

Bioinformatics

Bi1X-2010

Part I:

Public databases

Arbel Tadmor


Overview

- Obtaining sequence data from the internet
- Aligning two sequences
- Using the biologist's google: BLAST

We'll use hemoglobin as a case study

Related reading: Ch. 6 of Stryer (Biochemistry, 7th edition)

First stop: wiki



WIKIPEDIA
The Free Encyclopedia

navigation

- Main page
- Contents
- Featured content
- Current events
- Random article

search

Go Search

interaction

- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

toolbox

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link
- Cite this page

languages

- العربية
- বাংলা

article discussion edit this page history

Hemoglobin

From Wikipedia, the free encyclopedia

Hemoglobin (also spelled **haemoglobin** and abbreviated **Hb** or **Hgb**) is the **iron-containing oxygen-transport metalloprotein** in the **red blood cells** of **vertebrates**,^[1] and the tissues of some **invertebrates**.

In **mammals**, the protein makes up about 97% of the red blood cell's dry content, and around 35% of the total content (including water)^[citation needed]. Hemoglobin transports oxygen from the **lungs** or **gills** to the rest of the body (i.e. the tissues) where it releases the oxygen for cell use.

Hemoglobin has an oxygen binding capacity between 1.36 and 1.37 ml O₂ per gram of hemoglobin,^[2] which increases the total **blood oxygen capacity** seventyfold.^[3]

Hemoglobin is also found in outside red blood cells and their progenitor lines. Other cells that contain hemoglobin include the A9 dopaminergic neurons in the **substantia nigra**, **macrophages**, **alveolar cells**, and **mesangial cells** in the kidney. In these tissues, hemoglobin has a non-oxygen-carrying function as an **antioxidant** and a regulator of **iron metabolism**.^[4]

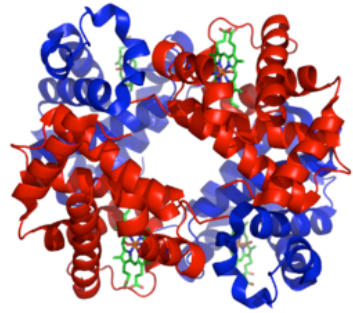
Contents [hide]

- Research history
- Genetics
- Synthesis
- Structure
- Iron's oxidation state in oxyhemoglobin
- Ligand binding
 - Cooperative
 - Competitive
 - Allosteric
- Types in humans
- Degradation in vertebrate animals
- Role in disease
- Diagnostic uses
- Analogues in non-vertebrate organisms
- Other oxygen-binding proteins
- Presence in nonerythroid cells
- In history, art and music

Try Beta Log in / create account

Hemoglobin, human, adult

(heterotetramer, (αβ)₂)



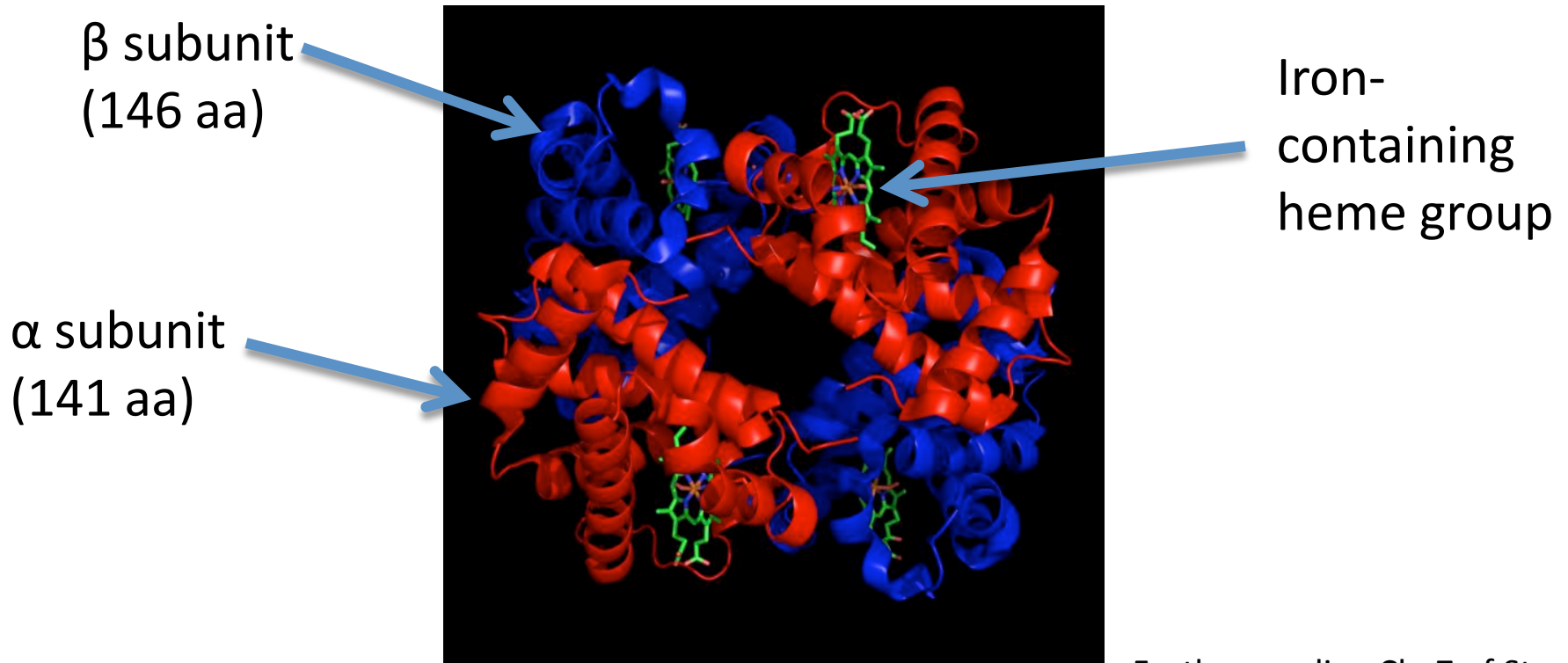
Structure of human hemoglobin. The protein's α and β subunits are in red and blue, and the iron-containing heme groups in green. From PDB 1GZX [Proteopedia Hemoglobin](#)

Protein type	metalloprotein, globulin
Function	oxygen-transport
Cofactor(s)	heme (4)

Subunit Name	Gene	Chromosomal Locus
Hb-α1	HBA1	Chr. 16 p13.3 ↗
Hb-α2	HBA2	Chr. 16 p13.3 ↗
Hb-β	HBB	Chr. 11 p15.5 ↗

Structure of human hemoglobin

- In adults hemoglobin is a tetramer: $\alpha_2\beta_2$
- Each subunit contains a non-protein heme group (that holds an iron)
- The iron binds to an oxygen (shifting absorbance from blue to red)
- Multiple subunits give rise to cooperatively in oxygen binding and unbinding allowing this protein to release more oxygen in the tissues (making it a good transporter).





WIKIPEDIA
The Free Encyclopedia

navigation

- Main page
- Contents
- Featured content
- Current events
- Random article

search

interaction

- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

toolbox

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link
- Cite this page

languages

- العربية
- বাংলা

[article](#) [discussion](#) [edit this page](#) [history](#)

Hemoglobin

From Wikipedia, the free encyclopedia

Hemoglobin (also spelled **haemoglobin** and abbreviated **Hb** or **Hgb**) is the [iron-containing oxygen-transport metalloprotein](#) in the [red blood cells of vertebrates](#),^[1] and the tissues of some [invertebrates](#).

In [mammals](#), the protein makes up about 97% of the red blood cell's dry content, and around 35% of the total content (including water)^[*citation needed*]. Hemoglobin transports oxygen from the [lungs](#) or [gills](#) to the rest of the body (i.e. the tissues) where it releases the oxygen for cell use.

Hemoglobin has an oxygen binding capacity between 1.36 and 1.37 ml O₂ per gram of hemoglobin,^[2] which increases the total [blood oxygen capacity](#) seventyfold.^[3]

Hemoglobin is also found in outside red blood cells and their progenitor lines. Other cells that contain hemoglobin include the A9 dopaminergic neurons in the [substantia nigra](#), [macrophages](#), [alveolar cells](#), and [mesangial cells](#) in the kidney. In these tissues, hemoglobin has a non-oxygen-carrying function as an [antioxidant](#) and a regulator of [iron metabolism](#).^[4]

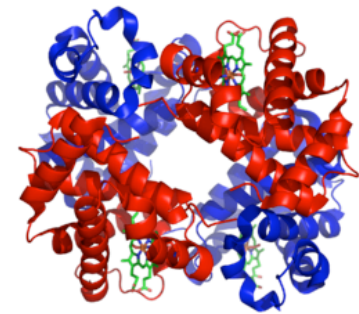
Contents [hide]

- 1 Research history
- 2 Genetics
- 3 Synthesis
- 4 Structure
- 5 Iron's oxidation state in oxyhemoglobin
- 6 Ligand binding
 - 6.1 Cooperative
 - 6.2 Competitive
 - 6.3 Allosteric
- 7 Types in humans
- 8 Degradation in vertebrate animals
- 9 Role in disease
- 10 Diagnostic uses
- 11 Analogues in non-vertebrate organisms
- 12 Other oxygen-binding proteins
- 13 Presence in nonerythroid cells
- 14 In history, art and music

Let's focus on the $\alpha 1$ subunit of hemoglobin. This gene is called **HBA1**

Try Beta [Log in / create account](#)

Hemoglobin, human, adult (heterotetramer, ($\alpha\beta$)₂)



Structure of human hemoglobin. The protein's α and β subunits are in red and blue, and the iron-containing heme groups in green. From PDB 1GZX [Proteopedia Hemoglobin](#)

Protein type	metalloprotein, globulin
Function	oxygen-transport
Cofactor(s)	heme (4)

Subunit Name	Gene	Chromosomal Locus
Hb- $\alpha 1$	HBA1	Chr. 16 p13.3
Hb- $\alpha 2$	HBA2	Chr. 16 p13.3
Hb- β	HBB	Chr. 11 p15.5

On which chromosome is this gene?

Let's look up this gene

- Go to the PubMed website: *google pubmed* (<http://www.ncbi.nlm.nih.gov/PubMed/>)
- Search for **HBA1** under **gene** category

NCBI Resources How To

PubMed.gov
U.S. National Library of Medicine
National Institutes of Health

Search: Gene
Limits Advanced search Help

HBA1 Search Clear

click

Welcome to PubMed
PubMed comprises more than 19 million citations for biomedical articles from MEDLINE and life science journals. Citations may include links to full-text articles from PubMed Central or publisher web sites.

Using PubMed
[PubMed Quick Start](#)
[New and Noteworthy](#)
[PubMed Tutorials](#)
[Full Text Articles](#)
[PubMed FAQs](#)

PubMed Tools
[Single Citation Matcher](#)
[Batch Citation Matcher](#)
[Clinical Queries](#)
[Topic-Specific Queries](#)

More Resources
[MeSH Database](#)
[Journals Database](#)
[Clinical Trials](#)
[E-Utilities](#)
[LinkOut](#)

We'll start with human hemoglobin

NCBI Entrez Gene

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search Gene for hba1 Go Clear Save Search

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Sort by Relevance Send to

All: 25 Current Only: 24 Genes Genomes: 15 SNP GeneView: 10

Items 1 - 20 of 25 Page 1 of 2 Next

1: HBA1 *click* Order cDNA clone, Links
Official Symbol HBA1 and Name: hemoglobin, alpha 1 [*Homo sapiens*]
Other Aliases: CD31, MGC126895, MGC126897
Other Designations: alpha one globin; alpha-1 globin; alpha-1-globin; alpha-globin; hemoglobin alpha 1 globin chain; hemoglobin alpha chain; hemoglobin alpha-1 chain; hemoglobin subunit alpha
Chromosome: 16 Location: 16p13.3
Annotation: Chromosome 16 NC_000016.9 (226679..227520)
MIM: 141800
GeneID: 3039

human →

2: HBA1 Links
hemoglobin, alpha 1 [*Bos taurus*]
Chromosome: 25
Annotation: Chromosome 25 NC_007326.3 (672689..673498)
GeneID: 100140149

cattle →

3: HBA1 Links
hemoglobin, alpha 1 [*Bos taurus*]
Chromosome: 29
GeneID: 281221
This record was replaced with GeneID: 100140149

4: hba1 Order cDNA clone, Links
Official Symbol hba1 and Name: hemoglobin, alpha 1 [*Xenopus (Silurana) tropicalis*]
Other Aliases: cd31, hba-a
Other Designations: X.tropicalis alpha globin; alpha globin adult; alpha-globin; hemoglobin alpha chain; hemoglobin subunit alpha
GeneID: 394454

frog →

5: hba1 Order cDNA clone, Links
Official Symbol hba1 and Name: hemoglobin, alpha 1 [*Xenopus laevis*]
Other Aliases: hba2
Other Designations: alpha-2-globin chain; hemoglobin alpha-2 chain; hemoglobin alpha-minor chain; hemoglobin subunit alpha-2
GeneID: 397869

etc.

RefSeq accession number XX_#

More about the RefSeq database here... <http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=handbook&part=ch18>
<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC539979/pdf/gki025.pdf>

Which sequence should we pick?

Genomic? mRNA? Protein?



Entrez Gene

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search Gene for Go Clear

Limits Preview/Index History Clipboard Details

Display Full Report Send to

1: HBA1 hemoglobin, alpha 1 [*Homo sapiens*]
GeneID: 3039 updated 10-Apr-2010

Summary

Official Symbol HBA1

Official Full Name hemoglobin, alpha 1

Primary source [HGNC:4823](#)

See related [Ensembl:ENSG00000206172](#); [HPRD:00784](#); [MIM](#)

Gene type protein coding

RefSeq status REVIEWED

Organism *Homo sapiens*

Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia; Eumammalia; Placentalia; Primates; Haplorrhini; Catarrhini; Hominidae; Homo

Also known as CD31; MGC126895; MGC126897; HBA1

Summary The human alpha globin gene cluster located on chromosome 16p13.3 consists of five genes: alpha-2 (HBA2), alpha-1 (HBA1), alpha-1 pseudogene (HBA1P), alpha-2 pseudogene (HBA2P), and alpha-1 pseudogene 2 (HBA1P2). The alpha-1 (HBA1) coding sequences are identical to the alpha-2 (HBA2) coding sequences, and the introns, but they differ significantly over the beta chains constitute HbA, which in normal adult hemoglobin chains combine with delta chains to constitute HbA1, which in normal adult hemoglobin chains combine with delta chains to constitute HbA1c, which is the major form of hemoglobin in the blood. Alpha thalassemia is a group of inherited disorders characterized by a deficiency of alpha-globin chains, resulting in a reduced number of red blood cells and anemia. Alpha thalassemia is caused by deletions of one or more of the alpha-globin genes. Alpha thalassemia is well as deletions of both HBA2 and HBA1; some nondeletion alpha thalassemias have also been reported. [provided by RefSeq]

Genomic view of HBA1

Genomic regions, transcripts, and products

(plus) Go to [reference sequence details](#) [Try our new Sequence Viewer](#)

Genomic context

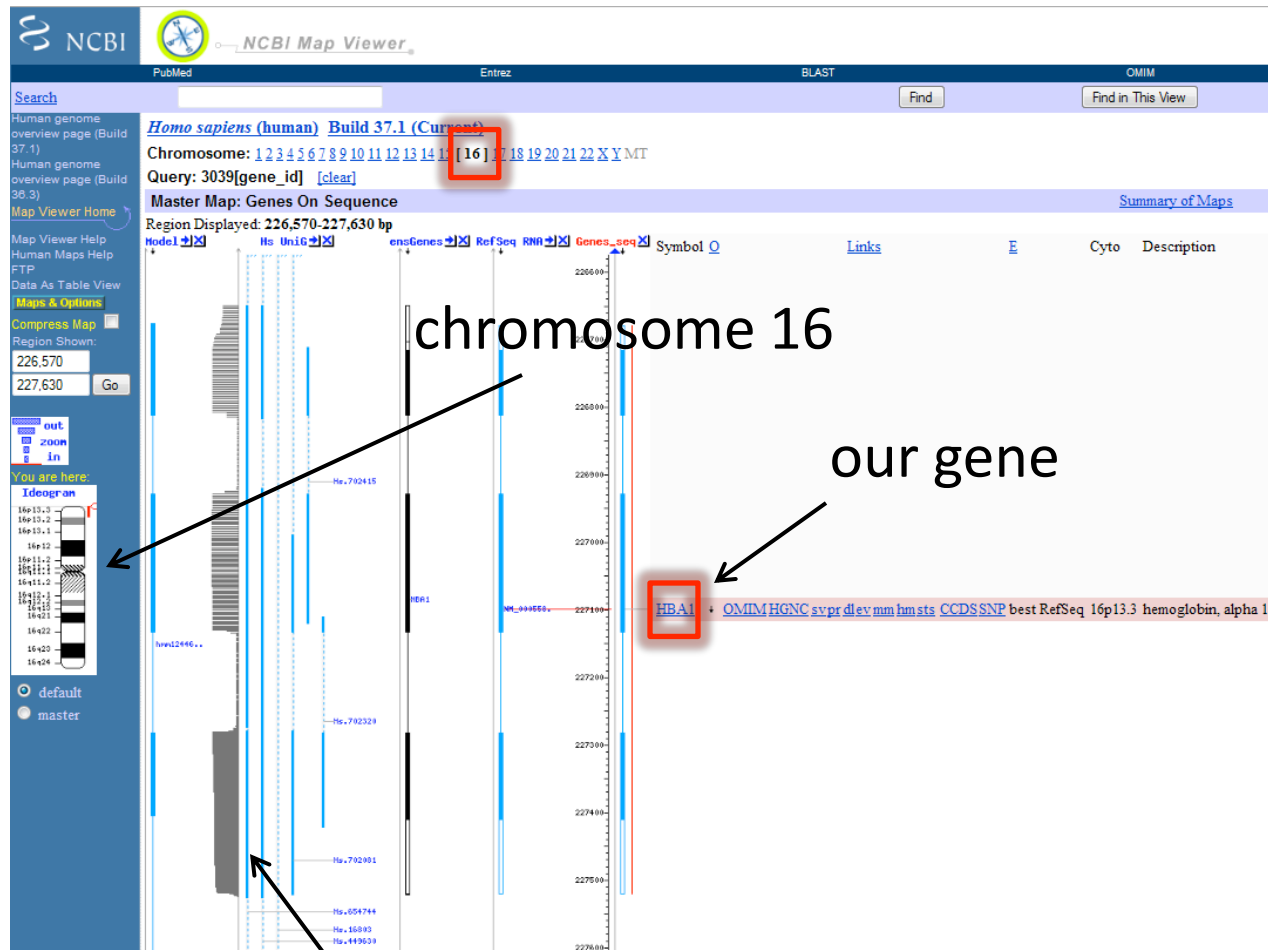
chromosome: 16; Location: 16p13.3

[See HBA1 in MapViewer](#)

click

Genomic context of HBA1

Where HBA1 falls on chromosome 16...



transcriptome

Genomic context

Genomic regions, transcripts, and products

(plus) Go to [reference sequence details](#)

[Try our new Sequence Viewer](#)

NC_000016.9

[226679 ▶] 5' [227520 ▶] 3'

[NM_000558.3](#) [NP_000549.1](#) [CCDS10399.1](#)

■ - coding region ■ - untranslated region

Genomic context

chromosome: 16; Location: 16p13.3

[See HBA1 in MapViewer](#)

[215973 ▶] [279449 ▶]

HBM ▶ HBA2 ▶ HBA1 ▶ LUC7L ◀

Bibliography

click

What is the length of ...

- The genomic sequence (NC_...)?
- The mRNA sequence (NM_...)?
- The protein sequence (NP_...)?

RefSeq codes:

NC_123456 Genomic Mixed Complete genomic molecules including genomes, chromosomes, organelles, plasmids.

NM_123456 mRNA Mixed Transcript products; mature messenger RNA (mRNA) transcripts. NP_123456

NP_123456789 Protein Mixed Protein products; primarily full-length precursor products but may include some partial proteins and mature peptide products.

Hint: Double click on gene then Right click → properties...

The screenshot displays the NCBI Nucleotide database interface. At the top, the search bar contains 'Nucleotide' and 'for'. Below the search bar, the format is set to 'GenBank'. The main content area shows the 'NC_000016.9 (90,354,753 bases)' sequence. A zoomed-in view of the '226,174 : 228,025 (1,852 bases shown, positive strand)' is visible. In this view, a gene model is shown with a red box around it. A red arrow points to the red box with the text 'double click'.

Genomic vs. mRNA vs. protein sequences

HBA1
total range: NC_000016.9 (226,679..227,520)
total length: 842

gene 226679..227520
/gene="HBA1"
/note="Derived by automated computational analysis using
gene prediction method: BestRefseq."
/db_xref="GeneID:3039"
/db_xref="HGNC:4823"
/db_xref="MIM:141800"

Genomic

NM_000558.3: mRNA-hemoglobin, alpha 1
total range: NC_000016.9 (226,679..227,520)
total length: 842
processed length: 576
mRNA product length: 576

mRNA join(226679..226810,226928..227132,227282..227520)
/gene="HBA1"
/product="hemoglobin, alpha 1"
/note="Derived by automated computational analysis using
gene prediction method: BestRefseq."
/transcript_id="NM_000558.3"
/db_xref="GeneID:3039"
/db_xref="GI:14456711"

mRNA

NP_000549.1: alpha 1 globin
total range: NC_000016.9 (226,716..227,410)
total length: 695
processed length: 429
protein product length: 142

CDS join(226716..226810,226928..227132,227282..227410)
/gene="HBA1"
/note="Derived by automated computational analysis using
gene prediction method: BestRefseq."
/codon_start=1
/product="alpha 1 globin"
/protein_id="NP_000549.1"
/db_xref="CCDS:CCDS10399.1"
/db_xref="GeneID:3039"
/db_xref="GI:4504347"

protein

Genomic vs. mRNA vs. protein sequences

HBA1
total range: NC_000016.9 (226,679..227,520)
total length: 842

gene 226679..227520
/gene="HBA1"
/note="Derived by automated computational analysis using
gene prediction method: BestRefseq."
/db_xref="GeneID:3039"
/db_xref="HGNC:4823"
/db_xref="MIM:141800"

Genomic

NM_000558.3: mRNA-hemoglobin, alpha 1
total range: NC_000016.9 (226,679..227,520)
total length: 842
processed length: 576
mRNA product length: 576

mRNA join(226679..226810,226928..227132,227282..227520)
/gene="HBA1"
/product="hemoglobin, alpha 1"
/note="Derived by automated computational analysis using
gene prediction method: BestRefseq."
/transcript_id="NM_000558.3"
/db_xref="GeneID:3039"
/db_xref="GI:14456711"

mRNA

Question 1

Why is the protein product length
 $142 \times 3 = 426\text{bp}$ shorter than the
protein processed length = 429bp?

NP_000549.1: alpha 1 globin
total range: NC_000016.9 (226,716..227,410)
total length: 695

processed length: 429
protein product length: 142

CDS join(226716..226810,226928..227132,227282..227410)
/gene="HBA1"
/note="Derived by automated computational analysis using
gene prediction method: BestRefseq."
/codon_start=1
/product="alpha 1 globin"
/protein_id="NP_000549.1"
/db_xref="CCDS:CCDS10399.1"
/db_xref="GeneID:3039"
/db_xref="GI:4504347"

protein

Genomic vs. mRNA vs. protein sequences

HBA1
total range: NC_000016.9 (226,679..227,520)
total length: 842

gene 226679..227520
/gene="HBA1"
/note="Derived by automated computational analysis using
gene prediction method: BestRefseq."
/db_xref="GeneID:3039"
/db_xref="HGNC:4823"
/db_xref="MIM:141800"

Genomic

NM_000558.3: mRNA-hemoglobin, alpha 1
total range: NC_000016.9 (226,679..227,520)
total length: 842
processed length: 576
mRNA product length: 576

mRNA join(226679..226810,226928..227132,227282..227520)
/gene="HBA1"
/product="hemoglobin, alpha 1"
/note="Derived by automated computational analysis using
gene prediction method: BestRefseq."
/transcript_id="NM_000558.3"
/db_xref="GeneID:3039"
/db_xref="GI:14456711"

mRNA

Question 1

Why is the protein product length
 $142 \times 3 = 426\text{bp}$ shorter than the
protein processed length = 429bp?

The stop codon was removed

NP_000549.1: alpha 1 globin
total range: NC_000016.9 (226,716..227,410)
total length: 695

processed length: 429
protein product length: 142

CDS join(226716..226810,226928..227132,227282..227410)
/gene="HBA1"
/note="Derived by automated computational analysis using
gene prediction method: BestRefseq."
/codon_start=1
/product="alpha 1 globin"
/protein_id="NP_000549.1"
/db_xref="CCDS:CCDS10399.1"
/db_xref="GeneID:3039"
/db_xref="GI:4504347"

protein

Genomic vs. mRNA vs. protein sequences

HBA1
total range: NC_000016.9 (226,679..227,520)
total length: 842

gene 226679..227520
/gene="HBA1"
/note="Derived by automated computational analysis using
gene prediction method: BestRefseq."
/db_xref="GeneID:3039"
/db_xref="HGNC:4823"
/db_xref="MIM:141800"

Genomic

NM_000558.3: mRNA-hemoglobin, alpha 1
total range: NC_000016.9 (226,679..227,520)
total length: 842
processed length: 576

mRNA product length: 576

mRNA join(226679..226810,226928..227132,227282..227520)
/gene="HBA1"
/product="hemoglobin, alpha 1"
/note="Derived by automated computational analysis using
gene prediction method: BestRefseq."
/transcript_id="NM_000558.3"
/db_xref="GeneID:3039"
/db_xref="GI:14456711"

mRNA

Question 2

Why is the mRNA length after splicing (576bp) longer than the protein processed length (429bp)?

NP_000549.1: alpha 1 globin
total range: NC_000016.9 (226,716..227,410)
total length: 695

processed length: 429

protein product length: 142

CDS join(226716..226810,226928..227132,227282..227410)
/gene="HBA1"
/note="Derived by automated computational analysis using
gene prediction method: BestRefseq."
/codon_start=1
/product="alpha 1 globin"
/protein_id="NP_000549.1"
/db_xref="CCDS:CCDS10399.1"
/db_xref="GeneID:3039"
/db_xref="GI:4504347"

protein

Genomic vs. mRNA vs. protein sequences

HBA1
total range: NC_000016.9 (226,679..227,520)
total length: 842

gene 226679..227520
/gene="HBA1"
/note="Derived by automated computational analysis using
gene prediction method: BestRefseq."
/db_xref="GeneID:3039"
/db_xref="HGNC:4823"
/db_xref="MIM:141800"

Genomic

NM_000558.3: mRNA-hemoglobin, alpha 1
total range: NC_000016.9 (226,679..227,520)
total length: 842
processed length: 576

mRNA product length: 576

mRNA join(226679..226810,226928..227132,227282..227520)
/gene="HBA1"
/product="hemoglobin, alpha 1"
/note="Derived by automated computational analysis using
gene prediction method: BestRefseq."
/transcript_id="NM_000558.3"
/db_xref="GeneID:3039"
/db_xref="GI:14456711"

mRNA

Question 2

Why is the mRNA length after splicing (576bp) longer than the protein processed length (429bp)?

The protein sequence is the sequence between the start and stop codons. The mRNA includes an additional untranslated 5' region and an untranslated 3' regions

NP_000549.1: alpha 1 globin
total range: NC_000016.9 (226,716..227,410)
total length: 695

processed length: 429

protein product length: 142

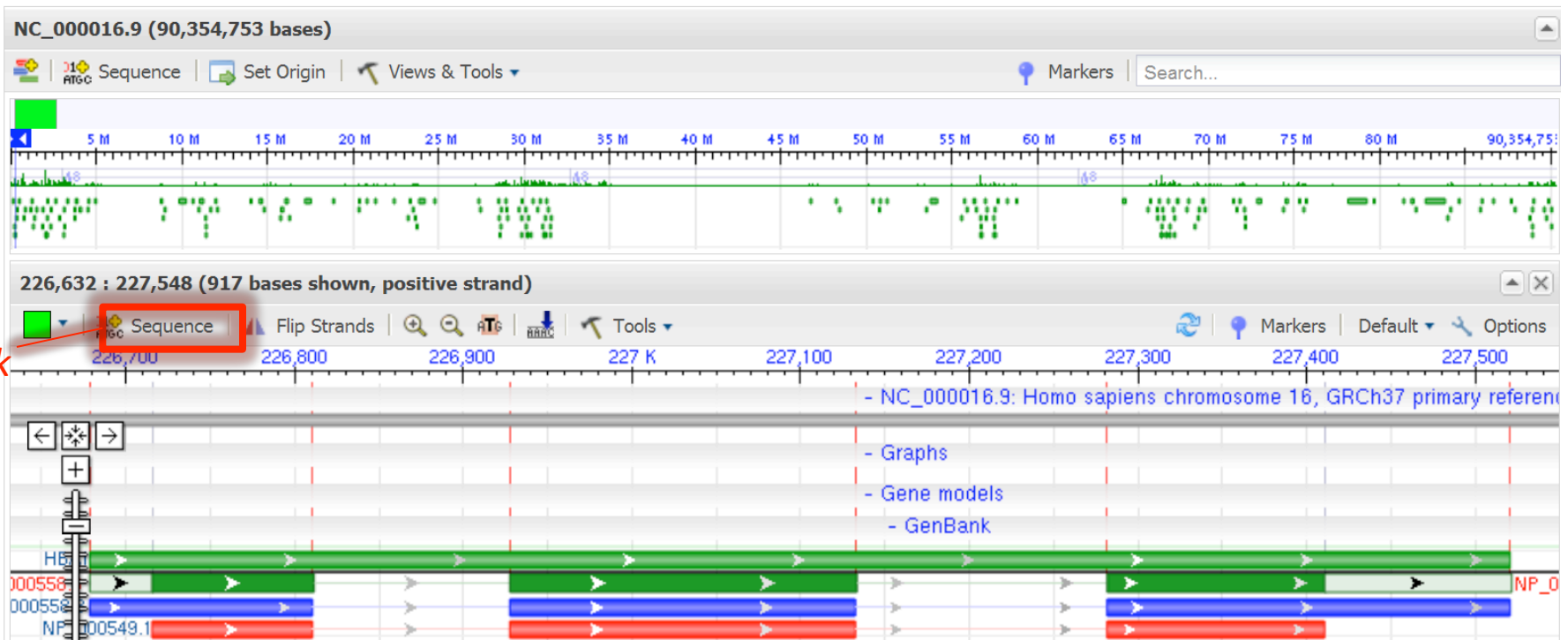
CDS join(226716..226810,226928..227132,227282..227410)
/gene="HBA1"
/note="Derived by automated computational analysis using
gene prediction method: BestRefseq."
/codon_start=1
/product="alpha 1 globin"
/protein_id="NP_000549.1"
/db_xref="CCDS:CCDS10399.1"
/db_xref="GeneID:3039"
/db_xref="GI:4504347"

protein

Let's look at the sequence

Homo sapiens chromosome 16, GRCh37 primary reference assembly

[Link To This Page](#) | [Help](#) | [Feedback](#) | [Printer-Friendly Page](#)



Let's look at the sequence

Untranslated 5' region of the mRNA (blue)

ATG Start codon Exon (red)

5' splice site (GU)

Intron (green)

3' splice site (AG)

Homo sapiens chromosome 16, GRCh37 primary reference assembly

[Link To This Page](#) | [Help](#) | [Feedback](#) | [Printer-Friendly Page](#)

Sequence View (positive strand)

NC_000016.9 (90,354,7...)

HBA2/NM_000517.4/NP_000508.1

226101 GCTGCAGGAAGCGAGGCTGGAGAGCAAGAGGGGCTCTGCGCAGAAATCTTTTGTGTTCTATGGGCCAAGGGCTCCGGGTGCCCGCATT

226191 CCTCTCGCCCCAGGATTTGGGCGAAGCCCTCCGGCTCGCACTCGCTCGCCCGTGTGTTCCCGATCCCGCTCGATCGATGCCGCTCCAG

226281 CGCGTGCCAGGCCGGGGCGGGGTGCGGGCTCACTTCTCCCTCGCTAGGACGCTCCGGCGCCGAAAGGAAAAGGTGGCGCTCGCTC

226371 CGGGGTGCAAGAGCCGACAGCCGACCCCAACCGCCGGCCCGCCAGCAGCCCGCTACCGCCCTGCCCGCGGGCCGCGGATGGCGG

226461 GAGTGGAGTGGCGGGTGGAGGGTGGAGACGTCCTGCCCCCGCCCGCGTGCACCCCAAGGGGAGGCCGACCCCGCGCGCCGGCCCGCG

226551 CAGGCCCGCCCGGACTCCCCTGCGGTCCAGGCCGCGCCCGGGCTCCGCGCCAGCCAATGAGCGCCGCCCCGCGCGCCGCCCCCGC

226641 GCCCAAGCATAAACCCCTGGCGCGCTCGCGGCCCGGCACTCTTCTGGTCCCCACAGACTCAGAGAGAACCACCATGCTGCTGCTCCTG

226731 CCGACAAGACCAACGTC AAGGCCGCGCTGGGTAAGGTCCGCGCGCACGCTGGCGAGTATGGTGCAGGAGGCCCTGGAGACCGT GAGGCTCCC

A D K T N V K A A W G K V G A H A G E Y G A E A L E R

226821 TCCCTGCTCCGACCCCGGCTCCTGCGCCCGCCGACCCACAGGCCACCCCTCAACCGTCTGCGCCCGGACCCAAACCCACCCCTCACT

226911 CTGCTTCTCCCGCAGATGTTCTGCTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCAGGTTA

M F L S F P T T K T Y F P H F D L S H G S A Q V

227001 AGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCGGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACC

K G H G K K V A D A L T N A V A H V D D M P N A L S A L S D

227091 TGCACGCGCACAGCTTCCGGTGGACCCGGTCAACTTCAAGCTGAGCGCGGGCCGGGAGCGATCTGGGTTCGAGGGGCGAGATGGCGCCT

L H A H K L R V D P V N F K

227181 TCCTCGCAGGGCAGAGGATCACGCGGGTTCGGGAGGTGTAGCGCAGGCGCGGCTGCGGGCCTGGGCCCTCGGCCCACTGACCCCTCTT

227271 CTCTGCACAGCTCCTAAGCCAATGCTGCTGGTGAACCTGGCCGCCCACTCCCGCCGAGTTACCCCTCGGGTGCACGCTCCCTGGGA

L L S H C L L V T L A A H L P A E F T P A V H A S L D

227361 CAAGTTCCTGGCTTCTGTGAGCACCGTCTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTGGCCATGCTTCTGGCCCTTGGGCCTC

K F L A S V S T V L T S K V K

227451 CCCCAGCCCTCCTCCCTTCTGCAACCGTACCCCGTGGTCTTTGAAATAAGTCTGAGTGGGCGGCAGCCCTGTGTGCTGCTGAGTTT

227541 TTCCCTCAGCAACCGTCCAGGCATGGGCTGGAGCAGCTGGGATACACATGGCTAGAACCTCTCTGAGCTGGATAGGGTAGGAAA

227631 AGGCAGGGGCGGGAGGAGGGATGGAGGAGGAAAGTGGAGCCACCAGAGTCCAGCTGGAAAAACGCTGGACCCTAGAGTGCTTTGAG

227721 GATGCATTTGCTCTTTCCCGATTTTATTCCAGACTTTTTCAGATCAATGACAGTTTGTGAAATAATGAATTTATCCATCTTTACGTT

227811 TCTGGGCACTCTGTCCCAAGAAGTGGCTGGCTTCTGCTGGGAGCACTGCTGTTTCCAGAGGTCCTCCACATATGGGTGGTGGGTA

227901 GGTACAGAGAAGTCCACTCCAGCATGGCTGATTGATCCCTCATCGTTCCCACTAGTCTCCGTAAACCTCCAGATACAGGCACAGCT

227991 AGATCAATCAGGGGTGCGGGTGCACCTGACAGGCCCAAGCAATTCAATAGGGGCTCTACTTTACCCCCAGGTCACCCAGAAATGCTC

228081 ACACACCAGACACTGACGCCCTGGGGCTGTCAAGATCGGGCTTTGTCTGGGCCAGCTCAGGGCCAGCTCAGCACCCTCAGCTC

click

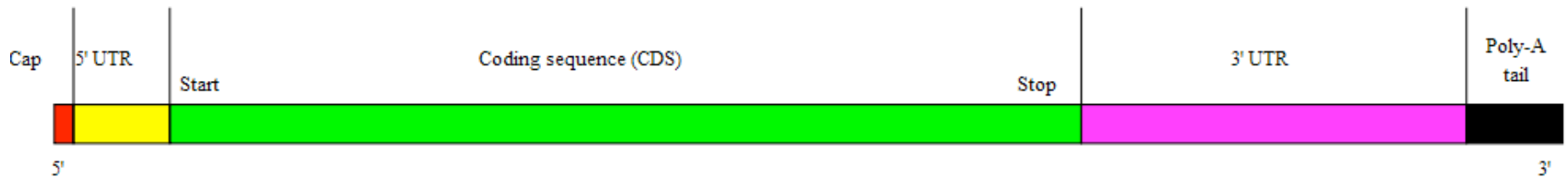
UAA stop codon

Untranslated 3' region of the mRNA (blue)

Total length mRNA = blue + red = 576 bp
 Total length protein = red = 429 bp

Typical Eukaryotic mRNA

The structure of a typical human protein coding mRNA including the untranslated regions (UTRs)



We would like to find similar proteins in nature

Which sequence should we use for the search?

- Genomic?
- mRNA ?
- Protein?

OK let's grab the aa sequence of the protein

The image shows a workflow on the NCBI website to retrieve a protein sequence. It starts with the Entrez Gene page for HBA1 (hemoglobin, alpha 1) in Homo sapiens. A red arrow points from the 'Summary' section to the 'mRNA and Protein(s)' section in the RefSeq page. In the RefSeq page, a red box highlights the protein entry 'NP_000549.1 hemoglobin subunit alpha' with a red arrow and the word 'click'. A second red arrow points from this box to the 'FASTA' link in the 'Format' dropdown menu of the protein's detail page. A final red arrow points from the 'FASTA' link to a box containing the amino acid sequence.

Entrez Gene Summary:

- GeneID: 3039
- Official Symbol: HBA1
- Official Full Name: hemoglobin, alpha 1
- Primary source: HGNC:4823
- See related: Ensembl:ENSG00000206172; HPRD:
- Gene type: protein coding
- RefSeq status: REVIEWED
- Organism: *Homo sapiens*
- Lineage: Eukaryota; Meta; Chordata; Cr; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
- Also known as: CD31; MGC126895; MGC126897; HBA1
- Summary: The human alpha globin gene cluster located on chromosome 16 spans about 30 kb and in 5'- zeta - pseudozeta - mu - pseudoalpha-1 - alpha-2 - alpha-1 - theta - 3'. The alpha-1 (HBA1) coding sequences are identical. These genes differ slightly over the 5' untranslated regions. Two alpha chains constitute HbA, which in normal adult life comprises about 97% of the total hemoglobin. The remaining 3% of adult hemoglobin is HbA2, which with HbF (fetal hemoglobin) constitutes the remaining 3% of adult hemoglobin. Alpha thalassemias result from deletions of each of the alpha genes. Alpha thalassemias have also been reported as deletions of both HBA2 and HBA1; some nondeletion alpha thalassemias have also been reported [provided by RefSeq]

RefSeq mRNA and Protein(s):

- 1. **NM_000558.3** - **NP_000549.1** hemoglobin subunit alpha
- Source sequence: [FASTA](#) [Graphics](#)
- Consensus CDS: [CCDS_3399.1](#)
- UniProtKB/Swiss-Prot: [P69905](#)
- Related Ensembl: [ENSP0000022421](#), [ENST00000320868](#), [ENST00000397797](#)
- Conserved Domains (1): [summary](#)
- Location: chr16:1101040-1101137
- Blast Score: 1362
- Description: globin; Globins are heme proteins, which bind and transport oxygen. This family summarizes a diverse set of homologous protein domains, including: (1) tetrameric vertebrate hemoglobins, which are the major protein component of erythrocytes and transport oxygen...

RefSeqs of Annotated Genomes: Build 37.1

The following sections contain reference sequences that belong to the following sections:

- Genomic
- mRNA and Protein(s)
- Protein

Protein Detail Page:

- Search: Protein
- Format: GenPept | **FASTA** | Graphics | More Formats
- NCBI Reference Sequence: NP_000549.1
- hemoglobin subunit alpha [*Homo sapiens*]
- Links: [Comments](#) | [Features](#) | [Sequence](#)

NCBI Reference Sequence: NP_000549.1

hemoglobin subunit alpha [*Homo sapiens*]

```
>gi|4504347|ref|NP_000549.1| hemoglobin subunit alpha [Homo sapiens]
MVLSPADKTNVKAAWGKVGAHAGEYGAELERMFLSFPTTKTYFPHFDLSHGSQAQVKGHGKQVADALNTNA
VAHVDDMPNALSALSDLHAHKLRLVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLQKFLASVSTVLTSLK
YR
```

000549 142 aa line
hemoglobin subunit alpha [*Homo sapiens*].
000549
000549.1 GI:4504347
SEQ: accession NM_000558.3

and the nt sequence of the protein

The image shows a workflow for finding the protein sequence from a gene entry on NCBI. It starts with the Entrez Gene page for HBA1 (GeneID: 3039), which lists various sources like HGNC:4823 and Ensembl. A red arrow points from the 'Summary' section to the 'mRNA and Protein(s)' section in the RefSeq page. In the RefSeq page, the entry 'NM_000558.3 - NP_000549.1 hemoglobin subunit alpha' is highlighted. A red box around 'NP_000549.1' is labeled 'click'. Another red box around the 'CDS' link is also labeled 'click'. A third red box around the 'FASTA' format option in the 'Format' dropdown is labeled 'click'. The final step shows the resulting FASTA sequence for the CDS region (bases 38 to 466).

Entrez Gene Summary:

- Official Symbol: HBA1
- Official Full Name: hemoglobin, alpha 1
- Primary source: HGNC:4823
- See related: Ensembl:ENSG00000206172; HPRD:
- Gene type: protein coding
- RefSeq status: REVIEWED
- Organism: *Homo sapiens*
- Lineage: Eukaryota; Meta; Chordata; Cr
- Also known as: CD31; MGC126895; MGC126897; HBA1

RefSeq mRNA and Protein(s):

- 1. **NM_000558.3** - **NP_000549.1** hemoglobin subunit alpha

FASTA Sequence:

```
>gi|14456711|38-466 Homo sapiens hemoglobin, alpha 1 (HBA1), mRNA
ATGGTGGCTCTCTCTGCCGACAAGACCAAGCTCAAGCGCGCTGGGTAAAGTCGGCGGCACGCTGGCG
AGTATGGTGGCGAGGCCCTGGAGAGGATGTTCTCTCTCTCCACCACCAAGACCTACTTCCCGCACT
CGACCTGAGCCACGCTCTGCCAGGTTAAGGGCCACGSCAAGAAGTGGCCGACGCGTACCAACGCC
GTGGCGCAGTGGACGACATGCCAACCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGG
TGGACCCGGTCAACTTCAAGCTCTAAGCCACTGCTGTGGTACCCTGGCCGCCACCTCCCGCCGA
GTTACCCCTGGGGTGCACGCTCCTGGACAAGTCTCTGGCTCTGTGAGCACCGTGCTGACCTCCAA
TACCGTTAA
```

Nucleotide versus amino acid sequences

> NP_000549.1 Homo sapiens nt (429)

ATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGG
 CCCTGGAGAGGATGTTCTGTCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCAGGTTAAGGGCCA
 CGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCAACGCGCTGTCCGCCCTGAGCGACCTGCAC
 GCGCACAAGCTTCGGGTGGACCCGGTCAACTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCTGGCCGCCACCTCCCCGCCGAGTTC
 ACCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGTGACCTCAAATACCGT**TAA**

> NP_000549.1 Homo sapiens aa (142)

MVLSPADKTNVKAAWGKVGHAHAGEYGAELERMFLSPTTKTYFPHFDSLHGSAQVKGHGKKVADALNAVAHVDDMPNALSALSDLHA
 HKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR

		Second letter							
		U		C		A		G	
U	UUU	Phenylalanine	UCU	Serine	UAU	Tyrosine	UGU	Cysteine	U
	UUC		UCC		UAC		UGC		C
	UUA	Leucine	UCA		UAA	Stop codon	UGA	Stop codon	A
	UUG		UCG		UAG		UGG	Tryptophan	G
C	CUU	Leucine	CCU	Proline	CAU	Histidine	CGU	Arginine	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Glutamine	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Isoleucine	ACU	Threonine	AAU	Asparagine	AGU	Serine	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lysine	AGA	A	
	AUG	Methionine; start codon	ACG		AAG		AGG	Arginine	G
G	GUU	Valine	GCU	Alanine	GAU	Aspartic acid	GGU	Glycine	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glutamic acid	GGA		A
	GUG		GCG		GAG		GGG		G

Nucleotide versus amino acid sequences

> NP_000549.1 Homo sapiens nt (429)

ATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGG
CCCTGGAGAGGATGTTCTGTCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCAGGTTAAGGGCCA
CGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCAACGCGCTGTCCGCCCTGAGCGACCTGCAC
GCGCACAAGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCACCTCCCCGCCGAGTTC
ACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCAAATACCGT**TAA**

> NP_000549.1 Homo sapiens aa (142)

MVLSPADKTNVKAAWGKVGHAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNVAHVDDMPNALSALSDLHA
HKLRVDPVNFKLLSHCLLVTLAAHLPAEFPAVHASLDKFLASVSTVLTSKYR

- Which sequence should we use to search with, the amino acid sequence or the nucleotide sequence?

Nucleotide versus amino acid sequences

> NP_000549.1 Homo sapiens nt (429)

```
ATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGG  
CCCTGGAGAGGATGTTCTGTCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCA  
CGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCAACGCGCTGTCCGCCCTGAGCGACCTGCAC  
GCGCACAAGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCACCTCCCCGCCGAGTTC  
ACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCAAATACCGTTAA
```

> NP_000549.1 Homo sapiens aa (142)

```
MVLSPADKTNVKAAWGKVGVAHAGEYGAELERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKQVADALTNVAHVDDMPNALSALSDLHA  
HKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
```

- Which sequence should we use to search with, the amino acid sequence or the nucleotide sequence?

It depends on your goal, but generally to find homologs, aa sequences is the way to go:

- Selection pressure on amino acid sequence is much stronger than on nt sequence
- Two random nt sequences share 25% (50% with gaps) identical characters whereas amino acids sequences share only 5% (10-15% with gaps), improving the signal to noise considerably <we'll see this later...>