possono essere consultati via web.
- Ogni database è caratterizzato da un «**elemento centrale**» attorno al quale viene costruito il singolo «***record***» in formato HTML.
  Ad esempio, elemento centrale per le banche dati di sequenze di acidi nucleici è la sequenza nucleotidica di DNA o di RNA; elemento centrale di una banca dati di riviste scientifiche è l'articolo scientifico che descrive i risultati di una ricerca.

Un pò di storia …
le banche dati

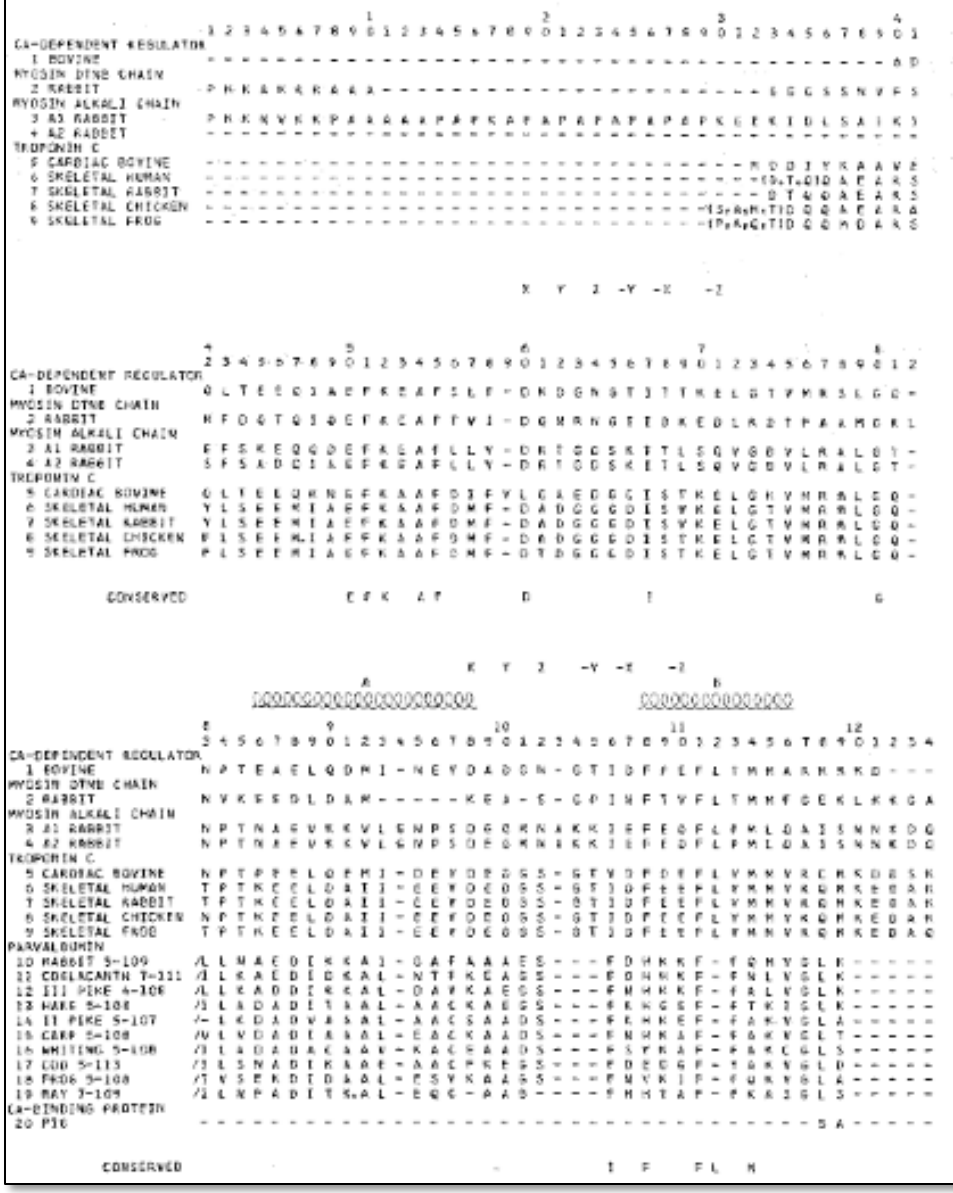






- ☑ La prima banca dati biologica risale ai tempi di Margareth Dayhoff, biochimica e studiosa dell'evoluzione che nel 1956 effettuò studi di analisi delle sequenze proteiche.
- ☑ Raccolse le sequenze di 65 proteine in un atlante secondo gruppi di proteine omologhe.
- ☑ Alla fine degli anni 70 fu trasformato in versione elettronica nella banca dati NBRF (National Biomedical Research Foundation).
- ☑ Il suo lavoro divenne la base per SwissProt nel 1985 fondato da Amos Bairoch
- ☑ Agli inizi degli anni 80 il laboratorio Europeo di Biologia Molecolare (EMBL) fonda la EMBL-datalibrary di sequenze di DNA e RNA. Il primo release dell'81 conteneva 519 records

TROPONIN C SUPERFAMILY — ALIGNMENT 35





- ☑ Negli Stati Uniti veniva fondato da Walter Goad un archivio simile da cui si è poi originato GenBank, il cui primo release fu reso pubblico nel 1982

- ☑ Nell'1986 veniva realizzata la banca dati Giapponese DDBJ ora residente al National Institute of Genetics.

- ☑ Viene stabilito un accordo tra DDBJ, GenBank e EMBLdatalibrary per trasferire costantemente i dati attraverso i tre centri per cui le banche dati risultano praticamente identiche anche se il formato è leggermente differente.

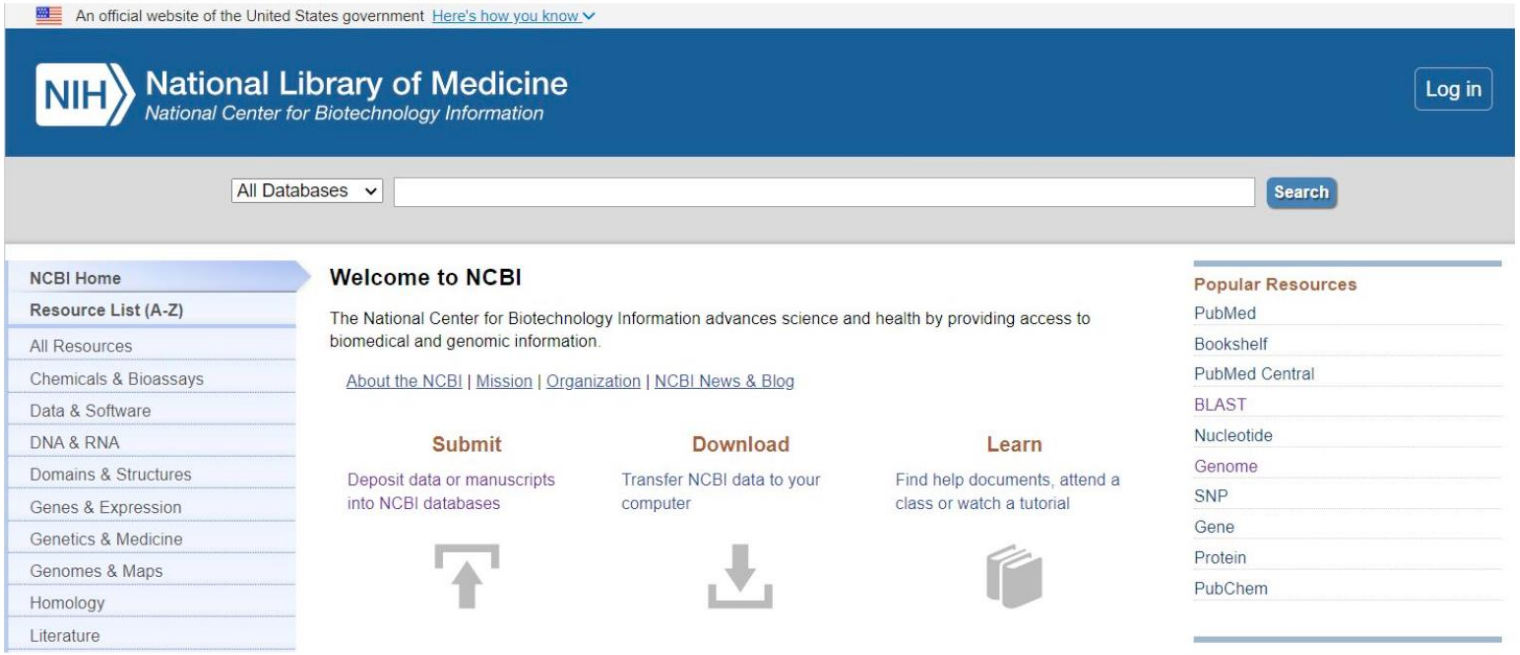
mirroring

- ☑ Negli anni 80 si ha la realizzazione delle prime banche dati specializzate come PROSITE o PDB. Ad oggi esistono molte centinaia di database specializzati accessibili via rete.

- ☑ Assieme ai database nascono sistemi integrati per il recupero dei dati, i cosidetti **sistemi di "data retrieval"**:

- ☑ **Il sistema SRS**, sviluppato da Thure Etzold il cui prototipo era il risultato della sua tesi di laurea. SRS è stato fino a qualche anno fa uno dei sistemi di ricerca più usati, ora dismesso.

- ☑ **Il sistema ENTREZ**, sviluppato al National Center of Biotecnology Information dell'NIH.

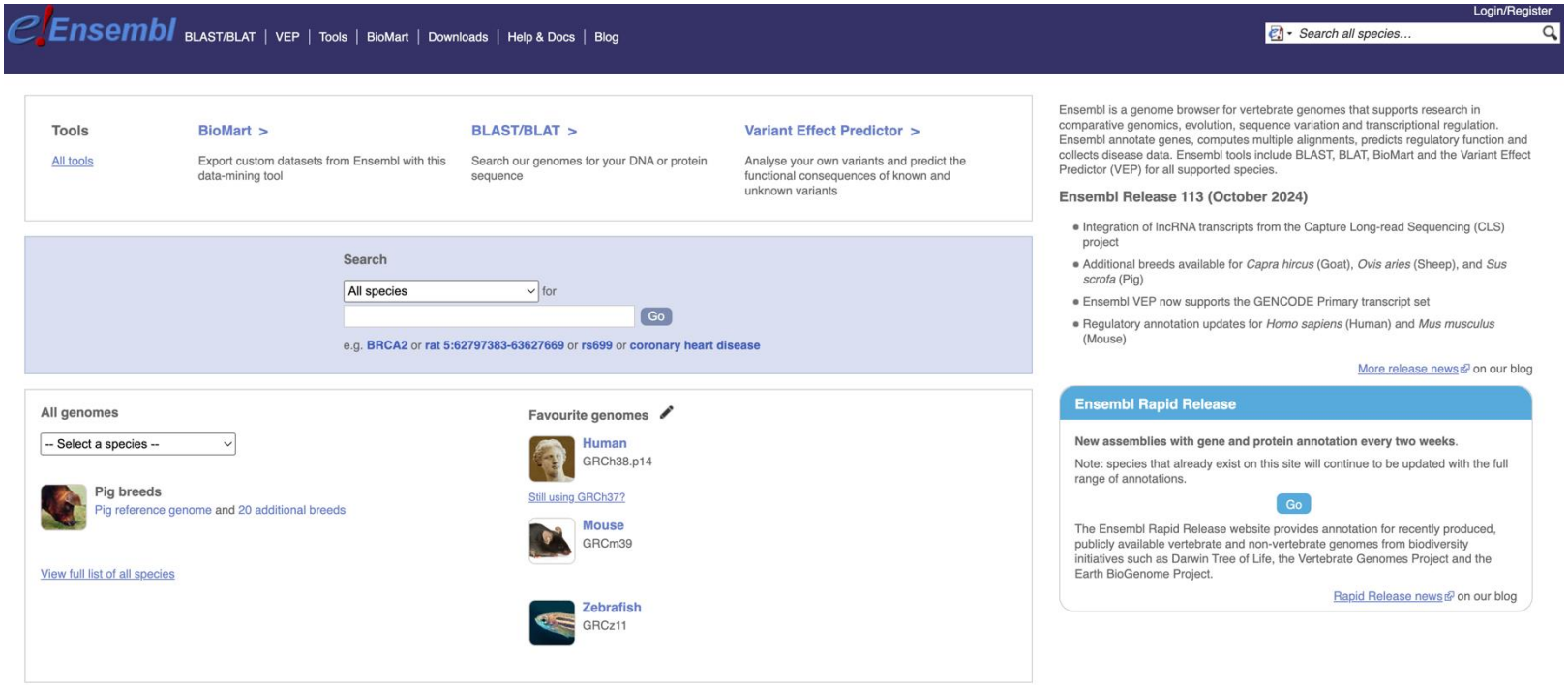


- ☑ Assieme ai database nascono sistemi integrati per il recupero dei dati, i cosidetti **sistemi di "data retrieval"**:

- ☑ **Il sistema SRS**, sviluppato da Thure Etzold il cui prototipo era il risultato della sua tesi di laurea. SRS è stato fino a qualche anno fa uno dei sistemi di ricerca più usati, ora dismesso.

- ☑ **Il sistema ENSEMBL**, sviluppato **al European Bioinformatics Institute (EMBL-EBI)**

EMBL-EBI



Struttura e organizzazione di database

Database diversi organizzano i dati in modi differenti

1. database **Flat-File** il tipo piu' semplice di database
2. database **relazionali** (Oracle, mySQl) permettono la composizione di *interrogazioni* o *query* che collegano piu' tabelle, stabilendo delle "relazioni" tra i contenuti delle singole tabelle.
3. database **Object Oriented**, le informazioni sono gestite come "oggetti" con varie "proprietà", invece che records con campi.

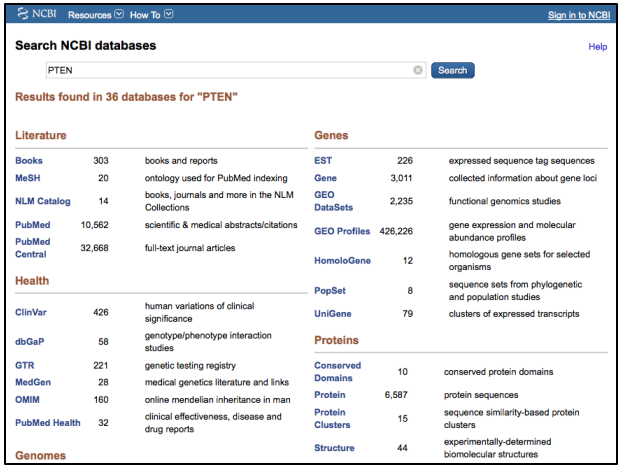
Tutti i sistemi di database impiegano delle interfacce,
API (Application Programming Interface) per accedere ai dati e modificarli

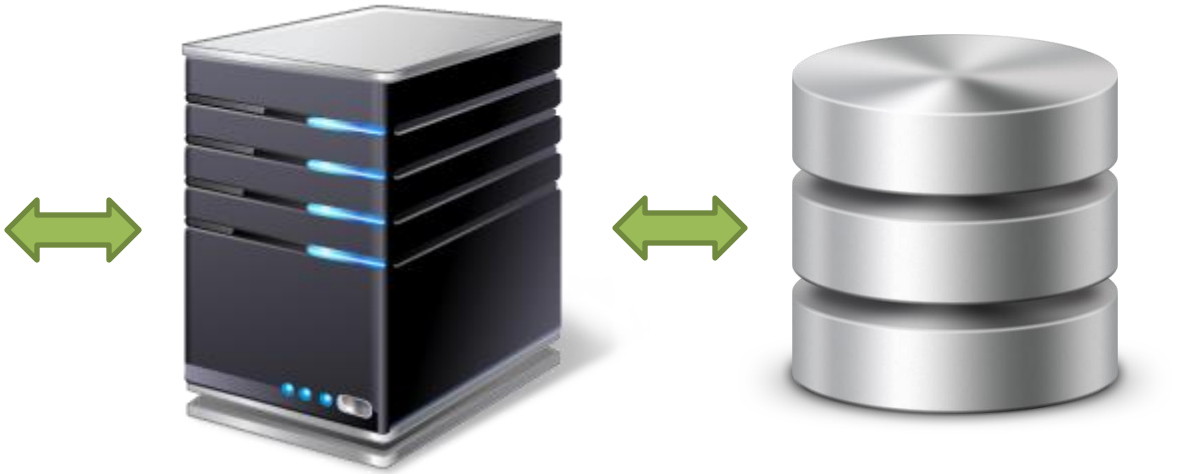
Web Interfaces

Browser

Web Server

Database





Free Text Query

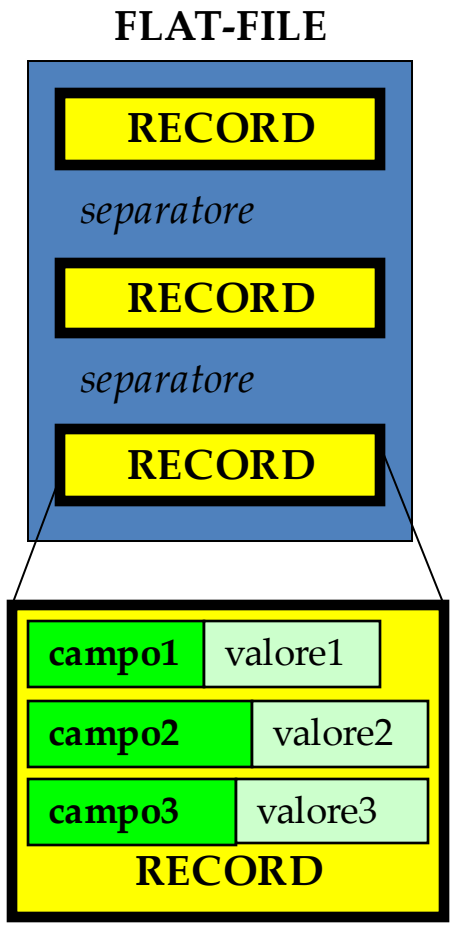
Database Query

List / Table

Result Set

FLAT-FILE

È un file sequenziale nel quale ogni classe di informazione è riportata su una o più linee consecutive identificate da un codice caratterizzante gli attributi annotati nella linea stessa.



All'interno del file esistono delle sequenze di caratteri che permettono di individuare i singoli record.

I record rappresentano l'unità di memorizzazione del database.

All'interno di un record esistono delle parole chiave che permettono di individuare i campi di quel record.

In genere esiste un campo (ID, identificatore, chiave…) che identifica univocamente i record del DB (non possono esistere 2 record con lo stesso valore nel campo chiave!)

Swissprot Record

```
ID   TNFA_HUMAN STANDARD; PRT; 233 AA.
AC   P01375;
DT   21-JUL-1986 (REL. 01, CREATED)
DT   21-JUL-1986 (REL. 01, LAST SEQUENCE UPDATE)
DT   01-FEB-1995 (REL. 31, LAST ANNOTATION UPDATE)
DE   TUMOR NECROSIS FACTOR PRECURSOR (TNF-ALPHA) (CACHECTIN).
GN   TNFA.
OS   HOMO SAPIENS (HUMAN).
OC   EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA; MAMMALIA;
OC   EUTHERIA; PRIMATES.
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 87217060.
RA   NEDOSPASOV S.A., SHAKHOV A.N., TURETSKAYA R.L., METT V.A.,
RA   AZIZOV M.M., GEORGIEV G.P., KOROBKO V.G., DOBRYNIN V.N.,
RA   FILIPPOV S.A., BYSTROV N.S., BOLDYREVA E.F., CHUVPILO S.A.,
RA   CHUMAKOV A.M., SHINGAROVA L.N., OVCHINNIKOV Y.A.;
RL   COLD SPRING HARB. SYMP. QUANT. BIOL. 51:611-624(1986).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 85086244.
RA   PENNICA D., NEDWIN G.E., HAYFLICK J.S., SEEBURG P.H., DERYNCK R.,
RA   PALLADINO M.A., KOHR W.J., AGGARWAL B.B., GOEDDEL D.V.;
RL   NATURE 312:724-729(1984).
RN   [3]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 85137898.
RA   SHIRAI T., YAMAGUCHI H., ITO H., TODD C.W., WALLACE R.B.;
RL   NATURE 313:803-806(1985).
RN   [4]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 86016093.
RA   NEDWIN G.E., NAYLOR S.L., SAKAGUCHI A.Y., SMITH D.H.,
RA   VARRETT-NEDWIN J., PENNICA D., GOEDDEL D.V., GRAY P.W.;
RL   NUCLEIC ACIDS RES. 13:6361-6373(1985).
RN   [5]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 85142190.
```

Codice identificativo

Data di sottomissione

Descrizione

Gene Name

Referenze

Codice iniziale

Indica il tipo di informazione disponibile nel campo

GenBank Record

Codice identificativo

Descrizione

Codice identificativo

Organismo a cui appartiene

Dettagli dell'articolo scientifico di riferimento

Homo sapiens colipase (CLPS), transcript variant 1, mRNA

NCBI Reference Sequence: NM_001832.3

FASTA Graphics

Go to:

```
LOCUS       NM_001832               578 bp    mRNA    linear   PRI 12-MAY-2018
DEFINITION  Homo sapiens colipase (CLPS), transcript variant 1, mRNA.
ACCESSION   NM_001832
VERSION     NM_001832.3
KEYWORDS    RefSeq.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 578)
  AUTHORS   Zhang G, He P, Tan H, Budhu A, Gaedcke J, Ghadimi BM, Ried T,
            Yfantis HG, Lee DH, Maitra A, Hanna N, Alexander HR and Hussain SP.
  TITLE     Integration of metabolomics and transcriptomics revealed a fatty
            acid network exerting growth inhibitory effects in human pancreatic
            cancer
  JOURNAL   Clin. Cancer Res. 19 (18), 4983-4993 (2013)
   PUBMED   23918603
  REMARK    GeneRIF: Low CLPS expression is associated with pancreatic cancer.
REFERENCE   2  (bases 1 to 578)
  AUTHORS   Xiao X, Ferguson MR, Magee KE, Hale PD, Wang Y and Lowe ME.
  TITLE     The Arg92Cys colipase polymorphism impairs function and secretion
            by increasing protein misfolding
  JOURNAL   J. Lipid Res. 54 (2), 514-521 (2013)
```

GenBank Record

Altre informazioni utili

```
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
            reference sequence was derived from BQ271978.1 and BP324927.1.
            On Nov 19, 2011 this sequence version replaced NM_001832.2.

            Summary: The protein encoded by this gene is a cofactor needed by
            pancreatic lipase for efficient dietary lipid hydrolysis. It binds
            to the C-terminal, non-catalytic domain of lipase, thereby
            stabilizing an active conformation and considerably increasing the
            overall hydrophobic binding site. The gene product allows lipase to
            anchor noncovalently to the surface of lipid micelles,
            counteracting the destabilizing influence of intestinal bile salts.
            This cofactor is only expressed in pancreatic acinar cells,
            suggesting regulation of expression by tissue-specific elements.
            Three transcript variants encoding different isoforms have been
            found for this gene. [provided by RefSeq, Nov 2011].

            Transcript Variant: This variant (1) encodes the longest isoform
            (1).

            Publication Note:  This RefSeq record includes a subset of the
            publications that are available for this gene. Please see the Gene
            record to access additional publications.

            ##Evidence-Data-START##
            Transcript exon combination :: BP324927.1, BC007061.1 [ECO:0000332]
            RNAseq introns              :: single sample supports all introns
                                           SAMEA2151405, SAMEA2153347
                                           [ECO:0000348]
            ##Evidence-Data-END##
            COMPLETENESS: full length.
PRIMARY     REFSEQ_SPAN         PRIMARY_IDENTIFIER PRIMARY_SPAN       COMP
            1-21                BQ271978.1         1-21
            22-578              BP324927.1         1-557
FEATURES             Location/Qualifiers
     source          1..578
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /chromosome="6"
                     /map="6p21.31"
     gene            1..578
                     /gene="CLPS"
                     /note="colipase"
                     /db_xref="GeneID:1208"
                     /db_xref="HGNC:HGNC:2085"
                     /db_xref="MIM:120105"
     exon            1..140
```

le coordinate cromosomiche

il gene

gli esoni

la traduzione in proteina

```
                     /translation="MEKILILLLVALSVAYAAPGPRGIIINLENGELCMNSAQCKSNC
                     CQHSSALGLARCTSMASENSECSVKTLYGIYYKCPCERGLTCEGDKTIVGSITNTNFG
                     ICHDAGRSKQ"
                     57..107
                     /gene="CLPS"
                     /inference="COORDINATES: ab initio prediction:SignalP:4.0"
                     108..392
                     /gene="CLPS"
                     /product="colipase isoform 1 proprotein"
     mat_peptide     123..392
                     /gene="CLPS"
                     /product="colipase isoform 1"
     exon            141..263
                     /gene="CLPS"
                     /inference="alignment:Splign:2.1.0"
     exon            264..559
                     /gene="CLPS"
                     /inference="alignment:Splign:2.1.0"
     STS             282..426
                     /gene="CLPS"
                     /standard_name="RH46790"
                     /db_xref="UniSTS:38953"
     STS             376..518
                     /gene="CLPS"
                     /standard_name="D6S2128"
                     /db_xref="UniSTS:41237"
     STS             404..518
                     /gene="CLPS"
                     /standard_name="D6S2063"
                     /db_xref="UniSTS:9000"
     regulatory      536..541
                     /regulatory_class="polyA_signal_sequence"
                     /gene="CLPS"
     polyA_site      559
                     /gene="CLPS"
ORIGIN
        1 cccatcacca cagctggctc tgtctcccgc cacccacacc agctgtccca ctcaccatgg
       61 agaagatcct gatcctcctg cttgtcgccc tctctgtggc ctatgcagct cctggccccc
      121 gggggatcat tatcaacctg gagaacggtg agctctgcat gaatagtgcc cagtgtaaga
      181 gcaattgctg ccagcattca agtgcgctgg gcctggcccg ctgcacatcc atggccagcg
      241 agaacagcga gtgctctgtc aagacgctct atgggattta ctacaagtgt ccctgtgagc
      301 gtggcctgac ctgtgaggga gacaagacca tcgtgggctc catcaccaac accaactttg
      361 gcatctgcca tgacgctgga cgctccaagc agtgagactg cccacccact cccacaccta
      421 gcccagaatg ctgtaggcca ctaggcgcag gggcatctct cccctgctcc agcgcatctc
      481 ccgggctggc cacctccttg accagcatat ctgttttctg attgcgctct tcacaattaa
      541 aggcctcctg caaaccttta aaaaaaaaaa aaaaaaaa
//
```

la sequenza nucleotidica

GenBank Record

```
     proprotein      108..392
                     /gene="CLPS"
                     /product="colipase isoform 1 proprotein"
     mat_peptide     123..392
                     /gene="CLPS"
                     /product="colipase isoform 1"
     exon            141..263
                     /gene="CLPS"
                     /inference="alignment:Splign:2.1.0"
     exon            264..559
                     /gene="CLPS"
                     /inference="alignment:Splign:2.1.0"
     STS             282..426
                     /gene="CLPS"
                     /standard_name="RH46790"
                     /db_xref="UniSTS:38953"
     STS             376..518
                     /gene="CLPS"
                     /standard_name="D6S2128"
                     /db_xref="UniSTS:41237"
     STS             404..518
                     /gene="CLPS"
                     /standard_name="D6S2063"
                     /db_xref="UniSTS:9000"
     regulatory      536..541
                     /regulatory_class="polyA_signal_sequence"
                     /gene="CLPS"
     polyA_site      559
                     /gene="CLPS"
ORIGIN
        1 cccatcacca cagctggctc tgtctcccgc cacccacacc agctgtccca ctcaccatgg
       61 agaagatcct gatcctcctg cttgtcgccc tctctgtggc ctatgcagct cctggccccc
      121 gggggatcat tatcaacctg gagaacggtg agctctgcat gaatagtgcc cagtgtaaga
      181 gcaattgctg ccagcattca agtgcgctgg gcctggcccg ctgcacatcc atggccagcg
      241 agaacagcga gtgctctgtc aagacgctct atgggattta ctacaagtgt ccctgtgagc
      301 gtggcctgac ctgtgaggga gacaagacca tcgtgggctc catcaccaac accaactttg
      361 gcatctgcca tgacgctgga cgctccaagc agtgagactg cccacccact cccacaccta
      421 gcccagaatg ctgtaggcca ctaggcgcag gggcatctct cccctgctcc agcgcatctc
      481 ccgggctggc cacctccttg accagcatat ctgttttctg attgcgctct tcacaattaa
      541 aggcctcctg caaaccttta aaaaaaaaaa aaaaaaaa
//
```

Notare che cliccando sulle descrizioni (exon, per esempio) queste parti vengono automaticamente evidenziate nella sequenza riportata a piè di pagina, un metodo molto immediato per avere un'idea della posizione delle varie parti nella sequenza.

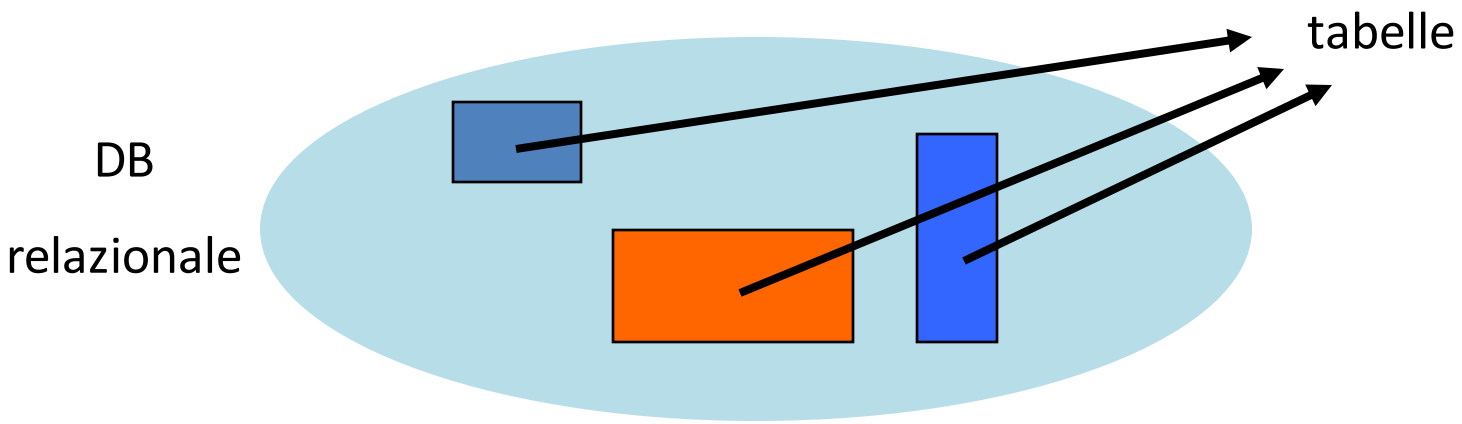
Un altro esempio di Flat File: Fasta File

>gi|110640215|ref|YP_667943.1| bifunctional aspartokinase I/homeserine dehydrogenase I [Escherichia coli 536]
MRVLKFGGTSVANAERFLRVADILESNARQGQVATVLSAPAKITNHLVAMIEKTISGQDALPNISDAERI
FAELLTGLAAAQPGFPLAQLKTFVDQEFAQIKHVLHGISLLGQCPDSINAALICRGEKMSIAIMAGVLEA
RGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAASRIPADHMVLMAGFTAGNEKGELVVLGRNGSDYS
AAVLAACLRADCCEIWTDVDGVYTCDPRQVPDARLLKSMSYQEAMELSYFGAKVLHPRTITPIAQFQIPC
LIKNTGNPQAPGTLIGASRDEDELPVKGISNLNNMAMFSVSGPGMKGMVGMAARVFAAMSRARISVVLIT
QSSSEYSISFCVPQSDCVRAERAMQEEFYLELKEGLLEPLAVTERLAIISVVGDGMRTLRGISAKFFAAL
ARANINIVAIAQGSSERSISVVVNNDDATTGVRVTHQMLFNTDQVIEVFVIGVGGVGGALLEQLKRQQSW
LKNKHIDLRVCGVANSKALLTNVHGLNLENWQEELAQAKEPFNLGRLIRLVKEYHLLNPVIVDCTSSQAV
ADQYADFLREGFHVVTPNKKANTSSMDYYHQLRYAAEKSRRKFLYDTNVGAGLPVIENLQNLLNAGDELM
KFSGILSGSLSYIFGKLDEGMSFSEATTLAREMGYTEPDPRDDLSGMDVARKLLILARETGRELELADIE
IEPVLPAEFNAEGDVAAFMANLSQLDDLFAARVAKARDEGKVLRYVGNIDEDGVCRVKIAEVDGNDPLFK
VKNGENALAFYSHYYQPLPLVLRGYGAGNDVTAAGVFADLLRTLSWKLGV
>gi|110640216|ref|YP_667944.1| homoserine kinase [Escherichia coli 536]
MVKVYAPASSANMSVGFDVLGAAVTPVDGALLGDVVTVEAAETFSLNNLGRFADKLPSEPRENIVYQCWE
RFCQELGKQIPVAMTLEKNMPIGSGLGSSACSVVAALMAMNEHCGKPLNDTRLLALMGELEGRISGSIHY
DNVAPCFLGGMQLMIEENDIISQQVPGFDEWLWVLAYPGIKVSTAEARAILPAQYRRQDCIAHGRHLAGF
IHACYSRQLELAAKLMKDVIAEPYRERLLPGFRQARQAVAEIGAVASGISGSGPTLFALCDKPDTAQRVA
DWLGKNYLQNQEGFVHICRLDTAGARVLEN

Il pregio principale dei flat-file è **la semplicità di gestione**, controbilanciata però **dall'incapacità di gestire accesso concorrente** e dalla mancanza di indicizzazione dei dati, che **non consentono interrogazioni sequenziali**.

Database Relazionale

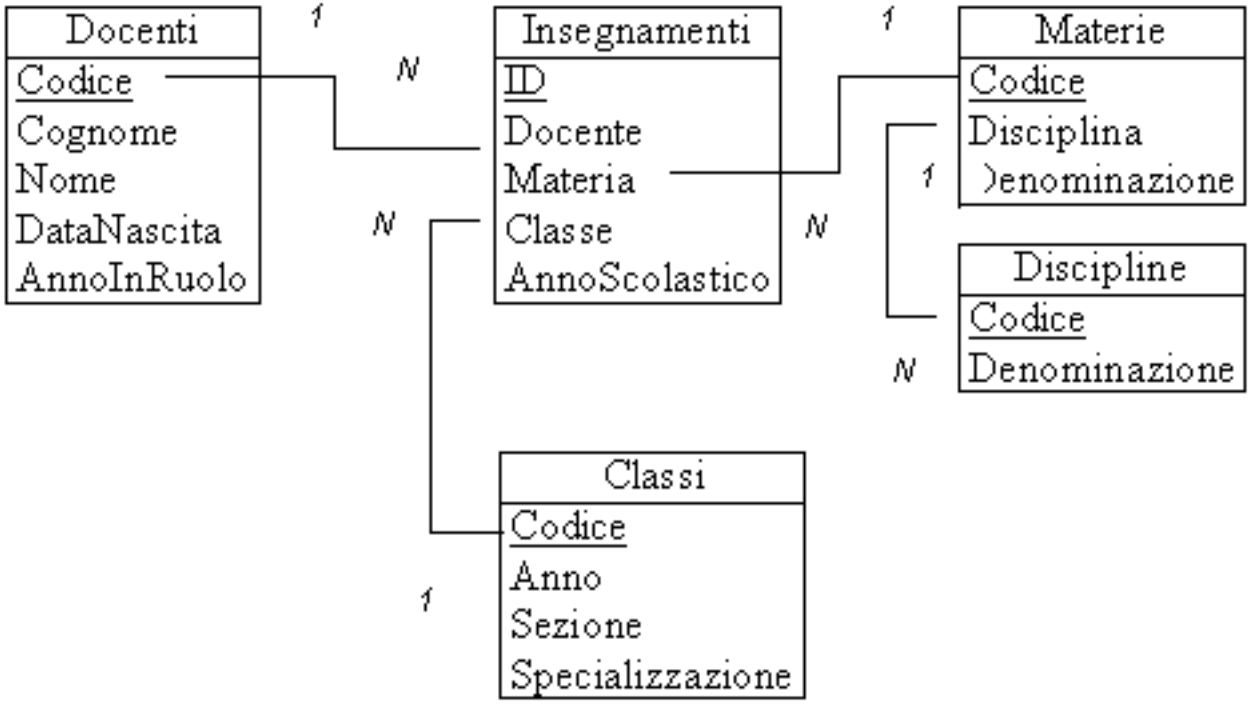
Un **DATABASE RELAZIONALE** è un insieme di TABELLE (table), in origine chiamate relazioni, con un qualche collegamento logico tra di esse. Una tabella di un DB relazionale è l'equivalente di un flat-file



La progettazione di un DB è guidata dal mondo reale. In genere si vuole creare un modello di qualcosa (banca, traffico aereo, sistema biologico…).

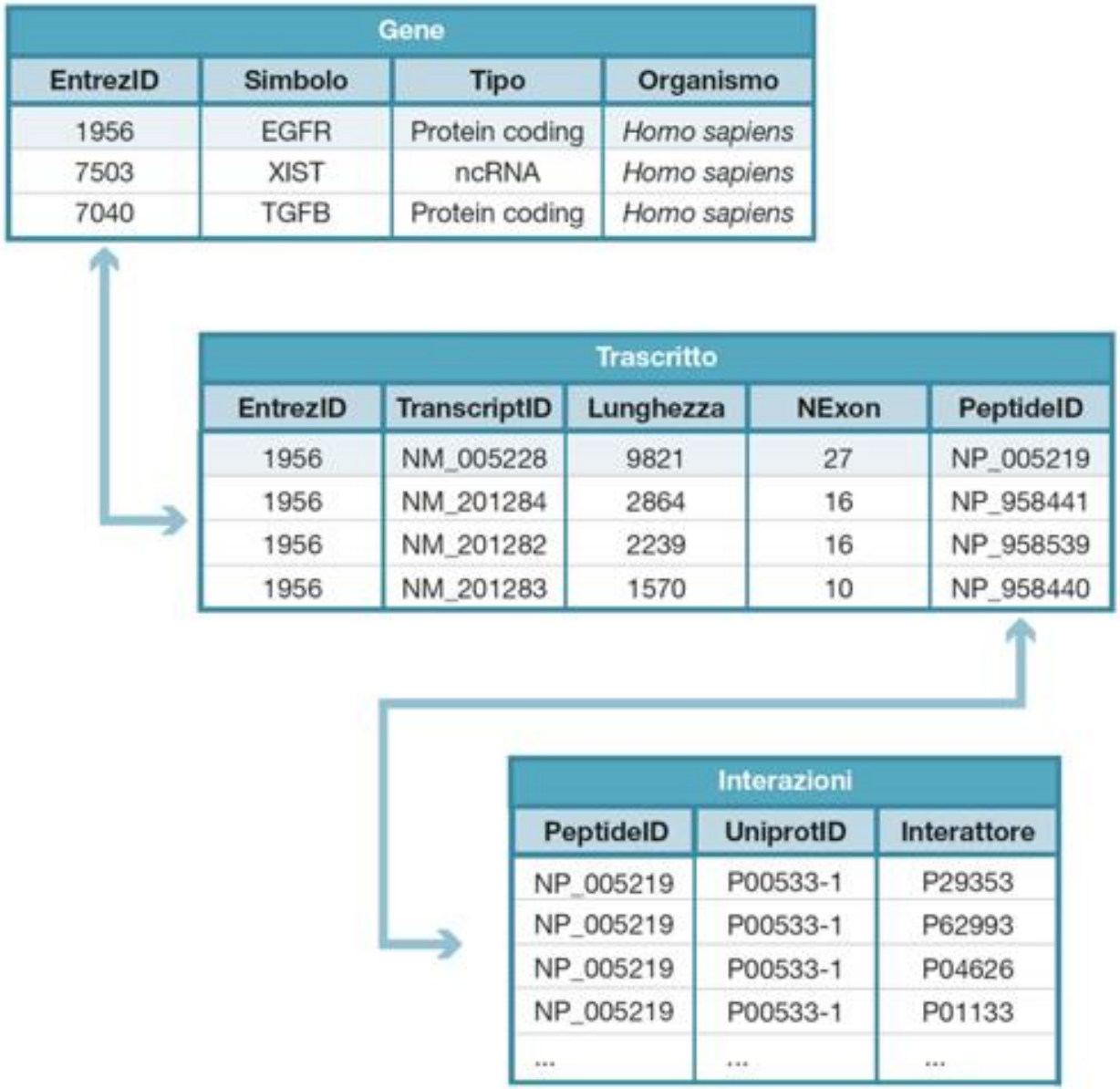
Si tratta di capire quali entità devono essere rappresentate ed in genere si costruisce una tabella per ciascuna di esse.

DATABASE RELAZIONALE: esempio



Ogni tabella ha **una o più chiavi**, che determinano l'**unicità** del campo, e il **collegamento** tra le varie tabelle.

Figura 2.2
Esempio di database relazionale formato da tre tabelle (Gene, Trascritto e Interazioni), ognuna con la sua chiave (rispettivamente EntrezID, TranscriptID e PeptideID + Interattore) e con relazioni uno-a-molti.



Gene

EntrezID	Simbolo	Tipo	Organismo
1956	EGFR	Protein coding	*Homo sapiens*
7503	XIST	ncRNA	*Homo sapiens*
7040	TGFB	Protein coding	*Homo sapiens*

Trascritto

EntrezID	TranscriptID	Lunghezza	NExon	PeptideID
1956	NM_005228	9821	27	NP_005219
1956	NM_201284	2864	16	NP_958441
1956	NM_201282	2239	16	NP_958539
1956	NM_201283	1570	10	NP_958440

Interazioni

PeptideID	UniprotID	Interattore
NP_005219	P00533-1	P29353
NP_005219	P00533-1	P62993
NP_005219	P00533-1	P04626
NP_005219	P00533-1	P01133
...	...	...

INTERROGAZIONI DELLE BANCHE DATI

① **RICERCHE TESTUALI (QUERY)**

Restituiscono i record di un database che soddisfano i criteri richiesti

② **RICERCHE PER SIMILARITÀ**

Restituiscono le sequenze di un database più simili ad una sequenza fornita come query. Non sono delle vere e proprie query in quanto richiedono l'esecuzione di programmi anche piuttosto complessi (ad esempio BLAST o FASTA).

Es. Trovare la sequenza nucleotidica del gene 'telethonin' è una ricerca testuale o di similarità?

INTERROGAZIONI DELLE BANCHE DATI

① **RICERCHE TESTUALI (QUERY)**

Restituiscono i record di un database che soddisfano i criteri richiesti

② **RICERCHE PER SIMILARITÀ**

Restituiscono le sequenze di un database più simili ad una sequenza fornita come query. Non sono delle vere e proprie query in quanto richiedono l'esecuzione di programmi anche piuttosto complessi (ad esempio BLAST o FASTA).

Es. Trovare la sequenza nucleotidica del gene 'telethonin' è una ricerca testuale o di similarità?

Ricerca Testuale

Le interrogazioni (**queries**) mirano ad identificare un insieme di voci (**entries**) sulla base di specifiche caratteristiche o di criteri di similarità con una sequenza o una struttura campione.

I software di ricerca identificano le entry con contenuto pertinente alla query.

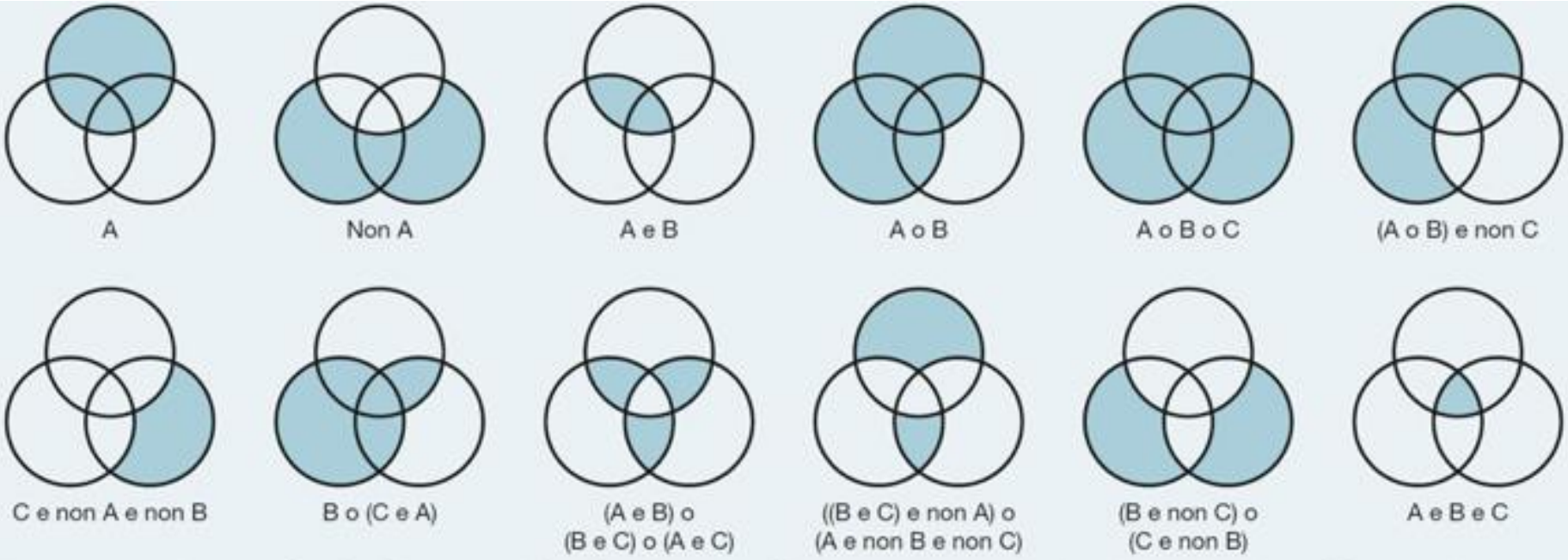
Utilizzo di **OPERATORI BOOLEANI** per specificare diverse combinazioni logiche dei termini di indicizzazione. Rappresentano lo strumento che fa la differenza fra una ricerca bibliografica efficace o meno, permettendo di combinare in vario modo piu` concetti nella stessa ricerca.

Alcuni esempi:

horse **OR** human
dehydrogenase **NOT** alcohol : (lattato o malato deidrogenasi)
(horse **OR** human) **AND** (dehydrogenase **NOT** alcohol)

Programmi diversi possono richiedere sintassi diverse :

(horse | human) **&** (dehydrogenase **!** alcohol)



Rappresentazione grafica degli operatori booleani tramite gli insieme di Venn. Dati tre insiemi A, B e C di intersezione non nulla, le operazioni tra insiemi usando NOT, AND e OR ci permettono di identificare ogni sottoinsieme degli insiemi dati. Gli operatori booleani poi possono essere combinati tra loro.

Tipologie di Database

I principali **sistemi di archiviazione** di dati biologici si dividono a seconda delle tipologie di elementi centrali che contengono. Si possono dividere in database di sequenze, strutturali e funzionali. I principali database di sequenze contengono:

Bibliografia

Sequenze di acidi nucleici

- DNA genomico
- Filogenesi, tassonomia
- Trascritti (mRNA)
- RNA-seq

Sequenze proteiche

- Sequenziamento diretto
- Studi di espressione proteica
- Cristallografia e determinazione delle strutture secondarie e terziarie

Dati grezzi

- Sequence Read Archive (SRA)
- Gene Expression Omnibus (GEO)
- European MassBank

PubMed

Ricerca bibliografica

Ricerca bibliografica

Le interrogazioni di tipo testuale ricercando termini che siano simili o uguali ad uno dato.

Partendo dalla più semplice tipologia di ricerca di tipo testuale vediamo un esempio di ricerca bibliograbica, e cioè di un **articolo scientifico** che tratti un argomento oggetto di interesse.

Ma prima di vedere come trovare un articolo scientifico nel database corrispondente verifichiamo di conoscerne la **forma** ed il **contenuto**. Esistono variate tipologie di bibliografia scientifica, le pirincipali sono: **articoli, lettere e review**. Iniziamo a vedere la struttura di un "research article" ovvero articolo di ricerca originale.

Bibliografia

Struttura di un articolo scientifico

- Rivista dove è pubblicato (nome, data di pubblicazione, volume, pagina)
- Titolo
- Autori
- Abstract (Riassunto dell'articolo)

Introduzione

Materiali e metodi

Risultati e Discussione

Conclusioni (opzionali)

- Ringraziamenti/Fondi di ricerca
- Reference (Bibliografia)

N.B. Le modalità con cui si esegue una ricerca bibliografica sono di esempio per una qualsiasi ricerca testuale o *query*.

ARTICLES

Titolo

Genome sequencing in microfabricated high-density picolitre reactors

Autori

Marcel Margulies[1]*, Michael Egholm[1]*, William E. Altman[1], Said Attiya[1], Joel S. Bader[1], Lisa A. Bemben[1], Jan Berka[1], Michael S. Braverman[1], Yi-Ju Chen[1], Zhoutao Chen[1], Scott B. Dewell[1], Lei Du[1], Joseph M. Fierro[1], Xavier V. Gomes[1], Brian C. Godwin[1], Wen He[1], Scott Helgesen[1], Chun He Ho[1], Gerard P. Irzyk[1], Szilveszter C. Jando[1], Maria L. I. Alenquer[1], Thomas P. Jarvie[1], Kshama B. Jirage[1], Jong-Bum Kim[1], James R. Knight[1], Janna R. Lanza[1], John H. Leamon[1], Steven M. Lefkowitz[1], Ming Lei[1], Jing Li[1], Kenton L. Lohman[1], Hong Lu[1], Vinod B. Makhijani[1], Keith E. McDade[1], Michael P. McKenna[1], Eugene W. Myers[2], Elizabeth Nickerson[1], John R. Nobile[1], Ramona Plant[1], Bernard P. Puc[1], Michael T. Ronan[1], George T. Roth[1], Gary J. Sarkis[1], Jan Fredrik Simons[1], John W. Simpson[1], Maithreyan Srinivasan[1], Karrie R. Tartaro[1], Alexander Tomasz[3], Kari A. Vogt[1], Greg A. Volkmer[1], Shally H. Wang[1], Yong Wang[1], Michael P. Weiner[4], Pengguang Yu[1], Richard F. Begley[1] & Jonathan M. Rothberg[1]

The proliferation of large-scale DNA-sequencing projects in recent years has driven a search for alternative methods to reduce time and cost. Here we describe a scalable, highly parallel sequencing system with raw throughput significantly greater than that of state-of-the-art capillary electrophoresis instruments. The apparatus uses a novel fibre-optic slide of individual wells and is able to sequence 25 million bases, at 99% or better accuracy, in one four-hour run. To achieve an approximately 100-fold increase in throughput over current Sanger sequencing technology, we have developed an emulsion method for DNA amplification and an instrument for sequencing by synthesis using a pyrosequencing protocol optimized for solid support and picolitre-scale volumes. Here we show the utility, throughput, accuracy and robustness of this system by shotgun sequencing and *de novo* assembly of the *Mycoplasma genitalium* genome with 96% coverage at 99.96% accuracy in one run of the machine.

Abstract (riassunto)

DNA sequencing has markedly changed the nature of biomedical research and medicine. Reductions in the cost, complexity and time required to sequence large amounts of DNA, including improvements in the ability to sequence bacterial and eukaryotic genomes, will have significant scientific, economic and cultural impact. Large-scale sequencing projects, including whole-genome sequencing, have usually required the cloning of DNA fragments into bacterial vectors, amplification and purification of individual templates, followed by Sanger sequencing[1] using fluorescent chain-terminating nucleotide analogues[2] and either slab gel or capillary electrophoresis. Current estimates put the cost of sequencing a human genome between $10 million and $25 million[3]. Alternative sequencing methods have been described[4–6]; however, no technology has displaced the use of bacterial vectors and Sanger sequencing as the main generators of sequence information.

Here we describe an integrated system whose throughput routinely enables applications requiring millions of bases of sequence information, including whole-genome sequencing. Our focus has been on the co-development of an emulsion-based method[7–11] to isolate and amplify DNA fragments *in vitro*, and of a fabricated substrate and instrument that performs pyrophosphate-based sequencing (pyrosequencing[5,12]) in picolitre-sized wells.

In a typical run we generate over 25 million bases with a Phred quality score of 20 or better (predicted to have an accuracy of 99% or higher). Although this Phred 20 quality throughput is significantly higher than that of Sanger sequencing by capillary electrophoresis, it is currently at the cost of substantially shorter reads and lower average individual read accuracy. [illegible] esis sequencing systems [illegible] information from each of [illegible] accuracy of 99.4% in 1 h, or [illegible] all of the bases having Ph[illegible] characterize the performan[illegible] is possible to assemble ba[illegible] short reads by sequencing a known bacterial genome, *Mycoplasma genitalium* (580,069 bases), and comparing our shotgun sequencing and *de novo* assembly with the results originally obtained for this genome[13]. The results of shotgun sequencing and *de novo* assembly of a larger bacterial genome, that of *Streptococcus pneumoniae*[14] (2.1 megabases (Mb)), are presented in Supplementary Table 4.

Introduzione

Emulsion-based sample preparation

We generate random libraries of DNA fragments by shearing an entire genome and isolating single DNA molecules by limiting dilution (Supplementary Methods). Specifically, we randomly fragment the entire genome, add specialized common adapters to the fragments, capture the individual fragments on their own beads and, within the droplets of an emulsion, clonally amplify the individual fragment (Fig. 1a, b). Unlike in current sequencing technology, our [illegible] handling of

[1]454 Life Sciences Corp., 20 Commercial Street, Branford, Connecticut 06405, USA. [2]U[illegible] [illegible]icrobiology, The Rockefeller University, New York, New York 10021, USA. [4]The Rothberg Institute for [illegible], USA. *These authors contributed equally to this work.

© 2005 Nature

individual clones; the templates are handled in bulk within the emulsions[9–11].

Sequencing in fabricated picolitre-sized reaction vessels

We perform sequencing by synthesis simultaneously in open wells of a fibre-optic slide using a modified pyrosequencing protocol that is designed to take advantage of the small scale of the wells. The fibre-optic slides are manufactured by slicing of a fibre-optic block that is obtained by repeated drawing and fusing of optic fibres. At each iteration [illegible] fibres decrease as they are hexagon[illegible] using cross-sectional sizes. Each fib[illegible]ber and surrounded by 2–3 μm [illegible] creates reaction wells approximately 55 μm in depth with a centre-to-centre distance of 50 μm (Fig. 1c), resulting in a calculated well size of 75 pl and a well density of 480 wells mm^{-2}. The slide, containing approximately 1.6 million wells[15], is loaded with beads and mounted in a flow chamber designed to create a 300-μm high channel, above the well openings, through which the sequencing reagents flow (Fig. 2a, b). The unetched base of the slide is in optical contact with a second fibre-optic imaging bundle bonded to a charge-coupled device (CCD) sensor, allowing the capture of emitted photons from the bottom of each individual well (Fig. 2c; see also Supplementary Methods).

We developed a three-bead system, and optimized the components to achieve high efficiency on solid support. The combination of picolitre-sized wells, enzyme loading uniformity allowed by the small beads and enhanced solid support chemistry enabled us to develop a method that extends the useful read length of sequencing-by-synthesis to 100 bases (Supplementary Methods).

In the flow chamber cyclically delivered reagents flow perpendicularly to the wells. This configuration allows simultaneous extension reactions on template-carrying beads within the open wells and relies on convective and diffusive transport to control the addition or removal of reagents and by-products. The timescale for diffusion into and out of the wells is on the order of 10 s in the current configuration and is dependent on well depth and flow channel height. The timescales for the signal-generating enzymatic reactions are on the order of 0.02–1.5 s (Supplementary Methods). The current reaction is dominated by mass transport effects, and improvements based on faster delivery of reagents are possible. Well depth was selected on the basis of a number of competing requirements: (1) wells need to be deep enough for the DNA-carrying beads to remain in the wells in the presence of convective transport past the wells; (2) they must be sufficiently deep to provide adequate isolation against diffusion of by-products from a well in which incorporation is taking place to a well where no incorporation is occurring; and (3) they must be shallow enough to allow rapid diffusion of nucleotides into the wells and rapid washing out of remaining nucleotides at the end of each flow cycle to enable high sequencing throughput and reduced reagent use. After the flow of each nucleotide, a wash containing apyrase is used to ensure that nucleotides do not remain in any well before the next nucleotide being introduced.

Base calling of individual reads

Nucleotide incorporation is detected by the associated release of inorganic pyrophosphate and the generation of photons[5,12]. Wells containing template-carrying beads are identified by detecting a known four-nucleotide 'key' sequence at the beginning of the read

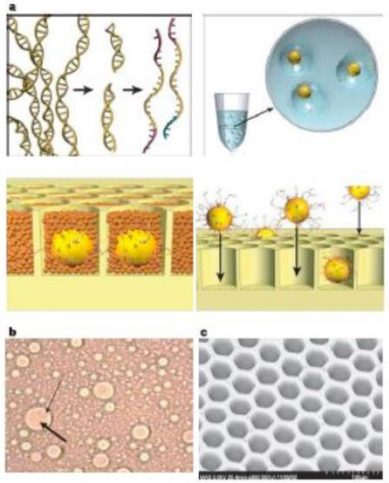


Figure 1 | Sample preparation. **a**, Genomic DNA is isolated, fragmented, ligated to adapters and separated into single strands (top left). Fragments are bound to beads under conditions that favour one fragment per bead, the beads are captured in the droplets of a PCR-reaction-mixture-in-oil emulsion and PCR amplification occurs within each droplet, resulting in beads each carrying ten million copies of a unique DNA template (top right). The emulsion is broken, the DNA strands are denatured, and beads carrying single-stranded DNA clones are deposited into wells of a fibre-optic slide (bottom right). Smaller beads carrying immobilized enzymes required for pyrophosphate sequencing are deposited into each well (bottom left). **b**, Microscope photograph of emulsion showing droplets containing a bead and empty droplets. The thin arrow points to a 28-μm bead; the thick arrow points to an ap[illegible] micrograph of [illegible] and wells befor[illegible]

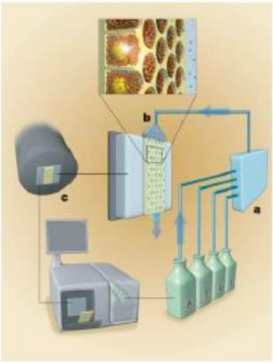


Figure 2 | Sequencing instrument. The sequencing instrument consists of the following major subsystems: a fluidic assembly (a), a flow chamber that includes the well-containing fibre-optic slide (b), a CCD camera-based imaging assembly (c), and a computer that provides the necessary user interface and instrument control.

Risultati

blishing Group

(Supplementary Methods). Raw signals are background-subtracted, normalized and corrected. The normalized signal intensity at each nucleotide flow, for a particular well, indicates the number of nucleotides, if any, that were incorporated. This linearity in signal is preserved to at least homopolymers of length eight (Supplementary Fig. 6). In sequencing by synthesis a very small number of templates on each bead lose synchronism (that is, either get ahead of, or fall behind, all other templates in sequence[16]). The effect is primarily due to leftover nucleotides in a well (creating 'carry forward') or to incomplete extension. Typically, we observe a carry forward rate of 1–2% and an incomplete extension rate of 0.1–0.3%. Correction of these shifts is essential because the loss of synchronism is a cumulative effect that degrades the quality of sequencing at longer read lengths. We have developed algorithms, based on detailed models of the underlying physical phenomena, that allow us to determine, and correct for, the amounts of carry forward and incomplete extension occurring in individual wells (Supplementary Methods). Figure 3 shows the processed result, a 113-bases-long read generated in the *M. genitalium* run discussed below. To assess sequencing performance and the effectiveness of the correction algorithms, independently of artefacts introduced during the emulsion-based sample preparation, we created test fragments with difficult-to-sequence stretches of identical bases of increasing length (homopolymers) (Supplementary Methods and Supplementary Fig. 4). Using these test fragments, we have verified that at the individual read level we achieve base call accuracy of approximately 99.4%, at read lengths in excess of 100 bases (Table 1).

High-quality reads and consensus accuracy

Before base calling or aligning reads, we select high-quality reads without relying on a priori knowledge of the genome or template being sequenced (Supplementary Methods). This selection is based on the observation that poor-quality reads have a high proportion of signals that do not allow a clear distinction between a flow during which no nucleotide was incorporated and a flow during which one or more nucleotide was incorporated. When base calling individual reads, errors can occur because of signals that have ambiguous values (Supplementary Fig. 5). To improve the usability of our reads, we also developed a metric that allows us to estimate *ab initio* the quality (or probability of correct base call) of each base of a read, analogous to the Phred score[17] used by current Sanger sequencers (Supplementary Methods and Supplementary Fig. 8).

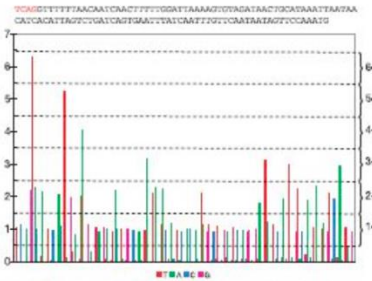


Figure 3 | Flowgram of a 113-bases read from an *M. genitalium* run. Nucleotides are flowed in the order T, A, C, G. The sequence is shown above the flowgram. The signal value intervals corresponding to the various homopolymers are indicated on the right. The first four bases (in red, above the flowgram) constitute the 'key' sequence, used to identify wells containing a DNA-carrying bead.

Higher quality sequence can be achieved by taking advantage of the high over sampling that our system affords and building a consensus sequence. Sequences are aligned to one another using the signal strengths at each nucleotide flow, rather than individual base calls, to determine optimal alignment (Supplementary Methods). The corresponding signals are then averaged, after which base calling is performed. This approach greatly improves the accuracy of the sequence (Supplementary Fig. 7) and provides an estimate of the quality of the consensus base. We refer to that quality measure as the *Z*-score—it is a measure of the spread of signals in all the reads at one location and the distance between the average signal and the closest base-calling threshold value. In both re-sequencing and *de novo* sequencing, as the minimum *Z*-score is raised the consensus accuracy increases, while coverage decreases; approximately half of the excluded bases, as the *Z*-score is increased, belong to homopolymers of length four and larger. Sanger sequencers usually require a depth of coverage at any base of three or more in order to achieve a consensus accuracy of 99.99%. To achieve a minimum of threefold coverage of 95% of the unique portions of a typical genome requires approximately seven- to eightfold over sampling. Owing to our higher error rate, we have observed that comparable consensus accuracies, over a similar fraction of a genome, are achieved with a depth of coverage of four or more, requiring approximately ten to twelve times over sampling.

Mycoplasma genitalium

Mycoplasma genomic DNA was fragmented and prepared into a sequencing library as described above. (This was accomplished by a single individual in 4 h.) After emulsion polymerase chain reaction (PCR) and bead deposition onto a 60 × 60 mm² fibre-optic slide, a process which took one individual 6 h, 42 cycles of four nucleotides were flowed through the sequencing system in an automated 4-h run of the instrument. The results are summarized in Table 2. In order to m… ach high quality r… sing flow-space m… in assessing the a… uencing quality, o… ference genome were included. Because this process excludes repeat regions (parts of the genome for which corresponding flowgrams are 70% similar to one another), the selected reads did not cover the genome completely. Figure 4a illustrates the distribution of read lengths for this run. The average read length was 110 bases, the resulting over sample 40-fold, and 84,011 reads (27.4%) were perfect. Figure 4b summarizes the average error as a function of base position. Coverage of non-repeat regions was consistent with the sample preparation and emulsion not being biased (Supplementary Fig. 8). At the individual read level, we observe an insertion and deletion error rate of approximately 3.3%; substitution errors have a much lower rate, on the order of 0.5%. When using these reads without any *Z*-score restriction, we covered 99.94% of the genome in ten contiguous regions with a consensus accuracy of 99.97%. The error rate in homopolymers is significantly reduced in the consensus sequence (Supplementary Fig. 7). Of the bases not covered by this consensus

Table 1 | Summary of sequencing statistics for test fragments

Size of fibre-optic slide	60 × 60 mm²
Run time/number of cycles	243 min/42
Test fragment reads	497,893
Average read length (bases)	108
Number of bases in test fragments	53,705,267
Bases with a Phred score of 20 and above	47,181,792
Individual read insertion error rate	0.44%
Individual read deletion error rate	0.15%
Individual read substitution error rate	0.004%
All errors	0.60%

© 2005 Nature Publishing Group

Discussione

sequence (366 bases), all belonged to excluded repeat regions. Setting a minimum *Z*-score equal to 4, coverage was reduced to 98.1% of the genome, while consensus accuracy increased to 99.996%. We further demonstrated the reproducibility of the system by repeating the whole-genome sequencing of *M. genitalium* an additional eight times, achieving a 40-fold coverage of the genome in each of the eight separate instrument runs (Supplementary Table 3).

We assembled the *M. genitalium* reads from a single run into 25 contigs with an average length of 22.4 kb. One of these contigs was misassembled due to a collapsed tandem repeat region of 60 bases, and was corrected by hand. The original sequencing of *M. genitalium* resulted in 28 contigs before directed sequencing used for finishing the sequence[13]. Our assembly covered 96.54% of the genome and attained a consensus accuracy of 99.96%. Non-resolvable repeat regions amount to 3% of the genome: we therefore covered 99.5% of the unique portions of the genome. Sixteen of the breaks between contigs were due to non-resolvable repeat regions, two were due to missed overlapping reads (our read filter and trimmer are not perfect and the algorithms we use to perform the pattern matching of flowgrams occasionally miss valid overlaps) and the remainder to thin read coverage. Setting a minimum *Z*-score of 4, coverage was reduced to 95.27% of the genome (98.2% of the resolvable part of the genome) with the consensus accuracy increasing to 99.994%.

Discussion

We have demonstrated the simultaneous acquisition of hundreds of thousands of sequence reads, 80–120 bases long, at 96% average accuracy in a single run of the instrument using a newly developed *in vitro* sample preparation methodology and sequencing technology. With Phred 20 as a cutoff, we show that our instrument is able to produce over 47 million bases from test fragments and 25 million bases from genomic libraries. We used test fragments to de-couple our sample preparation methodology from our sequencing technology. The decrease in single-read accuracy from 99.4% for test fragments to 96% for genomic libraries is primarily due to a lack of clonality in a fraction of the genomic templates in the emulsion, and is not an inherent limitation of the sequencing technology. Most of the remaining errors result from a broadening of signal distributions, particularly for large homopolymers (seven or more), leading to ambiguous base calls. Recent work on the sequencing chemistry and algorithms that correct for crosstalk between wells suggests that the signal distributions will narrow, with an attendant reduction in errors and increase in read lengths. In preliminary experiments with genomic libraries that also include improvements in the emulsion protocol, we are able to achieve, using 84 cycles, read lengths of 200 bases with accuracies similar to those demonstrated here for 100 bases. On occasion, at 168 cycles, we have generated individual reads that are 100% accurate over greater then 400 bases.

Using *M. genitalium* we demonstrate that short fragments a priori do not prohibit the *de novo* assembly of bacterial genomes. In fact, the larger over sampling afforded by the throughput of our system resulted in a draft sequence having fewer contigs than with Sanger reads, with substantially less effort. By taking advantage of the over sampling, consensus accuracies greater then 99.96% were achieved for this genome. Further quality filtering of the assembly results in the selection of a consensus sequence with accuracy exceeding 99.99% while incurring only a minor loss of genome coverage. Comparable results were seen when we shotgun sequenced and *de novo* assembled the 2.1-Mb genome of *Streptococcus pneumoniae*[14] (Supplementary Table 4). The *de novo* assembly of genomes more complex than bacteria, including mammalian genomes, may require the development of methods similar to those developed for Sanger sequencing, to prepare and sequence paired end libraries that can span repeats in these genome. To facilitate the use of paired end libraries we have developed methods to sequence, in an individual well, from both ends of genomic template, and plan to add paired end read capabilities to our assembler (Supplementary Methods).

Future increases in throughput, and a concomitant reduction in cost per base, may come from the continued miniaturization of the fibre-optic reactors, allowing more sequence to be produced per unit area—a scaling characteristic similar to that which enabled the

Table 2 | Summary statistics for *M. genitalium*

Sequencing summary	
Number of instrument runs	1
Size of fibre-optic slide	60 × 60 mm²
Run time/number of cycles	243 min/42
High quality reads	306,178
Average read length (bases)	110
Number of bases in high quality reads	33,655,553
Bases with a Phred score of 20 and above	26,753,540
Re-sequencing	
Reads mapped to single locations	238,066
Number of bases in mapped reads	27,687,747
Individual read insertion error rate	1.67%
Individual read deletion error rate	1.60%
Individual read substitution error rate	0.68%
Re-sequencing consensus	
Average over sampling	× 40
Coverage, all (*Z* ≥ 4)	99.9% (98.2%)
Consensus accuracy, all (*Z* ≥ 4)	99.97% (99.996%)
Consensus insertion error rate, all (*Z* ≥ 4)	0.02% (0.003%)
Consensus deletion error rate, all (*Z* ≥ 4)	0.01% (0.002%)
Consensus substitution error rate, all (*Z* ≥ 4)	0.001% (0.0003%)
Number of contigs	10
De novo assembly	
Coverage, all (*Z* ≥ 4)	96.54% (95.27%)
Consensus accuracy, all (*Z* ≥ 4)	99.96% (99.994%)
Number of contigs	25
Average contig size (kb)	22.4

The individual read error rates are referenced to the total number of bases in mapped reads.

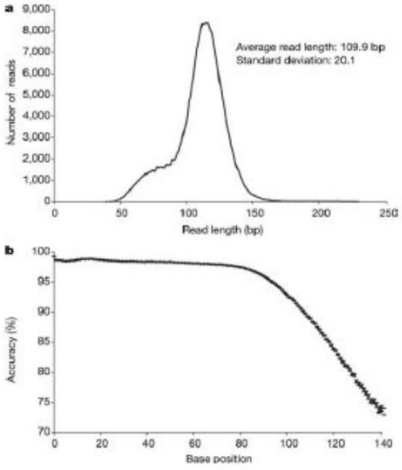


Figure 4 | *M. genitalium* data. a, Read length distribution for the 306,178 high-quality reads of the *M. genitalium* sequencing run. This distribution reflects the base composition of individual sequencing templates. **b**, Average read accuracy, at the single read level, as a function of base position for the 238,066 mapped reads of the same run.

© 2005 Nature Publishing Group

prediction of significant improvements in the integrated circuit at the start of its development cycle[18].

METHODS

Emulsion-based clonal amplification. The simultaneous amplification of fragments is achieved by isolating individual DNA-carrying beads in separate ~100-μm aqueous droplets (on the order of 2×10^8 ml^{-1}) made through the creation of a PCR-reaction-mixture-in-oil emulsion. (Fig. 1b; see also Supplementary Methods). The droplets act as separate microreactors in which parallel DNA amplifications are performed, yielding approximately 10^7 copies of a template per bead; 800 μl of emulsion containing 1.5 million beads are prepared in a standard 2-ml tube. Each emulsion is aliquoted into eight PCR tubes for amplification. After PCR, the emulsion is broken to release the beads, which include beads with amplified, immobilized DNA template and empty beads (Supplementary Methods). We then enrich for template-carrying beads (Supplementary Methods). Typically, about 30% of the beads will have DNA, producing 450,000 template-carrying beads per emulsion reaction. The number of emulsions prepared depends on the size of the genome and the expected number of runs required to achieve adequate over sampling. The 580-kb *M. genitalium* genome, sequenced on one 60 × 60 mm^2 fibre-optic slide, required 1.6 ml of emulsion. A human genome, over sampled ten times, would require approximately 3,000 ml of emulsion.

Bead loading into picolitre wells. The enriched template-carrying beads are deposited by centrifugation into open wells (Fig. 1c), arranged along one face of a 60 × 60 mm^2 fibre-optic slide. The beads (diameter ~28 μm) are sized to ensure that no more than one bead fits in most wells (we observed that 2–5% of filled wells contain more than one bead). Loading 450,000 beads (from one emulsion preparation) onto each half of a 60 × 60 mm^2 plate was experimentally found to limit bead occupancy to approximately 35% of all wells, thereby reducing chemical and optical crosstalk between wells. A mixture of smaller beads that carry immobilized ATP sulphurylase and luciferase necessary to generate light from free pyrophosphate are also loaded into the wells to create the individual sequencing reactors (Supplementary Methods).

Image capture. A bead carrying 10 million copies of a template yields approximately 10,000 photons at the CCD sensor, per incorporated nucleotide. The generated light is transmitted through the base of the fibre-optic slide and detected by a large format CCD (4,095 × 4,096 pixels). The images are processed to yield sequence information [illegible] for all wells containing template-carrying [illegible] number of [illegible] from individ[illegible] optic slide's [illegible] software to [illegible] DNA-carryi[illegible] sphate solut[illegible] approximate[illegible] collected by the pixels covering a particular well are summed to generate a signal for that particular well at that particular nucleotide flow. Each image captured by the CCD produces 32 megabytes of data. In order to perform all of the necessary signal processing in real time, the control computer is fitted with an accessory board (Supplementary Methods), hosting a 6 million gate Field Programmable Gate Array (FPGA)[19,20].

***De novo* shotgun sequence assembler.** A *de novo* flow-space assembler was developed to capture all of the information contained in the original flow-based signal trace. It also addresses the fact that existing assemblers are not optimized for 80–120-bases reads, particularly with respect to memory management due to the increased number of sequencing reads needed to achieve equivalent genome coverage. (A completely random genome covered with 100-bases reads requires approximately 50% more reads to yield the same number of contiguous regions (contigs) as achieved with 700-bases reads, assuming the need for a 30-bases overlap between reads[21].) This assembler consists of a series of modules: the Overlapper, which finds and creates overlaps between reads; the Unitigger, which constructs larger contigs of overlapping sequence reads; and the Multi Aligner, which generates consensus calls and quality scores for the bases within each contig (Supplementary Methods). (The names of the software modules are based [illegible] developed pre-

Received 6 May; accepted 10 June 2005.
Published online 31 July 2005.

1. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA* 74, 5463–5467 (1977).
2. Prober, J. M. *et al.* A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* 238, 336–341 (1987).
3. NIH News Release. NHGRI seeks next generation of sequencing technologies. 14 October 2004 (http://www.genome.gov/12513210).
4. Nyren, P., Pettersson, B. & Uhlen, M. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Anal Biochem.* 208, 171–175 (1993).
5. Ronaghi, M. *et al.* Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* 242, 84–89 (1996).
6. Jacobson, K. B. *et al.* Applications of mass spectrometry to DNA sequencing. *GATA* 8, 223–229 (1991).
7. Bains, W. & Smith, G. C. A novel method for nucleic acid sequence determination. *J. Theor. Biol.* 135, 303–307 (1988).
8. Jett, J. H. *et al.* High-speed DNA sequencing: an approach based upon fluorescence detection of single molecules. *Biomol. Struct. Dynam.* 7, 301–309 (1989).
9. Tawfik, D. S. & Griffiths, A. D. Man-made cell-like compartments for molecular evolution. *Nature Biotechnol.* 16, 652–656 (1998).
10. Ghadessy, F. J., Ong, J. L. & Holliger, P. Directed evolution of polymerase function by compartmentalized self-replication. *Proc. Natl Acad. Sci. USA* 98, 4552–4557 (2001).
11. Dressman, D., Yan, H., Traverso, G., Kinzler, K. W. & Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl Acad. Sci. USA* 100, 8817–8822 (2003).
12. Ronaghi, M., Uhlen, M. & Nyren, P. A sequencing method based on real-time pyrophosphate. *Science* 281, 363–365 (1998).
13. Fraser, C. M. *et al.* The minimal gene complement of *Mycoplasma genitalium*. *Science* 270, 397–403 (1995).
14. Tettelin, H. *et al.* Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science* 293, 498–506 (2001).
15. Leamon, J. H. *et al.* A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis* 24, 3769–3777 (2003).
16. Ronaghi, M. Pyrosequencing sheds light on DNA sequencing. *Genome Res.* 11, 3–11 (2001).
17. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8, 175–185 (1998).
18. Moore, G. E. Cramming more components onto integrated circuits. *Electronics* 38 (8) (1965).
19. Mehta, K., Rajesh, V. A. & Veeraswamy, S. FPGA implementation of VXIbus interface hardware. *Biomed. Sci. Instrum.* 29, 507–513 (1993).
20. Fagin, B., Watt, J. G. & Gross, R. A special-purpose processor for gene sequence analysis. *Comput. Appl. Biosci.* 9, 221–226 (1993).
21. Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2, 231–239 (1988).
22. Myers, E. W. Toward simplifying and accurately formulating fragment assembly. *J Comput. Biol.* 2, 275–290 (1995).
23. Ogawa, T. *et al. Increased Productivity For Core Labs Using One Polymer and One Array Length for Multiple Applications.* Poster P108-T. ABRF '05: Biomolecular Technologies: Discovery to Hypotheses (Savannah, Georgia, 5–8 February 2005); also available as Applied Biosystems 3730x/DNA Analyzer Specification Sheet (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge P. Dacey and the support of the Operations groups of 454 Life Sciences. This research was supported in part by the US Department of Health and Human Services under NIH grants.

Author Information Sequences for *M. genitalium* and *S. pneumoniae* are deposited at DDBJ/EMBL/GenBank under accession numbers AAGA01000000 and AAGY01000000, respectively. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Correspondence and requests for materials should be addressed to J.M.R. (jrothberg@454.com).

© 2005 Nature Publishing Group

Materiali e metodi

Informazioni supplementari

References (citazioni)

ERRATUM

doi:10.1038/nature04774

Significant primordial star formation at redshifts $z \approx 3$–4

Raul Jimenez & Zoltan Haiman

Nature 440, 501-504 (2006)

In this Letter, the received and accepted dates were incorrect. They should read: 'Received 6 September 2005; accepted 4 January 2006.'

CORRIGENDUM

doi:10.1038/nature04726

Genome sequencing in microfabricated high-density picolitre reactors

Marcel Margulies, Michael Egholm, William E. Altman, Said Attiya, Joel S. Bader, Lisa A. Bemben, Jan Berka, Michael S. Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B. Dewell, Alex de Winter, James Drake, Lei Du, Joseph M. Fierro, Robin Forte, Xavier V. Gomes, Brian C. Godwin, Wen He, Scott Helgesen, Chun Heen Ho, Stephen K. Hutchison, Gerard P. Irzyk, Szilveszter C. Jando, Maria L. I. Alenquer, Thomas P. Jarvie, Kshama B. Jirage, Jong-Bum Kim, James R. Knight, Janna R. Lanza, John H. Leamon, William L. Lee, Steven M. Lefkowitz, Ming Lei, Jing Li, Kenton L. Lohman, Hong Lu, Vinod B. Makhijani, Keith E. McDade, Michael P. McKenna, Eugene W. Myers, Elizabeth Nickerson, John R. Nobile, Ramona Plant, Bernard P. Puc, Michael Reifler, Michael T. Ronan, George T. Roth, Gary J. Sarkis, Jan Fredrik Simons, John W. Simpson, Maithreyan Srinivasan, Karrie R. Tartaro, Alexander Tomasz, Kari A. Vogt, Greg A. Volkmer, Shally H. Wang, Yong Wang, Michael P. Weiner, David A. Willoughby, Pengguang Yu, Richard F. Begley & Jonathan M. Rothberg

Nature 437, 376-380 (2005)

In this paper, the second name of Chun Heen Ho was misspelled as He. In the earlier Corrigendum (*Nature* 439, 502; 2006) the surnames of Stephen K. Hutchison and Brian C. Godwin were misspelled as Hutchinson and Goodwin. In addition, in the Supplementary Information of the original paper, there were errors in the sequences of the DNA capture and enrichment primers, and in the formulation of emulsion oil. The Supplementary Information was updated on 4 May 2006.

CORRIGENDUM

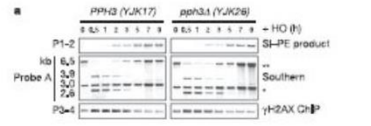
doi:10.1038/nature04772

A phosphatase complex that dephosphorylates γH2AX regulates DNA damage checkpoint recovery

Michael-Christopher Keogh, Jung-Ae Kim, Michael Downey, Jeffrey Fillingham, Dipanjan Chowdhury, Jacob C. Harrison, Megumi Onishi, Nira Datta, Sarah Galicia, Andrew Emili, Judy Lieberman, Xuetong Shen, Stephen Buratowski, James E. Haber, Daniel Durocher, Jack F. Greenblatt & Nevan J. Krogan

Nature 439, 497-501 (2006)

Figure 3a of this Letter contains an inadvertently duplicated panel set: those referring to '*pph3*Δ (YJK26)' are identical to '*PPH3* (YJK17)'. The corrected panels are shown here. Note that the quantification shown in Fig. 3b refers to the correct panels. Our results and conclusions are unaffected by this oversight.



CORRIGENDUM

doi:10.1038/nature04773

A high-mobility electron gas at the $LaAlO_3/SrTiO_3$ heterointerface

A. Ohtomo & H. Y. Hwang

Nature 427, 423-426 (2004)

In Fig. 2a and b of this Letter, the *y* axis should be labelled in units of ohms per square, not milli-ohms per square. This error does not occur elsewhere in the paper, and does not change any of our results or conclusions.

© 2006 Nature Publishing Group

Caso particolare: la Review

Si tratta di una pubblicazione contenente una **rassegna di articoli** pubblicati in precedenza su un determinato argomento.

L'idea è quella di fornire una panoramica delle conoscenze sull'argomento trattato fino al momento della pubblicazione della Review.

Bisogna sottolineare che, a differenza di un articolo di ricerca originale, **non presenterà nuovi risultati sperimentali** ma riporterà risultati prodotti da altri con le relative citazioni.

Il processo di referaggio

… apriamo una piccola parentesi

1. Sottomette il manoscritto
2. Il manoscritto viene valutato dall'Editor in Chief
3. Poi dall'Editor Associato (EA)
4. Se passa il primo filtro viene mandato a valutatori (referee) anonimi
5. Il manoscritto torna all'EA con i commenti:
   - Accept
   - Minor Revisions
   - Major Revisions
   - Reject
6. Se il manoscritto non viene rifiutato gli autori hanno di solito 6/9 mesi per rispondere ai commenti e fare nuovi esperimenti
7. Torna al punto 4 (repeat 2/3 times)



Valutazione dei giornali scientifici


NATURE

A WEEKLY ILLUSTRATED JOURNAL OF SCIENCE

THURSDAY, NOVEMBER 4, 1869

NATURE: APHORISMS BY GOETHE

NATURE! We are surrounded and embraced by her: powerless to separate ourselves from her, and powerless to penetrate beyond her.

Without asking, or warning, she snatches us up into her circling dance, and whirls on on until we are tired, and drop from her arms.

She is ever shaping new forms: what is, has never yet been; what has been, comes not again. Everything is new, and yet nought but the old.

We live in her midst and know her not. She is incessantly speaking to us, but betrays not her secret. We constantly act upon her, and yet have no power over her.

The one thing she seems to aim at is Individuality; yet she cares nothing for individuals. She is always building up and destroying; but her workshop is inaccessible.

Her life is in her children; but where is the mother? She is the only artist; working up the most uniform material into utter opposites; arriving, without a trace of effort, at perfection, at the most exact precision, though always veiled under a certain softness.

Each of her works has an essence of its own; each of her phenomena a special characterization; and yet their diversity is in unity.

She performs a play; we know not whether she sees it herself, and yet she acts for us, the lookers-on.

Incessant life, development, and movement are in her, but she advances not. She changes for ever and ever, and rests not a moment. Quietude is inconceivable to her, and she has laid her curse upon rest. She is firm. Her steps are measured, her exceptions rare, her laws unchangeable.

She has always thought and always thinks; though not as a man, but as Nature. She broods over an all-comprehending idea, which no searching can find out.

Mankind dwell in her and she in them. With all men she plays a game for love, and rejoices the more they win. With many, her moves are so hidden, that the game is over before they know it.

That which is most unnatural is still Nature; the stupidest philistinism has a touch of her genius. Whoso cannot see her everywhere, sees her nowhere rightly.

She loves herself, and her immutable eyes and affections are fixed upon herself. She has divided herself that she may be for ever delight. She causes an endless succession of new capacities for enjoyment to spring up, that her insatiable sympathy may be assuaged.

She rejoices in illusion. Whoso destroys it in himself and others, him she punishes with the severest tyranny. Whoso follows her in faith, him she takes as a child to her bosom.

Her children are numberless. To none is she altogether miserly; but she has her favourites, on whom she squanders much, and for whom she makes great sacrifices. Over greatness she spreads her shield.

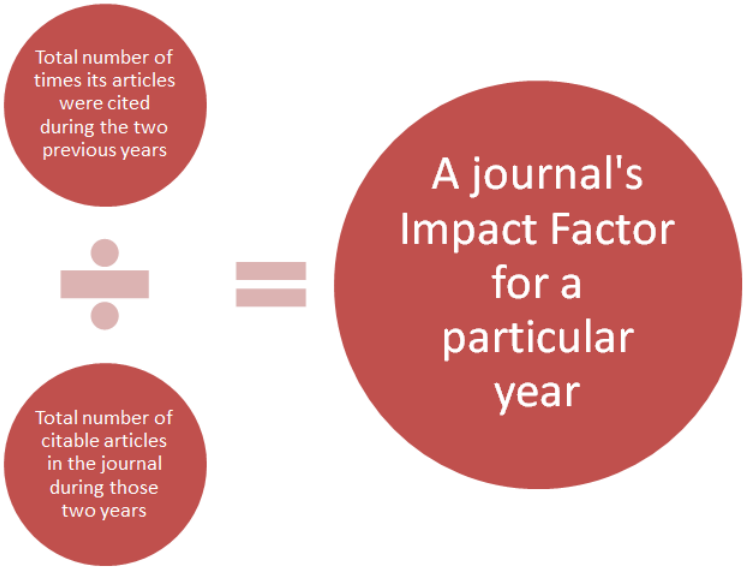
She tosses her creatures out of nothingness, and tells them not whence they came, nor whither they go. It is their business to run, she knows the road. Her mechanism has few springs—but they never wear out, are always active and manifold.

The spectacle of Nature is always new, for she is always renewing the spectators. Life is her most exquisite invention; and death is her expert contrivance to get plenty of life.

She wraps man in darkness, and makes him for ever long for light. She creates him dependent upon the earth, dull and heavy; and yet is always shaking him until he attempts to soar above it.

Impact Factor (IF) è usato per misurare l'importanza o il ranking delle riviste in uno specifico settore.

Giornali con IF alto sono considerate più prestigiose perché proporzionale a quanto la rivista viene letta e citata.



	Number of Categories	Number of Journals	Number of Citable Items
Agricultural Sciences	7	425	57,608
Arts & Humanities, Interdisciplinary	8	983	34,942
Biology & Biochemistry	34	3,971	750,109
Chemistry	21	2,375	697,416
Clinical Medicine	59	7,396	1,191,642
Computer Science	14	1,581	229,807
Economics & Business	21	3,309	254,792
Engineering	41	3,534	811,213
Environment/Ecology	13	1,690	315,254

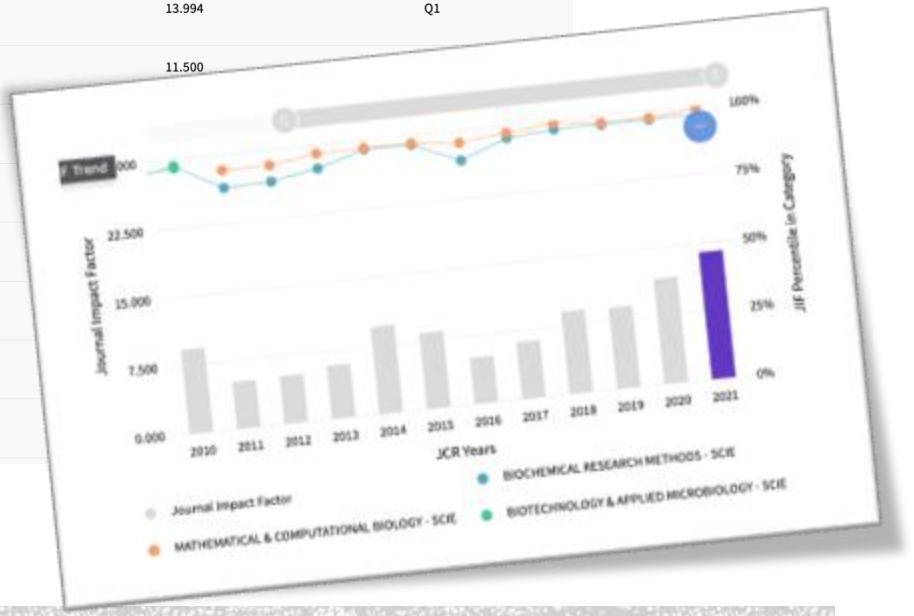
VALUTAZIONE DEI GIORNALI SCIENTIFICI

I giornali sono divisi in categorie (topics).

Un giornale può appartenere a più categorie

MATHEMATICAL & COMPUTATIONAL BIOLO... | SCIE | JCR Year: 2021

Journal name	ISSN	eISSN	Category	Total Citations	2021 JIF	JIF Quartile
BRIEFINGS IN BIOINFORMATICS	1467-5463	1477-4054	MATHEMATICAL & COMPUTATIONAL BIOLOGY - SCIE	15,191	13.994	Q1
Wiley Interdisciplinary Reviews-Computational Molecular Science	1759-0876	1759-0884	MATHEMATICAL & COMPUTATIONAL BIOLOGY - SCIE	9,679	11.500	
Research Synthesis Methods	1759-2879	1759-2887	MATHEMATICAL & COMPUTATIONAL BIOLOGY - SCIE	5,479		
IEEE Journal of Biomedical and Health Informatics	2168-2194	2168-2208	MATHEMATICAL & COMPUTATIONAL BIOLOGY - SCIE	11,304		
BIOINFORMATICS	1367-4803	1460-2059	MATHEMATICAL & COMPUTATIONAL BIOLOGY - SCIE	174,344		
COMPUTERS IN BIOLOGY AND MEDICINE	0010-4825	1879-0534	MATHEMATICAL & COMPUTATIONAL BIOLOGY - SCIE	14,531		
BIOSTATISTICS	1465-4644	1468-4357	MATHEMATICAL & COMPUTATIONAL BIOLOGY - SCIE	5,158		
Current Bioinformatics	1574-8936	2212-392X	MATHEMATICAL & COMPUTATIONAL BIOLOGY - SCIE	1,505		



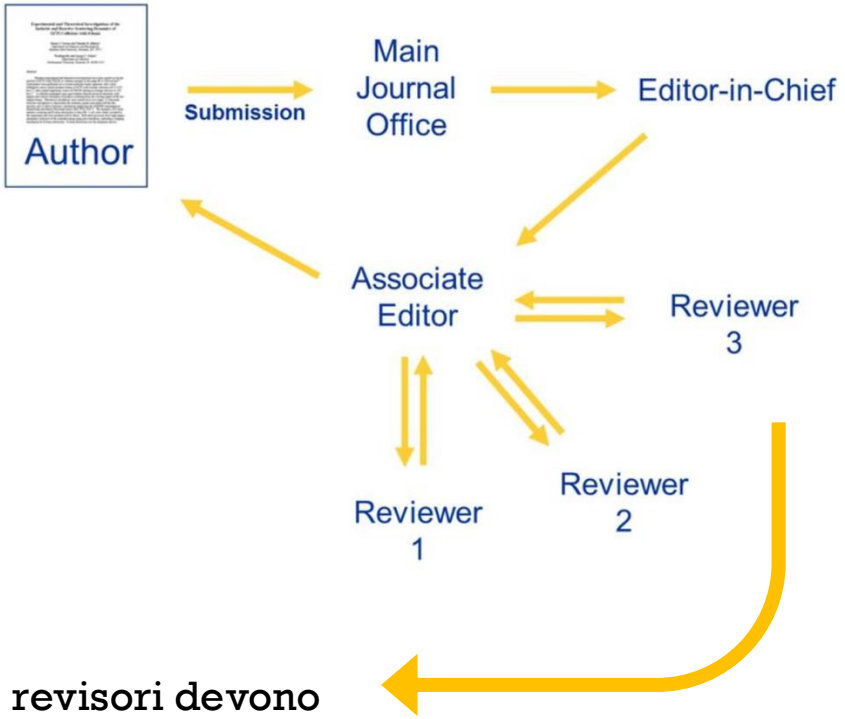
VALUTAZIONE DEI GIORNALI SCIENTIFICI

All'interno di ogni categoria i giornali sono ordinati usati delle metriche basate sulle citazioni

Il processo di referaggio

… apriamo una piccola parentesi

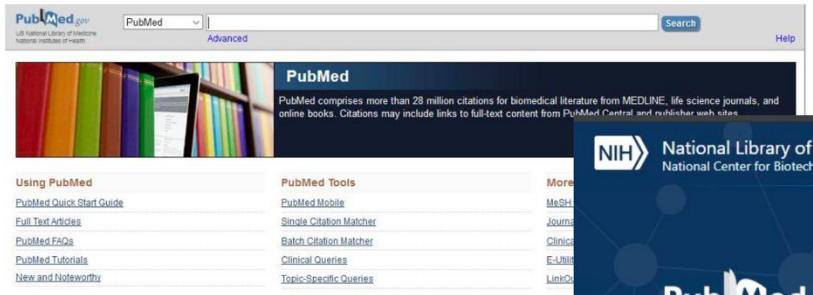
1. Sottomette il manoscritto
2. Il manoscritto viene valutato dall'Editor in Chief
3. Poi dall'Editor Associato (EA)
4. Se passa il primo filtro viene mandato a valutatori (referee) anonimi
5. Il manoscritto torna all'EA con i commenti:
   - Accept
   - Minor Revisions
   - Major Revisions
   - Reject
6. Se il manoscritto non viene rifiutato gli autori hanno di solito 6/9 mesi per rispondere ai commenti e fare nuovi esperimenti
7. Torna al punto 4 (repeat 2/3 times)

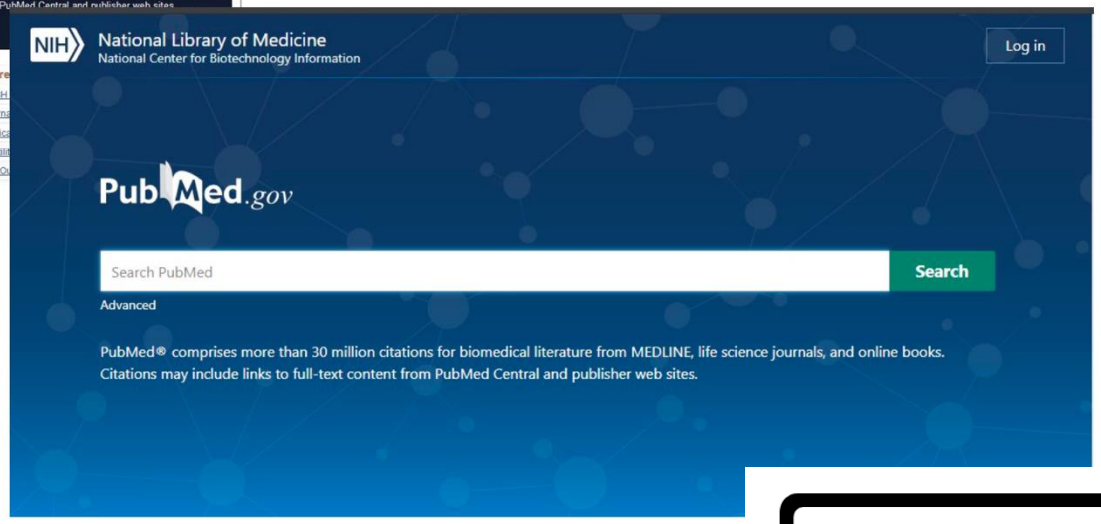


I revisori devono accertarsi che i dati siano correttamente depositati nei database

Ricerche Bibliografiche

Il database di **libero accesso PubMed ->** http://www.ncbi.nlm.nih.gov/pubmed





Consente l'accesso alle citazioni della **letteratura biomedica** e fornisce i link ad altre risorse biomolecolari di Entrez. Permette di trovare rapidamente una particolare **pubblicazione** di cui conosciamo anche solo alcune informazioini (e.g. nome di un autore o parola chiave)

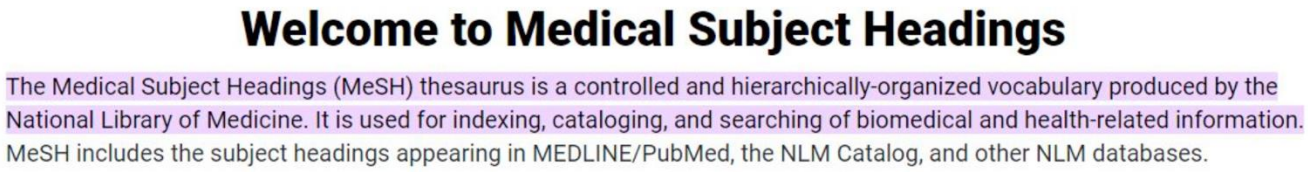
SCAN ME

PUBMED

PubMed contiene oltre 33 milioni di citazioni relative alla letteratura biomedica che provengono da diverse fonti, tra cui **MEDLINE** (che è parte di PubMed), riviste scientifiche nel campo delle scienze della vita e libri online.

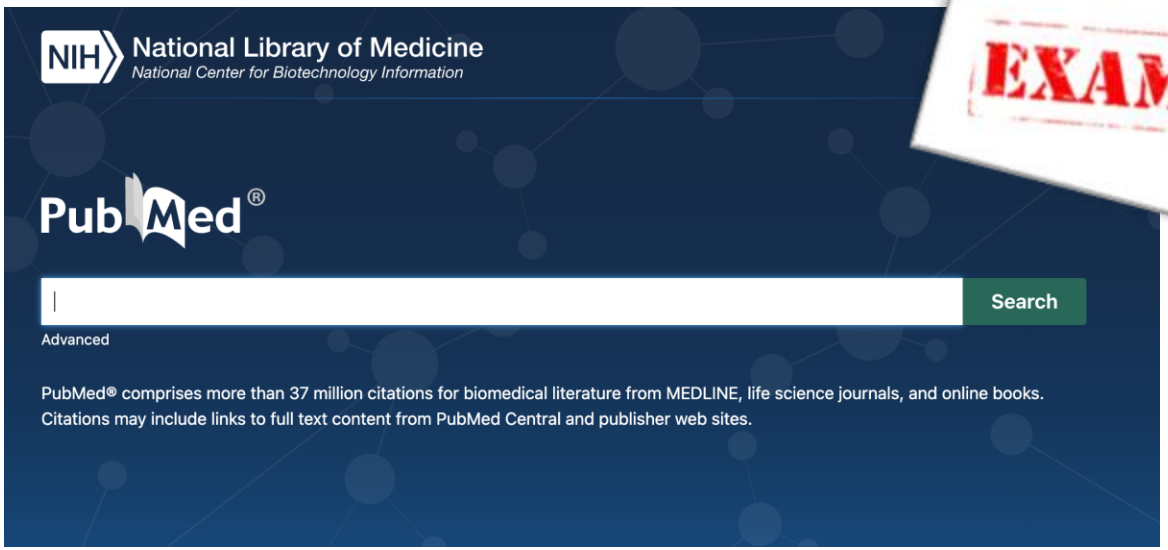
L'interfaccia grafica di ricerca è molo semplice e permette interrogazioni testuali di vario tipo.

I dati, tratti da oltre 5.200 riviste scientifiche specializzate, sono classificati in campi che permettono la strutturazione e indicizzati per permettere la ricerca mirata delle informazioni tramite il vocabolario controllato Medical Subject Headings o **MeSH**.



Dizionario dei sinonimi (thesaurus) che include **termini** descritti in **strutture gerarchiche** che permettono di effettuare ricerche a vari livelli di specificità.
Utilizzato per indicizzare gli articoli provenienti da più di 5200 riviste scientifiche di Medline e PubMed, più catalogare libri, documenti, audiovisivi ecc.
Ogni referenza bibliografica è associata con un insieme di **elementi del dizionario MeSH che ne descrivono il contenuto**.

SCAN ME



EXAMPLE

Cercate il lavoro che abbiamo citato prima sulla balena **minke** usando le parole chiave:

`«Minke whale genome»`

1. provate prima sul campo di testo generale
2. Provate cliccando su «advanced» e mettete le parole chiave indicando il campo di riferimento «title»
3. Fate la stessa cosa mettendo le parole chiave «`Minke whale AND genome`»
4. Cosa è cambiato e perché ?