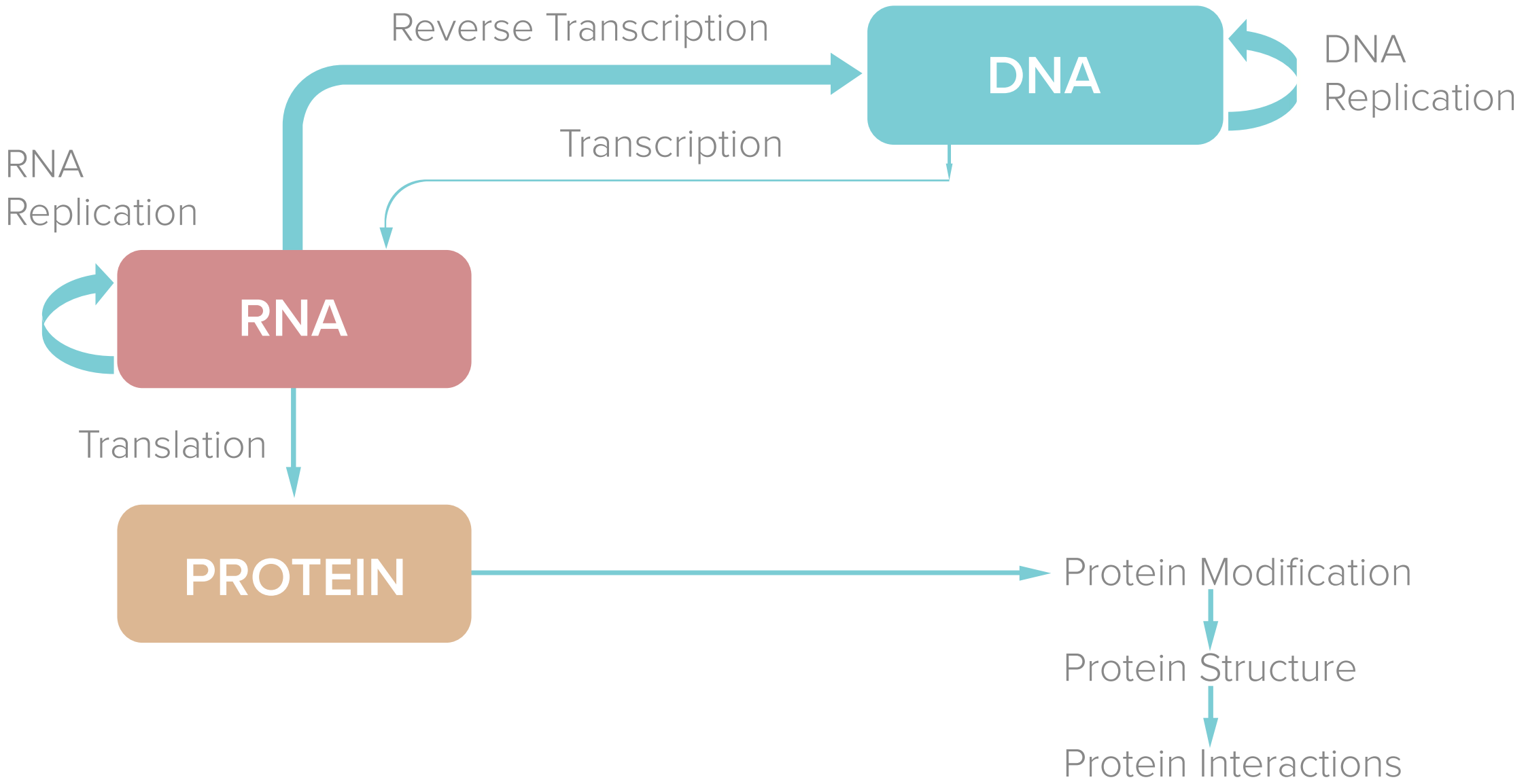
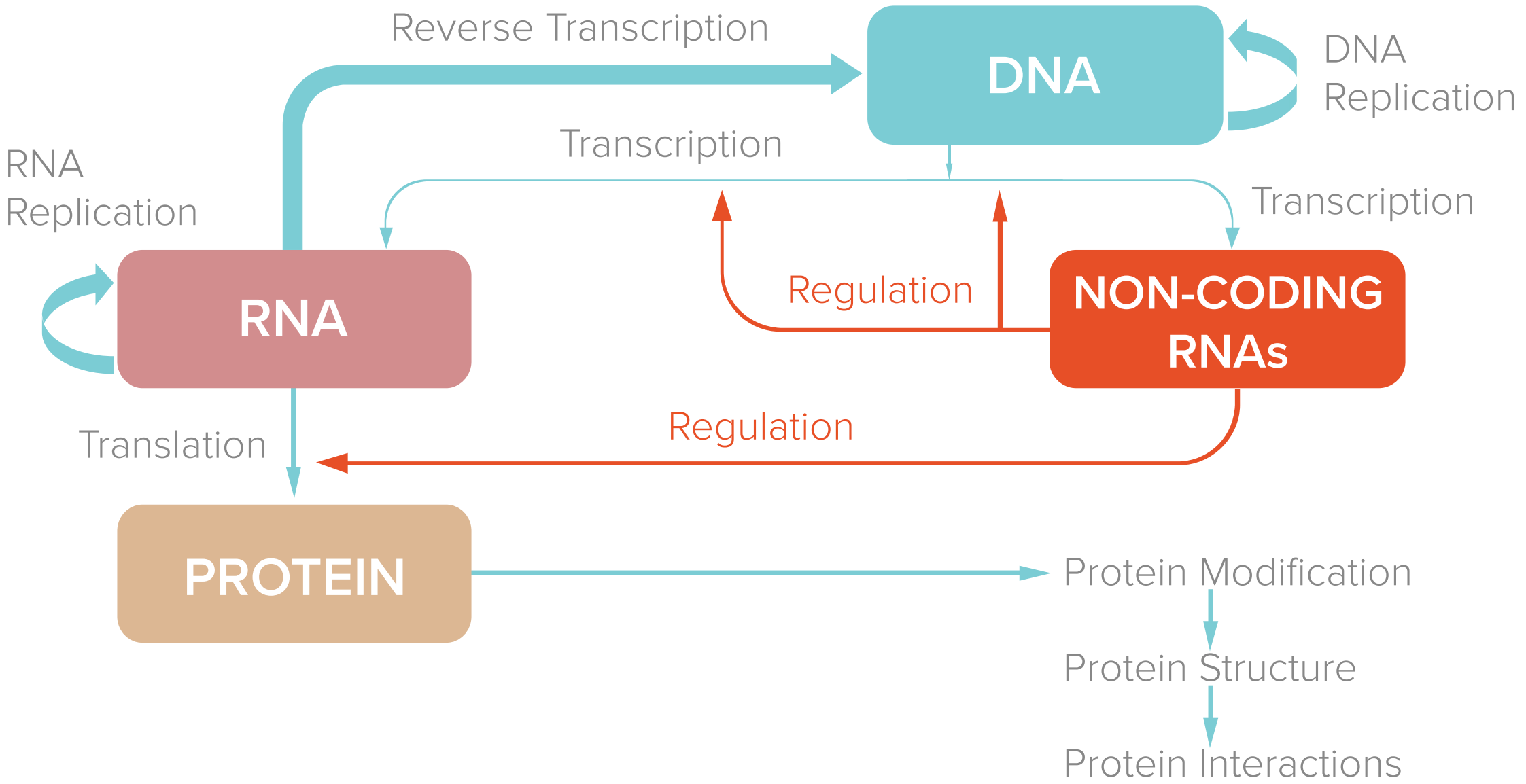
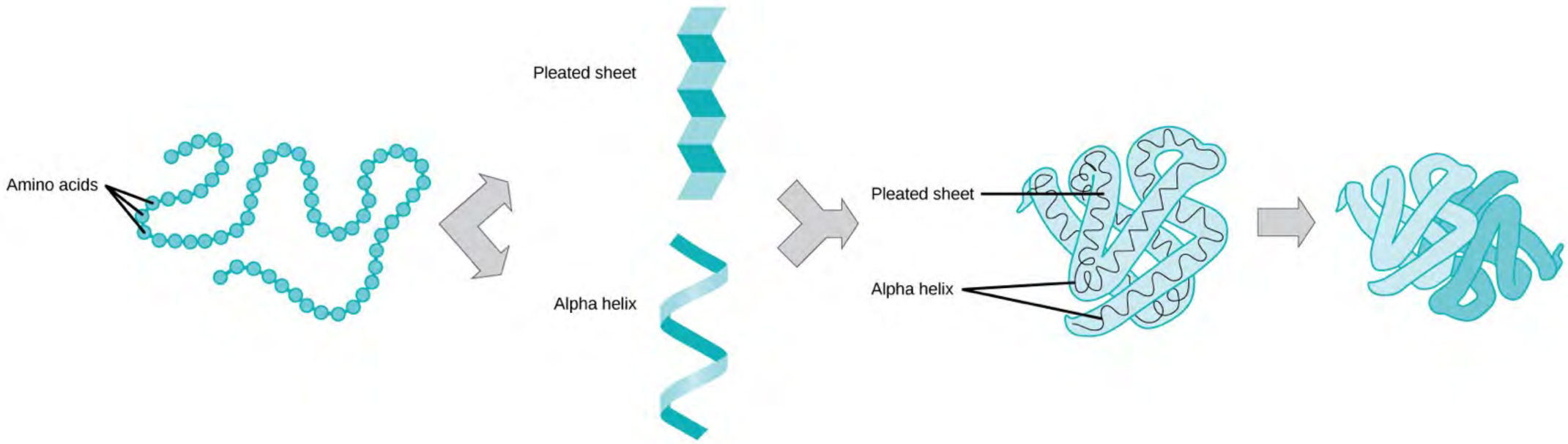
Introduction to Proteins and Proteomics Data

- Proteins
- Protein Sequence Data
- Protein Interaction Data
- Mass Spectrometry Proteomics


Reverse Transcription
DNA
DNA
Replication
Transcription
RNA
Replication
RNA
Translation
PROTEIN
Protein Modification
Protein Structure
Protein Interactions


Reverse Transcription
DNA
DNA Replication
Transcription
RNA Replication
Transcription
RNA
Regulation
NON-CODING RNAs
Regulation
Translation
PROTEIN
Protein Modification
Protein Structure
Protein Interactions



Primary protein structure

Sequence of a chain of amino acids

Secondary protein structure

Hydrogen bonding of the peptide backbone causes the amino acids to fold into a repeating patterna

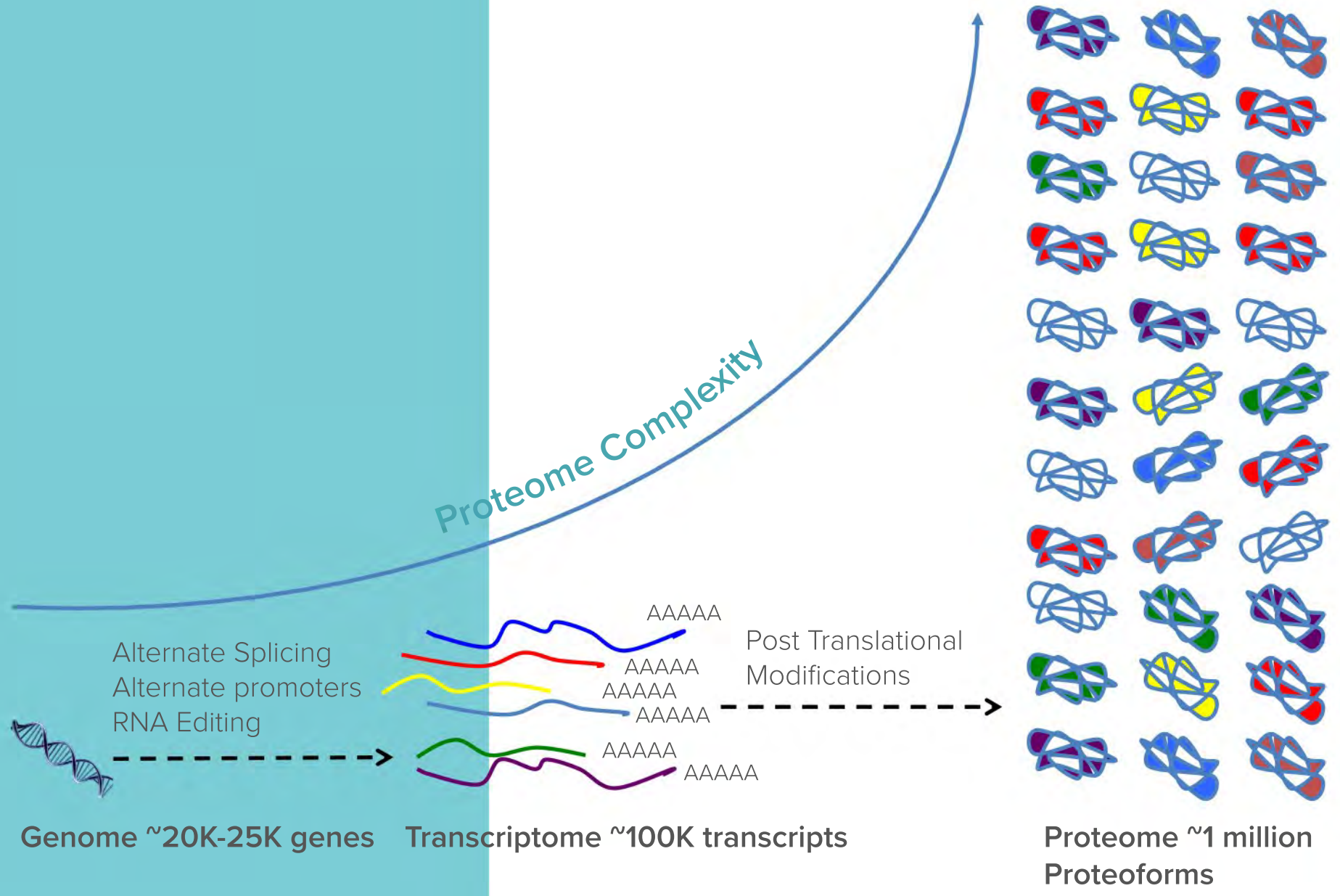
Tertiary protein structure

Three-dimensional folding pattern of a protein due to side chain interactions

Quaternary protein structure

Protein consisting of more than one amino acid chain

Source: “Protein Structure.” Boundless Biology. Retrieved 11 May. 2016 from https://www.boundless.com/biology/textbooks/boundless-biology-textbook/biological-macromolecules-3/proteins-56/protein-structure-304-11437/


Proteome Complexity
Alternate Splicing
Alternate promoters
RNA Editing
AAAAA
AAAAA
AAAAA
AAAAA
AAAAA
AAAAA
Post Translational
Modifications
Genome ~20K-25K genes
Transcriptome ~100K transcripts
Proteome ~1 million
Proteoforms

Sources of Protein Data 1

Primary AA Translations from the International Nucleotide Sequence Database Collaboration (INSDC) Archives

NCBI in USA (http://www.ncbi.nlm.nih.gov/genbank/)
EMBL-EBI in Europe (http://www.ebi.ac.uk/ena)
DDBJ in Japan (http://www.ddbj.nig.ac.jp)

Curated Primary and Secondary AA data

Universal Protein Resource (UniProt) (http://www.uniprot.org)
RefSeq (http://www.ncbi.nlm.nih.gov/refseq/)

3-D Protein Structure

One of the world wide PDB resources (http://www.wwpdb.org)

UniProt

UniProtKB | Advanced Search

BLAST Align Retrieve/ID mapping Help Contact

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

UniProtKB

UniProt Knowledgebase

Swiss-Prot (551,193)
Manually annotated and reviewed.

TrEMBL (62,148,086)
Automatically annotated and not reviewed.

UniRef

Sequence clusters

UniParc

Sequence archive

Proteomes

Supporting data

Literature citations
Taxonomy
Subcellular locations
Cross-ref. databases
Diseases
Keywords

News

Forthcoming changes
Planned changes for UniProt

UniProt release 2016_05
Slow/White and the 6 DWORFs | Cross-references to SIGNOR | Change of UniProt website job identifiers

UniProt release 2016_04
Small changes, big effects | New UniProt JAPI | Change of the UniProt RDF files distribution

News archive

Getting started

YouTube

Text search
Our basic text search allows you to search all the resources available

BLAST
Find regions of similarity between your sequences

Sequence alignments
Align two or more protein sequences using the Clustal Omega program

UniProt data

Download latest release
Get the UniProt data

Statistics
View Swiss-Prot and TrEMBL statistics

How to cite us
The UniProt Consortium

Submit your data
Submit your sequences and annotation updates

Protein spotlight

The Shape Of Harm

May 2016

Sometimes we are forced to see things differently. But it is never easy because we are creatures of habit and, like it or not, shackled by what we were first led to believe. This is exactly what happened with the prion. Prions are proteins whose shape can change under certain conditions, and in so doing be at the heart of fatal diseases...

Protein **Cytochrome c**

Gene **CYCS**

Organism *Homo sapiens (Human)*

Status Reviewed - Annotation score: ◉◉◉◉◉ - Experimental evidence at protein level[i]

Function[i]

Electron carrier protein. The oxidized form of the cytochrome c heme group can accept an electron from the heme group of the cytochrome c1 subunit of cytochrome reductase. Cytochrome c then transfers this electron to the cytochrome oxidase complex, the final protein carrier in the mitochondrial electron-transport chain.
Plays a role in apoptosis. Suppression of the anti-apoptotic members or activation of the pro-apoptotic members of the Bcl-2 family leads to altered mitochondrial membrane permeability resulting in release of cytochrome c into the cytosol. Binding of cytochrome c to Apaf-1 triggers the activation of caspase-9, which then accelerates apoptosis by activating other caspases.

Sites

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Binding site[i]	15 – 15	1	Heme (covalent)			
Binding site[i]	18 – 18	1	Heme (covalent)			
Metal binding[i]	19 – 19	1	Iron (heme axial ligand)			
Metal binding[i]	81 – 81	1	Iron (heme axial ligand)			

Pathology & Biotech

Involvement in disease

Thrombocytopenia 4 (THC4) 1 Publication

The disease is caused by mutations affecting the gene represented in this entry.
Disease description: Thrombocytopenia is defined by a decrease in the number of platelets in circulating blood, resulting in the potential for increased bleeding and decreased ability for clotting.
See also OMIM:612004

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Natural variant	42 – 42	1	G → S in THC4; increases the pro-apoptotic function by triggering caspase activation more efficiently than wild-type; does not affect the redox function. 1 Publication		VAR_044450	

Keywords - Disease

Disease mutation

Organism-specific databases

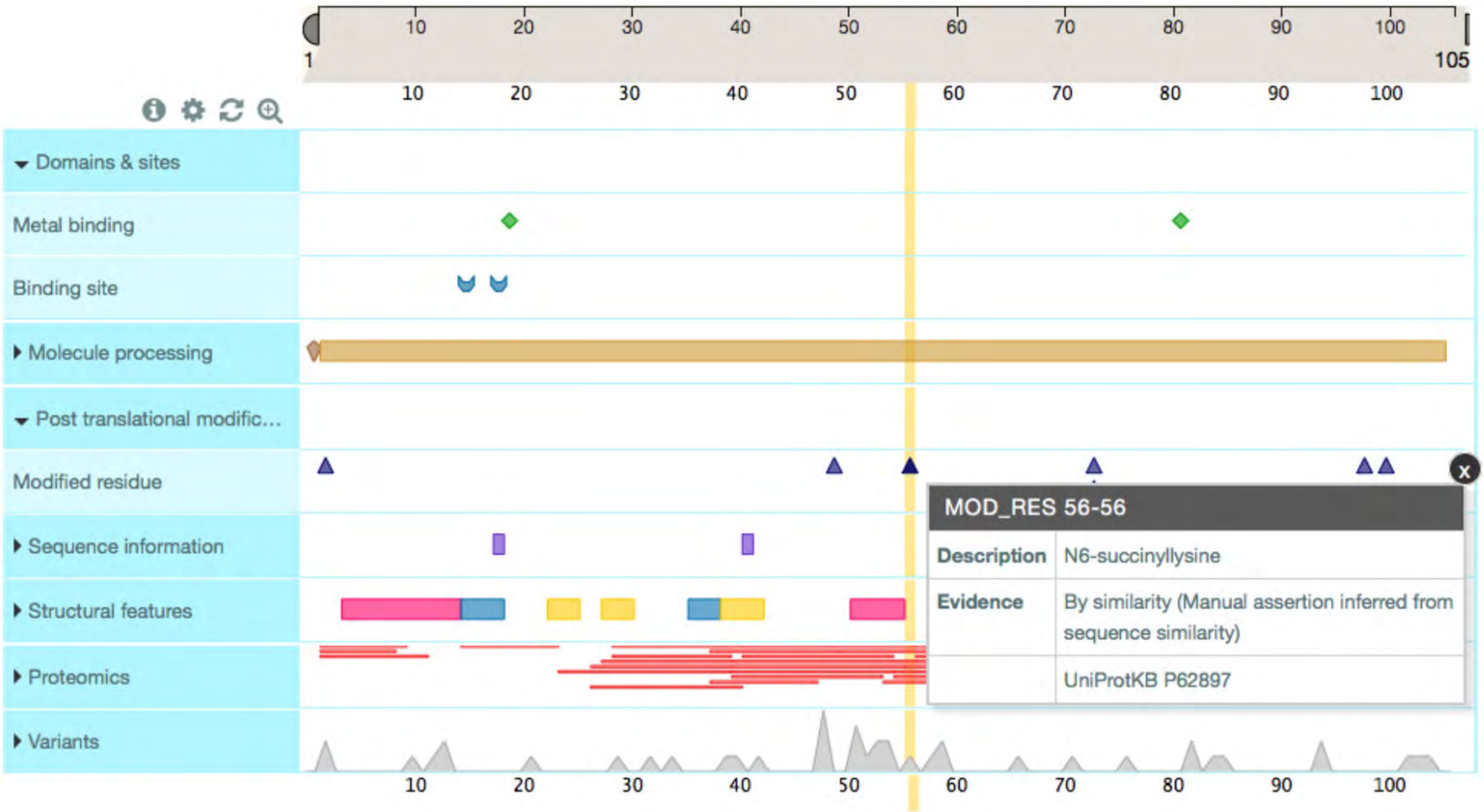
MalaCards	CYCS.
MIM	612004. phenotype.
Orphanet	168629. Autosomal thrombocytopenia with normal platelets.
PharmGKB	PA134981636.

Publications[1]

Show 'large scale' publications »

1. **"The human somatic cytochrome c gene: two classes of processed pseudogenes demarcate a period of rapid molecular evolution."**
   Evans M.J., Scarpulla R.C.
   Proc. Natl. Acad. Sci. U.S.A. 85:9625-9629(1988) [PubMed] [Europe PMC] [Abstract]
   Cited for: NUCLEOTIDE SEQUENCE [GENOMIC DNA].
2. **"The amino acid sequence of human heart cytochrome c."**
   Matsubara H., Smith E.L.
   J. Biol. Chem. 237:3575-3576(1962) [PubMed] [Europe PMC] [Abstract]
   Cited for: PROTEIN SEQUENCE OF 2-105, ACETYLATION AT GLY-2.
   Tissue: Heart.
3. **"Human heart cytochrome c. Chymotryptic peptides, tryptic peptides, and the complete amino acid sequence."**
   Matsubara H., Smith E.L.
   J. Biol. Chem. 238:2732-2753(1963) [PubMed] [Europe PMC] [Abstract]
   Cited for: PROTEIN SEQUENCE OF 2-105.
   Tissue: Heart.
4. **"Cytochrome c in the apoptotic and antioxidant cascades."**
   Skulachev V.P.
   FEBS Lett. 423:275-280(1998) [PubMed] [Europe PMC] [Abstract]
   Cited for: REVIEW ON ROLE IN APOPTOSIS.
5. **"Solution structure of reduced recombinant human cytochrome c."**
   Jeng W.-Y., Shiu J.-H., Tsai Y.-H., Chuang W.-J.
   Submitted (FEB-2003) to the PDB data bank
   Cited for: STRUCTURE BY NMR.
6. **"A mutation of human cytochrome c enhances the intrinsic apoptotic pathway but causes only thrombocytopenia."**
   Morison I.M., Cramer Borde E.M.C., Cheesman E.J., Cheong P.L., Holyoake A.J., Fichelson S., Weeks R.J., Lo A., Davies S.M.K., Wilbanks S.M., Fagerlund R.D., Ludgate M.W., da Silva Tatley F.M., Coker M.S.A., Bockett N.A., Hughes G., Pippig D.A., Smith M.P., Capron C., Ledgerwood E.C.
   Nat. Genet. 40:387-389(2008) [PubMed] [Europe PMC] [Abstract]
   Cited for: VARIANT THC4 SER-42, IDENTIFICATION BY MASS SPECTROMETRY, CHARACTERIZATION OF VARIANT THC4 SER-42, X-RAY CRYSTALLOGRAPHY (2.75 ANGSTROMS) OF VARIANT THC4 SER-42 AND WILD TYPE.

+Additional computationally mapped references.


1
105
10 20 30 40 50 60 70 80 90 100
Domains & sites
Metal binding
Binding site
Molecule processing
Post translational modific...
Modified residue
Sequence information
Structural features
Proteomics
Variants
MOD_RES 56-56
Description
N6-succinyllysine
Evidence
By similarity (Manual assertion inferred from sequence similarity)
UniProtKB P62897

UniProtKB - Q9VXE3 (Q9VXE3_DROME)

Basket

Display

- Entry
- Feature viewer
- Feature table

None

- Function
- Names & Taxonomy
- Subcell. location
- Pathol./Biotech
- PTM / Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence
- Cross-references
- Publications
- Entry information
- Miscellaneous
- Similar proteins

BLAST | Align | Format | Add to basket | History | Feedback | Help video | Other tutorials and videos

Protein **Rho-associated protein kinase**

Gene **Rok**

Organism *Drosophila melanogaster (Fruit fly)*

Status Unreviewed - Annotation score: ◉◉◉◉◉ - Experimental evidence at protein level[i]

Function[i]

Protein kinase which is a key regulator of actin cytoskeleton and cell polarity. UniRule annotation

Automatic assertion according to rules[i]
PIRNR:PIRNR037568

Catalytic activity[i]
ATP + a protein = ADP + a phosphoprotein. UniRule annotation SAAS

Cofactor[i]
Mg^{2+} UniRule annotation

Enzyme regulation[i]
Activated by RHOA binding. Inhibited by Y-27632. UniRule annotation

Sites

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Binding site[i]	116 – 116	1	ATP UniRule annotation			
Active site[i]	209 – 209	1	Proton acceptor UniRule annotation			

UniRule: UR000112939

Source Rule PIRNR: PIRNR037568

View all proteins annotated by this rule | Remove highlights

If a protein meets these conditions... [i]

Common conditions

Matches PIR Superfamily signature PIRSF037568
sequence length = 1354 - 1401
taxon = Metazoa
fragment ≠ the sequence is fragmented.

... then these annotations are applied [i]

Protein name [i]

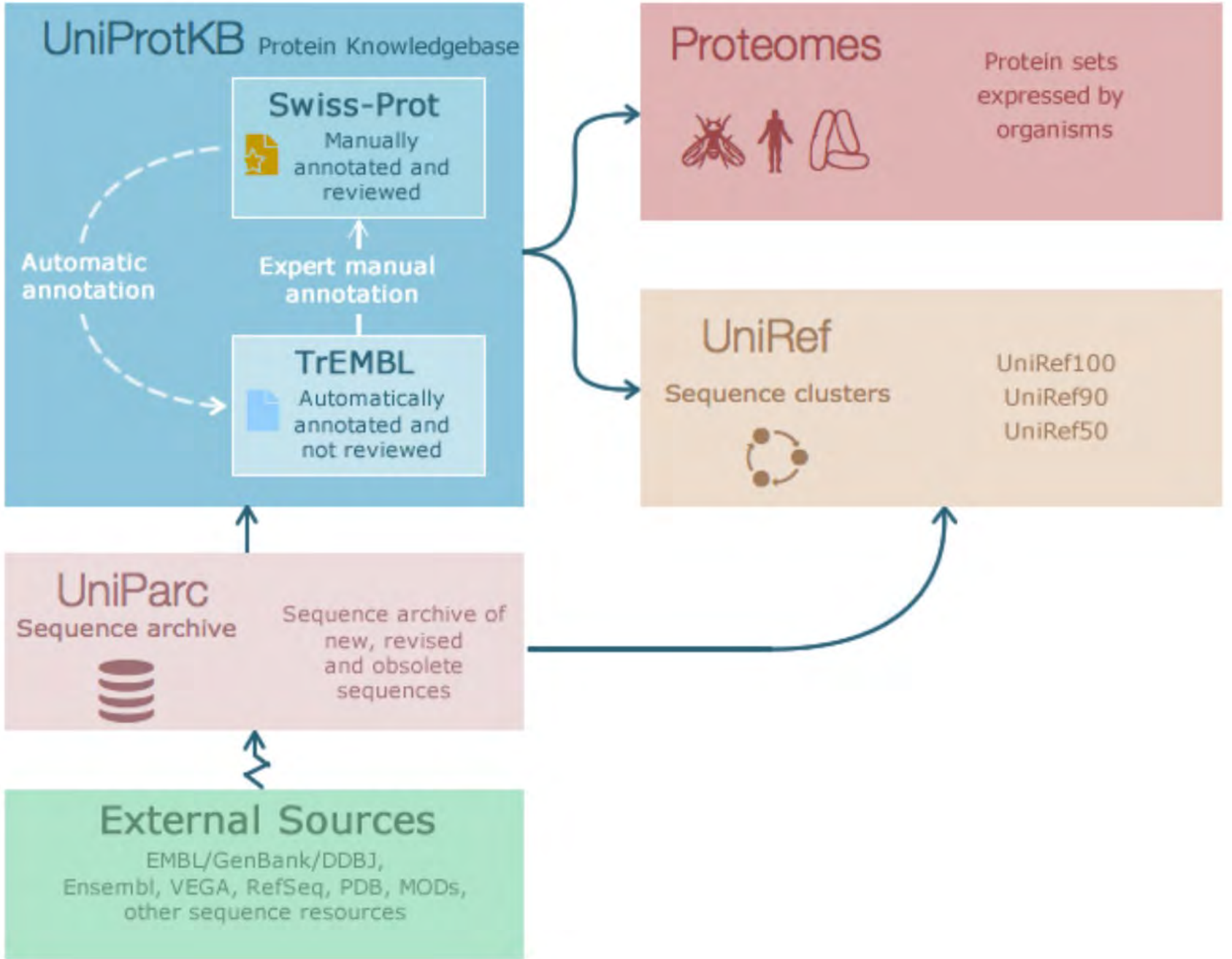
Recommended name:
Rho-associated protein kinase (EC:2.7.11.1)

Catalytic activity [i]

ATP + a protein = ADP + a phosphoprotein.

Function [i]

Protein kinase which is a key regulator of actin cytoskeleton and cell polarity.


UniProtKB Protein Knowledgebase
Swiss-Prot
Manually annotated and reviewed
Automatic annotation
Expert manual annotation
TrEMBL
Automatically annotated and not reviewed
Proteomes
Protein sets expressed by organisms
UniRef
Sequence clusters
UniRef100
UniRef90
UniRef50
UniParc
Sequence archive
Sequence archive of new, revised and obsolete sequences
External Sources
EMBL/GenBank/DDBJ, Ensembl, VEGA, RefSeq, PDB, MODs, other sequence resources

How to Handle Big Data?
One Approach: Make it A little Smaller

Filter out data you do not need

Need to understand your use cases well. If you trash it and need it later it maybe a problem.

Remove redundancy

Duplicate values can be stored once not thousands or millions of times.

Cluster or group your data by some algorithm

First protein data reduction was done the 1970s to save precious disk and memory space.

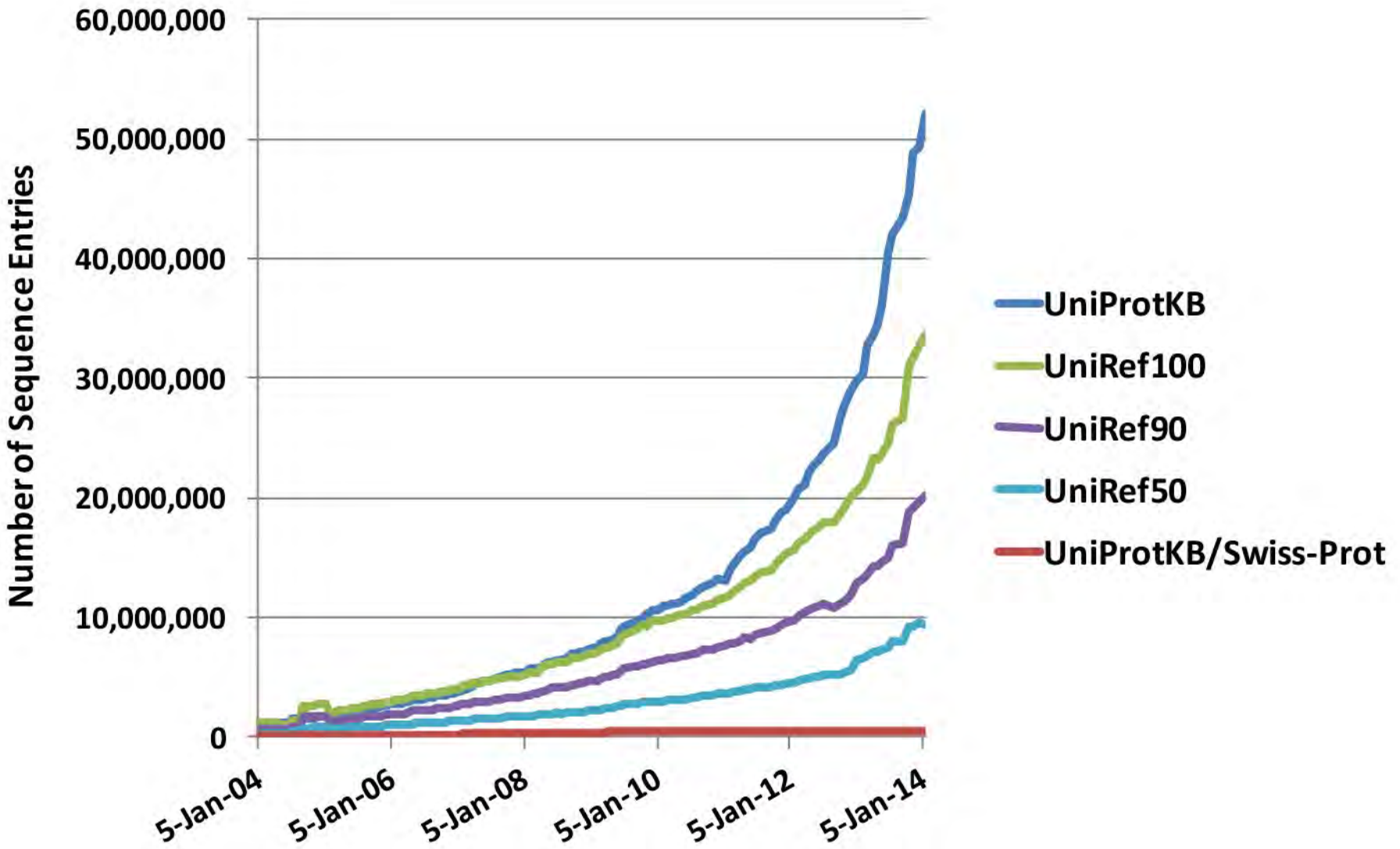
Three letter AA code > One letter

>sp|P99999|CYC_HUMAN Cytochrome c OS=Homo sapiens GN=CYCS PE=1 SV=2
MetGlyAspValGluLysGlyLysLysIlePheIleMetLysCysSerGlnCysHisThr-
ValGluLysGlyGlyLysHisLysThrGlyProAsnLeuHisGlyLeuPheGlyArgLysThrGlyGlnAlaProGlyTyrSerTyrThrAlaAla
AsnLysAsnLysGlyIleIleTrpGlyGluAspThrLeuMetGluTyrLeuGluAsnProLysLysTyrIleProGlyThrLysMetIlePheValGl
yIleLysLysLysGluGluArgAlaAspLeuIleAlaTyrLeuLysLysAlaThrAsnGlu

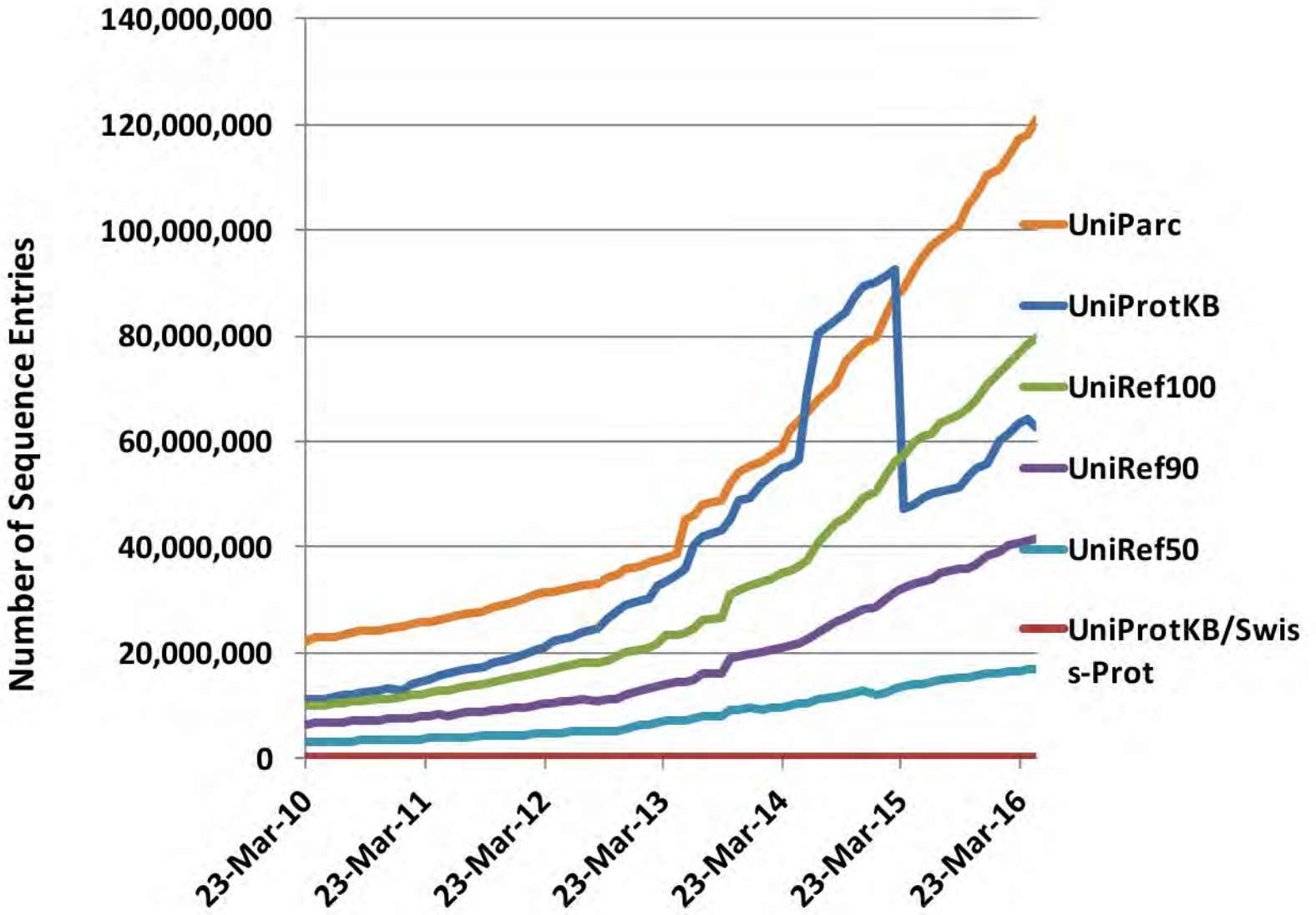
>sp|P99999|CYC_HUMAN Cytochrome c OS=Homo sapiens GN=CYCS PE=1 SV=2
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIW
GEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE

375 > 172 characters 54% reduction yeah!

Growth of UniProt



Growth of UniProt Databases



Proteomes results

About Proteomes

Filter by

5,448 Reference proteomes

49,897 Other proteomes

Superkingdom

50,350 Bacteria

3,446 Viruses

443 Archaea

1,105 Eukaryota

Map to

UniProtKB

UniParc (for redundant proteomes)

Demo

Help video

Download | Columns

1 to 250 of 55,345 Show 250

Show only non-redundant proteomes?

Proteome ID	Organism	Organism ID	Protein count
UP000001807	**Borrelia burgdorferi (strain ATCC 35210 / B31 / CIP 102532 / DSM 4680)**	224326	1290
UP000005640	**Homo sapiens (Human)**	9606	70615
UP000000806	**Mycobacterium leprae (strain TN)**	272631	1603
UP000000808	**Mycoplasma pneumoniae (strain ATCC 29342 / M129)**	272634	687
UP000000812	**Xylella fastidiosa (strain 9a5c)**	160492	2772
UP000002481	**Staphylococcus aureus (strain Mu50 / ATCC 700699)**	158878	2714
UP000000804	**Helicobacter pylori (strain J99 / ATCC 700824) (Campylobacter pylori J99)**	85963	1488
UP000000576	**Xanthomonas axonopodis pv. citri (strain 306)**	190486	4354
UP000000300	**Streptococcus thermophilus JIM 8232**	1051074	2139

Overview

Proteome name	Gallus gallus Red jungle fowl - Reference proteome
Proteins	17,719
Proteome ID[i]	UP000000539
Strain	Red jungle fowl
Taxonomy	9031 - Gallus gallus
Last modified	April 17, 2016
Genome assembly	GCA_000002315.3


© farm4.static.flickr.com

The red jungle fowl is a herbivorous and insectivorous member of the Phasianidae family. It is the closest wild ancestor of the domesticated chicken, its subspecies. After domestication had taken place 6,000-8,000 years ago in Asia, this species spread all over the world. The chicken is an important agricultural animal and a source of meat and eggs. In biomedical research it serves as a model organism to study various aspects of virology, oncogenesis, immunology, and especially embryogenesis. Since it provides an evolutionary link between mammals and other groups of vertebrates, sequencing its genome is of a significant importance.

The chicken genome has 39 chromosomes containing 1.1 Gb with 20,000-23,000 protein-coding genes. The reference proteome is derived from the genome sequence published in 2004.

Components[i]

Download | View all proteins

Component name	Genome Accession(s)	Proteins
Chromosome 1	CM000093	2120
Chromosome 2	CM000094	1381
Chromosome 3	CM000095	1211