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

Try Beta | Log in / create account

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.

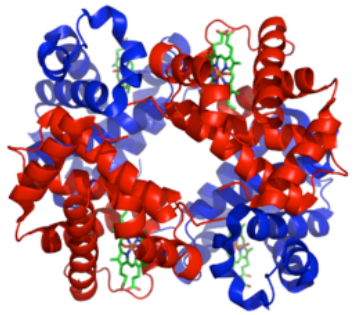
Hemoglobin has an oxygen binding capacity between 1.36 and 1.37 ml O_2 per gram of hemoglobin,[2] which increases the total blood oxygen capacity seventyfold.[3]

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

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)

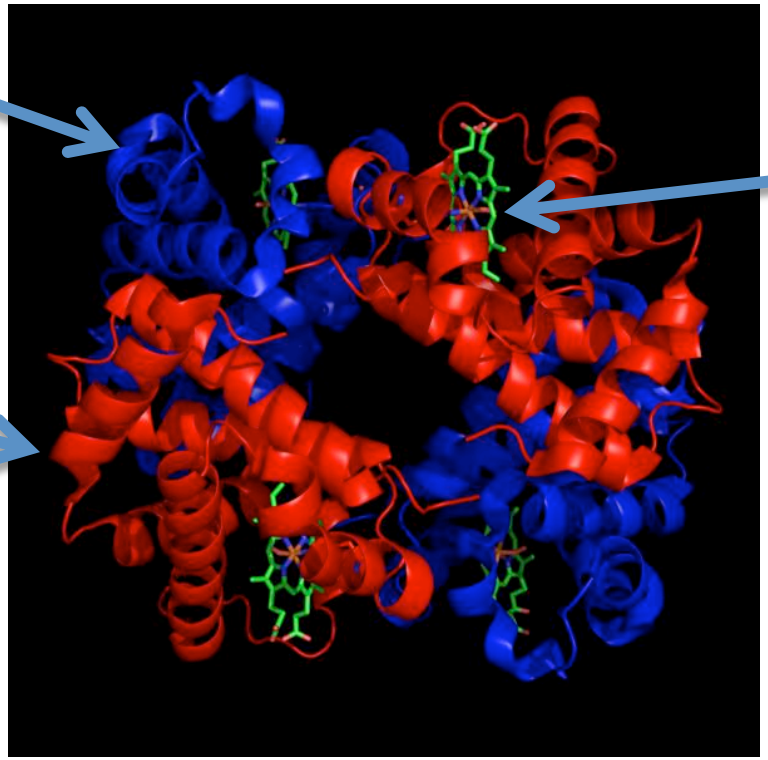
Subunit Name	Gene	Chromosomal Locus
Hb-α1	HBA1	Chr. 16 p13.3
Hb-α2	HBA2	Chr. 16 p13.3
Hb-β	HBB	Chr. 11 p15.5

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.

β subunit (146 aa)

α subunit (141 aa)



Iron-containing heme group

Further reading Ch. 7 of Stryer

Try Beta Log in / create account

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.

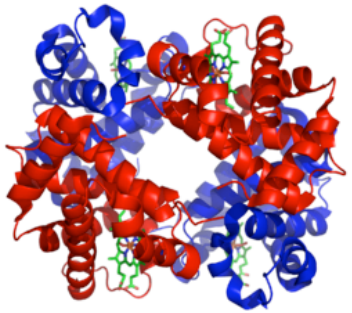
Hemoglobin has an oxygen binding capacity between 1.36 and 1.37 ml O_2 per gram of hemoglobin,[2] which increases the total blood oxygen capacity seventyfold.[3]

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

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)

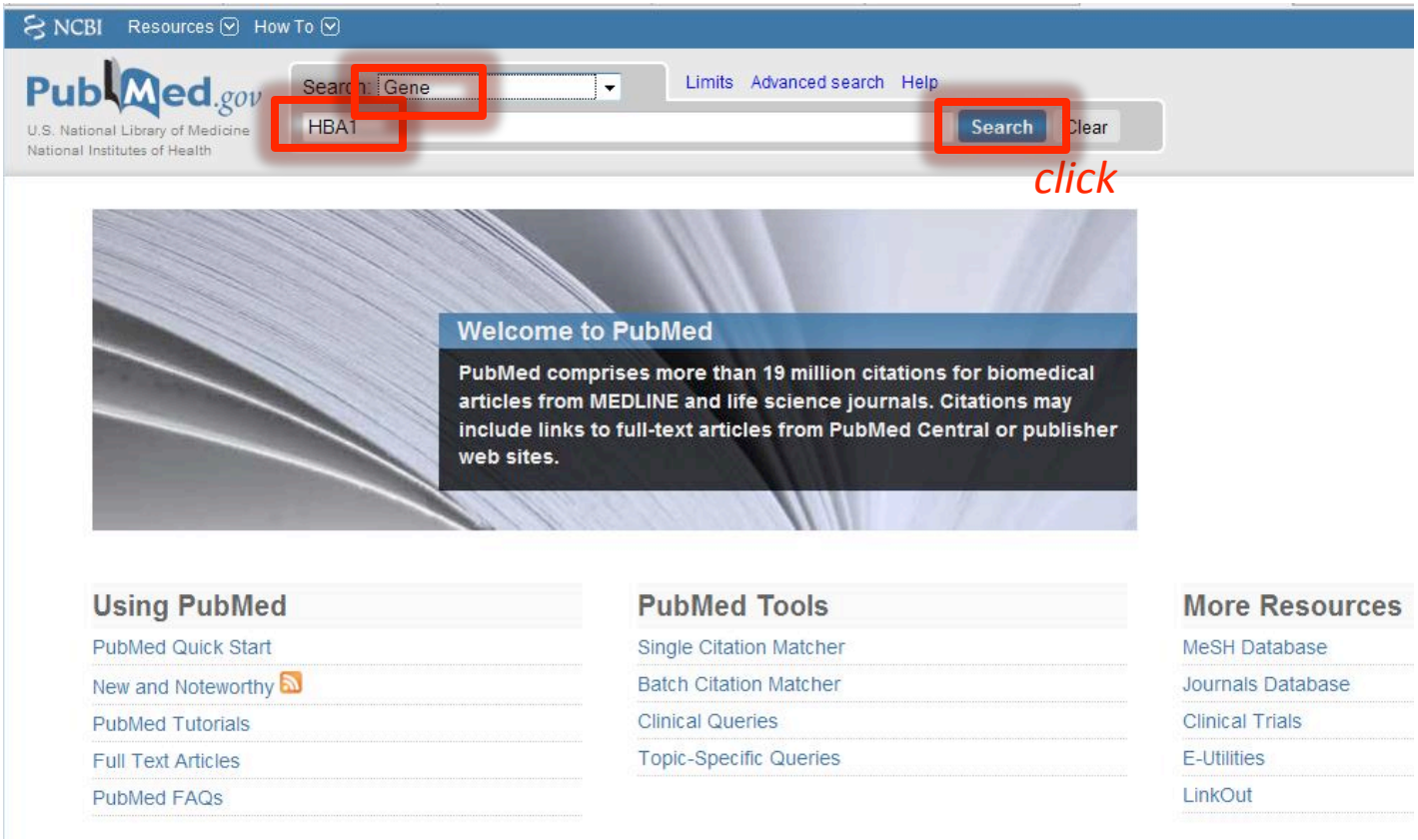
Subunit	Gene	Chromosomal Locus
Hb-α1	HBA1	Chr. 16 p13.3
Hb-α2	HBA2	Chr. 16 p13.3
Hb-β	HBB	Chr. 11 p15.5

Let's focus on the α1 subunit of hemoglobin. This gene is called **HBA1**

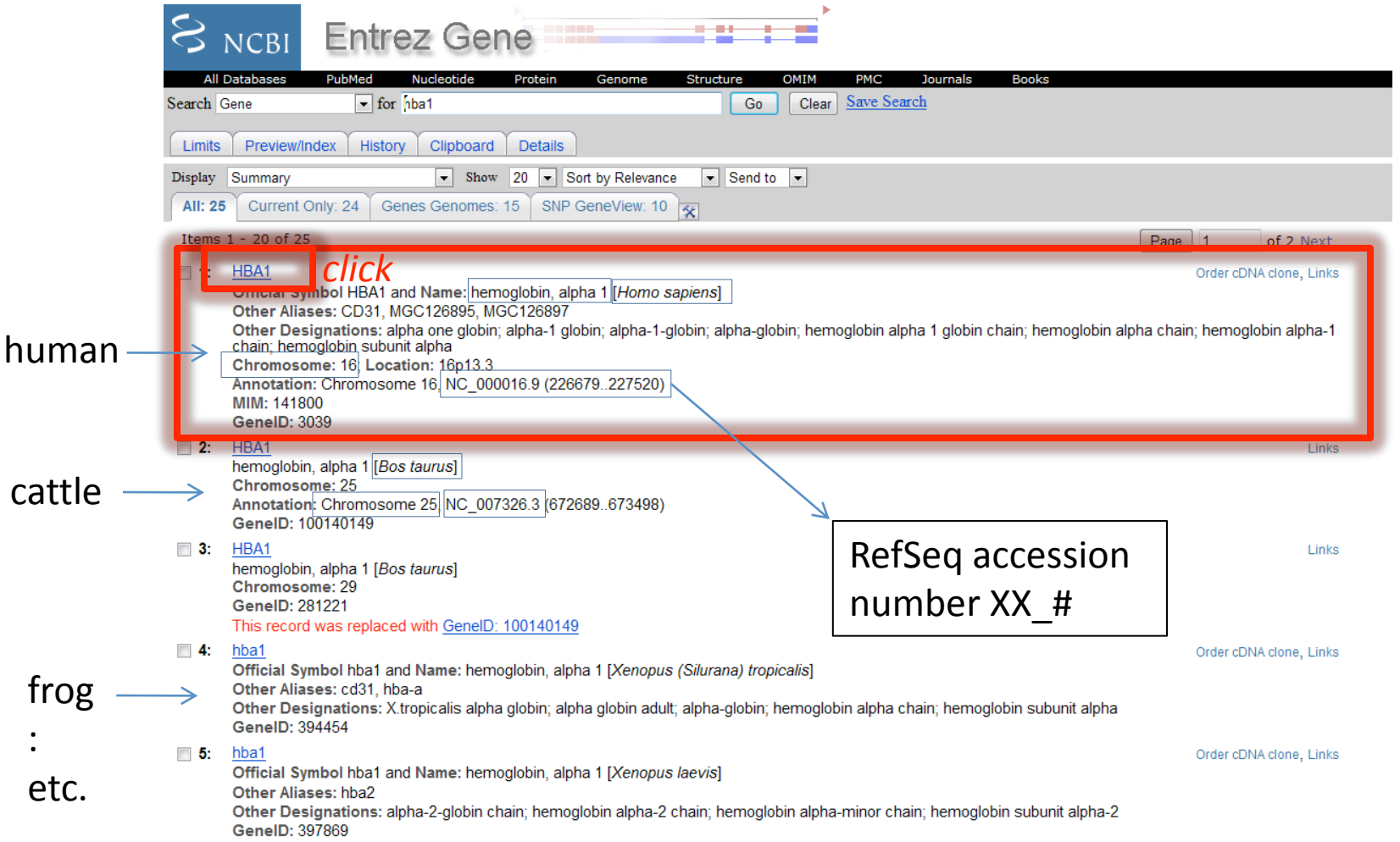
On which chromosome is this gene?

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



We'll start with human hemoglobin



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?

NCBI 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

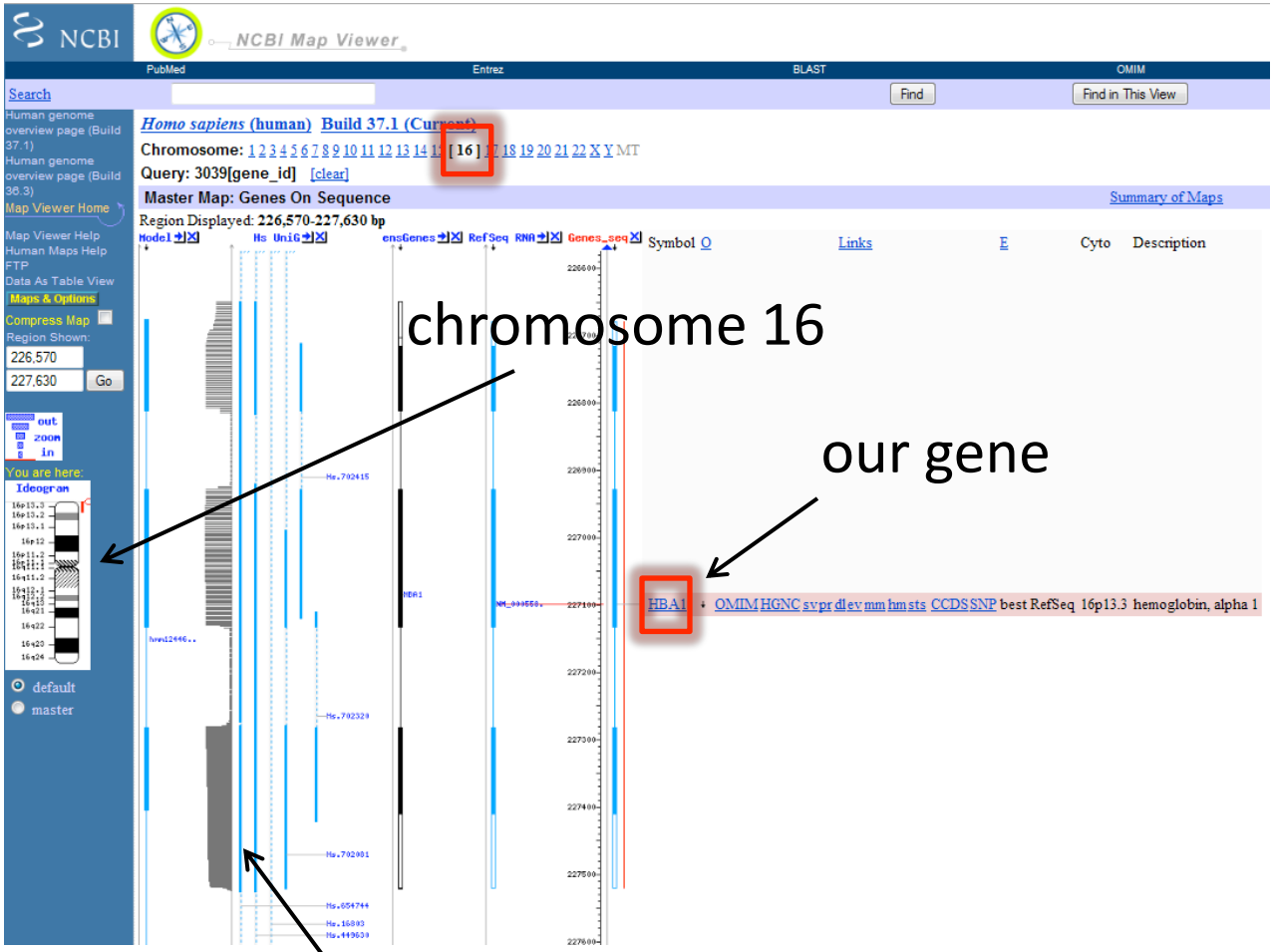
Official Symbol	HBA1
Official Full Name	hemoglobin, alpha 1
Primary source	HGNC:4823
See related	Ensembl:ENSG00000206172; HPRD:00784; MIM
Gene type	protein coding
RefSeq status	REVIEWED
Organism	*Homo sapiens*
Lineage	*Eukaryota; Metazoa; Chordata; Craniata; Vert Euarchontoglires; Primates; Haplorrhini; Catarr*
Also known as	CD31; MGC126895; MGC126897; HBA1
Summary	The human alpha globin gene cluster located o 5'- zeta - pseudozeta - mu - pseudoalpha-1 - alpha-1 (HBA1) coding sequences are identical and the introns, but they differ significantly ov beta chains constitute HbA, which in normal ad chains combine with delta chains to constitute remaining 3% of adult hemoglobin. Alpha thalas well as deletions of both HBA2 and HBA1; some nondeletion alpha thalassemias have also bee reported. [provided by RefSeq]

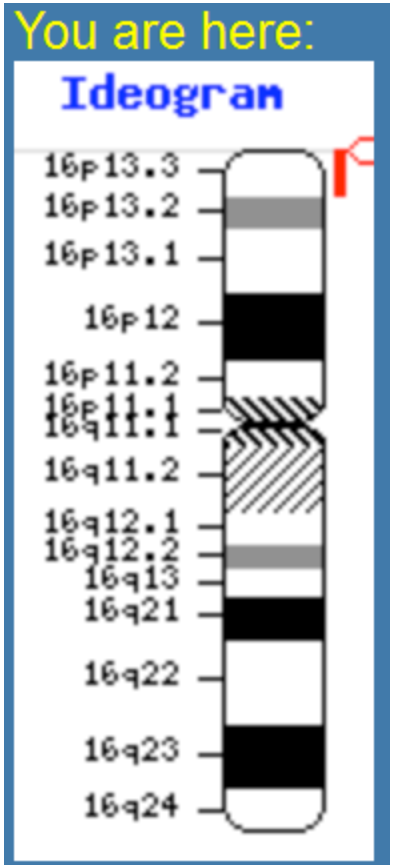
Genomic view of HBA1

click

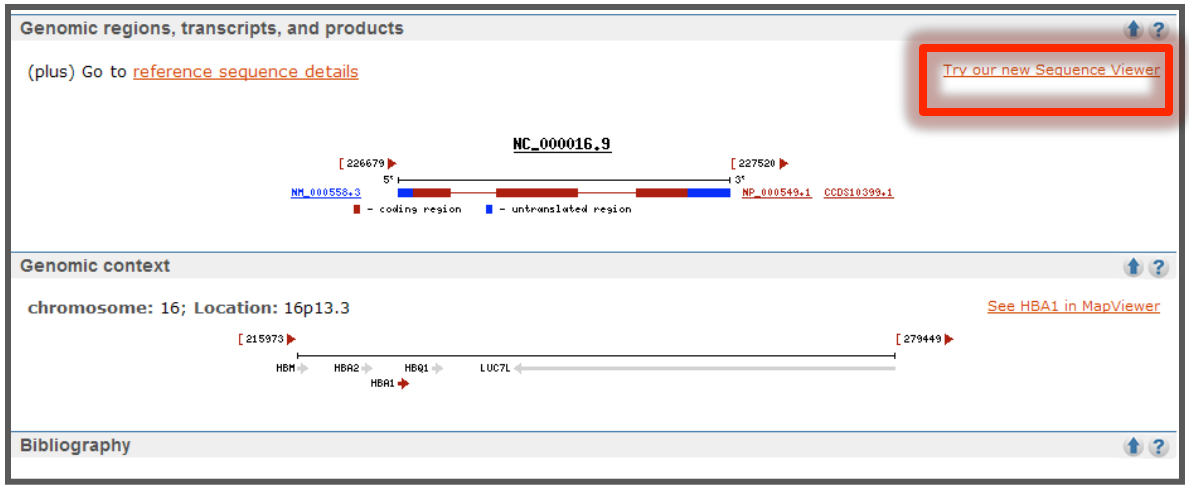
Genomic context of HBA1

Where HBA1 falls on chromosome 16…





Genomic context



click

What is the length of …

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

Hint: Double click on gene then Right click→ ***properties…***

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.

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 <u>product</u> length 142x3 = 426bp shorter than the protein <u>processed</u> 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 <u>product</u> length 142x3 = 426bp shorter than the protein <u>processed</u> 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

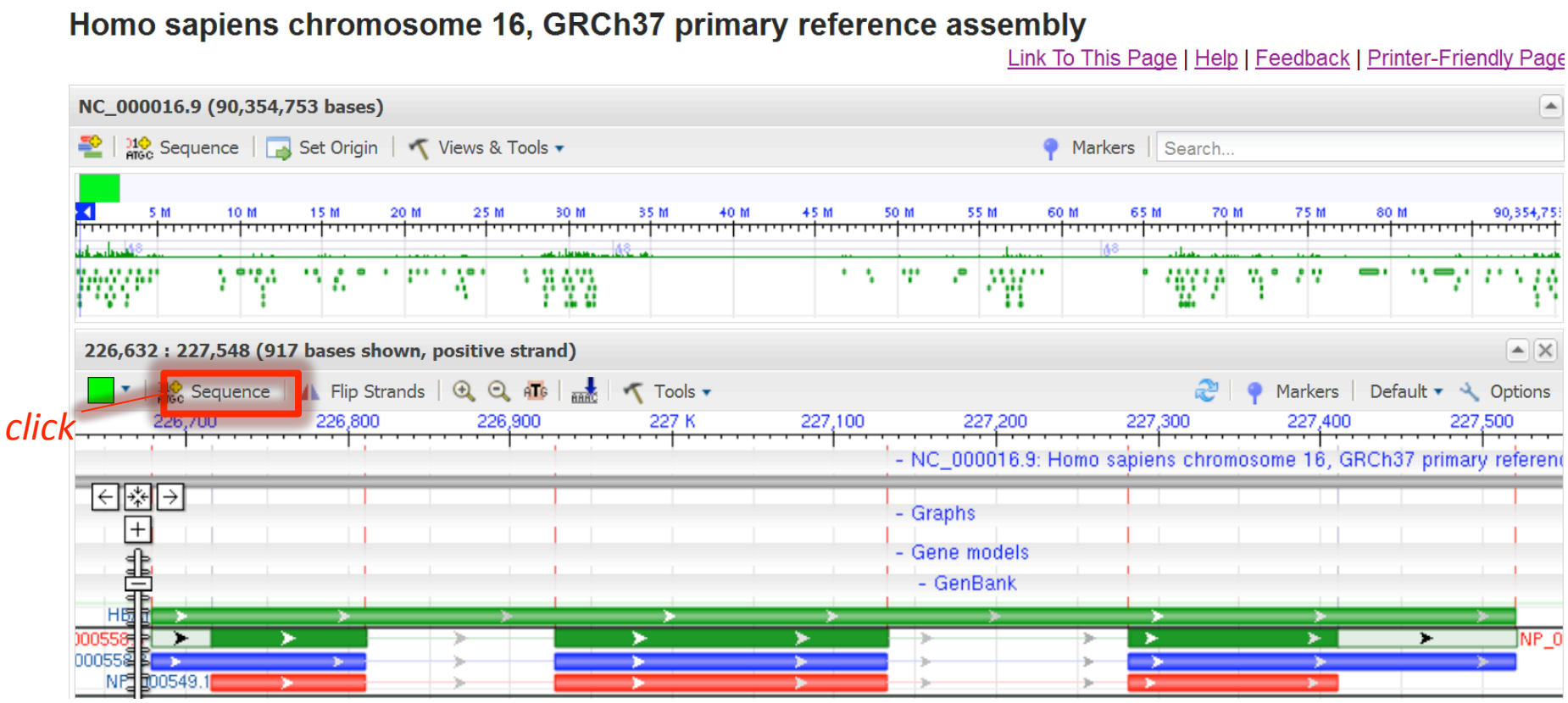
Question 2
Why is the mRNA length after splicing (576bp) longer than the protein processed length (429bp)?
The protein sequence is the sequence between the start and stop codons. The mRNA includes an additional untransalted 5' region and an untransalted 3' regions

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

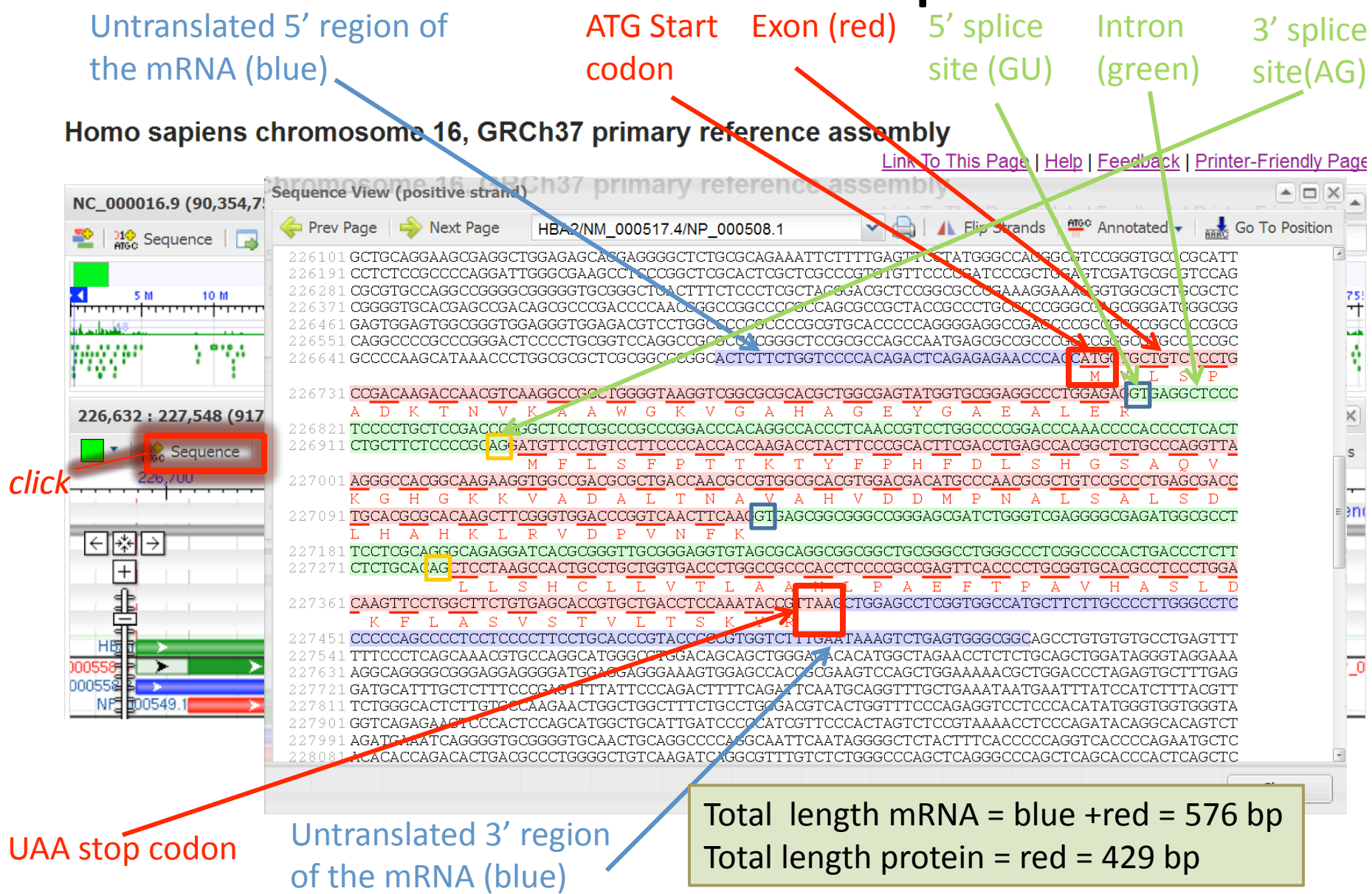
CDS join(226716..226810,226928..227132,227282..227410)
/gene="HBA1"
/note="Derived by automated computational analysis using
gene prediction method: BestRefseq."
/codon_start=1
/product="alpha 1 globin"
/protein_id="NP_000549.1"
/db_xref="CCDS:CCDS10399.1"
/db_xref="GeneID:3039"
/db_xref="GI:4504347"

protein

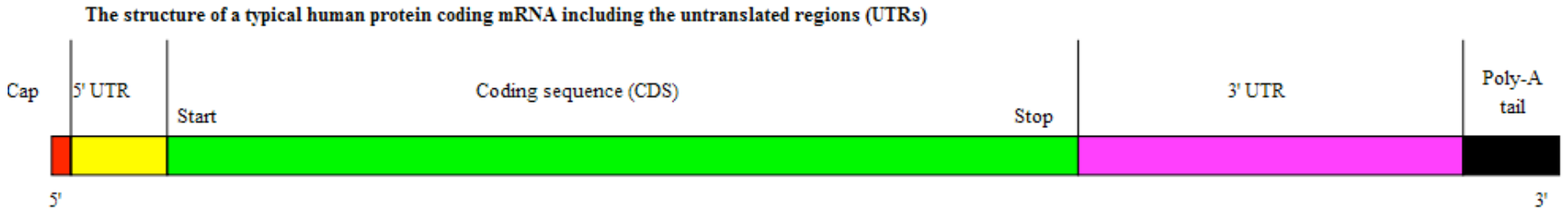
Let’s look at the sequence



Let's look at the sequence



Typical Eukaryotic mRNA

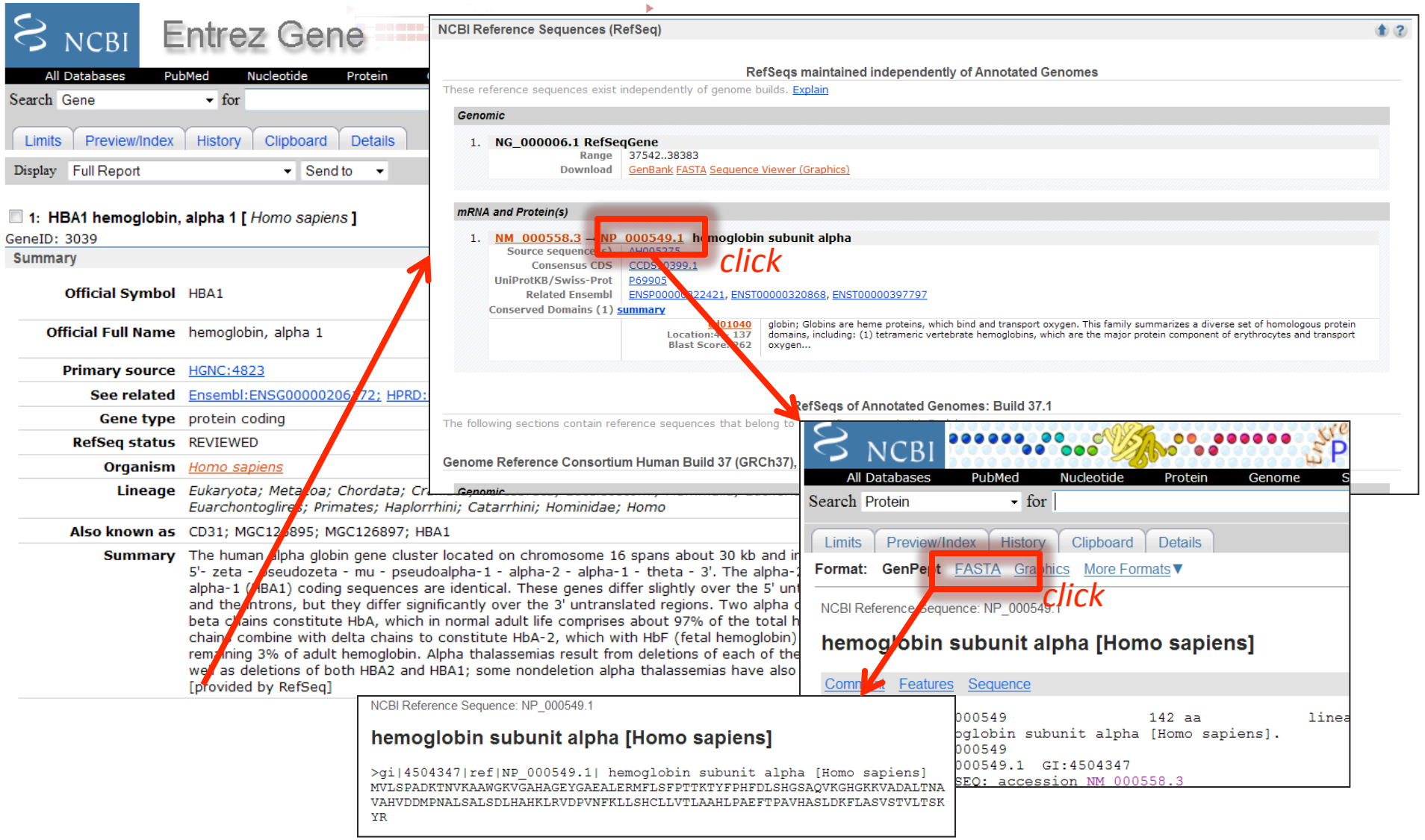


We would like to find similar proteins in nature

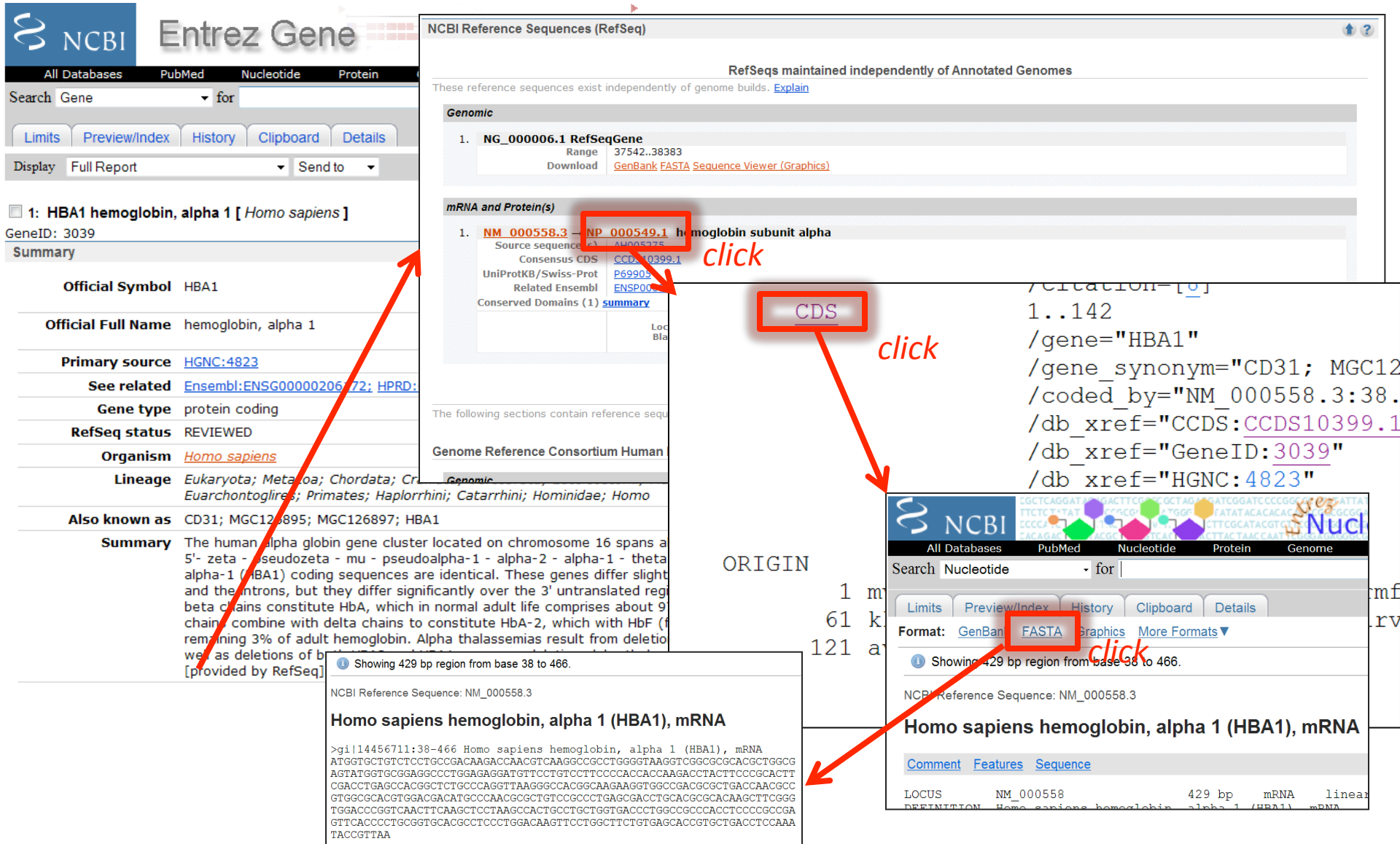
Which sequence should we use for the search?

- Genomic?
- mRNA ?
- Protein?

OK let's grab the aa sequence of the protein



and the nt sequence of the protein



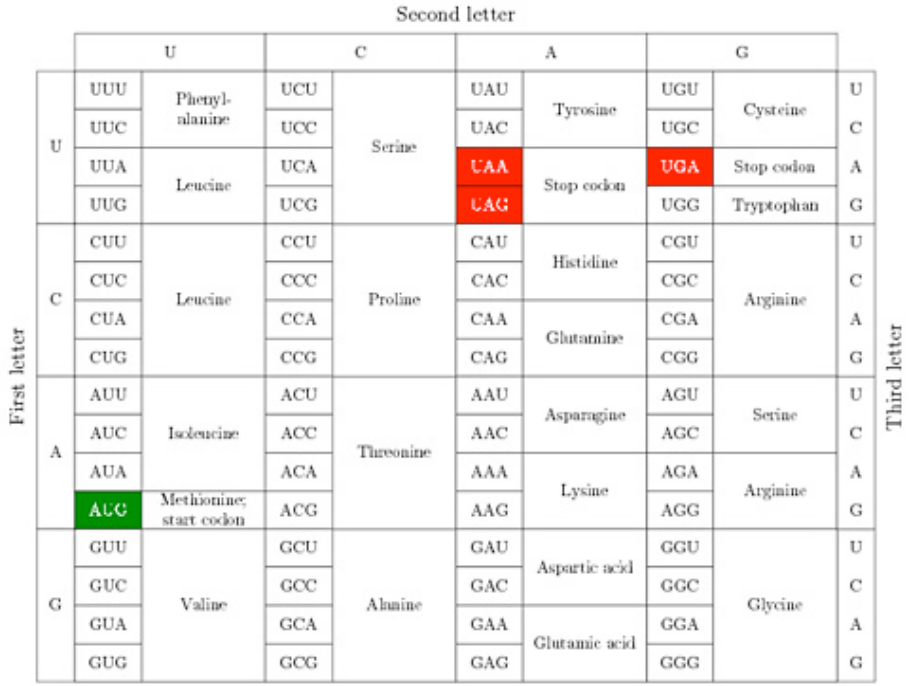
Nucleotide versus amino acid sequences

> NP_000549.1 Homo sapiens nt (429)

ATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGG
CCCTGGAGAGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCA
CGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCAC
GCGCACAAGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTC
ACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGT**TAA**

> NP_000549.1 Homo sapiens aa (142)

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHA
HKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR



Nucleotide versus amino acid sequences

> NP_000549.1 Homo sapiens nt (429)

ATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGG
CCCTGGAGAGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCA
CGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCAC
GCGCACAAGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTC
ACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGT**TAA**

> NP_000549.1 Homo sapiens aa (142)

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHA
HKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR

- 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
CCCTGGAGAGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCA
CGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCAC
GCGCACAAGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTC
ACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGT**TAA**

> NP_000549.1 Homo sapiens aa (142)

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHA
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…>