

Basi di dati

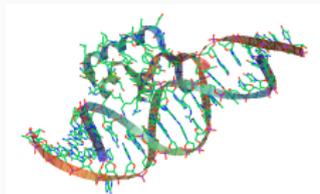
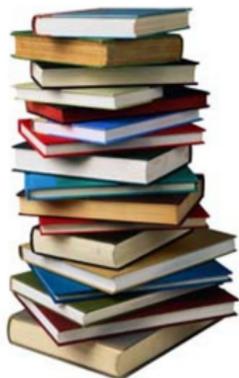
Capitolo 9 del testo

Alberto Policriti



5 Dicembre, 2019

Informazioni



Trova il tuo hotel

1 Destinazione

2 Arrivo 26/12/2007 Partenza 04/01/2008

3 Camera 1 Room 1 Adulti 2 Bambini (0-17) 0

[Più opzioni di ricerca](#)

Vai

ALBUM IN YOGA ●○○ [Elenco completo](#)

Nessuno è solo Tiziano Ferro	Call Me Irresponsible Michael Bublé	Taking the Long Way Dixie Chicks	St. Elsewhere Gnarts Barkley
Good Girl Gone Bad Rihanna	Stop the Clocks Oasis	Rather Ripped (International...) Sonic Youth	Martyr - EP Depeche Mode

APPENA AGGIUNTI ●○○○○ [Elenco completo](#)

Respighi: Belkis, Queen of S... Geoffrey Simon & Phiharmon...	Corelli: Sonate Da Chiesa, O... London Baroque	Mozart: the Late Sonatas for... Chiara Banchini & Temensch...	Mi faccio in quattro Gigi D'Alessio

PubMed All Databases BLAST OMIM Books TaxBrowser Structure

Search for

- Raccolta delle informazioni
- Rappresentazione delle informazioni (dati)
- Organizzazione dei dati
- Conservazione dei dati
- Reperimento/interrogazione dei dati

... per il perseguimento degli scopi dell'organizzazione

Sistema informativo

Sistema informativo vs sistema informatico



The screenshot shows the Project Gutenberg website interface. At the top, there is a navigation bar with links for Home, Search, About, and Contact. Below this, the main heading reads "Free ebooks - Project Gutenberg". A search bar is prominently displayed. To the left, a sidebar contains a "search for books" section with a search input field and a "Go" button. Below the search bar, there are links for "Browse Catalog", "Bookshelves", "Main Page", "Categories", and "Contact Us". A section titled "Project Gutenberg appreciates your donation!" includes a "Donate" button and a link to "Why donate?". There is also a "Library" section with links for "Portugals", "Deutsch", and "Französisch".

The main content area features a "Some of the Latest Books" section with a row of book covers. Below this is a "Welcome" section with the following text:

Welcome

Project Gutenberg offers over 57,000 free eBooks. Choose among free epub books, free kindle books, download them or read them online. You will find the world's great literature here, with focus on older works for which copyright has expired. Thousands of volunteers digitized and diligently proofread the eBooks, for enjoyment and education.

No fee or registration is required. If you find Project Gutenberg useful, please consider a small [donation](#), to help Project Gutenberg digitize more books, maintain our online presence, and improve Project Gutenberg programs and offerings. Other ways to help include [donating more books](#), [recording audio books](#), or [reporting errors](#).

Below the welcome message is a "News" section with the following article:

Project Gutenberg Supports Net Neutrality

The Federal Communications Commission made a ruling to abandon major components of network neutrality. This legalizes "slow lanes" for network traffic that does not come from commercial partners of network providers. Sites whose content is free, and generates no revenue - such as Project Gutenberg - are at risk for downgraded speeds, access fees imposed by network providers, or ads interjected by network providers. We encourage Project Gutenberg readers to [contact Sen. L. Scott Brown](#) to express views on this important issue.

Another news item is titled "The Public Domain will grow again in 2019".

The Public Domain will grow again in 2019

In the US, annual copyright terms expiry is set to begin again in 2019 after a 95-year hiatus due to the Copyright Term Extension Act of 1998. On January 1, 2019, items published in 1923 will enter the public domain in the US. In the north hemisphere of Project Gutenberg, growth of the public domain on January 1 was an annual event. See Duke Law's "Public Domain Day" for a listing of many items that were scheduled to enter the public domain, but have yet to do so because of the 1998 extension. Some notable items scheduled to enter the public domain in 2019 include Felix Salten's "Bambi" and Khalil Gibran's "The Prophet".

At the bottom of the page, there is a "Site Map" section and a "Find eBooks" section with the following links:

- Book Search
- Recently added eBooks
- Most Frequently Downloaded eBooks and Top 100 eBooks this month
- Bookshelves of related eBooks
- New Books Feeds
- Browse Catalog: Browse and search, including full-text search.
- Offline Catalog: handy book listings to consult offline.
- Visit [gutenberg.org](#) for free eBooks by contemporary authors.

<http://www.gutenberg.org/>

Collezione strutturata di dati. . .

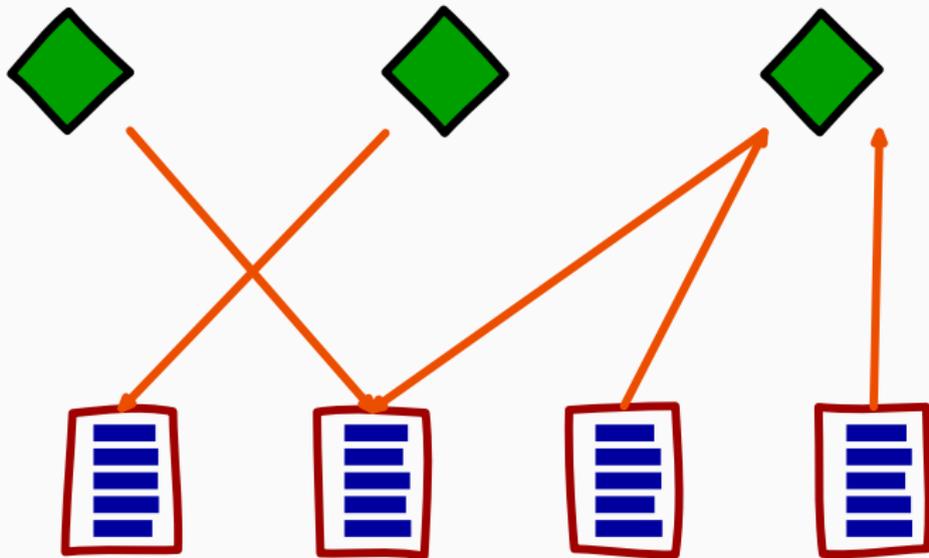
- di dimensioni arbitrarie
- persistente
- condivisa

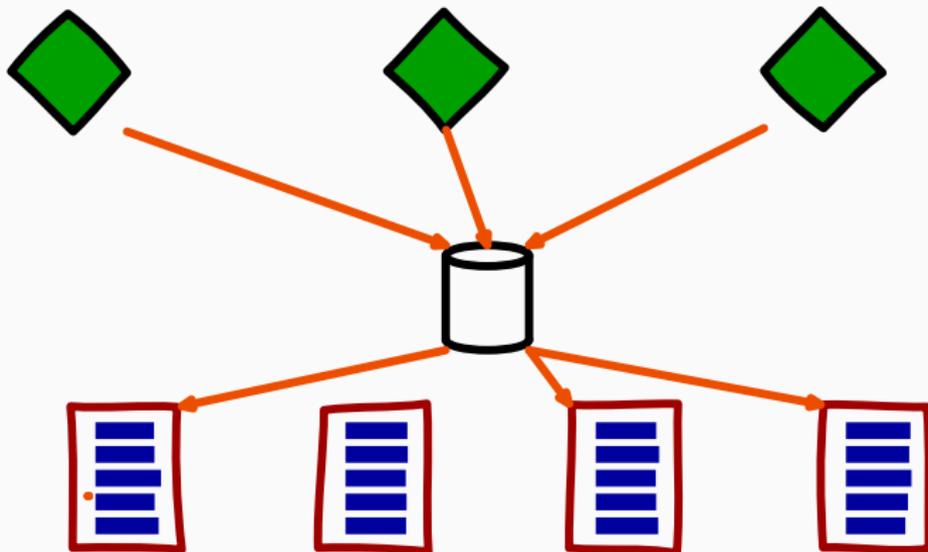
DataBase Management System (Sistema di gestione di basi di dati)

- strumento software (collezione di programmi)
- per la creazione e manipolazione di basi di dati
- di qualunque dimensione e per qualunque scopo

Esempi:

- Oracle
- PostgreSQL
- MySQL
- ...



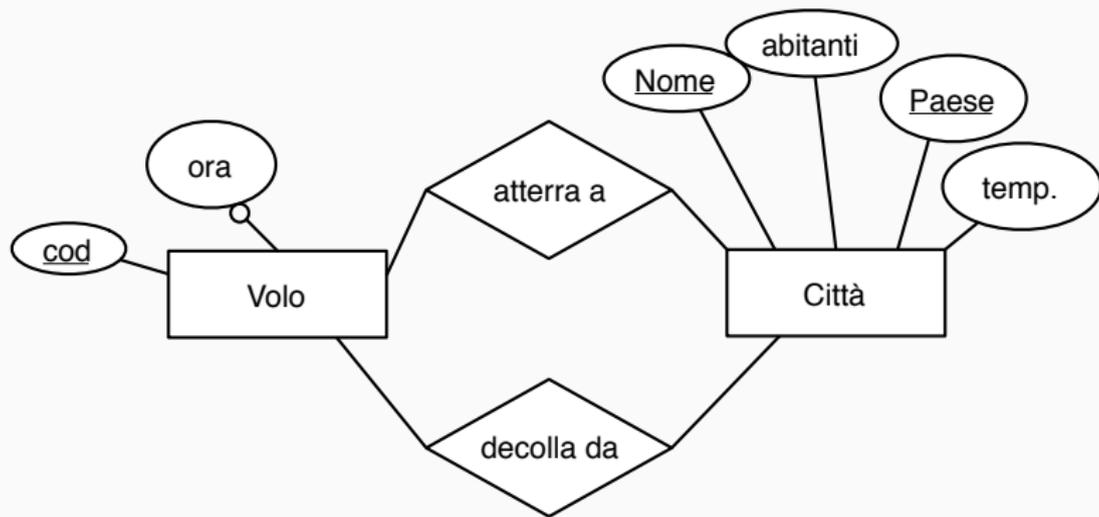


Collezione di concetti e regole per la descrizione dei dati, delle relazioni tra i dati e dei vincoli di consistenza sui dati.

- Proprietà statiche:
 - oggetti di informazione, entità
 - proprietà degli oggetti (attributi)
 - relazioni tra gli oggetti
 - vincoli su oggetti e relazioni
- Proprietà dinamiche:
 - operazioni su oggetti e relazioni
 - relazioni tra operazioni (transazioni)
 - vincoli sull'evoluzione degli oggetti e delle relazioni

- Modelli dei dati primitivi (ma vedi i ,“flat file” . . .)
 - basati sulla gestione diretta di file
 - Operazioni sui dati \equiv operazioni sui file
- Modelli dei dati classici
 - Fondamento dei DBMS attuali
 - Obiettivo: indipendenza dei dati
 - Modello relazionale
- Modelli dei dati semantici
 - Strumenti per la progettazione
 - Modello entità-relazione (ER)

Modello entità-relazione



Per una completa descrizione (con esempi) si veda:

<https://users.dimi.uniud.it/~massimo.franceschet/teatro-sql/index.html>

Il modello relazionale

- Storia:
 - Proposto da E. Codd nel 1970
 - Acquista popolarità negli anni Ottanta
 - Attualmente, il modello più diffuso
- Caratteristiche:
 - Semplicità: l'utente percepisce la base di dati come un insieme di **tabelle**
 - Le relazioni tra le tabelle sono implicitamente rappresentate dai valori
 - Linguaggi di manipolazione dichiarativi
 - specificano che risultato ottenere, non le modalità per ottenere il risultato)
 - Prospettiva algebrica: insieme di strutture dati e operatori
 - Prospettiva logica: SQL

$$R(A_1, \dots, A_n)$$

- **Nome** di relazione: R
- **Attributi**: A_1, \dots, A_n
- A ciascun A_i è associato un **dominio**.

Lo schema di una base di dati è un insieme di schemi di relazione

Esempio

- GENOMA(organismo, dimensione)
- SEQUENZA(accession, seq_grezza, specie)
- BLAST(database, data, risultato, seq_id)

Rappresentazione tabulare:

GENOMA

organismo dimensione

SEQUENZA

accession seq_grezza specie

BLAST

db data risultato seq_id

Istanza di base di dati

GENOMA

organismo	dimensione
Homo Sapiens	3000
Arabidopsis Thaliana	100

SEQUENZA

accession	seq_grezza	specie
1234	ATGCT...	Homo Sapiens
567	GTCCGT...	Arabidopsis
890	TGGGGA...	Homo Sapiens

BLAST

db	data	risultato	seq_id
----	------	-----------	--------

Relazioni e tabelle

Concetto relazionale:

- relazione
- attributo
- grado
- tupla
- cardinalità

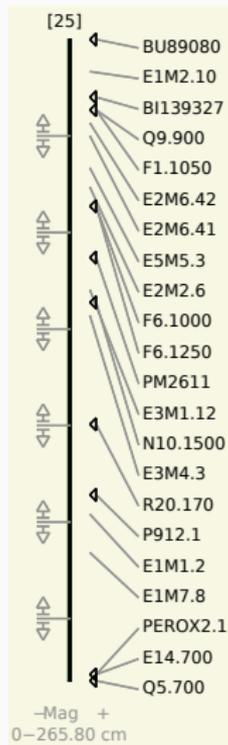
Equivalente informale:

- tabella
- colonna
- numero di colonne
- riga
- numero di righe

Differenze tra tabelle e relazioni:

- in una tabella le colonne e le righe sono ordinate
 - nelle relazioni non sussistono tali ordinamenti
- in una tabella possono esistere righe ripetute
 - una relazione è un insieme, pertanto non ci possono essere tuple ripetute

Mappe biologiche: schema



- Mappa(id, specie, nome, tipo, unità, start, stop)
- Marker(id, mappa, tipo, nome, start, stop)

Mappe: un'istanza

<u>id</u>	specie	nome	Mappa			
			tipo	unità	start	stop
1	F1	LG I	gen	cM	0	70.6
2	F1	LG II	gen	cM	0	122.4
3	D2	Chr IV	seq	bp	0	1437893

<u>id</u>	mappa	tipo	Marker		
			nome	start	stop
0	1	SSR	PM1234	36.8	NULL
1	1	SSR	PM3245	18.4	NULL
2	3	SSR	PM334	748876	748967
3	1	AFLP	E5M1.3	8.3	NULL

Proprietà, associate a uno schema di base di dati, che devono essere soddisfatte da tutte le istanze che rappresentano informazioni corrette della base di dati.

- Si specificano insieme allo schema della base di dati
- il DBMS verifica la consistenza dei dati rispetto ai vincoli
- Vincoli intrarelazionali
 - coinvolgono una o più tuple di una stessa relazione
- Vincoli interrelazionali
 - coinvolgono tuple di relazioni diverse

Mappa(id, specie, nome, tipo, u, start, stop)

Marker(id, mappa, tipo, nome, start, stop)

- I valori di start e stop devono essere non negativi
- Non ci possono essere id ripetuti
- Il valore di mappa in Marker deve corrispondere a qualche mappa
- I nomi dei marcatori devono essere sempre noti
- ...

- Tupla \equiv entità del mondo reale
- Identificabile mediante un sottoinsieme di valori
 - univoci
 - sempre noti
- **Chiavi**: sottoinsiemi minimali con tali caratteristiche
- **Chiave primaria**: scelta dal progettista tra le chiavi candidate

Esempio

Esame(matricola, corso, data, voto, lode)

Quali sono le chiavi? L'unica è {matricola, corso}

Visita(data, ora, stanza, medico, paziente, referto)

Quali sono le chiavi?

1. {data, ora, stanza}

- A una data ora di un certo giorno ci può essere solo una visita per stanza

2. {data, ora, medico}

- A una data ora di un certo giorno un medico non può fare più di una visita

3. {data, ora, paziente}

- A una data ora di un certo giorno un paziente non può essere sottoposto a più visite

Integrità referenziale

I valori di uno o più attributi di una relazione devono occorrere in attributi corrispondenti di un'altra relazione (e.g., specie → organismo)

Chiave esterna

GENOMA	
organismo	dimensione
Homo Sapiens	3000
Arabidopsis T.	100

SEQUENZA		
accession	seq_grezza	specie
1234	ATGCT...	Homo Sapiens
1235	GTCCGT...	Arabidopsis T.

1. Definire uno schema relazionale completo per rappresentare l'informazione relativa ai libri letti dai soci di una biblioteca.
2. Definire uno schema relazionale completo per rappresentare l'informazione relativa ai voli internazionali di una compagnia aerea. Ciascun volo è identificato da un codice e da una data, e decolla e atterra in determinati aeroporti (che devono essere sempre noti), che si trovano in determinate città, di cui interessa sapere il numero di abitanti. Si tenga conto che:
 - una città può avere diversi aeroporti;
 - città di stati diversi possono avere lo stesso nome.

Beh, perché non usare un foglio elettronico?

- File + programmi?!
- Assenza di linguaggi d'interrogazione \equiv limitate possibilità di estrarre le informazioni desiderate
- Inconsistenze, assenza di condivisione e concorrenza, etc. . .

Joint project to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes.

- Many databases, even for the same organism
- All data sets in the Ensembl system are stored in relational databases (MySQL)
- Data can be obtained by FTP
(<ftp://ftp.ensembl.org/pub/>)
- Data can be queried directly (the database schema is public)

```
mysql -u anonymous -h ensembl.db.ensembl.org
```

Ensembl | [BLAST/BLAT](#) | [VEP](#) | [Tools](#) | [BioMart](#) | [Downloads](#) | [Help & Docs](#) | [Blog](#)

Login/Register

Tools

[All tools](#)

BioMart >

Export custom datasets from Ensembl with this data-mining tool

BLAST/BLAT >

Search our genomes for your DNA or protein sequence

Variant Effect Predictor >

Analyse your own variants and predict the functional consequences of known and unknown variants

Search

All species for

e.g. BRCA2 or rat 5-82737363-63627648 or rs999 or coronary heart disease

All genomes

-- Select a species --

- View full list of all Ensembl species
- Edit your favourites

Favourite genomes

- Human GRC38.p13
Still using GRC307?
- Mouse GRCm38.p6
- Zebrafish GRCz11

Ensembl Release 98 (September 2019)

- eSNP152 for human with changes to repeat expansion/contraction representation
- Update to GENCODE 32 for human
- New beta PloS GWAS analysis pipeline tool online
- New genomes: 11 pig breeds with comparative analysis between them and other agricultural species
- New gene annotation: dog, cat, horse, rabbit, grey short-tailed opossum, marmoset and rhesus monkey
- New genomes: nine fish, one frog, five plants, one worm and one diatom

[More release news](#) on our blog

Other news from our blog

- 27 Nov 2019: Job: Software Engineer – Comparative Genomics
- 22 Nov 2019: Cool stuff the Ensembl VEP can do: external gene regulation annotation
- 14 Nov 2019: Changes in BioMart species support

Compare genes across species

Find SNPs and other variants for my gene

Gene expression in different tissues

Retrieve gene sequence

Find a Data Display

Use my own data in Ensembl

EMBL-EBI Ensembl creates, integrates and distributes reference datasets and analysis tools that enable genomics. We are based at [EMBL-EBI](#) and our software and data are freely available. Our [acknowledgements page](#) includes a list of current and previous funding bodies. [How to cite Ensembl](#) in your own publications.

Ensembl release 98 - September 2019 @ EMBL-EBI [Permanent link](#) - [View in archive](#)

About Us	Get help	Our sister sites	Follow us
About us	Using this website	Ensembl Bacteria	Blog
Contact us	Asking custom tracks	Ensembl Fungi	Twitter
Citing Ensembl	Downloading data	Ensembl Plants	Facebook
Privacy policy	Video tutorials	Ensembl Proteins	
Disclaimer	Variant Effect Predictor (VEP)	Ensembl Metascore	

PDB (<http://www.rcsb.org/pdb/home/home.do>) provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

- Single database (MySQL): 461 tables
- The database schema is public, but the db can be queried only through the web interface
- Complex queries through , “Advanced search”
- Experimental and theoretical models are kept distinct
- Integrates also data from Swiss-Prot, Genbank, PubMed.

Nucleic Acids Res. 2005 January 1; 33(Database Issue): D233–D237.

Published online 2004 December 17. doi: 10.1093/nar/gki057.

[Copyright](#) © 2005 Oxford University Press

The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema

Nita Deshpande,¹ Kenneth J. Address,¹ Wolfgang F. Bluhm,¹ Jeffrey C. Merino-Ott,¹ Wayne Townsend-Merino,¹ Qing Zhang,¹ Charlie Knezevich,¹ Lie Xie,¹ Li Chen,³ Zukang Feng,³ Rachel Kramer Green,³ Judith L. Flippen-Anderson,³ John Westbrook,³ Helen M. Berman,³ and Philip E. Bourne^{1,2*}

Entrez (<http://www.ncbi.nlm.nih.gov/Entrez/>) integrates

- the scientific literature
- DNA and protein sequence databases
- 3D protein structure and protein domain data
- population study datasets
- expression data
- assemblies of complete genomes
- taxonomic information

It is a retrieval system designed for searching its linked databases.



Search NCBI

Search

NCBI Databases

Literature

The World's largest repository of medical and scientific abstracts, full-text articles, books and reports

Bookshelf

Books and reports

MeSH

Ontology used for PubMed indexing

NLM Catalog

Books, journals and more in the NLM Collections

PubMed

Scientific and medical abstracts/citations

PubMed Central

Full-text journal articles

Genes

Gene sequences and annotations used as references for the study of orthologs structure, expression, and evolution

Gene

Collected information about gene loci

GEO DataSets

Functional genomics studies

GEO Profiles

Gene expression and molecular abundance profiles

HomoloGene

Homologous genes sets for selected organisms

PopSet

Sequence sets from phylogenetic and population studies

Proteins

Protein sequences, 3-D structures, and tools for the study of functional protein domains and active sites

Conserved Domains

Conserved protein domains

Identical Protein Groups

Protein sequences grouped by identity

Protein

Protein sequences

Protein Clusters

Sequence similarity-based protein clusters

Spangle

Functional categorization of proteins by domain architecture

Structure

Experimentally-determined biomolecular structures

Genomes

Genome sequence assemblies, large-scale functional genomics data, and source biological samples

Assembly

Genome assembly information

BioCollections

Museum, herbaria, and other biospecimens collections

BioProject

Biological projects providing data to NCBI

BioSample

Descriptions of biological source materials

Genome

Genome sequencing projects by organism

Nucleotide

DNA and RNA sequences

Probe

Sequence-based probes and primers

SRA

High-throughput sequence reads

Taxonomy

Taxonomic classification and nomenclature

Genetics

Heritable DNA variations, associations with human pathologies, and clinical diagnostics and treatments

ClinVar

Human variations of clinical significance

dbGaP

Genotype/phenotype interaction studies

dbSNP

Short genetic variations

dbVar

Genome structural variation studies

GTR

Genetic testing registry

MedGen

Medical genetics literature and links

OMIM

Online mendelian inheritance in man

Chemicals

Repository of chemical information, molecular pathways, and tools for bioactivity screening

BioSystems

Molecular pathways with links to genes, proteins and chemicals

PubChem BioAssay

Bioactivity screening studies

PubChem Compound

Chemical information with structures, information and links

PubChem Substance

Deposited substance and chemical information

PubMed is a service of the U.S. National Library of Medicine that includes over 17 million citations from MEDLINE and other life science journals for biomedical articles back to the 1950s.

PubMed Central is the U.S. National Institutes of Health (NIH) free digital archive of biomedical and life sciences journal literature.

- One of the literature databases of Entrez
- Number of records in PubMed: 17.505.726 (8/11/07, 12:40)
- Search can be done by author, title, journal
- Makes extensive use of XML technology

Functional Inactivation of EBV-Specific T-Lymphocytes in Nasopharyngeal Carcinoma: Implications for Tumor Immunotherapy.

[Li J](#), [Zeng XH](#), [Mo HY](#), [Rolén U](#), [Gao YF](#), [Zhang XS](#), [Chen QY](#), [Zhang L](#), [Zeng MS](#), [Li MZ](#), [Huang WL](#), [Wang XN](#), [Zeng YX](#), [Masucci MG](#).

State Key Laboratory of Oncology in Southern China, Cancer Center, Sun Yat-sen University, Guangzhou, China.

Nasopharyngeal carcinoma (NPC) is an Epstein-Barr virus (EBV) associated malignancy with high prevalence in Southern Chinese. In order to assess whether defects of EBV-specific immunity may contribute to the tumor, the phenotype and function of circulating T-cells and tumor infiltrating lymphocytes (TILs) were investigated in untreated NPC patients. Circulating naïve CD3(+)CD45RA(+) and CD4(+)CD25(-) cells were decreased, while activated CD4(+)CD25(+) T-cells and CD3(-)CD16(+) NK-cells were increased in patients compared to healthy donors. The frequency of T-cells recognizing seven HLA-A2 restricted epitopes in LMP1 and LMP2 was lower in the patients and remained low after stimulation with autologous EBV-carrying cells. TILs expanded in low doses of IL-2 exhibited an increase of CD3(+)CD4(+), CD3(+)CD45RO(+) and CD4(+)CD25(+) cells and 2 to 5 fold higher frequency of LMP1 and LMP2 tetramer positive cells compared to peripheral blood. EBV-specific cytotoxicity could be reactivated from the blood of most patients, whereas the TILs lacked cytotoxic activity and failed to produce IFN γ upon specific stimulation. Thus, EBV-specific rejection responses appear to be functionally inactivated at the tumor site in NPC.

A PubMed XML record

```
<PubmedArticle>
  <MedlineCitation Owner="NLM" Status="In-Data-Review">
    <PMID>17987110</PMID>
    <DateCreated>
      <Year>2007</Year>
      <Month>11</Month>
      <Day>07</Day>
    </DateCreated>
    <Article PubModel="Electronic">
      <Journal>
        <ISSN IssnType="Electronic">1932-6203</ISSN>
        <JournalIssue CitedMedium="Internet">
          <Volume>2</Volume>
          <Issue>11</Issue>
          <PubDate>
            <Year>2007</Year>
          </PubDate>
        </JournalIssue>
        <Title>PLoS ONE</Title>
        <ISOAbbreviation>PLoS ONE</ISOAbbreviation>
      </Journal>
      <ArticleTitle>Functional Inactivation of EBV-Specific T-Lymphocytes in Nasopharyngeal Carcinoma: Implications for Tumor Immunotherapy.</ArticleTitle>
      <Pageination>
        <MedlinePgn>e1122</MedlinePgn>
      </Pageination>
      <Abstract>
        <AbstractText>Nasopharyngeal carcinoma (NPC) is an Epstein-Barr virus (EBV) associated malignancy with high prevalence in Southern Chinese. In order to assess whether defects of EBV-specific immunity may contribute to the tumor, the phenotype and function of circulating T-cells and tumor infiltrating lymphocytes (TILs) were investigated in untreated NPC patients. Circulating naïve CD3(+)/CD45RA(+) and CD4(+)/CD25(-) cells were decreased, while activated CD4(+)/CD25(+) T-cells and CD3(-)/CD16(+) NK-cells were increased in patients compared to healthy donors. The frequency of T-cells recognizing seven HLA-A2 restricted epitopes in LMP1 and LMP2 was lower in the patients and remained low after stimulation with autologous EBV-carrying cells. TILs expanded in low doses of IL-2 exhibited an increase of CD3(+)/CD4(+), CD3(+)/CD45RO(+) and CD4(+)/CD25(+) cells and 2 to 5 fold higher frequency of LMP1 and LMP2 tetramer positive cells compared to peripheral blood. EBV-specific cytotoxicity could be reactivated from the blood of most patients, whereas the TILs lacked cytotoxic activity and failed to produce IFNgamma upon specific stimulation. Thus, EBV-specific rejection responses appear to be functionally inactivated at the tumor site in NPC.</AbstractText>
      </Abstract>
      <Affiliation>State Key Laboratory of Oncology in Southern China, Cancer Center, Sun Yat-sen University, Guangzhou, China.</Affiliation>
      <AuthorList CompleteYN="Y">
        <Author ValidYN="Y">
          <LastName>Li</LastName>
          <ForeName>Jiang</ForeName>
          <Initials>J</Initials>
        </Author>
      </AuthorList>
    </Article>
  </MedlineCitation>
</PubmedArticle>
```