



# BGGN 213

## Foundations of Bioinformatics

Barry Grant  
UC San Diego

<http://thegrantlab.org/bggn213>

# Recap From Last Time:

- Bioinformatics is computer aided biology.
  - ▶ Deals with the collection, archiving, organization, and interpretation of a wide range of biological data.
- There are a large number of **primary**, **secondary** and **tertiary** bioinformatics databases (see [handout!](#)).
- The **NCBI** and **EBI** are major online bioinformatics service providers.
- Introduced **Gene**, **UniProt** and **PDB** databases as well as a number of 'boutique' databases including **PFAM** and **OMIM**.
- Introduced the notion of *controlled vocabularies* and *ontologies* via exploring **GO** annotations.
- Also covered: Course structure; Introductions, Software setup and **Database Vignette...**

## Example Vignette Questions:

- What chromosome location and what genes are in the vicinity of a given query gene? **NCBI GENE**
- What can you find out about molecular functions, biological processes, and prominent cellular locations? **EBI GO**
- What amino acid positions in the protein are responsible for ligand binding? **EBI UniProt**
- What variants of this gene are associated with gastric cancer and other human diseases? **NCBI OMIN**
- Are high resolution protein structures available to examine the details of these mutations? How might we explain their potential molecular effects? **RCSB PDB**
- What is known about the protein family, its species distribution, number in humans and residue-wise conservation? **EBI PFAM**

# TODAYS MENU

- More hands-on exploration of these databases and their associated tools (searching with a propose!)
- Major hands-on sections include:
  1. BLAST, GenBank and OMIM @ **NCBI** [~35 mins]
  2. GENE database @ **NCBI** [~15 mins]  
— BREAK —
  3. UniProt & Muscle @ **EBI** [~25 mins]
  4. PFAM, PDB & NGL [~30 mins]  
— BREAK —
  5. Optional extension exercises [~20 mins]
- ▶ Please do answer the last review question (**Q19**).
- ▶ We encourage discussion and exploration!



# SideNote: **Bioinformatics Databases**

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, Biolmage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAP, BSORF, BTKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, CCDC, CD4OLbase, CGAP, ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG, CyanoBase, dbCFC, dbEST, dbSTS, DDBJ, DGP, DictyDb, Picty\_cDB, DIP, DOGS, DOMO, DPD, DPInteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genlilesne, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB, HAEMB, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD, HIDB, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5 Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-U, MPDB, MRR, MutBase, MycDB, NDB, NRSub, O-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS-MODEL Repository, SWISS-PROT, TeIDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc ..... !!!!

# SideNote: **Bioinformatics Databases**

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCCP, Beanref, BiImage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVM, TKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, CAP, ChickGBASE, Colibri, COPE, CottonDB, bEST, dbSTS, DDBJ, DGP, DictyDb, Pi, CDC, ECGC, EC02DBASE, F, OTHER, FlyBase, F, Link, G, DB, HAEMB, H, vdb, HotMolecBase, K, ZRGbase, IMG, Kabat, KDNA, MHC, OM, Medline, Mendel, MEROPS, MGDB, MGI, MMAP, MJDB, MmtDB, Mol-R-U, MPDB, MRR, MutBase, Myc, O-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS-MODEL Repository, SWISS-PROT, TeIDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc ..... !!!!

**There are lots of Bioinformatics Databases**  
**For a annotated listing of major bioinformatics databases please see the online handout**  
**< [Major Databases.pdf](#) >**

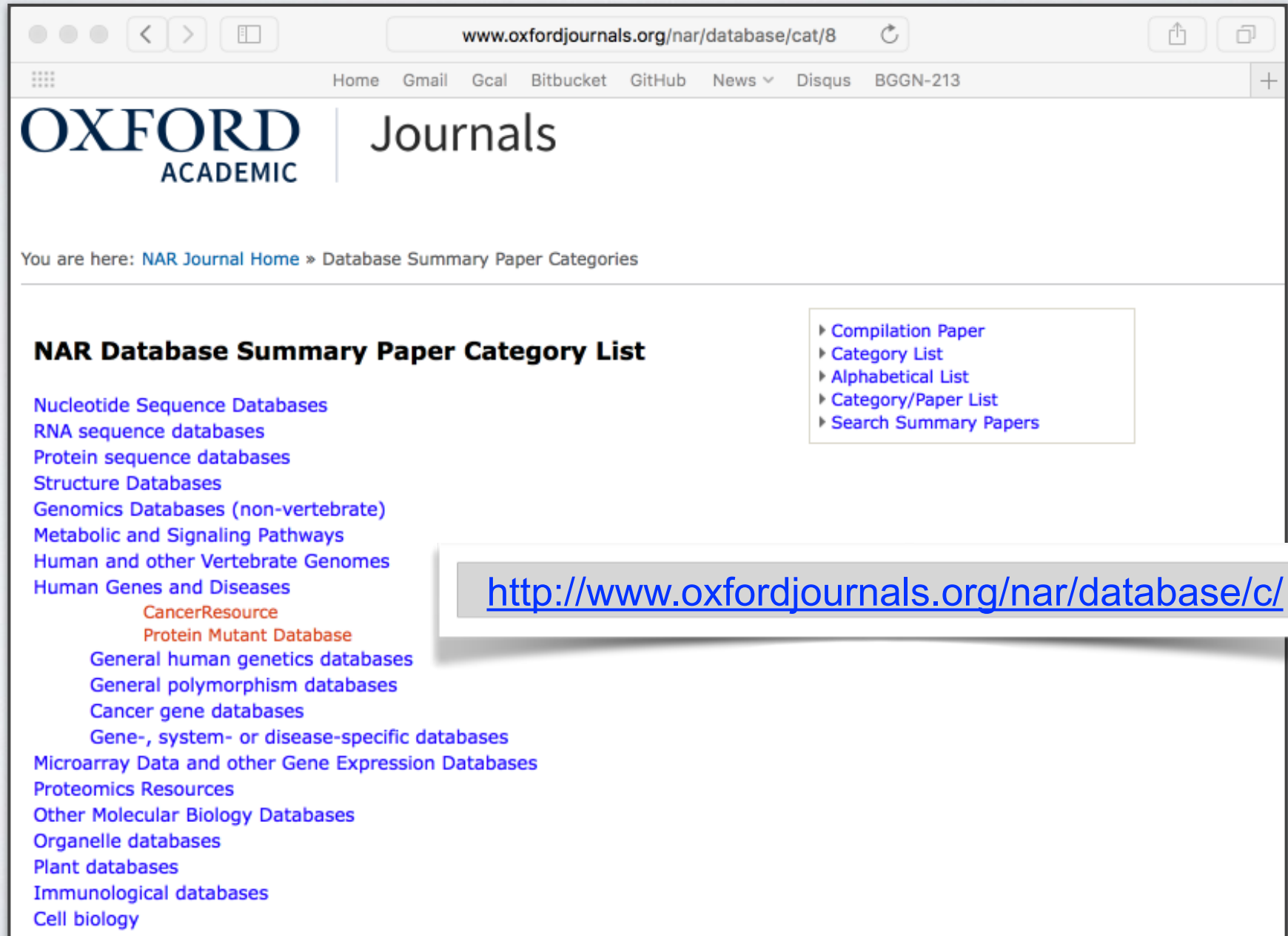
# Side-note: Databases come in all shapes and sizes



Databases can be of variable quality and often there are multiple databases with overlapping content.



# Finding Bioinformatics Databases



The screenshot shows a web browser window with the URL [www.oxfordjournals.org/nar/database/cat/8](http://www.oxfordjournals.org/nar/database/cat/8). The page header includes the Oxford Academic logo and the word "Journals". A breadcrumb trail indicates the current location: "You are here: NAR Journal Home » Database Summary Paper Categories".

## NAR Database Summary Paper Category List

- Nucleotide Sequence Databases
  - RNA sequence databases
  - Protein sequence databases
  - Structure Databases
  - Genomics Databases (non-vertebrate)
  - Metabolic and Signaling Pathways
  - Human and other Vertebrate Genomes
  - Human Genes and Diseases
    - CancerResource
    - Protein Mutant Database
  - General human genetics databases
  - General polymorphism databases
  - Cancer gene databases
  - Gene-, system- or disease-specific databases
- Microarray Data and other Gene Expression Databases
- Proteomics Resources
- Other Molecular Biology Databases
- Organelle databases
- Plant databases
- Immunological databases
- Cell biology

- ▶ [Compilation Paper](#)
- ▶ [Category List](#)
- ▶ [Alphabetical List](#)
- ▶ [Category/Paper List](#)
- ▶ [Search Summary Papers](#)

<http://www.oxfordjournals.org/nar/database/c/>



**GENBANK & REFSEQ:**  
NCBI'S NUCLEOTIDE SEQUENCE  
DATABASES

# What is GenBank?

- GenBank is NCBI's primary **nucleotide only** sequence database
  - ▶ Archival in nature - reflects the state of knowledge at time of submission
  - ▶ Subjective - reflects the submitter point of view
  - ▶ Redundant - can have many copies of the same nucleotide sequence
  - ▶ GenBank is actually three collaborating international databases from Europe, US and Japan



# GenBank sequence record

Homo sapiens kinesin family member 5A (KIF5A), mRNA - Nucleotide - NCBI

www.ncbi.nlm.nih.gov/nucore/NM\_004984.2

NCBI Resources How To Sign in to NCBI

Nucleotide (KIF5A) AND "Homo sapiens" Search

Display Settings: GenBank Send:

## Homo sapiens kinesin family member 5A (KIF5A), mRNA

NCBI Reference Sequence: NM\_004984.2

[FASTA](#) [Graphics](#)

Go to:

LOCUS NM\_004984 3897 bp mRNA linear PRI 10-JAN-2014

DEFINITION Homo sapiens kinesin family member 5A (KIF5A), mRNA.

**ACCESSION NM\_004984**

VERSION NM\_004984.2 GI:45446748

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 3897)

AUTHORS Kawaguchi,K.

TITLE Role of kinesin-1 in the pathogenesis of SPG10. a rare form of her

JOURNAL Neu

PUBMED 227

REMARK Gen

spa

Rev

REFERENCE 2

AUTHORS Pro

Boh

TITLE alpha-Synuclein oligomers impair neuronal microtubule-kinesin

interplay

JOURNAL J. Biol. Chem. 288 (30), 21742-21754 (2013)

PUBMED 23744071

Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

Articles about the KIF5A gene

[alpha-Synuclein oligomers impair neuronal microtubule-kinesin interplay \[J Biol Chem. 2013\]](#)

Peptide hormone metabolism

MHC class II antigen presentation

**GenBank flat file format** has defined fields including unique identifiers such as the **ACCESSION** number.

This same general format is used for other sequence database records too.



# Side node: **Database accession numbers**

Database **accession numbers** are strings of letters and numbers used as **identifying labels** for sequences and other data within databases

▶ Examples (all for retinol-binding protein, RBP4):

X02775 NT_030059	GenBank genomic DNA sequence Genomic contig	DNA
N91759.1 NM_006744	An expressed sequence tag (1 of 170) RefSeq DNA sequence (from a transcript)	RNA
NP_007635 AAC02945 Q28369 1KT7	RefSeq protein GenBank protein UniProtKB/SwissProt protein Protein Data Bank structure record	Protein
PMID: 12205585	PubMed IDs identify articles at NCBI/NIH	Literature

# GenBank sequence record

Homo sapiens kinesin family member 5A (KIF5A), mRNA - Nucleotide - NCBI

www.ncbi.nlm.nih.gov/nuccore/NM\_004984.2

Homo sapiens kinesin family member 5A (KIF5A), mRNA - Nucleotide - NCBI

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide (KIF5A) AND "Homo sapiens" Search Limits Advanced Help

Display Settings: GenBank Send:

## Homo sapiens kinesin family member 5A (KIF5A), mRNA

NCBI Reference Sequence: NM\_004984.2

[FASTA](#) [Graphics](#)

Go to:

LOCUS NM\_004984 3897 bp mRNA linear PRI 10-JAN-2014

DEFINITION Homo sapiens kinesin family member 5A (KIF5A), mRNA.

ACCESSION NM\_004984

VERSION NM\_004984.2 GI:45446748

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 3897)

AUTHORS Kawaguchi,K.

TITLE Role of kinesin-1 in the pathogenesis of SPG10, a rare form of hereditary spastic paraplegia

JOURNAL Neuroscientist 19 (4), 336-344 (2013)

PUBMED [22785106](#)

REMARK GeneRIF: A review of the mechanism of pathogenesis involved in spastic paraplegia type 10 when KIF5A is inactivated by mutations. Review article

REFERENCE 2 (bases 1 to 3897)

AUTHORS Prots,I., Veber,V., Brey,S., Campioni,S., Buder,K., Riek,R., Bohm,K.J. and Winner,B.

TITLE alpha-Synuclein oligomers impair neuronal microtubule-kinesin interplay

JOURNAL J. Biol. Chem. 288 (30), 21742-21754 (2013)

PUBMED [23744071](#)

Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

Articles about the KIF5A gene

[alpha-Synuclein oligomers impair neuronal microtubule-kinesin interplay \[J Biol Chem. 2013\]](#)

[Molecular motor KIF5A is essential for GABA\(A\) receptor transport, a \[Neuron. 2012\]](#)

[Systems-wide analysis of ubiquitylation dynamics reveals a key role \[Nat Cell Biol. 2012\]](#)

See all...

Pathways for the KIF5A gene

Peptide hormone metabolism

MHC class II antigen presentation

# GenBank sequence record

Homo sapiens kinesin family member 5A (KIF5A), mRNA - Nucleotide - NCBI

www.ncbi.nlm.nih.gov/nuccore/NM\_004984.2

NCBI Resources How To Sign in to NCBI

Nucleotide (KIF5A) AND "Homo sapiens" Search

Display Settings: GenBank Send: Change region shown

## Homo sapiens kinesin family member 5A (KIF5A), mRNA

NCBI Reference Sequence: NM\_004984.2

**FASTA** Graphics ← Can set different display formats here

Go to:

LOCUS NM\_004984 3897 bp mRNA linear PRI 10-JAN-2014

DEFINITION Homo sapiens kinesin family member 5A (KIF5A), mRNA.

ACCESSION NM\_004984

VERSION NM\_004984.2 GI:45446748

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 3897)

AUTHORS Kawaguchi,K.

TITLE Role of kinesin-1 in the pathogenesis of SPG10, a rare form of hereditary spastic paraplegia

JOURNAL Neuroscientist 19 (4), 336-344 (2013)

PUBMED [22785106](#)

REMARK GeneRIF: A review of the mechanism of pathogenesis involved in spastic paraplegia type 10 when KIF5A is inactivated by mutations. Review article

REFERENCE 2 (bases 1 to 3897)

AUTHORS Prots,I., Veber,V., Brey,S., Campioni,S., Buder,K., Riek,R., Bohm,K.J. and Winner,B.

TITLE alpha-Synuclein oligomers impair neuronal microtubule-kinesin interplay

JOURNAL J. Biol. Chem. 288 (30), 21742-21754 (2013)

PUBMED [23744071](#)

Analyze this sequence

- Run BLAST
- Pick Primers
- Highlight Sequence Features
- Find in this Sequence

Articles about the KIF5A gene

- [α-Synuclein oligomers impair neuronal microtubule-kinesin interplay \[J Biol Chem. 2013\]](#)
- [Molecular motor KIF5A is essential for GABA\(A\) receptor transport, a \[Neuron. 2012\]](#)
- [Systems-wide analysis of ubiquitylation dynamics reveals a key role for KIF5A \[Nat Cell Biol. 2012\]](#)

See all...

Pathways for the KIF5A gene

- Peptide hormone metabolism
- MHC class II antigen presentation



# FASTA sequence record

Homo sapiens kinesin family member 5A (KIF5A), mRNA - Nucleotide - NCBI

www.ncbi.nlm.nih.gov/nucleotide/45446748?report=fasta

NCBI Resources How To Sign in to NCBI

Nucleotide  Search

Limits Advanced Help

Display Settings: FASTA Send:

Change region shown

Customize view

## Homo sapiens kinesin family member 5A (KIF5A), mRNA

NCBI Reference Sequence: NM\_004984.2

[GenBank](#) [Graphics](#)

```
>gi|45446748|ref|NM_004984.2| Homo sapiens kinesin family member 5A (KIF5A), mRNA
ACGCCCAGGTCGCCCGCATCCCGCTGCCGACAGAGAGACAGCGCGCCCGCCCTGCTCCCCAGGCTT
CGCCCCGGCGCCCTCAACTCTGTCCCCAGAGACTGAGCACCTGTCTCCGCTCGGCCTCTGCTGAGAGC
CCTCTCTCTGGAGCACACACCACCCCTGCAGCCCAAGAAGAGTCCAGCCCCACCGGGCTACCACCAT
GGCGGAGACCAACAACGAATGTAGCATCAAGGTGCTCTGCCGATTCGGGCCCTGAACCAGGCTGAGATT
CTGCCGGGAGACAAGTTCATCCCCATTTTCCAAGGGGACGACAGCGCTGTTATTGGGGGGAAGCCATATG
TTTTGACCGTGTATTCCCCCAACAGACTCAAGAGCAAGTTTATCATGCATGTGCCATGCAGATTGT
CAAAGATGTCCTTGCTGGCTACAATGGCACCATTTTTGTCTATGGACAGACATCCTCAGGGAAAACACAT
ACCATGGAGGGAAAGCTGCACGACCCTCAGCTGATGGGAATCATTCTCGAATTGCCGAGACATCTTCA
ACCACATCTACTCCATGGATGAGAACCTTGAGTCCACATCAAGTTCCTTACTTTGAAATTTACCTGGA
CAAAATTCGTGACCTTCTGGATGTGACCAAGACAAATCTGTCCGTGCACGAGGACAGAACCAGGGTGCCA
TTTGTCAAGGGTTGACTGAACGCTTTGTGTCCAGCCCGAGGAGATTCTGGATGTGATTGATGAAGGGA
AATCAAATCGTCATGTGGCTGTCAACACATGAATGAACACAGCTCTCGGAGCCACAGCATCTTCTCAT
CAACATCAAGCAGGAGAACATGGAAACGGAGCAGAAGCTCAGTGGGAAGCTGTATCTGGTGAACCTGGCA
GGGAGTGAGAAGGTGACGAAGACTGGAGCAGAGGGAGCCGTGCTGGACGAGGCAAAGAATATCAACAAGT
CACTGTCAAGTCTGGCAATGTGATCTCCGACAGGCTGAGGGCACTAAAAGCTATGTTCCATATCGTGA
CAGCAAAATGACAAGGATTCACAGACTCTCTCGGGGAAACTGCCGAGACTATGTTTCTCTGTGTC
TCACCATCCAGTTATAATGATGCAGAGACCAAGTCCACCCTGATGTTTGGGACGCGGCAAAGACCATTA
AGAACACTGCCTCAGTAAATTTGGAGTTGACTGCTGAGCAGTGGAAAGAAGAAATATGAGAAGGAGAAGGA
GAAGACAAGGGCCAGAAGGAGACGATGGCAAGCTGGAGGCTGAGCTGAGCCGGTGGCGCAATGGAGAG
AATGTGCTTGAGACAGAGCGCTGGCTGGGGAGGAGGCCCTGGGAGCCGAGCTCTGTGAGGAGACCC
CTGTGAATGACAACCTCATCCATCGTGGTGGCATCGCGCCGAGGAGCGGCAGAAATACGAGGAGGAGAT
CCGCCGTCTCTATAAGCAGCTTGACGACAAGGATGATGAAATCAACCAACAAAGCCAACTCATAGAGAAG
CTCAAGCAGCAAAATGCTGGACAGGAAGAGCTGCTGGTGTCCACCCGAGGAGACAACGAGAAGGTCCAGC
GGGAGCTGAGCCACCTGCAATCAGAGAACGATGCCGCTAAGGATGAGGTGAAGGAAGTGTGACAGGCCCT
GGAGGAGCTGGCTGTGAACTATGACCAGAAGTCCCAGGAGGTGGAGGAGAAGAGCCAGCAGAACCAGCTT
CTGGTGGATGAGCTGTCTCAGAAGGTGGCCACCATTGCTGTCCCTGGAGCTGAGTGTGACAGCGGCTACAGG
AGGTCAGTGGACACCAGGAAAACGAATTGCTGAGGTGCTGAACGGGCTGATGAAGGATCTGAGCGAGTT
```

FASTA sequence files consist of records where each record begins with a “>” and header information on that same line. Each subsequent line of the record is sequence information.

**This format is commonly used by sequence analysis programs.**

Pathways for the KIF5A gene

- Peptide hormone metabolism
- MHC class II antigen presentation

# GenBank 'graphics' sequence record

NM\_004984.2: Homo sapiens kinesin family member 5A (KIF5A), mRNA

www.ncbi.nlm.nih.gov/nucleotide/45446748?report=graph

NM\_004984.2: Homo sapiens kinesin family member 5A (KIF5A), mRNA

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Search Limits Advanced Help

Display Settings: Graphics Send:

## Homo sapiens kinesin family member 5A (KIF5A), mRNA

NCBI Reference Sequence: NM\_004984.2

[GenBank](#) [FASTA](#)

[Link To This Page](#) [Feedback](#)

Genes - Exon

Genes

KIF5A

NP\_004975.2

KIF5A

KIF5A

ATP binding site Ic...

acetyla... Microtubule-binding

microtubule interact...

STS Markers

D12S1889

**Analyze this sequence**

- Run BLAST
- Pick Primers
- Highlight Sequence Features

**Articles about the KIF5A gene**

- $\alpha$ -Synuclein oligomers impair neuronal microtubule-kinesin interp [J Biol Chem. 2013]
- Molecular motor KIF5A is essential for GABA(A) receptor transport, a [Neuron. 2012]
- Systems-wide analysis of ubiquitylation dynamics reveals a key r [Nat Cell Biol. 2012]

[See all...](#)

**Pathways for the KIF5A gene**

- Peptide hormone metabolism
- MHC class II antigen presentation
- Dopaminergic synapse

[See all...](#)

**Reference sequence information**

- RefSeq alternative splicing
- See the other reference mRNA sequence splice variant for the KIF5A gene

# GenBank sequence record, cont.

Homo sapiens kinesin family member 5A (KIF5A), mRNA - Nucleotide - NCBI

www.ncbi.nlm.nih.gov/nucleotide/NM\_004984.2

Homo sapiens kinesin family member 5A (KIF5A), mRNA - Nucleotide - NCBI

NCBI Resources How To Sign in to NCBI

Nucleotide (KIF5A) AND "Homo sapiens" Search

Display Settings: GenBank Send:

## Homo sapiens kinesin family member 5A (KIF5A), mRNA

NCBI Reference Sequence: NM\_004984.2

[FASTA](#) [Graphics](#)

Go to:

LOCUS NM\_004984 3897 bp mRNA linear PRI 10-JAN-2014

DEFINITION Homo sapiens kinesin family member 5A (KIF5A), mRNA.

ACCESSION NM\_004984

VERSION NM\_004984.2 GI:45446748

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 3897)

AUTHORS Kawaguchi,K.

TITLE Role of kinesin-1 in the pathogenesis of SPG10, a rare form of hereditary spastic paraplegia

JOURNAL Neuroscientist 19 (4), 336-344 (2013)

PUBMED [22785106](#)

REMARK GeneRIF: A review of the mechanism of pathogenesis involved in spastic paraplegia type 10 when KIF5A is inactivated by mutations. Review article

REFERENCE 2 (bases 1 to 3897)

AUTHORS Prots,I., Veber,V., Brey,S., Campioni,S., Buder,K., Riek,R., Bohm,K.J. and Winner,B.

TITLE alpha-Synuclein oligomers impair neuronal microtubule-kinesin interplay

JOURNAL J. Biol. Chem. 288 (30), 21742-21754 (2013)

PUBMED [23744071](#)

Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

Articles about the KIF5A gene

[alpha-Synuclein oligomers impair neuronal microtubule-kinesin interplay \[J Biol Chem. 2013\]](#)

[Molecular motor KIF5A is essential for GABA\(A\) receptor transport, a \[Neuron. 2012\]](#)

[Systems-wide analysis of ubiquitylation dynamics reveals a key role \[Nat Cell Biol. 2012\]](#)

See all...

Pathways for the KIF5A gene

[Peptide hormone metabolism](#)

[MHC class II antigen presentation](#)



# GenBank sequence record, cont.

Homo sapiens kinesin family member 5A (KIF5A), mRNA - Nucleotide - NCBI

www.ncbi.nlm.nih.gov/nuccore/45446748?report=genbank&to=3897#feature\_45446748

Homo sapiens kinesin family member 5A (KIF5A), mRNA - Nucleotide - NCBI

FEATURES	Location/Qualifiers	OMIM
<a href="#">source</a>	1..3897 /organism="Homo sapiens" /mol_type="mRNA" /db_xref="taxon:9606" /chromosome="12" /map="12q13.13"	Probe Protein PubMed PubMed (RefSeq)
<a href="#">gene</a>	1..3897 /gene="KIF5A" /gene_synonym="D12S1889; MY050; NKHC; SPG10" /note="kinesin family member 5A" /db_xref="GeneID:3798" /db_xref="HGNC:6323" /db_xref="HPRD:09108" /db_xref="MIM:602821"	
<a href="#">exon</a>	1..337 /gene="KIF5A" /gene_synonym="D12S1889; MY050; NKHC; SPG10" /inference="alignment:Splign:1.39.8"	
<a href="#">misc_feature</a>	134..136 /gene="KIF5A" /gene_synonym="D12S1889; MY050; NKHC; SPG10" /note="upstream in-frame stop codon"	
<a href="#">CDS</a>	209..3307 /gene="KIF5A" /gene_synonym="D12S1889; MY050; NKHC; SPG10" /note="kinesin, heavy chain, neuron-specific; KIF5A variant protein; neuronal kinesin heavy chain; kinesin heavy chain neuron-specific 1" /codon_start=1 /product="kinesin heavy chain isoform 5A" /protein_id="NP_004975.2" /db_xref="GI:45446749" /db_xref="CCDS:CCDS8945.1" /db_xref="GeneID:3798" /db_xref="HGNC:6323" /db_xref="HPRD:09108" /db_xref="MIM:602821" /translation="MAETNNECSIKVLCRFRPLNQAEILRGDKFIPFQGGDSSVVIIG KPYVFDVRVFPNNTTQEQVYHACAMQIVKDVLAGYNGTIFAYGQTSSGKTHTEGKLDH PQLMGIIPRIARDIPNHIIYSMDENLEFHIKVSYPFEIYLDKIRDLLDVTKTNLSVHEDK NRVPFVKGCTERFVSSPEILDVIDEGKSNRHVAVTNMNEHSSRSHSIFLINIKQENM ETEQLSGKLYLVDLAGSEKVSKTGAEGAVLDEAKNINKSLGALGNVISALAEGTKSY VPYRDSKMTIRLQDLSLGGNCRRTTMFICCPSSYNDAETKSTLMFGQRAKTIKNTASVN	

The **FEATURES** section contains annotations including a conceptual translation of the nucleotide sequence.

**Recent activity**

- Turn Off Clear
- Homo sapiens kinesin family member 5A (KIF5A), mRNA Nucleotide
- (kinesin) AND "Homo sapiens"[porgn] (1351) Nucleotide
- kinesin (37064) Nucleotide

See more...

# GenBank sequence record, cont.

www.ncbi.nlm.nih.gov/nuccore/45446748?report=genbank&to=3897#sequence\_45446748

Homo sapiens kinesin family member 5A (KIF5A), mRNA - Nucleotide - NCBI

/gene\_synonym="D12S1889; MY050; NKHC; SPG10"  
/standard\_name="D12S1889"  
/db\_xref="UniSTS:48006"

ORIGIN

1	acgccaggt	cgccgcgcatc	ccgctgcccgc	aggagagaga	cagcgcgccc	cggccctgct
61	cccaggtt	cgccggggc	ccctcaactc	tgtcccaga	gactgagcac	ctgtcctccg
121	cctcggcctc	tgtgagagc	cctctcctct	ggagcacaca	ccaccctgc	agcccaagaa
181	gagtccagc	ccaagcccg	ctaccaccat	ggcggagacc	aacaacgaat	gtagcatcaa
241	ggtgctctgc	cgattccggc	ccctgaacca	ggctgagatt	ctcgggggag	acaagttcat
301	ccccatttcc	caaggggagc	acagcgtcgt	tattgggggg	aagccatatg	tttttgaccg
361	tgtattcccc	ccaacacga	ctcaagagca	agtttatcat	gcatgtgcca	tgcagattgt
421	caaagatgct	cttctggct	acaatggcac	catttttct	tatggacaga	catctcagg
481	gaaaacacat	accatggagg	gaaagctgca	cgaccctcag	ctgatgggaa	tcattctcgc
541	aattgcccga	gacatcttca	accacatcta	ctccatggat	gagaaccttg	agttccacat
601	caaggtttct	tactttgaaa	tttacctgga	caaaattcgt	gaccttctgg	atgtgaccaa
661	gacaaatctg	tccgtgcacg	aggacaagaa	ccgggtgcca	tttgtcaagg	gttgtactga
721	acgctttgtg	tccagcccgg	aggagattct	ggatgtgatt	gatgaagggg	aatcaaactg
781	tcatgtggct	gtcaccaaca	tgaatgaaca	cagctctcgg	agccacagca	tcttctcat
841	caacatcaag	caggagaaca	tggaaacgga	gcagaagctc	agtggaagc	tgtatctggt
901	ggacctggca	gggagtgaga	aggtcagcaa	gactggagca	gagggagccg	tgtgtgacga
961	ggcaagaat	atcaacaagt	cactgtcagc	tctgggcaat	gtgatctccg	cactggctga
1021	ggcactaaa	agctatgttc	catatcgtga	cagcaaatg	acaaggattc	tccaggactc
1081	tctcggggga	aactgcccga	cgactatgtt	catctgttgc	tcaccatcca	gttataatga
1141	tgcagagacc	aagtccacc	tgatgtttgg	gcagcgggca	aagaccatta	agaacactgc
1201	ctcagtaaat	ttggagtga	ctgctgagca	gtggaagaag	aaatatgaga	aggagaagga
1261	gaagacaaag	gcccagaagg	agacgattgc	gaagctggag	gctgagctga	gccggtggcg
1321	caatggagag	aatgtgcctg	agacagagcg	cctggctggg	gaggaggcag	ccctgggagc
1381	cgagctctgt	gaggagacc	ctgtgaatga	caactcatcc	atcgtggtgc	gcatcgcgcc
1441	cgaggagcgg	cagaaatcag	aggaggagat	ccgccgtctc	tataagcagc	ttgacgacaa
1501	ggatgatgaa	atcaaccaat	aaagccaact	catagagaag	ctcaagcagc	aatgtctgga
1561	ccaggaagag	ctgctggtct	ccaccggagg	agacaacgag	aaggtccagc	gggagctgag
1621	ccacctgcaa	tcagagaacg	atgccgctaa	ggatgaggtg	aaggaagtgc	tgcaggccct
1681	ggaggagctg	gctgtgaact	atgaccagaa	gtcccaggag	gtggaggaga	agagccagca
1741	gaaccagctt	ctggtggatg	agctgtctca	gaaggtggcc	accatgctgt	ccctggagtc
1801	tgagttgcag	cggctacagg	aggtcagtg	acaccagcga	aaacgaattg	ctgaggtgct
1861	gaacgggctg	atgaaggatc	tgagcagatt	cagtgctatt	gtgggcaacg	gggagattaa
1921	gctgccagtg	gagatcagtg	ggccatcga	ggaggagttc	actgtggccc	gactctacat
1981	cagcaaaatc	aaatcagaag	tcaagtctgt	ggtcaagcgg	tgccggcagc	tggagaacct
2041	ccaggtggag	tgtcaccgca	agatggaagt	gaccggggcg	gagctctcat	cctgccagct
2101	cctcatctct	cagcatgagg	ccaagatccg	ctcgttacg	gaatacatgc	agagcgtgga
2161	gctaaagaag	cggcaccctg	aagagtccca	tgactccttg	agcgatgagc	tggccaagct
2221	ccaggcccag	gaaactgtgc	atgaagtggc	cctgaaggac	aaggagcctg	acactcagga
2281	tgcagatgaa	gtgaagaagg	ctctggagct	gcagatggag	agtcaccggg	aggccatca
2341	ccggcagctg	gcccggctcc	gggacgagat	caacgagaag	cagaagacca	ttgatgagct

The actual sequence entry starts after the word **ORIGIN**

# RefSeq: NCBI's Derivative Sequence Database

- RefSeq entries are hand curated best representation of a transcript or protein (in their judgement)
- Non-redundant for a given species although alternate transcript forms will be included if there is good evidence

- Experimentally verified transcripts and proteins  
accession numbers begin with “NM\_” or “NP\_”
- Model transcripts and proteins based on bioinformatics predictions with little experimental support  
accession numbers begin with “XM\_” or “XP\_”
- RefSeq also contains contigs and chromosome records



**UNIPROT:**

THE PREMIER PROTEIN SEQUENCE  
DATABASE

# UniProt: Protein sequence database

UniProt is a comprehensive, high-quality resource of protein sequence and functional information

- UniProt comprises four databases:

## 1. UniProtKB (Knowledgebase)

Containing **Swiss-Prot** and **TrEMBL** components (these correspond to hand curated and automatically annotated entries respectively)

## 2. UniRef (Reference Clusters)

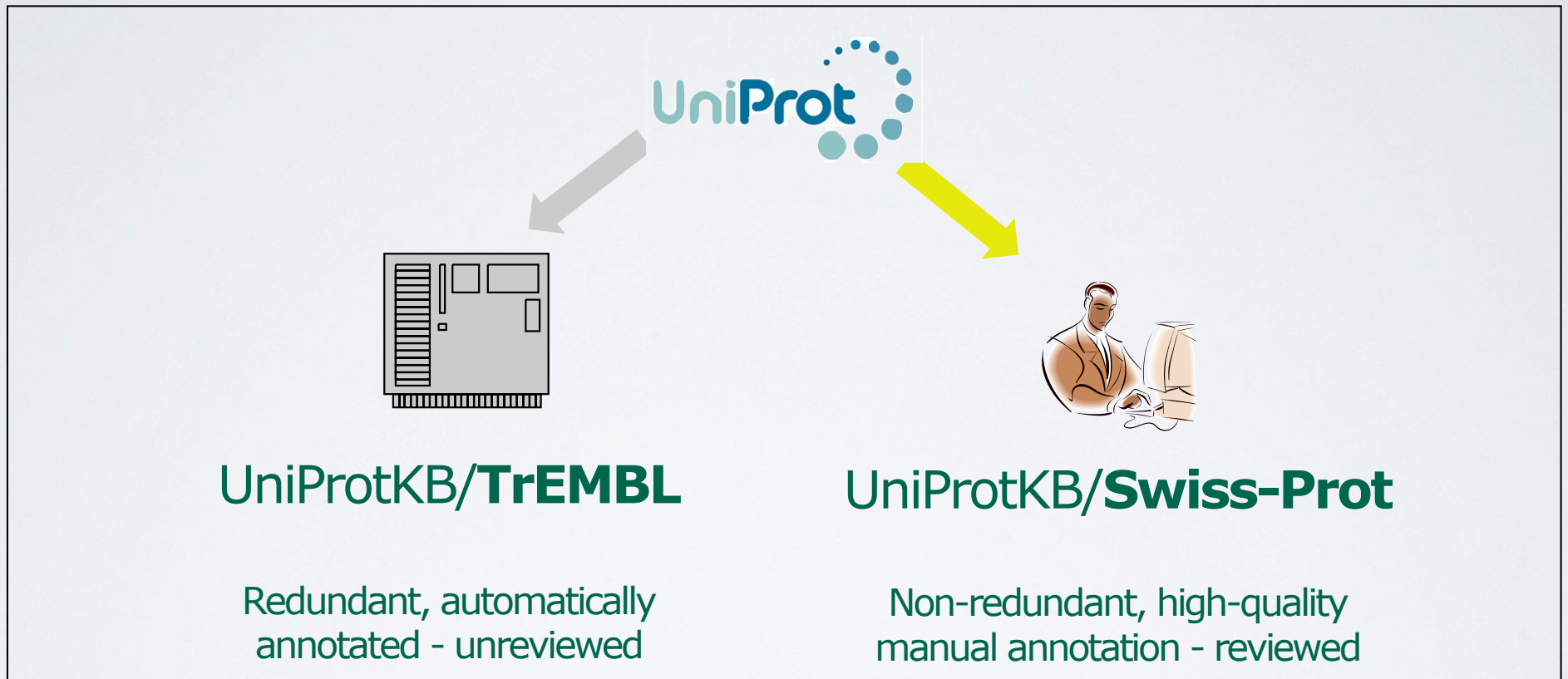
Filtered version of UniProtKB at various levels of sequence identity

e.g. UniRef90 contains sequences with a maximum of 90% sequence identity to each other

## 3. UniParc (Archive) with database cross-references to source.

## 4. UniMES (Metagenomic and Environmental Sequences)

# The two sides of UniProtKB



★ Unreviewed, UniProtKB/TrEMBL **Q9N0H9** (Q9N0H9\_EQUAS)

★ Reviewed, UniProtKB/Swiss-Prot **P38398** (BRCA1\_HUMAN)

Indicators of which part of UniProt an entry belongs to include the color of the stars and the ID



# The main information added to a UniProt/Swiss-Prot entry

- [1] "The quaking gene product necessary in embryogenesis and myelination combines features of RNA binding and signal transduction proteins." Ebersole T.A., Chen Q., Justice M.J., Artzt K. Nat. Genet. 12:260-265(1996) [PubMed: 8589716] [Abstract] Cited for: NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 3), INVOLVEMENT IN QKV, TISSUE SPECIFICITY, MUTAGENESIS OF...
- [2] "Genomic organization and alternative splicing of the quaking gene." Kondo T., Furuta T., Masunaga K., Ebersole T.A., Saitoh M., Wu J., Artzt K., Yamamura K., Abe K. Mamm. Genome 10:662-669(1999) [PubMed: 10384037] [Abstract] Cited for: NUCLEOTIDE SEQUENCE [GENOMIC DNA / MRNA] (ISOFORMS 2; 3; 4 AND 7), ALTERNATIVE SPLICING (ISOFORM 1), Strain: 129/J.

General annotation (Comments)	Title Title
Function	RNA-binding protein that plays a central role in myelination. Also required for visceral endoderm function and blood vessel development. Binds to the 5'-AACUAA(A)N(120)-UAA(A)'S RNA core sequence. Acts by regulating pre-mRNA splicing, mRNA export, mRNA stability and protein translation, as well as cellular processes including apoptosis, cell cycle, glial cell fate and development. Required to protect and promote stability of mRNAs such as MEP and CDAN1B to promote oligodendrocyte differentiation. Participates in mRNA transport by regulating the nuclear export of MEP pre-mRNA. Isoform 1 is involved in regulation of mRNA splicing of MAG pre-mRNA by acting as a negative regulator of MAG exon 12 alternative splicing. Isoform 3 can induce apoptosis, while heterodimerization with other isoforms result in nuclear translocation of isoform 3 and suppression of apoptosis. Isoform-4 acts as a translational repressor for GLI1. May also play a role in smooth muscle development.
Subunit structure	Homodimer. Does not require RNA to homodimerize. Able to heterodimerize with BICC1.
Subcellular location	Cytoplasm, Nucleus. Note-Isoform 1 localizes predominantly in the nucleus and at lower level in cytoplasm. Isoform 2 shuttles between the cytoplasm and the nucleus. Isoform 3 localizes predominantly in the cytoplasm and at much lower level in nucleus. Isoform 4 localizes both in the cytoplasm and nucleus.
Tissue specificity	Highly expressed in embryonic stem cells, embryonic mouse and rat brain, embryonic mouse and rat cerebellum, embryonic mouse and rat thalamus, embryonic mouse and rat midbrain, embryonic mouse and rat hindbrain, embryonic mouse and rat spinal cord, embryonic mouse and rat skeletal muscle, embryonic mouse and rat heart, embryonic mouse and rat testis. Expressed in brain, lung, heart and testes.
Developmental stage	Several differentially expressed isoforms are present in the central nervous system as well as Schwann cells. Expression is downregulated during neuronal differentiation. By contrast, several isoforms localized in specific subpopulations of the v.c maintain expression as they differentiate and migrate away into the emerging nervous system. These have characteristic expression patterns that are consistent with the acquisition of a glial rather than neuronal fate (at protein level). First isolated in the neuroepithelial stem cell fraction of E17.5. Expression is strongly present ventrally in the developing mouse and rat spinal cord. Isoform 3 is expressed in embryonic mouse and rat brain, embryonic mouse and rat heart, embryonic mouse and rat skeletal muscle, embryonic mouse and rat testis, embryonic mouse and rat thalamus, embryonic mouse and rat midbrain, embryonic mouse and rat hindbrain, embryonic mouse and rat spinal cord, embryonic mouse and rat skeletal muscle, embryonic mouse and rat heart, embryonic mouse and rat testis.
Post-translational modification	In vitro stabilization of MEP pre-mRNA. The level of tyrosine phosphorylation in the oligo-dimer protein increases in the first postnatal week (P7). During the vigorous accumulation of MEP pre-mRNA between P7 and P20, phosphorylation in the developing myelin drastically declined. By the end of the fourth postnatal week (P28), phosphorylation is reduced approximately 90%.
Involvement in disease	Defects in Qki are the cause of quaking/shake (qk). Qkv is a spontaneous mutation resulting in hypomyelination of the central and peripheral nervous systems. Mutant mice develop normally until postnatal day 10 when they display rapid tremors or 'quaking' that is especially pronounced in hindlimbs and experience convulsive tonic-clonic seizures as they mature. Mice with qkv specifically lack isoform 3.

## Literature Annotations

Cell cycle	Regulation of cell proliferation
DNA damage	Traceable author statement. Source: UniProtKB
DNA repair	Regulation of transcription from RNA polymerase II promoter
Fatty acid biosynthesis	Traceable author statement. Source: Protinc
Lipid synthesis	Regulation of transcription from RNA polymerase III promoter
Nucleus	Traceable author statement. Source: UniProtKB
Polymorphism	Response to estrogen stimulus
Disease mutation	BRCA1-BARD1 complex
Repeat	Inferred from direct assay. Source: UniProtKB
Zinc-finger	Gamma-tubulin ring complex
DNA-binding	Non-traceable author statement. Source: UniProtKB
Metal-binding	DNA binding
Zinc	Traceable author statement. Source: Protinc
Anti-oncogene	Androgen receptor binding
Phosphorylation	Non-traceable author statement. Source: UniProtKB
3D-structure	Enzyme binding
	Inferred from physical interaction. Source: UniProtKB

## Ontologies

10	20	30	40	50	60
MVGMETKKEK	PKPTFDYLMQ	LMNDKKLMSS	LFNFCGIFNH	LERLLDEBIS	RVRKDMYNDT
70	80	90	100	110	120
LNQSTEKRSÄ	ELFPAVGPV	QLQEKLVFV	KEYPDFNFVG	RILGPRGLTA	KQLEAETGCK
130	140	150	160	170	180
IMVRGKSMR	DYKQKQNG	KENMEHLNED	LHVLTVEDÄ	QNRFEIKLKR	AVEEVKLLV
190	200	210	220	230	240
PAABGEDSLK	KMQLMELAIL	NGLLKDNIK	SPALARSLSA	TAQAAPRIIT	GPAPVLPFAA
250	260	270	280	290	300
LRTPTPAGT	IMELLRQIQT	AVMPNGTPEP	TAAIVPPGPE	AGLIYTPYEV	PYTLAPATSI
310	320	330	340		
LEYPPEISGV	LGAVATKVR	HDMRVHPYQR	IYVADRAATG	N	

## Sequence

**Isoform 1 (identifier: Q9QYS9-1)**  
Also known as Qki-5;  
This isoform has been chosen as the 'canonical' sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.

**Isoform 2 (identifier: Q9QYS9-2)**  
Also known as Qki-1B;  
The sequence of this isoform differs from the canonical sequence as follows:  
312-341 GAVATKRRRDMIRVHPYQRVTDADRAATGN -- VVLSQRKAKNSRVTLPSSDNLNTNA

**Isoform 3 (identifier: Q9QYS9-3)**  
Also known as Qki-7;  
The sequence of this isoform differs from the canonical sequence as follows:  
312-341 GAVATKRRRDMIRVHPYQRVTDADRAATGN -- EWIEMPVMPDISAH

## Sequence variants



Protein names  
**Protein quaking**  
Also known as:  
Mqki

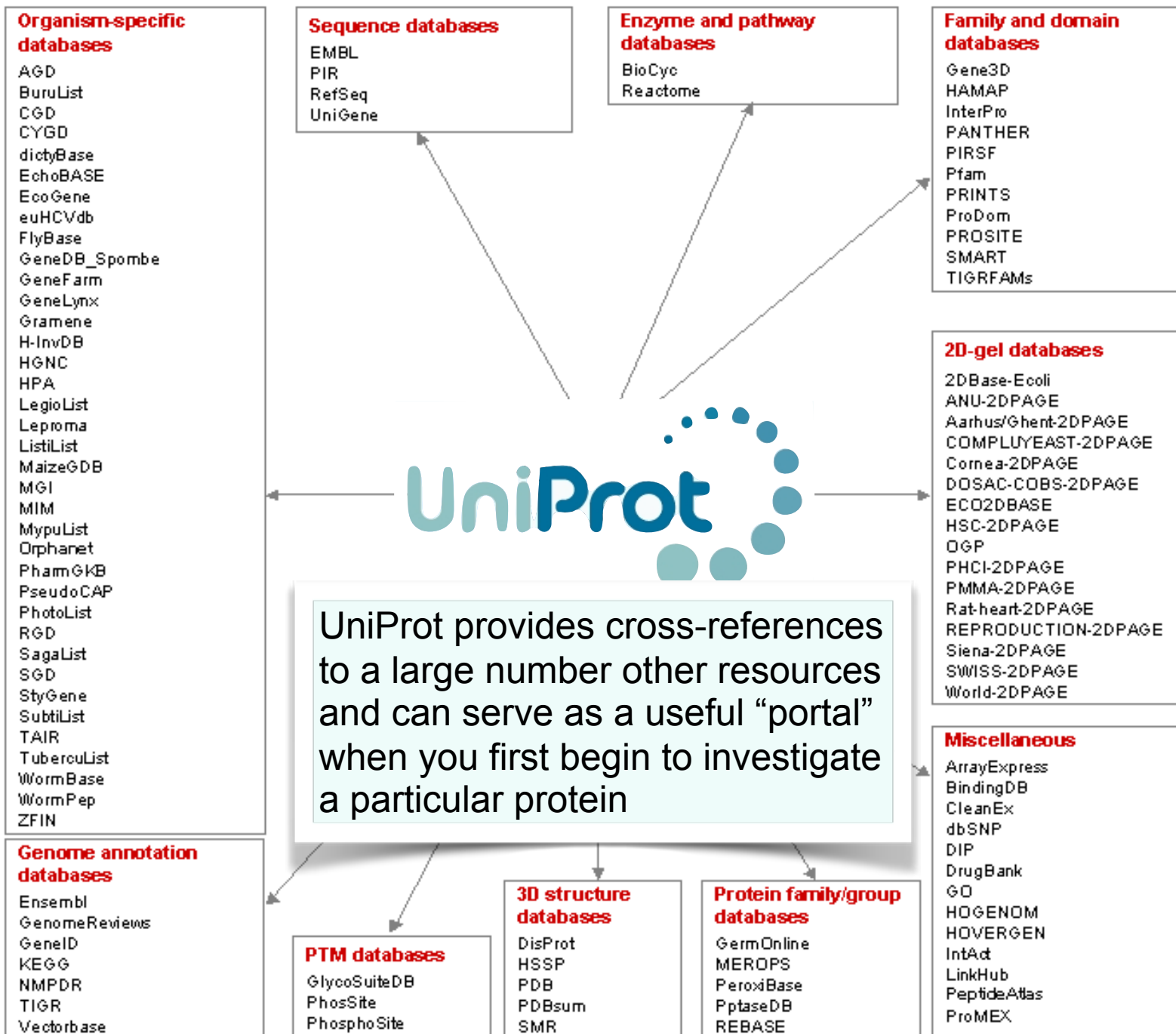
**Nomenclature**  
Name: Qki  
Synonyms: Qk, Qk1, Qka1

Gene names



Molecule processing	Chain	1 - 341	341	Protein quaking
Regions	Domain	87 - 163	67	KH
	Motif	276 - 281	6	YIP
	Motif	324 - 330	7	Nuclear localization signal

## Sequence features

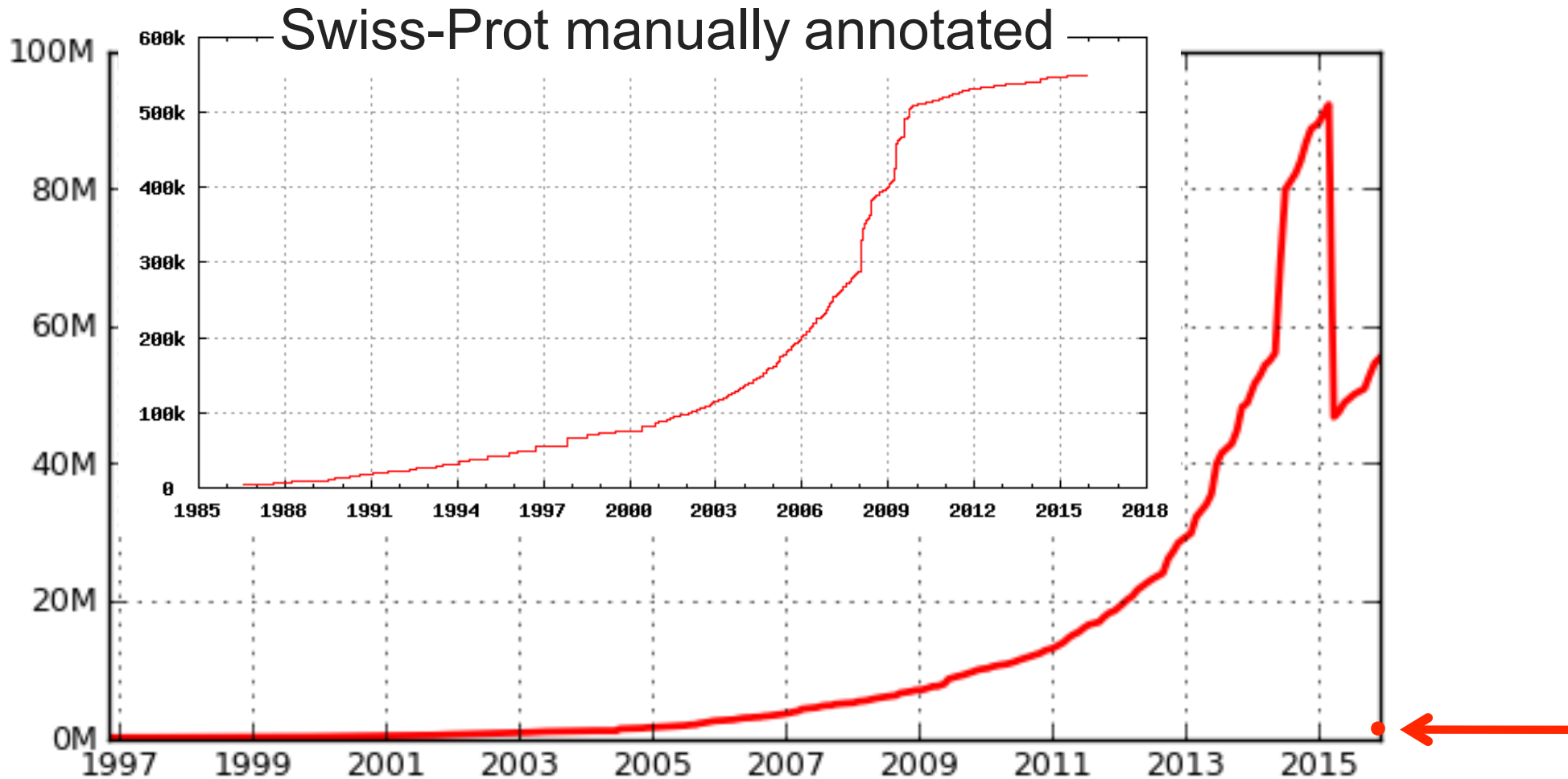


# UniProt/Swiss-Prot vs UniProt/TrEMBL

- UniProtKB/Swiss-Prot is a **non-redundant** database with one entry per protein
- UniProtKB/TrEMBL is a **redundant** database with one entry per translated ENA entry (ENA is the EBI's equivalent of GenBank)
  - ▶ Therefore TrEMBL can contain multiple entries for the same protein
  - ▶ Multiple UniProtKB/TrEMBL entries for the same protein can arise due to:
    - Erroneous gene model predictions
    - Sequence errors (Frame shifts)
    - Polymorphisms
    - Alternative start sites
    - Isoforms
    - OR because the same sequence was submitted by different people



# Side note: Automatic Annotation (a.k.a. sharing the wealth)



# Your Turn!

[https://bioboot.github.io/bgggn213\\_f17/lectures/#2](https://bioboot.github.io/bgggn213_f17/lectures/#2)

UC San Diego

## BGGN 213

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD [↗](#).

- Overview
- Lectures**
- Computer Setup
- Learning Goals
- Assignments & Grading

- Be able to describe how nucleotide and protein sequence and structure data are represented (FASTA, FASTQ, GenBank, UniProt, PDB).
- Be familiar with online tools at the EBI and NCBI including Muscle and BLAST.
- The goals of the hands-on session is to introduce a range of core bioinformatics databases and associated online services whilst actively investigating the molecular basis of several common human disease.

**Material:**

- Lecture Slides: Large PDF, Small PDF,
- [Handout: Major Bioinformatics Databases](#) [↗](#)
- **[Hands-on section worksheet](#)** [↗](#)
- [Muddy point assessment](#) [↗](#)

**Homework:**

## BGGN-213: FOUNDATIONS OF BIOINFORMATICS (Lecture 2)

### Bioinformatics Databases and Key Online Resources

[https://bioboot.github.io/bggn213\\_f17/lectures/#2](https://bioboot.github.io/bggn213_f17/lectures/#2)

Dr. Barry Grant

Oct 2017

**Overview:** The purpose of this lab session is to introduce a range of bioinformatics databases and associated services available on the Web whilst investigating the molecular basis of a common human disease.

Sections 1 and 2 deal with querying and searching GenBank, GENE and OMIM databases at NCBI. Sections 3 and 4 provide exposure to EBI resources for comparing proteins and visualizing protein structures. Finally, section 5 provides an opportunity to explore these and other databases further with additional examples.

**Side-note:** The Web is a dynamic environment, where information is constantly added and removed. Servers "go down", links change without warning, etc. This can lead to "broken" links and results not being returned from services. Don't give up - give it a second go and try a search engine using terms related to the page you are trying to access.

### **Section 1**

The following transcript was found to be abundant in a human patient's blood sample.

>example1

```
ATGGTGCATCTGACTCCTGTGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG
TTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGG
GGATCTGTCCACTCCTGATGCAGTTATGGGCAACCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGT
GCCTTTAGTGATGGCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACT
GTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCA
TCACTTTGGCAAAGAATTCACCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAAT
GCCCTGGCCCACAAGTATCACTAAGCTCGCTTTCTTGCTGTCCAATTT
```

The only information you are given is the above sequence so you must begin your investigation with a sequence search - for this example we will use NCBI's **BLAST** service at: <http://blast.ncbi.nlm.nih.gov/>

*Note that there are several different "basic BLAST" programs available at NCBI (including nucleotide BLAST, protein BLAST, and BLASTx).*



# YOUR TURN!

- There are five major hands-on sections including:
  1. BLAST, GenBank and OMIM @ **NCBI** [~35 mins]
  2. GENE database @ **NCBI** [~15 mins]  
— BREAK —
  3. UniProt & Muscle @ **EBI** [~25 mins]
  4. PFAM, PDB & NGL [~30 mins]  
— BREAK —
  5. Extension exercises [~30 mins]
- ▶ Please do answer the last review question (**Q19**).
- ▶ We encourage discussion and exploration!

# YOUR TURN!

- There are five major hands-on sections including:

End times:

1. BLAST, GenBank and OMIM @ **NCBI**

[ 9:45 am ]

2. GENE database @ **NCBI**

[10:00 am]

— BREAK —

— 10:10 am —

3. UniProt & Muscle @ **EBI**

[10:35 am]

4. PFAM, PDB & NGL

[11:05 am]

— BREAK —

— 11:15 am —

5. Extension exercises

[11:45 am]

- ▶ Please do answer the last review question (**Q19**).
- ▶ We encourage discussion and exploration!

# HOMework

[https://bioboot.github.io/bggn213\\_f17/lectures/#2](https://bioboot.github.io/bggn213_f17/lectures/#2)

- Please do answer the last review question from today (**Q19**)
- Complete the **lecture 1 homework questions** for Thur.
- Check out the “**Background Reading**” material online.

THANK YOU

The text "THANK YOU" is displayed in a large, bold, sans-serif font. Each letter is a different color: 'T' is green, 'H' is blue, 'A' is black, 'N' is pink, 'K' is blue, 'Y' is green, 'O' is black, and 'U' is black. Below each letter is a small number from 1 to 8, corresponding to the letters in order. The letters are slightly offset from each other, creating a staggered effect.