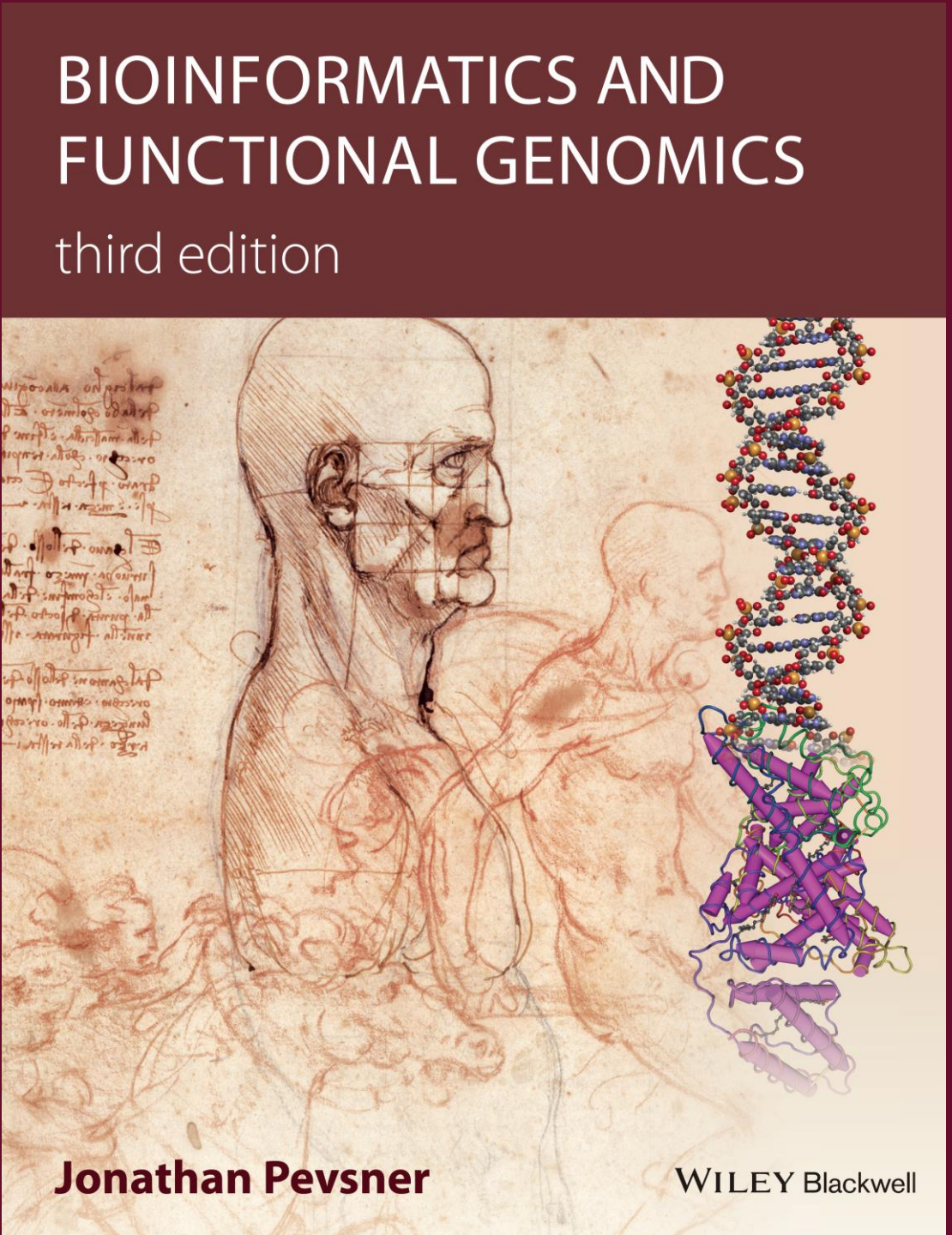
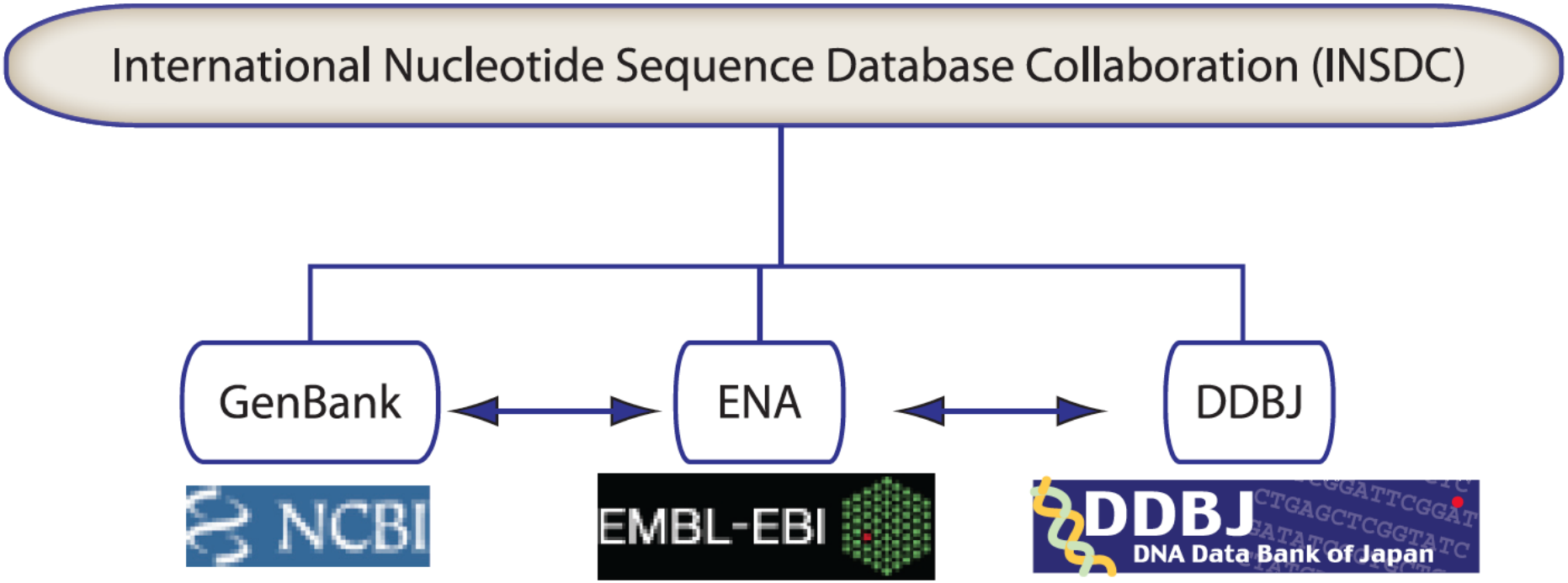


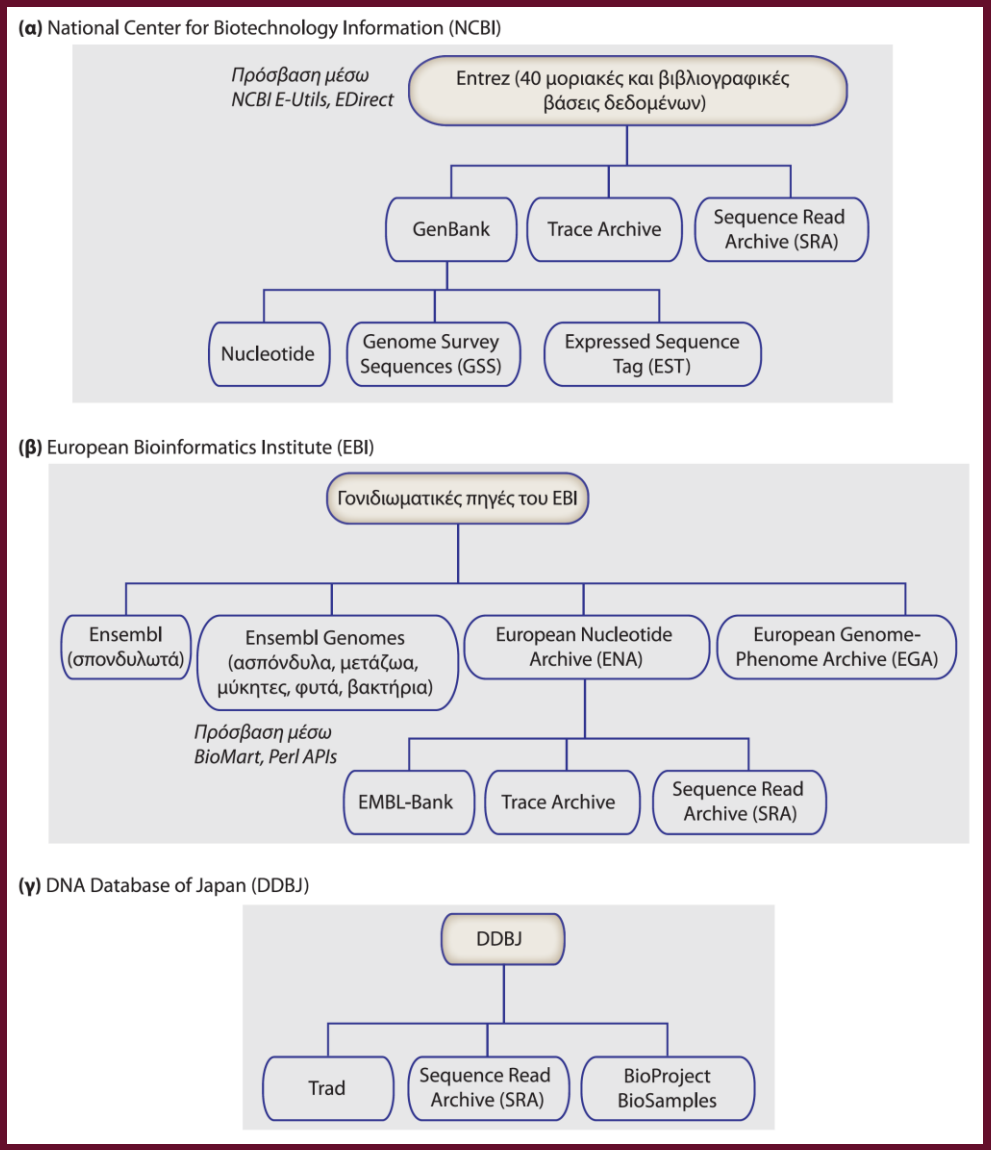
Κεφάλαιο 2

Πρόσβαση σε δεδομένα βιολογικών αλληλουχιών και σχετικές πληροφορίες

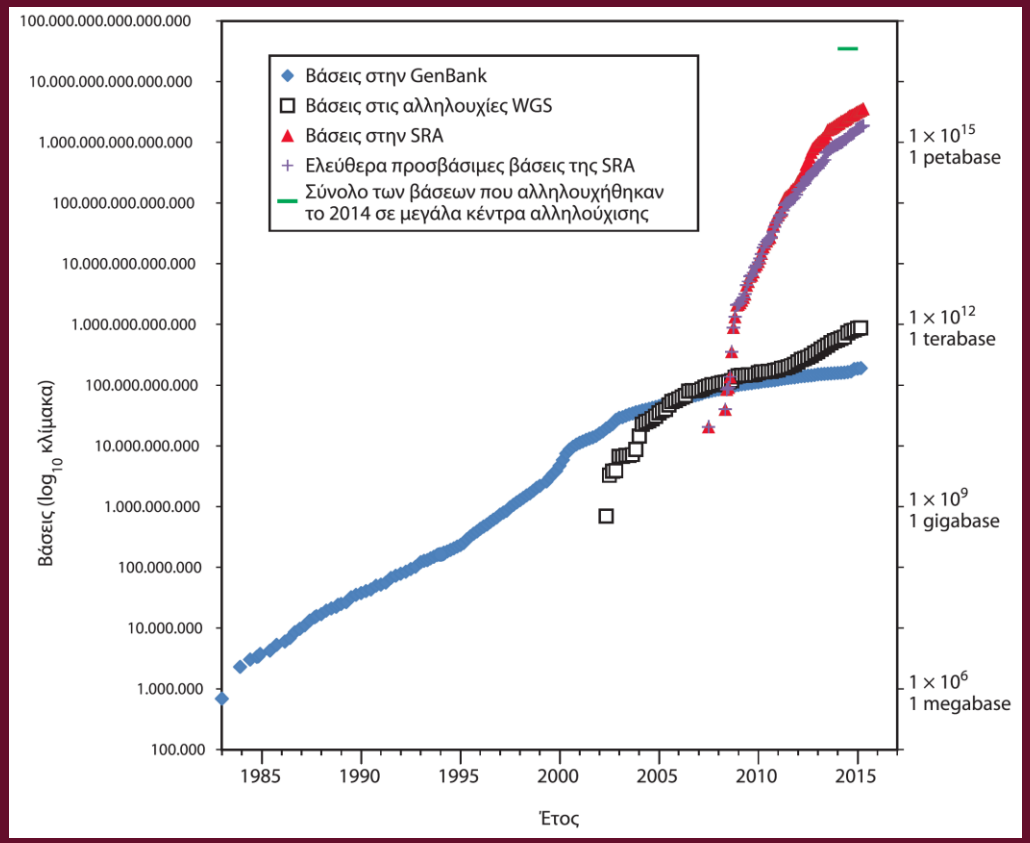
Ακαδημαϊκές Εκδόσεις



Εικόνα 2.1 Οι βάσεις δεδομένων GenBank, EMBL-Bank και DDBJ συντονίζονται από την INSDC.



Εικόνα 2.2 Οι τρεις κύριες βάσεις που αποθηκεύουν δεδομένα νουκλεοτιδικών αλληλουχιών μοιράζονται τις ίδιες αλληλουχίες DNA. (α) Το NCBI στεγάζει την GenBank ως μέρος της Entrez, η οποία περιλαμβάνει 40 βάσεις με μοριακά και βιβλιογραφικά δεδομένα. Στο Trace Archive αποθηκεύονται αλληλουχίες που έχουν διαβαστεί με αλληλούχιση κατά Sanger, ενώ στην SRA αποθηκεύονται δεδομένα αλληλούχισης επόμενης γενιάς. Η GenBank περιλαμβάνει ξεχωριστές ενότητες για αλληλουχίες διερεύνησης γονιδιώ-ματος, για ετικέτες εκφραζόμενης αλληλουχίας και για τις υπόλοιπες νουκλεοτιδικές αλληλουχίες. (β) Στις πηγές του EBI περιλαμβάνονται η Ensembl (που εστιάζεται στο γονιδίωμα των σπονδυλωτών), η Ensembl Genomes (που περιλαμβάνει γονι-διώματα από ένα ευρύ φάσμα ειδών εκτός των σπονδυλωτών), το ENA (European Nucleotide Archive) και το EGA (European Genome-Phenome Archive). Η EMBL-Bank του ENA περιλαμβάνει τα ίδια πρωτογενή δεδομένα αλληλουχιών με την GenBank της Entrez. Παρόμοια δεδομένα περιέχονται και στο Trace Archive και στην SRA. (γ) Και η DDBJ περιλαμβάνει την SRA. Η ενότητα «Trad» της DDBJ μοιράζεται σε καθημερινή βάση τα ίδια πρωτογενή δεδομένα με την GenBank και την EMBL-Bank. Όλες αυτές οι βάσεις δεδομένων είναι προσβάσιμες μέσω του διαδικτύου ή μέσω προγραμμάτων όπως η σουίτα προγραμμάτων EDirect, που επιτρέπει την πρόσβαση μέσω της γραμμής εντολών στις βάσεις δεδομένων της Entrez.



Εικόνα 2.3 Η αύξηση του αριθμού των καταχωρισμένων αλληλουχιών DNA σε αποθετήρια δεδομένων. Τα δεδομένα για την GenBank (μπλε ρόμβοι) περιλαμβάνουν τις εκδόσεις 3 (Δεκέμβριος 1982) έως και 206 (Φεβρουάριος 2015). Με τα λευκά τετράγωνα παρουσιάζονται οι αλληλουχίες WGS, που άρχισαν να κατατίθενται το 2002. Τα δεδομένα για την SRA από το NCBI περιλαμβάνουν το σύνολο των βάσεών της (κόκκινα τρίγωνα) ή μόνο τις βάσεις της που είναι ελεύθερα προσβάσιμες (μοβ σύμβολα +). Τα δεδομένα αυτά βασίζονται σε στοιχεία της GenBank (http://www.ncbi.nlm.nih.gov/Genbank) και της SRA (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?). Ο συνολικός αριθμός των βάσεων DNA που εκτιμάται πως αλληλουχήθηκαν σε μεγάλα κέντρα αλληλούχισης το 2014 υποδεικνύεται με μια πράσινη παύλα (~40 petabases). Αν επιπλέον ληφθεί υπόψιν και ο όγκος των δεδομένων αλληλούχισης που παράγουν οι ιδιωτικές εταιρείες οι οποίες πραγματοποιούν υψηλής απόδοσης αλληλούχιση, η εκτίμηση αυτή θα αυξηθεί σημαντικά. Σύμφωνα με το NCBI, για την SRA οι 3,5 × 10^{15} βάσεις στην τρέχουσα έκδοση (Μάρτιος 2015) αντιστοιχούν σε 2,3 × 10^{15} bytes δεδομένων.

Πίνακας 2.1 Μονάδες μέτρησης για τα ζεύγη βάσεων του DNA.

Ζεύγη βάσεων	Μονάδα	Συντομογραφία	Παράδειγμα
1	1 base pair	1 bp	
1.000	1 kilobase pair	1 kb	Μέγεθος της κωδικής περιοχής ενός τυπικού γονιδίου
1.000.000	1 megabase pair	1 Mb	Μέγεθος ενός τυπικού βακτηριακού γονιδιώματος
10^9	1 gigabase pair	1 Gb	Το ανθρώπινο γονιδίωμα αποτελείται από 3 δισεκατομμύρια ζεύγη βάσεων
10^{12}	1 terabase pair	1 Tb	
10^{15}	1 petabase pair	1 Pb	

Πίνακας 2.2 Εύρος του μεγέθους των αρχείων και τυπικά παραδείγματα.

Μέγεθος	Συντομογραφία	Αριθμός bytes	Παραδείγματα
Bytes	–	1	1 byte είναι 8 bits και είναι ο χώρος αποθήκευσης που καταλαμβάνει στον υπολογιστή ένας χαρακτήρας κειμένου (γράμμα, αριθμός ή σύμβολο)
Kilobytes	1 kb	10^3	Μέγεθος ενός αρχείου κειμένου με μέχρι 1.000 χαρακτήρες
Megabytes	1 Mb	10^6	Μέγεθος ενός αρχείου κειμένου με 1 εκατομμύριο χαρακτήρες
Gigabytes	1 Gb	10^9	600 Gb: Μέγεθος της GenBank (μη συμπιεσμένα αρχεία) ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt (WebLink 2.84)
Terabytes	1 Tb	10^{12}	385 Tb: Διαδικτυακή πύλη της Βιβλιοθήκης του Κογκρέσου των Ηνωμένων Πολιτειών (http://www.loc.gov/webarchiving/faq.html) (WebLink 2.85) 464 Tb: Μέγεθος των δεδομένων που δημιουργήθηκαν από το πρότζεκτ 1.000 Genomes (http://www.1000genomes.org/ faq/how-much-disk-space-used-1000-genomes-project) (WebLink 2.86)
Petabytes	1 PB	10^{15}	1 Pb: Μέγεθος των δεδομένων του πρότζεκτ TCGA (The Cancer Genome Atlas) 5 Pb: Μέγεθος των δεδομένων της SRA του NCBI 15 Pb: Μέγεθος των δεδομένων που παράγονται κάθε χρόνο από το τμήμα φυσικής του CERN (κοντά στη Γενεύη) (http://home.web.cern.ch/about/computing) (WebLink 2.87)
Exabytes	1 Eb	10^{18}	2,5 Eb: Μέγεθος των δεδομένων που παράγονται κάθε μέρα παγκοσμίως (Lampitt, 2014)

Πίνακας 2.3 Ταξινομικές μονάδες στην GenBank.

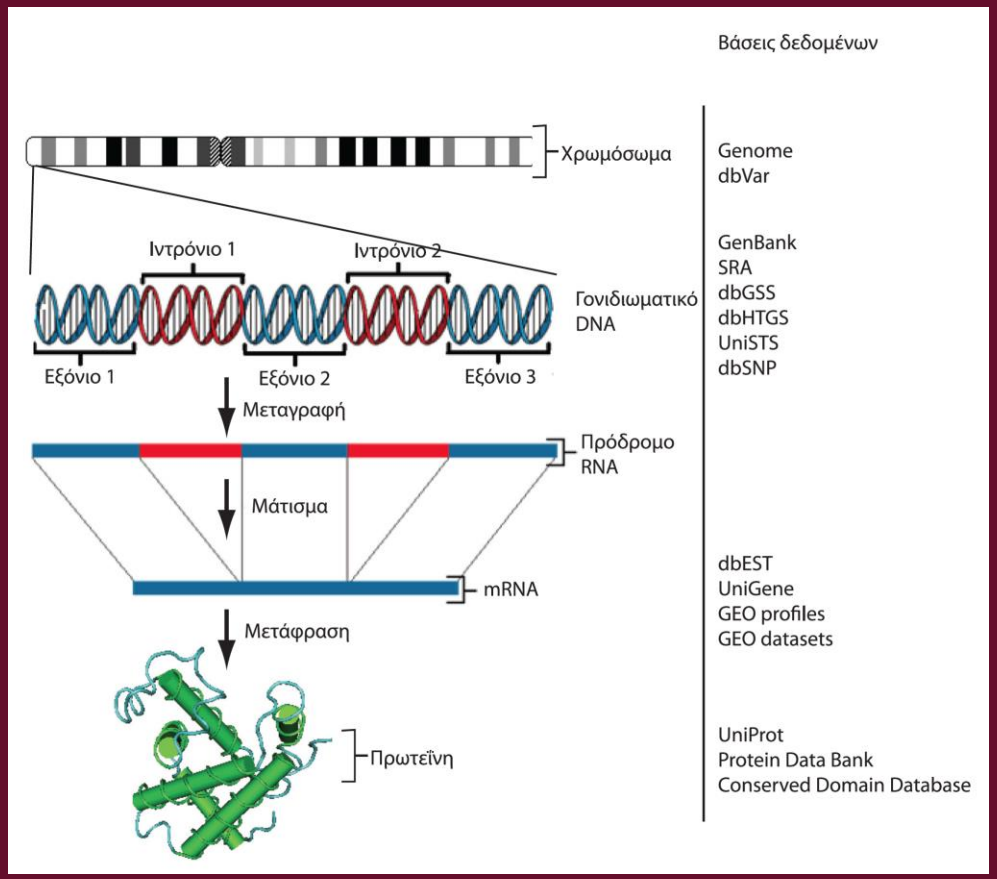
Κατηγορίες	Ανώτερες ταξινομικές μονάδες	Γένη	Είδη	Κατώτερες ταξινομικές μονάδες	Σύνολο
Αρχαία	143	140	525	0	808
Βακτήρια	1.370	2.611	13.331	819	18.131
Ευκαρυώτες	20.443	67.606	297.207	22.608	407.864
Μύκητες	1.550	4.620	29.450	1.128	36.748
Μετάζωα	14.670	45.517	145.044	11.428	216.659
Φύκη	2.622	14.680	113.529	9.789	140.620
Ιοί	618	442	2.349	0	3.409
Όλες οι ταξινομικές μονάδες	22.603	70.806	313.443	23.427	430.279

Πηγή: GenBank, NCBI, http://www.ncbi.nlm.nih.gov/Taxonomy/txstat.cgi.

Πίνακας 2.4 Οι 10 πιο αλληλουχημένοι οργανισμοί στην GenBank.

Αριθμός καταχωρίσεων	Αριθμός νουκλεοτιδίων	Είδος	Κοινή ονομασία
20.614.460	17.575.474.103	*Homo sapiens*	Άνθρωπος
9.724.856	9.993.232.725	*Mus musculus*	Ποντικός
2.193.460	6.525.559.108	*Rattus norvegicus*	Αρουραίος
2.203.159	5.391.699.711	*Bos taurus*	Αγελάδα
3.967.977	5.079.812.801	*Zea mays*	Καλαμπόκι
3.296.476	4.894.315.374	*Sus scrofa*	Χοίρος
1.727.319	3.128.000.237	*Danio rerio*	Ψάρι ζέβρα
1.796.154	1.925.428.081	*Triticum aestivum*	Σιτάρι
744.380	1.764.995.265	*Solanum lycopersicum*	Ντομάτα
1.332.169	1.617.554.059	*Hordeum vulgare subsp. vulgare*	Κριθάρι

Πηγή: GenBank, NCBI, ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt (GenBank έκδοση 194.0).



Εικόνα 2.4 Είναι δυνατόν να δούμε τους τύπους των δεδομένων που αποθηκεύονται στις διάφορες βάσεις (δεξιά στήλη) από τη σκοπιά του κεντρικού δόγματος της βιολογίας. Σύμφωνα με αυτό, το γονιδιωματικό DNA (που οργανώνεται σε χρωμοσώματα) περιλαμβάνει τα γονίδια που μεταγράφονται σε πρόδρομο αγγελιοφόρο RNA, το οποίο υφίσταται επεξεργασία ώστε να ωριμάσει και να μετατραπεί σε mRNA που κατόπιν μεταφράζεται σε πρωτεΐνη. Η πρωτεϊνική δομή είναι από την καταχώριση 1HBS (βλ. λογισμικό Cn3D, Κεφάλαιο 13). Για να μάθετε περισσότερα σχετικά με αυτές τις βάσεις δεδομένων, αναζητήστε την *αλφαβητική λίστα πηγών* [Resource List (A-Z)] στην αρχική σελίδα του NCBI.

Πίνακας 2.5 Οι δέκα οργανισμοί για τους οποίους έχουν αλληλουχηθεί οι περισσότερες EST. Στις δημόσιες βάσεις δεδομένων βρίσκονται καταχωρισμένες περισσότερες από 41 εκατομμύρια EST, που προέρχονται από πολλές χιλιάδες βιβλιοθήκες κλώνων cDNA οι οποίες έχουν δημιουργηθεί από μία ποικιλία οργανισμών.

Οργανισμός	Κοινή ονομασία	Αριθμός EST
Homo sapiens	Άνθρωπος	8.704.790
Mus musculus + domesticus	Ποντικός	4.853.570
Zea mays	Καλαμπόκι	2.019.137
Sus scrofa	Χοίρος	1.669.337
Bos taurus	Αγελάδα	1.559.495
Arabidopsis thaliana	Αραβίδοψη	1.529.700
Danio rerio	Ψάρι ζέβρα	1.488.275
Glycine max	Σόγια	1.461.722
Triticum aestivum	Σιτάρι	1.286.372
Xenopus (Silurana) tropicalis	Βάτραχος	1.271.480

Πηγή: NCBI, http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html (dbEST έκδοση 130101).

Πίνακας 2.6 Τα 19 φύλα και οι 142 οργανισμοί που εκπροσωπούνται στη UniGene.

Φύλο	Αριθμός ειδών	Παράδειγμα
Chordata	42	*Equus caballus* (horse)
Echinodermata	2	*Strongylocentrotus purpuratus* (purple sea urchin)
Arthropoda	19	*Apis mellifera* (honey bee)
Mollusca	2	*Aplysia californica* (California sea hare)
Annelida	2	*Alvinella pompejana*
Nematoda	2	*Caenorhabditis elegans* (nematode)
Platyhelminthes	3	*Schistosoma mansoni*
Porifera	1	*Amphimedon queenslandica*
Cnidaria	3	*Nematostella vectensis* (starlet sea anemone)
Ascomycota	5	*Neurospora crassa*
Basidiomycota	1	*Filobasidiella neoformans*
Codonosigidae	1	*Monosiga ovata*
Streptophyta	50	*Zea mays* (maize)
Chlorophyta	2	*Chlamydomonas reinhardtii*
Apicomplexa	1	*Toxoplasma gondii*
Bacillariophyta	1	*Phaeodactylum tricornutum*
Oomycetes	2	*Phytophthora infestans* (potato late blight agent)
Dictyosteliida	1	*Dictyostelium discoideum* (slime mold)
Ciliophora	2	Paramecium tetraurelia

Πηγή: UniGene, NCBI (Απρίλιος 2013).

(α)

UGID:914190 UniGene Hs.523443 *Homo sapiens* (human) HBB Order cDNA clone, Links

Hemoglobin, beta (HBB)

Human protein-coding gene HBB. Represented by 2363 ESTs from 234 cDNA libraries. Corresponds to reference sequence NM_000518.4. [UniGene 914190 - Hs.523443]

SELECTED PROTEIN SIMILARITIES

Comparison of cluster transcripts with RefSeq proteins. The alignments can suggest function of the cluster.

	Best Hits and Hits from model organisms	Species	Id(%)	Len(aa)
XP_508242.1	PREDICTED: hemoglobin subunit beta isoform 2	*P. troglodytes*	100.0	146
NP_000509.1	HBB gene product	*H. sapiens*	100.0	146
NP_001188320.1	hemoglobin subunit beta-1-like	*M. musculus*	83.7	146
NP_001091375.1	uncharacterized protein LOC100037217	*X. laevis*	61.9	146
NP_571095.1	ba1 gene product	*D. rerio*	52.7	147
	Other hits (2 of 21) [Show all]	Species	Id(%)	Len(aa)
NP_001157900.1	HBB gene product	*M. mulatta*	95.9	146
NP_001162318.1	HBB gene product	*P. anubis*	95.2	146

GENE EXPRESSION

Tissues and development stages from this gene's sequences survey gene expression. Links to other NCBI expression resources.

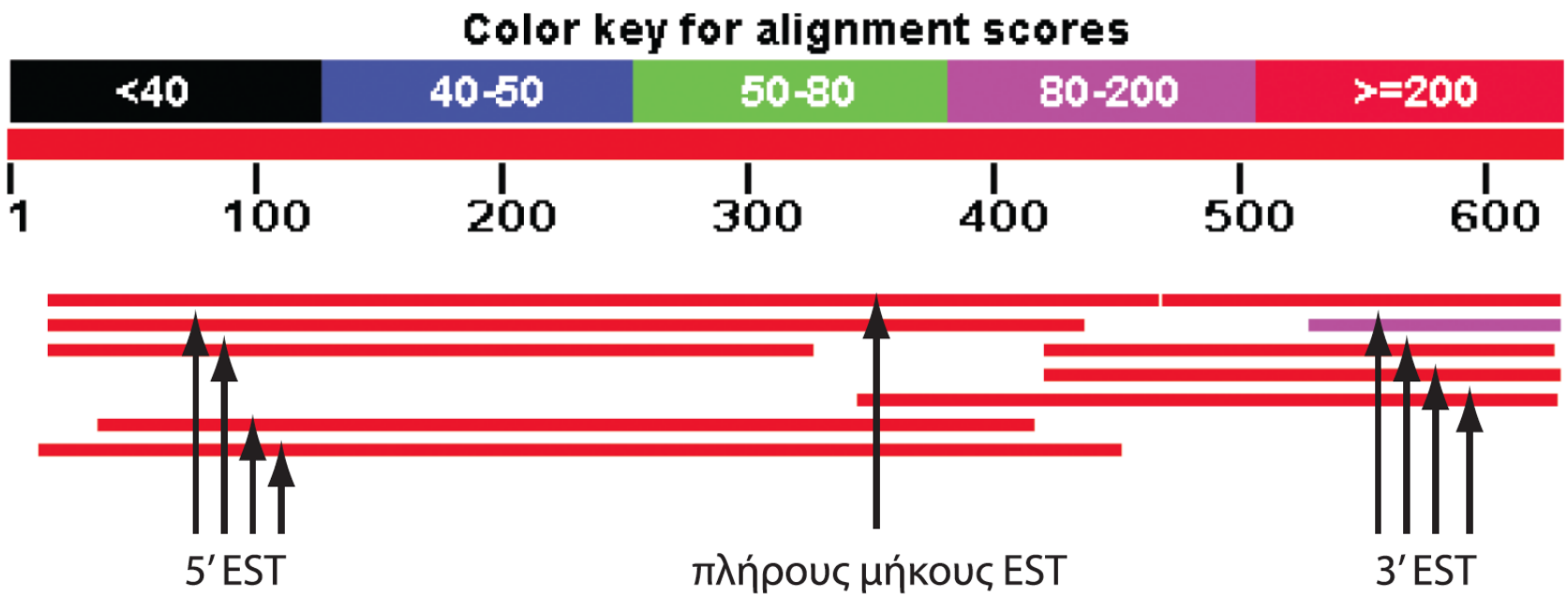
EST Profile: Approximate expression patterns inferred from EST sources. [Show more entries with profiles like this]

GEO Profiles: Experimental gene expression data (Gene Expression Omnibus).

cDNA Sources: blood; mixed; muscle; placenta; bone marrow; lung; brain; spleen; pancreas; connective tissue; pharynx; eye; ovary; uterus; liver; bone; heart; prostate; mammary gland; kidney; uncharacterized tissue; skin; adipose tissue; intestine; stomach; umbilical cord; adrenal gland; nerve; vascular; thymus; testis; embryonic tissue; pituitary gland; parathyroid; ganglia; thyroid; lymph node; pineal gland; ear

Εικόνα 2.5 Η βάση δεδομένων UniGene περιλαμβάνει ομάδες ετικετών εκφραζόμενης αλλη-λουχίας (EST) από τον άνθρω-πο και από μια μεγάλη ποικιλία άλλων ευκαρυωτών. (α) Στην καταχώριση της UniGene για την ανθρώπινη HBB (β-σφαι-ρίνη) αναφέρεται ότι 2.363 EST του γονιδίου έχουν ταυτο-ποιηθεί από 234 διαφορετικές βιβλιοθήκες κλώνων cDNA. Η UniGene οργανώνει τις αλλη-λουχίες (δηλαδή τις EST) σε ομάδες με βάση συγκεκριμένους βαθμούς ομοιότητας σε επίπεδο πρωτεΐνης. Επίσης, συνοψίζει τα δεδομένα για τη χωροχρο-νική έκφραση των γονιδίων.

(β)



Εικόνα 2.5 (β) Οι EST χαρτογραφούνται σε ένα συγκεκριμένο γονίδιο και στοιχίζονται με την αλληλουχία του και μεταξύ τους. Ο αριθμός των EST που αποτελούν μια ομάδα της UniGene κυμαίνεται από 1 έως και πάνω από 1.000. Κατά μέσο όρο υπάρχουν 100 EST ανά ομάδα. Μερικές φορές (αν και δε θα έπρεπε), ξεχωριστές ομάδες της UniGene αντιστοιχούν σε διαφορετικές περιοχές ενός γονιδίου (ιδιαίτερα όταν πρόκειται για μεγάλα γονίδια). Εδώ η αλληλουχία του mRNA του ανθρώπινου *HBB* (NM_000518.4) στοιχίστηκε στο BLAST (Κεφάλαιο 4) με εννέα EST που επιλέχθηκαν από τις >2.000 EST του *HBB*. Τέσσερις από αυτές προέρχονται από την 5΄ και τέσσερις από την 3΄ περιοχή του γονιδίου, ενώ μία καλύπτει το πλήρες μήκος του. Οι αριθμοί καταχώρισης είναι AA985606.1, AA910627.1, AI089557.1, AI150946.1, R25417.1, R27238.1, R27242.1, R27252.1, R31622.1, R32259.1.

Search for Homo sapiens as complete name lock Go Clear

Display 0 levels using filter: none

Homo sapiens

Taxonomy ID: 9606
Genbank common name: **human**
Inherited blast name: **primates**
Rank: species
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 2 (Vertebrate Mitochondrial)
Other names:

common name: **man**

authority: **Homo sapiens Linnaeus, 1758**

Lineage(full)

cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Euarchontoglires; Primates; Haplorrhini; Simiiformes; Catarrhini; Hominoidea; Hominidae; Homininae; Homo

Entrez records

Database name	Subtree links	Direct links
Nucleotide	10,217,570	10,217,541
Nucleotide EST	8,704,803	8,704,803
Nucleotide GSS	1,729,196	1,727,870
Protein	696,378	696,243
Structure	20,041	20,041
Genome	1	1
Popset	22,687	22,687
SNP	63,228,028	63,228,028
Domains	12	12
GEO Datasets	475,213	475,213
UniGene	130,045	130,045
UniSTS	328,844	328,844
PubMed Central	11,154	11,148
Gene	43,470	43,433
HomoloGene	18,473	18,473
SRA Experiments	53,471	53,469
Probe	24,258,933	24,258,933
Assembly	25	25
Bio Project	13,443	13,442
Bio Sample	812,246	812,243
Bio Systems	2,518	2,518
dbVar	2,517,546	2,517,546
Epigenomics	4,186	4,186
GEO Profiles	27,034,750	27,034,750
Protein Clusters	13	13
Taxonomy	3	1

Εικόνα 2.6 Η καταχώριση για τον Homo sapiens στον περιηγητή ταξινομικής του NCBI εμφανίζει πληροφορίες για την ταξινόμησή του, καθώς και μια ποικιλία διαδικτυακών συνδέσμων που οδηγούν σε αρχεία της Entrez. Μέσω των συνδέσμων αυτών μπορείτε να μεταβείτε σε έναν κατάλογο πρωτεϊνών, γονιδίων, αλληλουχιών DNA, δομών ή άλλων τύπων δεδομένων για τον συγκεκριμένο οργανισμό. Αυτό μπορεί να αποτελέσει μια χρήσιμη στρατηγική για την εύρεση μιας πρωτεΐνης ή ενός γονιδίου ενός συγκεκριμένου οργανισμού (ενός είδους ή ενός υποείδους), καθώς δεν περιλαμβάνονται δεδομένα από οποιονδήποτε άλλο οργανισμό.

beta globin | Search

About 75,478 search results for "beta globin"

Literature

Books	339	books and reports
MeSH	4	ontology used for PubMed indexing
NLM Catalog	10	books, journals and more in the NLM Collections
PubMed	8,827	scientific & medical abstracts/citations
PubMed Central	18,185	full-text journal articles

Health

ClinVar	163	human variations of clinical significance
dbGaP	1,368	genotype/phenotype interaction studies
GTR	18	genetic testing registry
MedGen	13	medical genetics literature and links
OMIM	119	online mendelian inheritance in man
PubMed Health	21	clinical effectiveness, disease and drug reports

Genomes

Assembly	0	genomic assembly information
BioProject	19	biological projects providing data to NCBI
BioSample	21	descriptions of biological source materials
Clone	32,086	genomic and cDNA clones
dbVar	214	genome structural variation studies
Epigenomics	24	epigenomic studies and display tools
Genome	351	genome sequencing projects by organism
GSS	3	genome survey sequences
Nucleotide	3,276	DNA and RNA sequences
Probe	125	sequence-based probes and primers
SNP	789	short genetic variations
SRA	13	high-throughput DNA and RNA sequence read archive
Taxonomy	0	taxonomic classification and nomenclature catalog

Genes

EST	2,042	expressed sequence tag sequences
Gene	113	collected information about gene loci
GEO DataSets	148	functional genomics studies
GEO Profiles	3,828	gene expression and molecular abundance profiles
HomoloGene	4	homologous gene sets for selected organisms
PopSet	59	sequence sets from phylogenetic and population studies
UniGene	41	clusters of expressed transcripts

Proteins

Conserved Domains	8	conserved protein domains
Protein	2,316	protein sequences
Protein Clusters	0	sequence similarity-based protein clusters
Structure	404	experimentally-determined biomolecular structures

Chemicals

BioSystems	283	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	45	bioactivity screening studies
PubChem Compound	0	chemical information with structures, information and links
PubChem Substance	186	deposited substance and chemical information

Εικόνα 2.7 Η μηχανή αναζήτησης της Entrez (που είναι προσβάσιμη από τον διαδικτυακό τόπο του NCBI) παρέχει αποτελέσματα από 40 διαφορετικές βάσεις δεδομένων του NCBI. Για πολλά γονίδια και πρωτεΐνες υπάρχουν χιλιάδες αριθμοί καταχώρισης. Το πρότζεκτ RefSeq είναι ιδιαίτερα σημαντικό στην προσπάθεια προσδιορισμού της καλύτερης αντιπροσωπευτικής αλληλουχίας τόσο των φυσιολογικών (μη μεταλλαγμένων) μεταγράφων που παράγονται από ένα γονίδιο όσο και των αντίστοιχων πρωτεϊνών.

Πίνακας 2.7 Μορφή των αριθμών καταχώρισης στη RefSeq. Υπάρχουν 22 διαφορετικοί τύποι αριθμών καταχώρισης για τις αλληλουχίες αναφοράς (RefSeq). Προσαρμοσμένος από http://www.ncbi.nlm.nih.gov/refseq/about/.

Μόριο	Μορφή καταχώρισης	Περιγραφή
Πλήρες γονιδίωμα	NC_123456	Πλήρη γονιδιωματικά μόρια (χρωμοσώματα, πλασμίδια) πυρηνικών και οργανιδιακών γονιδιωμάτων
Γονιδιωματικό DNA	NW_123456 or NW_123456789	Μη πλήρη γονιδιωματικά συναρμολογήματα (assemblies)
Γονιδιωματικό DNA	NZ_ABCD12345678	Δεδομένα αλληλούχισης τυχαίας προσπέλασης ολικού γονιδιώματος
Γονιδιωματικό DNA	NT_123456	Μη πλήρη γονιδιωματικά συναρμολογήματα (δεδομένα από αλληλούχιση τμημάτων κλωνοποιημένων σε τεχνητά βακτηριακά χρωμοσώματα και/ή αλληλούχιση τυχαίας προσπέλασης ολικού γονιδιώματος)
mRNA	NM_123456 or NM_123456789	Μετάγραφα που κωδικοποιούν πρωτεΐνες (mRNA)
Πρωτεΐνη	NP_123456 or NM_123456789	Πρωτεϊνικά μόρια (κυρίως πλήρους μήκους)
RNA	NR_123456	Μη κωδικά μετάγραφα (π.χ. δομικά RNA, μεταγραφόμενα ψευδογονίδια)

Πίνακας 2.8 Αριθμοί καταχώρισης των RefSeq που αφορούν την ανθρώπινη β-σφαιρίνη. Προσαρμοσμένος από http://www.ncbi.nlm.nih.gov/refseq/about/.

Μόριο	Καταχώριση	Μέγεθος	Περιγραφή
DNA	NC_000011.9	135. 006.516 bp	Γονιδιωματικό συναρμολόγημα (χρωμόσωμα 11)
DNA	NG_000007.3	81.706 bp	Γονιδιωματικός τόπος
DNA	NM_000518.4	626 bp	mRNA
Πρωτεΐνη	NP_000509.1	147 αμινοξέα	Πρωτεΐνη

NCBI Resources How To pevsner My NCBI Sign Out

Gene | Gene | beta globin | Search

Save search Advanced Help

Show additional filters

Gene sources
Genomic

Categories
Alternatively spliced
NEWENTRY
Protein-coding
Pseudogene

Sequence content
CCDS
Ensembl
RefSeq
RefSeqGene

Status
Current only

Chromoso... locations
Select ...

Clear all

Show additional filters

Display Settings: Tabular, 20 per page, Sorted by Relevance

Send to:

Results: 1 to 20 of 113

<< First < Prev Page 1 of 6 Next > Last >>

Name/Gene ID	Description	Location	Aliases
HBB ID: 3043	hemoglobin, beta [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (5225466..5227071, complement)	CD113t-C, **beta-globin**
hbg1 ID: 394453	hemoglobin, gamma A [*Xenopus (Silurana) tropicalis* (western clawed frog)]	NW_004668244.1 (60116737..60118249)	**beta-globin**, hbb1, hbga, hbgr, hsggl1
hbg1 ID: 734881	hemoglobin, gamma A [*Xenopus laevis* (African clawed frog)]		**beta-globin**, hbb1, hbga, hbgr, hsggl1
Hbb-bh1 ID: 15132	hemoglobin Z, beta-like embryonic chain [*Mus musculus* (house mouse)]	Chromosome 7, NC_000073.6 (103841638..103843162, complement)	betaH1
HBG2 ID: 396485	hemoglobin, gamma G [*Gallus gallus* (chicken)]	Chromosome 1, NC_006088.3 (193724299..193725801)	HBB, HBD, HBE1

Filters: Manage Filters

Top Organisms [Tree]
Homo sapiens (39)
Mus musculus (27)
Rattus norvegicus (6)
Danio rerio (6)
Bos taurus (5)
All other taxa (30)
More...

Find related data
Database: Select
Find items

Search details
beta globin[All Fields]
Search
See more...

Recent activity

Εικόνα 2.8 Αποτελέσματα της αναζήτησης με τον όρο «beta globin» στη βάση δεδομένων Gene του NCBI. Παρέχονται πληροφορίες για ποικίλους οργανισμούς, όπως ο *Homo sapiens*, ο *Mus musculus* και διάφορα είδη βατράχων. Οι διαδικτυακοί σύνδεσμοι πα-ρέχουν πρόσβαση σε πληροφορίες για τη β-σφαιρίνη από διάφορες άλλες βάσεις δεδομένων.

NCBI Resources How To pevsner My NCBI Sign Out

Gene Gene Search

Limits Advanced Help

Display Settings: Full Report Send to:

HBB hemoglobin, beta [*Homo sapiens* (human)]

Gene ID: 3043, updated on 16-Apr-2013

Summary

Official Symbol	HBB provided by HGNC
Official Full Name	hemoglobin, beta provided by HGNC
Primary source	HGNC:4827
See related	Ensembl:ENSG00000244734; HPRD:00786; MIM:141900; Vega:OTTHUMG00000066678
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	CD113t-C; beta-globin
Summary	The alpha (HBA) and beta (HBB) loci determine the structure of the 2 types of polypeptide chains in adult hemoglobin, Hb A. The normal adult hemoglobin tetramer consists of two alpha chains and two beta chains. Mutant beta globin causes sickle cell anemia. Absence of beta chain causes beta-zero-thalassemia. Reduced amounts of detectable beta globin causes beta-plus-thalassemia. The order of the genes in the beta-globin cluster is 5'-epsilon -- gamma-G -- gamma-A -- delta -- beta--3'. [provided by RefSeq, Jul 2008]

Genomic context

Location: 11p15.5 See HBB in Epigenomics, MapViewer

Sequence: Chromosome: 11; NC_000011.9 (5246696..5248301, complement)

Chromosome 11 - NC_000011.9

[5198951 ▶ [5264822 ▶

OR52Z1 OR51V1 HBB HBD HBBP1

Table of contents

Summary
Genomic context
Genomic regions, transcripts, and products
Bibliography
Phenotypes
Interactions
Pathways
General gene information
Markers, Related pseudogene(s), Homology, Gene Ontology
General protein information
Reference sequences
Related sequences
Additional links

Related information

Order cDNA clone
3D structures
BioAssay
BioAssay, by Protein Target
BioProjects
BioSystems
Books
CCDS
ClinVar
Conserved Domains

Εικόνα 2.9 Τμήμα της καταχώρισης για την ανθρώπινη β-σφαιρίνη στη βάση δεδομένων Gene του NCBI. Παρέχονται πληροφορίες σχετικά με τη δομή του γονιδίου και τη χρωμοσωμική του θέση, καθώς και μια περίληψη για την πρωτεΐνη. Επίσης, κάνοντας κλικ στον σύνδεσμο «Reference sequences» από τον πίνακα περιεχομένων (table of contents, επάνω δεξιά), παρέχονται αριθμοί καταχώρισης RefSeq (δε φαίνονται). Στον πίνακα περιεχομένων υπάρχουν επίσης διαδικτυακοί σύνδεσμοι που παρέχουν πρόσβαση σε διάφορες βάσεις δεδομένων, όπως η PubMed, η OMIM, η UniGene (Κεφάλαιο 10), μια βάση με δεδομένα που αφορούν πολυμορφισμούς (dbSNP), η HomoloGene (βάση με πληροφορίες για ομόλογα γονίδια, Κεφάλαιο 12) και η Ensembl (Κεφάλαιο 8).

NCBI Resources How To pevsner My NCBI Sign Out

Protein Protein Search

Limits Advanced Help

Display Settings: GenPept Send to:

hemoglobin subunit beta [Homo sapiens]

NCBI Reference Sequence: NP_000509.1

FASTA Graphics

Go to:

```
LOCUS       NP_000509                147 aa            linear   PRI 17-APR-2013
DEFINITION  hemoglobin subunit beta [Homo sapiens].
ACCESSION   NP_000509
VERSION     NP_000509.1  GI:4504349
DBSOURCE    REFSEQ: accession NM_000518.4
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 147)
  AUTHORS   Lacerra,G., Prezioso,R., Musollino,G., Piluso,G., Mastrullo,L. and
            De Angioletti,M.
  TITLE     Identification and molecular characterization of a novel 55-kb
            deletion recurrent in southern Italy: the Italian (G) gamma((A)
            gammadeltabeta) degrees -thalassemia
  JOURNAL   Eur. J. Haematol. 90 (3), 214-219 (2013)
   PUBMED   23281611
```

Change region shown

Customize view

Analyze this sequence

Run BLAST

Identify Conserved Domains

Highlight Sequence Features

Find in this Sequence

Protein 3D Structure

Human Zeta-2 Beta-2-s Hemoglobin

PDB: 3W4U

Source: Homo sapiens

Method: X-Ray

Diffraction Resolution: 1.95 Å

See all 196 structures...

```
     CDS             1..147
                     /gene="HBB"
                     /gene_synonym="beta-globin; CD113t-C"
                     /coded_by="NM_000518.4:51..494"
                     /db_xref="CCDS:CCDS7753.1"
                     /db_xref="GeneID:3043"
                     /db_xref="HGNC:4827"
                     /db_xref="HPRD:00786"
                     /db_xref="MIM:141900"
ORIGIN
        1 mvhltpeeks avtalwgkvn vdevggealg rllvvypwtq rffesfgdls tpdavmgnpk
       61 vkahgkkvlg afsdglahld nlkgtfatls elhcdklhvd penfrllgnv lvcvlahhfg
      121 keftppvqaa yqkvvagvan alahkyh
//
```

Εικόνα 2.10 Η καταχώριση στη βάση δεδομένων Protein του NCBI για την ανθρώπινη β-σφαιρίνη. Αποτελεί τυπικό παράδειγμα καταχώρισης για μια πρωτεΐνη στην Protein. Άνω τμήμα της εικόνας: Παρέχονται σημαντικές πληροφορίες για την πρωτεΐνη, όπως το μήκος της (147 αμινοξέα), ο αριθμός καταχώρισής της (NP_000509.1), ο οργανισμός από τον οποίο προέρχεται (*H. sapiens*), βιβλιογραφικές αναφορές και σχόλια σχετικά με τη λειτουργία των σφαιρινών. Δεξιά εμφανίζονται διαδικτυακοί σύνδεσμοι που παρέχουν πρόσβαση σε άλλες βάσεις δεδομένων. Στο επάνω αριστερό μέρος της ιστοσελίδας, ο σύνδεσμος επιλογή απεικόνισης (display option) επιτρέπει την προβολή εναλλακτικών μορφών αυτής της καταχώρισης. Μία από αυτές παρέχει την αλληλουχία της πρωτεΐνης σε μορφή FASTA (Εικόνα 2.11). Κάτω τμήμα της εικόνας: Δίνεται η αλληλουχία της πρωτεΐνης (σύμφωνα με τον κώδικα συμβολισμού κάθε αμινοξέος με ένα γράμμα), όμως εδώ δεν είναι σε μορφή FASTA.

NCBI Resources How To

Protein

Protein

Limits Advanced

Display Settings: FASTA

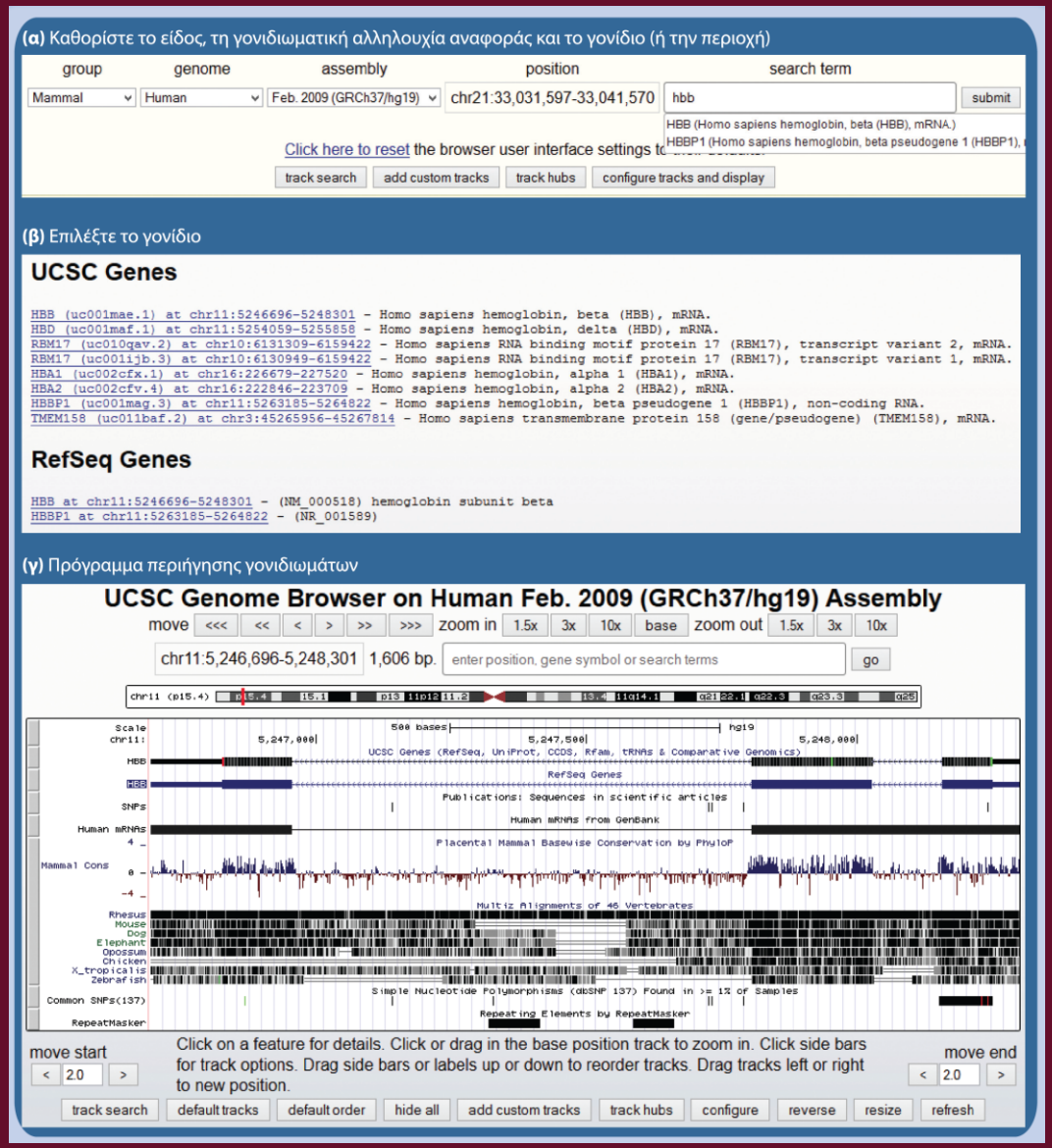
hemoglobin subunit beta [Homo sapiens]

NCBI Reference Sequence: NP_000509.1

GenPept Graphics

```
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
```

Εικόνα 2.11 Μία από τις προβολές των καταχωρίσεων των πρωτεϊνών δείχνει την αλληλουχία τους σε μορφή FASTA. Στη μορφή FASTA η πρώτη σειρά που ξεκινά με το σύμβολο > αποτελεί την επικεφαλίδα και από την επόμενη σειρά ξεκινά η αλληλουχία (είτε μιας πρωτεΐνης, οπότε κάθε αμινοξύ συμβολίζεται με ένα γράμμα, είτε ενός τμήματος DNA, οπότε για τα νουκλεοτίδια χρησιμοποιούνται τα γράμματα G, A, T και C). Η μορφή FASTA χρησιμοποιείται για την εισαγωγή δεδομένων σε διάφορα λογισμικά που θα παρουσιάσουμε όταν θα μιλήσουμε για τη στοίχιση κατά ζεύγη (Κεφάλαιο 3), το BLAST (Κεφάλαιο 4), την αλληλούχιση επόμενης γενιάς (Κεφάλαιο 9) και την πρωτεωμική (Κεφάλαιο 12).



Εικόνα 2.12 Ο περιηγητής γονιδιωμάτων του UCSC. (α) Επιλέξτε ανάμεσα σε δεκάδες οργανισμούς (κυρίως σπονδυ-λωτά) και γονιδιωματικές αλληλουχίες αναφοράς και κατόπιν εισαγάγετε έναν όρο αναζήτησης, όπως «beta globin», ή έναν αριθμό καταχώρισης ή μια χρωμοσωμική θέση. (β) Κάνοντας κλικ στο κουμπί Submit, εμφανίζεται μια λί-στα με γνωστά γονίδια και καταχωρίσεις RefSeq. (γ) Αν κάνετε κλικ στον σύν-δεσμο της RefSeq του HBB, ανοίγει ένα παράθυρο το οποίο εμφανίζει τα 1.606 ζεύγη βάσεων του γονιδίου της β-σφαιρίνης στο ανθρώπινο χρωμόσωμα 11. Εμφανίζονται επίσης διάφορες ενότητες (γραμμές) υπομνηματισμού που περιλαμ-βάνουν καταχωρίσεις RefSeq και προβλε-πόμενα γονίδια από την Ensembl. Τα εξόνια αναπαρίστανται με μαύρα πλαί-σια, ενώ η κατεύθυνση της μεταγραφής υποδεικνύεται με βέλη [από τα δεξιά προς τα αριστερά, με κατεύθυνση προς το άκρο (προς το τελομερές) του κοντού βραχίονα του χρωμοσώματος 11].

Πίνακας 2.9 Αριθμοί καταχώρισης στην Ensembl. Οι ανθρώπινες καταχωρίσεις λαμβάνουν το πρόθεμα ENS. Άλλα προθέματα είναι τα ENSBTA (αγελάδα *Bos taurus*), ENSMUS (ποντικός *Mus musculus*), ENSRNO (αρουραίος *Rattus norvegicus*) και FB (μύγα των φρούτων *Drosophila melanogaster*).

Γράμμα που ακολουθεί το πρόθεμα του είδους	Στοιχείο	Παράδειγμα από την ανθρώπινη β-σφαιρίνη
E	Εξόνιο	ENSE00001829867
FM	Πρωτεϊνική οικογένεια	ENSFM00250000000136
G	Γονίδιο	ENSG00000244734
GT	Γονιδιακό δέντρο	ENSGT00650000093060
P	Πρωτεΐνη	ENSP00000333994
R	Ρυθμιστικό στοιχείο	ENSR00000557622
T	Μετάγραφο	ENST00000335295

Πηγή: Ensembl έκδοση 76, Flicek *et al.* (2014). Αναδημοσιεύεται κατόπιν αδείας της Ensembl.

(α)

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Feb. 2009 (GRCh37/hg19) ← 1

group: Genes and Gene Prediction Tracks **track:** RefSeq Genes add custom tracks track hubs

table: refGene describe table schema

region: ○ genome ○ ENCODE Pilot regions ◉ position chr11:5240001-5300000 lookup define regions ← 2

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

correlation: create

↑ 3

output format: all fields from selected table Send output to ☐ Galaxy ☐ GREAT ← 4

output file: (leave blank to keep output in browser)

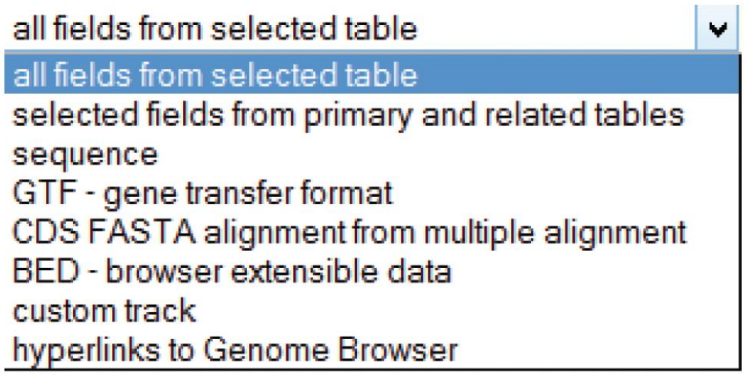
file type returned: ◉ plain text ○ gzip compressed

get output summary/statistics ← 5

To reset **all** user cart settings (including custom tracks), click here.

Εικόνα 2.13 Ο περιηγητής γονιδιωμάτων του UCSC προσφέρει ένα συμπληρωματικό πρόγραμμα περιήγησης πινάκων που είναι εξίσου χρήσιμο. (α) Ο περιηγητής πινάκων περιλαμβάνει επιλογές για να διαλέξετε τον ταξινομικό κλάδο, το γονιδίωμα και τη γονιδιωματική αλληλουχία αναφοράς, π.χ. GRCh37/hg19 (βέλος 1). Είναι δυνατόν να επιλέξετε «group» (π.χ. Genes and Gene Prediction Tracks), μια ενότητα (γραμμή) πληροφοριών (track), για παράδειγμα τη RefSeq, και να επικεντρωθείτε σε μια περιοχή που σας ενδιαφέρει (βέλος 2). Σημειώστε ότι στο πλαίσιο που υποδεικνύεται με το βέλος 3 μπορείτε να εισαγάγετε χρωμοσωμικές συντεταγμένες ή ένα όνομα γονιδίου (π.χ. hbb) και να επιλέξετε «lookup» ώστε να εισαχθούν αυτόματα οι γονιδιωματικές συντεταγμένες του. Στη συνέχεια, πραγματοποιήστε μια επιλογή στο πεδίο «output format» (βέλος 4). Κάντε κλικ στο «summary statistics» (βέλος 5) ώστε να δείτε συνοπτικά πόσα στοιχεία περιλαμβάνονται στα αποτελέσματά σας ή επιλέξτε «get output» για να λάβετε τα πλήρη αποτελέσματα.

(β)



(γ)

```
chr11   5246695 5248301 NM_000518   0   -   5246827 5248251 0   3   261,223,142,   0,1111,1464,
chr11   5254058 5255858 NM_000519   0   -   5254193 5255663 0   3   264,223,287,   0,1162,1513,
chr11   5263184 5264822 NR_001589   0   -   5264822 5264822 0   3   293,223,143,   0,1151,1495,
chr11   5269501 5271087 NM_000559   0   -   5269588 5271034 0   3   216,223,145,   0,1096,1441,
chr11   5274420 5276011 NM_000184   0   -   5274506 5275958 0   3   215,223,145,   0,1101,1446,
chr11   5289579 5291373 NM_005330   0   -   5289698 5291120 0   3   248,223,345,   0,1104,1449,
```

Εικόνα 2.13 (β) Παραδείγματα τύπων μορφοποίησης των αποτελεσμάτων (output format). Αφού κάνετε την επιλογή σας από το πτυσσόμενο μενού, συνήθως ανοίγει μια νέα ιστοσελίδα που προσφέρει πρόσθετες επιλογές [π.χ. η αλληλουχία (sequence) μπορεί να είναι αυτή του DNA ή της πρωτεΐνης, ένα αρχείο BED μπορεί να περιλαμβάνει ένα ολόκληρο γονίδιο ή τα κωδικά εξόνια κ.ά.]. (γ) Παράδειγμα αρχείου BED. Αυτά τα αρχεία είναι ευπροσάρμοστα και είναι δυνατόν να χρησιμοποιηθούν για ποικίλες περαιτέρω αναλύσεις, για παράδειγμα από λογισμικά αλληλούχισης επόμενης γενιάς (βλ. Κεφάλαιο 9).

Πίνακας 2.10 Μορφές αρχείων που υποστηρίζουν η Ensembl και/ή το UCSC. Δίνονται δύο ορισμοί του GtF (από την Ensembl και το UCSC).

Μορφή αρχείου	Ορισμός	Τυπικό μέγεθος αρχείου
BAM		Οποιοδήποτε μέγεθος, συχνά εκατομμύρια σειρές
BED	Browser extensible data	Οποιοδήποτε μέγεθος, συχνά δεκάδες ή χιλιάδες ή και εκατομμύρια σειρές
BedGraph		Οποιοδήποτε μέγεθος
bigBed		
GFF/GTF	General feature format, General transfer format, Gene transfer format	Οποιοδήποτε μέγεθος
MAF		
PSL		Οποιοδήποτε μέγεθος
WIG	Wiggle	Οποιοδήποτε μέγεθος
BAM	Binary alignment/map	Πολύ μεγάλο
BigWig		Πολύ μεγάλο
VCF	Variant call format	Πολύ μεγάλο

Πίνακας 2.11 Μέτρηση που βασίζεται στο 0 ή στο 1.

Πηγή	Σύστημα	Σύνδεσμος
Python	Μέτρηση που βασίζεται στο 0	
Περιηγητής του UCSC σε BED ή άλλη μορφή	Μέτρηση που βασίζεται στο 0	
Δεδομένα του UCSC σε BED ή άλλη μορφή	Μέτρηση που βασίζεται στο 0	
Αρχεία BAM (Κεφάλαιο 9)	Μέτρηση που βασίζεται στο 0	http://samtools.πηγήforge.net/SAM1.pdf (WebLink 2.88)
Ensembl	Μέτρηση που βασίζεται στο 1	http://www.ensembl.org/Help/Faq?id=286 (WebLink 2.89)
Περιηγητής UCSC σε μορφή συντεταγμένων	Μέτρηση που βασίζεται στο 1	http://genome.ucsc.edu/FAQ/FAQtracks.html (WebLink 2.90)
BLAST (Κεφάλαιο 4)	Μέτρηση που βασίζεται στο 1	
Αρχεία GFF (Κεφάλαιο 9)	Μέτρηση που βασίζεται στο 1	
Αρχεία VCF (Κεφάλαιο 9)	Μέτρηση που βασίζεται στο 1	http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20 Format/vcf-variant-call-format-version-41 (WebLink 2.91)

Πηγή: http://alternateallele.blogspot.com/2012/03/genome-coordinate-cheat-sheet.html (WebLink 2.92).