How to use the web for bioinformatics

Molecular Technologies

October 8, 2010

Ethan Strauss

ethan.strauss@promega.com

274-4330 X 1171

http://www.q7.com/~ethan

Objectives

At the end of this session you should be able to:

- Find sequences in Genbank
- Find additional information about those sequences
  - Peer reviewed citations
  - Single Nucleotide Polymorphisms
  - Etc.
- Design primers to amplify a specific region of DNA
- Understand BLAST

How to find all those dang URLs!

http://q7.com/~ethan/molbio/

Outline

- NCBI Databases
  - Nucleotide
  - Pubmed
  - Gene
  - SNP
- BLAST
- Primer BLAST
- Translation and other Utilities

NCBI Databases

- Nucleotide
  - Huge database of known DNA sequences
- Protein
  - Huge database of known protein sequences
- SNP
  - Database of Single Nucleotide Polymorphisms
- PubMed
  - Extensive database of scientific journal articles.
  - Many articles have full text free online
- Gene
  - Aggregation of data for genes

Nucleotide (& protein)

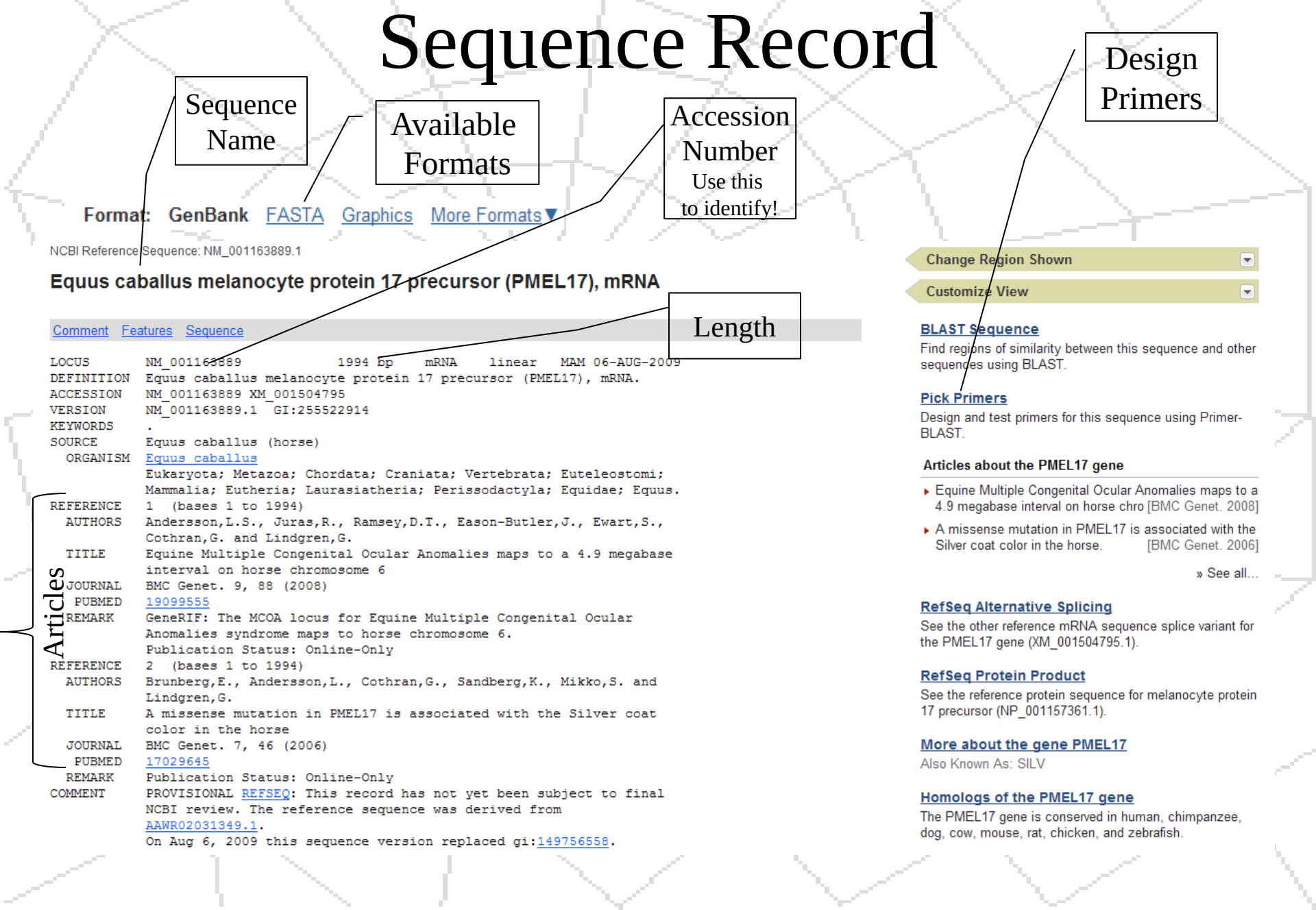
- NCBI has a nucleotide database (Genbank)
  - Synchronized with EMBL and DDBJ
- And a protein database
- Both have RefSeq designations
  RefSeq sequences are the most reliable

Sequences Databases
Sequence Records

- LOCUS Number Size Type Topology Division Date
- DEFINITION - Name of the Sequence
- ACCESSION - *Unique Id number* best for communication
- VERSION - Other numbers which are associated
- KEYWORDS
- SOURCE – What was it isolated from
- ORGANISM - More taxonomic detail
- REFERENCE - Paper or papers about the sequence
  - AUTHORS
  - TITLE
  - JOURNAL
- FEATURES - A complete list of all of the features of a sequence. Can be very extensive and useful!
- ORIGIN – The actual Sequence!

http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=58533118

Sequence Record



Format: GenBank FASTA Graphics More Formats▼

NCBI Reference Sequence: NM_001163889.1

Equus caballus melanocyte protein 17 precursor (PMEL17), mRNA

Comment Features Sequence

```
LOCUS       NM_001163889            1994 bp    mRNA    linear   MAM 06-AUG-2009
DEFINITION  Equus caballus melanocyte protein 17 precursor (PMEL17), mRNA.
ACCESSION   NM_001163889 XM_001504795
VERSION     NM_001163889.1  GI:255522914
KEYWORDS    .
SOURCE      Equus caballus (horse)
  ORGANISM  Equus caballus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Laurasiatheria; Perissodactyla; Equidae; Equus.
REFERENCE   1  (bases 1 to 1994)
  AUTHORS   Andersson,L.S., Juras,R., Ramsey,D.T., Eason-Butler,J., Ewart,S.,
            Cothran,G. and Lindgren,G.
  TITLE     Equine Multiple Congenital Ocular Anomalies maps to a 4.9 megabase
            interval on horse chromosome 6
  JOURNAL   BMC Genet. 9, 88 (2008)
   PUBMED   19099555
  REMARK    GeneRIF: The MCOA locus for Equine Multiple Congenital Ocular
            Anomalies syndrome maps to horse chromosome 6.
            Publication Status: Online-Only
REFERENCE   2  (bases 1 to 1994)
  AUTHORS   Brunberg,E., Andersson,L., Cothran,G., Sandberg,K., Mikko,S. and
            Lindgren,G.
  TITLE     A missense mutation in PMEL17 is associated with the Silver coat
            color in the horse
  JOURNAL   BMC Genet. 7, 46 (2006)
   PUBMED   17029645
  REMARK    Publication Status: Online-Only
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from
            AAWR02031349.1.
            On Aug 6, 2009 this sequence version replaced gi:149756558.
```

Change Region Shown

Customize View

BLAST Sequence
Find regions of similarity between this sequence and other sequences using BLAST.

Pick Primers
Design and test primers for this sequence using Primer-BLAST.

Articles about the PMEL17 gene

- Equine Multiple Congenital Ocular Anomalies maps to a 4.9 megabase interval on horse chro [BMC Genet. 2008]
- A missense mutation in PMEL17 is associated with the Silver coat color in the horse. [BMC Genet. 2006]

» See all...

RefSeq Alternative Splicing
See the other reference mRNA sequence splice variant for the PMEL17 gene (XM_001504795.1).

RefSeq Protein Product
See the reference protein sequence for melanocyte protein 17 precursor (NP_001157361.1).

More about the gene PMEL17
Also Known As: SILV

Homologs of the PMEL17 gene
The PMEL17 gene is conserved in human, chimpanzee, dog, cow, mouse, rat, chicken, and zebrafish.

Sequence Record
Continued

Features

```
            1853-1994           AAWR02031349.1     72977-73118         c
FEATURES             Location/Qualifiers
     source          1..1994
                     /organism="Equus caballus"
                     /mol_type="mRNA"
                     /db_xref="taxon:9796"
                     /chromosome="6"
                     /map="6"
     gene            1..1994
                     /gene="PMEL17"
                     /gene_synonym="SILV"
                     /note="melanocyte protein 17 precursor"
                     /db_xref="GeneID:100033885"
     CDS             6..1994
                     /gene="PMEL17"
                     /gene_synonym="SILV"
                     /note="premelanosomal protein"
                     /codon_start=1
                     /product="melanocyte protein 17 precursor"
                     /protein_id="NP_001157361.1"
                     /db_xref="GI:255522915"
                     /db_xref="GeneID:100033885"
                     /translation="MDPVLRRCLLHMAVMGTLLAVGATEGPRDRNWLGVSRQLRTKAW
                     NRQLYPEWTEIQGPDCWRGGQVALKISNDGPTLVGANASFSIALHFPESQKVLPDGQV
                     IWANNTIINGSQVWRGQPVYPQEPDDSCIFPDGGACPSGPLSQRRSFVYVWKTWGQYW
                     QVLGGPVSGLSIGTGTAALGTHSMEVTVYHRRGSQSYVPLAHSRSAFTITDQVPFSVS
                     VSQLQALDGRNKHFLKNQPLTFALRLHDPSGYLAGADLSYTWDFGDSTGTLISRALVV
                     THTYLESGPVTAQVVLQAAIPLTSCGSSPVPGTTGGYVPTAEAPGTTAGQVPTADVVN
                     TTPGQVPTAEPSRTTAVQVPTTEVISTTPVQVPTAEDIGTTPEQVSTSEFLGTTLAEM
                     PTAEAKGMTPEVSTPEPSGTTVAQVTATELVETTAGEAPTPEPEGPDASPFMPTEGTT
                     GSQSPLLDGTATLFLVKRQVPLDCVLYRYGSFSLTLDIVQGIESAEILQAVPSSEGDA
                     FELTVSCQGGLPKEACMDISSPGCQPPAQRLCQPVTPNPACQLVLHQVLKSGSGTYCL
                     NVSLADANSLAMVSTQLVMPGQEAGLGQAPLFVGILLVLMAMVLASLIYRRRLMKQGS
                     ALPLPQLPHGRTHWLRLPWVFRSSPFGESSPLLSGQQV"
     STS             798..996
                     /gene="PMEL17"
                     /gene_synonym="SILV"
                     /standard_name="SILV"
                     /db_xref="UniSTS:504027"
ORIGIN
        1 agaggatgga tccagtgctc agaagatgcc ttctccatat ggctgtgatg ggtactcttc
       61 tggctgtggg ggccacagaa ggacccagag accggaactg gctcggtgtc tcaaggcagc
      121 tcaggactaa agcctggaac aggcagctgt atccagagtg gacagaaatt caggggcctg
      181 actgttggag aggtggccag gtggccctga agatcagtaa tgatgggcct acactggttg
      241 gggcaaatgc ctccttctct attgccttgc acttccctga aagccaaaag gtgctgccgg
```

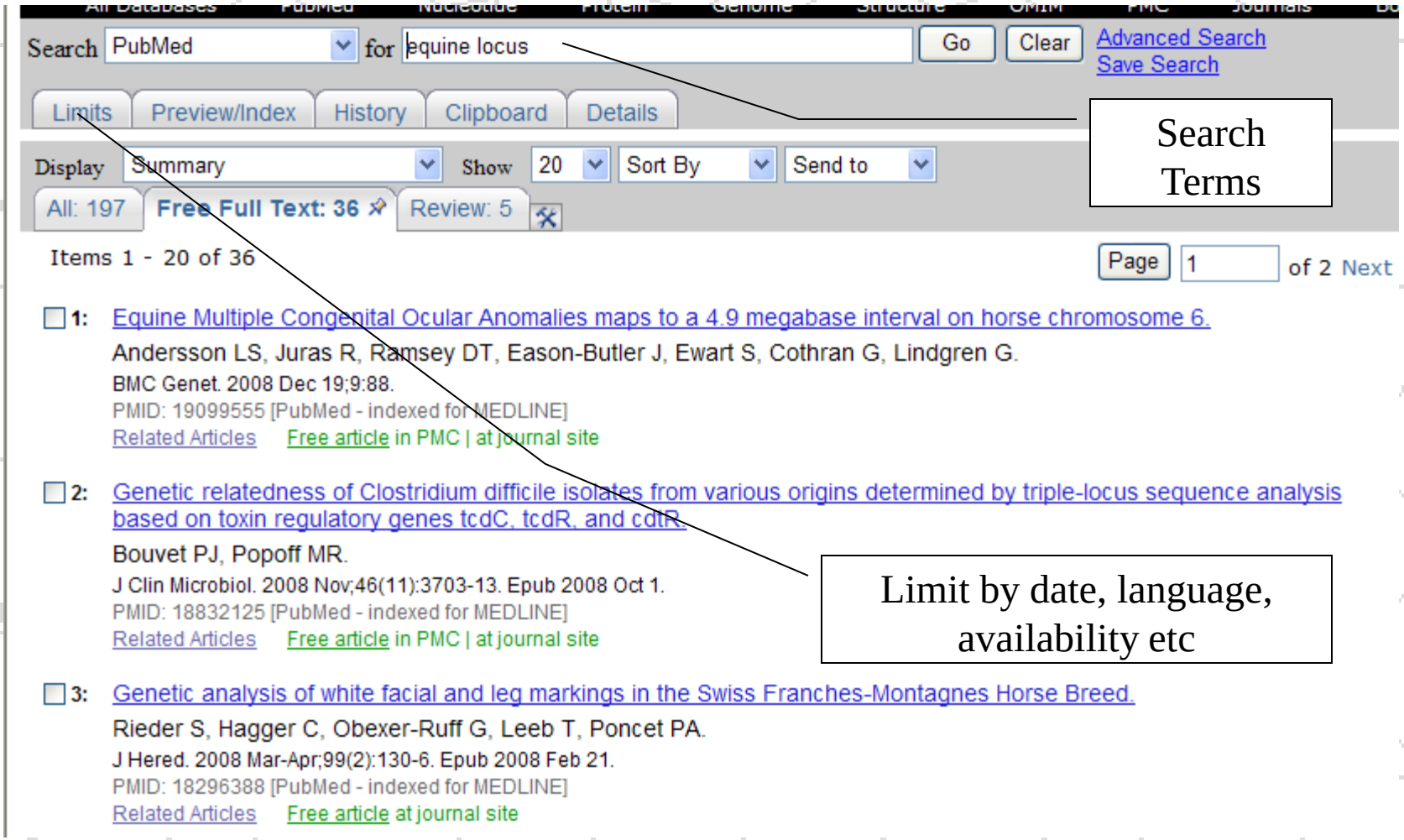
All links from this record

- Components(Core)
- Full text in PMC
- Gene
- Protein
- PubMed
- PubMed (RefSeq)
- PubMed (Weighted)
- Taxonomy
- Related Sequences
- UniSTS

The Sequence

Links to related Information in Other databases

PubMed



Search Terms

Limit by date, language, availability etc

PubMed
Article detail

Article source

Link to full text

J Gynecol Oncol. 2010 Sep;21(3):181-5. Epub 2010 Sep 28.

High expression of mTOR is associated with radiation resistance in cervi

Kim MK, Kim TJ, Sung CO, Choi CH, Lee JW, Kim BG, Bae DS.

Department of Obstetrics and Gynecology, Samsung Medical Center, Sungkyunkwan University School of Medicine

Abstract

OBJECTIVE: Mammalian target of rapamycin (mTOR) is known to promote cell proliferation, survival, a whether mTOR expression was associated with survival and the response to radiation in patients with

METHODS: After reviewing 119 patients treated by primary radiotherapy for stage IIB-IVA cervical cance recurrence or radiation failure after primary radiation therapy were selected. For each case, two contro paraffin-embedded tissues, the cytoplasmic expression of phosphorylated-mTOR (p-mTOR) was eval intensity (intensity score [IS] 0-3) and proportion (proportion score [PS] 0-100). The progression free su recurrence by imaging studies or biopsy. The staining distribution and PFS were compared between t Mann-Whitney U-test, Fisher's exact test, and Cox proportional hazards regression model.

RESULTS: The p-mTOR cytoplasmic expression was significantly associated with a poor response to cytoplasmic expression of p-mTOR was associated with a worse outcome (p=0.02). The hazard ratio f for mTOR PS (p<0.05 for both), indicating that the degree of p-mTOR staining correlated with the recur

CONCLUSION: High expression of p-mTOR was associated with radiation resistance; therefore p-mT patients with cervical cancer.

PMID: 20922141 [PubMed - in process] Free Article

JGO ASCO FREE Journal of Gynecologic Oncology

Related citations

Predictive and prognostic role of activated mammalian target of rapamyc [Oncol Rep. 2006]

Final Report on Carcinogens Background Document for [Rep Carcinog Backgr Doc. 2010]

Morphoproteomic evidence of constitutively activated and overex [Int J Clin Exp Pathol. 2009]

EMMPRIN expression as a prognostic factor in radiotherapy of cervical [Clin Cancer Res. 2008]

Review Overview of resistance to systemic therapy in patients with [Adv Exp Med Biol. 2007]

See reviews...

See all...

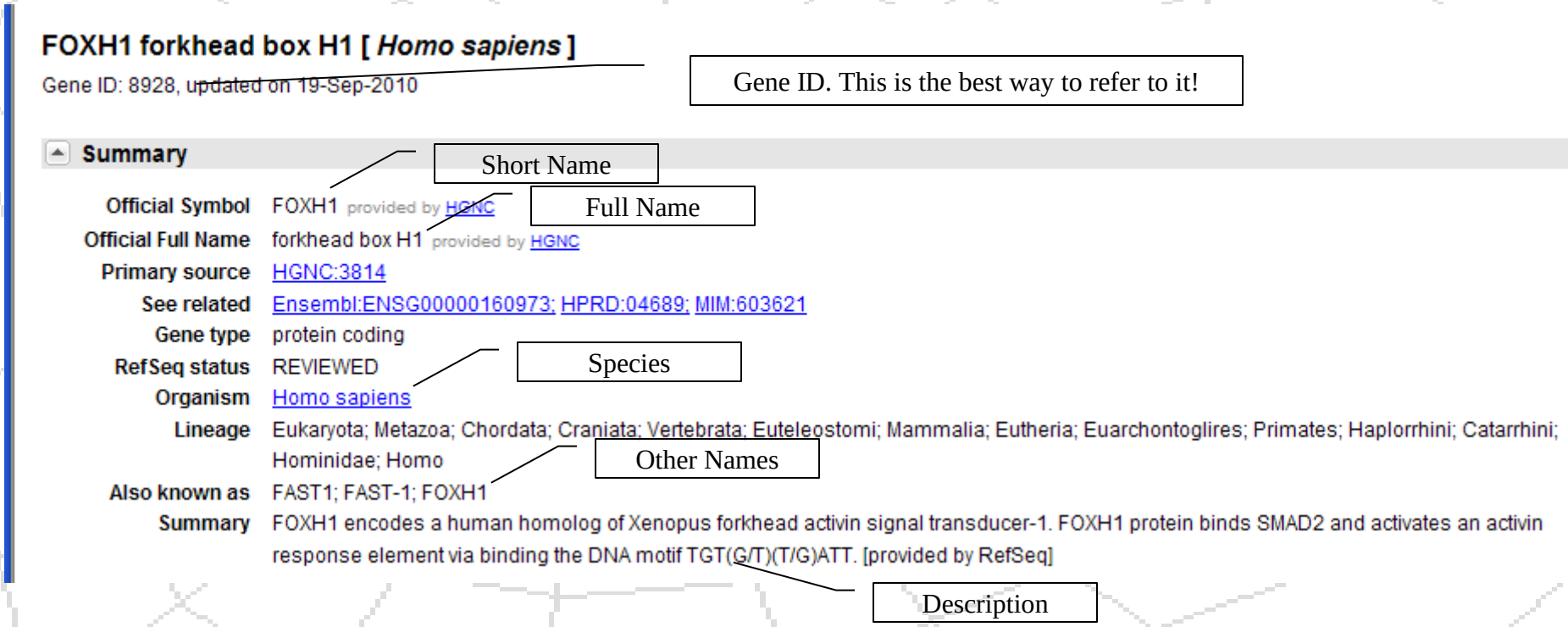
All links from this record

Related Citations

Free in PMC

Article abstract

Gene



- Additional Sections
  - Links (including SNP)
  - Sequences (Genomic, RNA, Protein)
  - Publications
  - Interactions
  - etc

Hands On

Perform a search for a nucleotide sequence

- What is the accession number?
- Notice what it looks like in FASTA format.
- Copy the sequence and save in a document.

Find an article associated with it and look at it in PubMed.

Find a Gene associated with it and look at it in Gene

Note: Not all sequences have associated articles or Genes.

Single Nucleotide Polymorphisms

(SNPs)

Differences in single nucleotides within the population

Can be linked to specific phenotypes such as heritable disease or drug responses

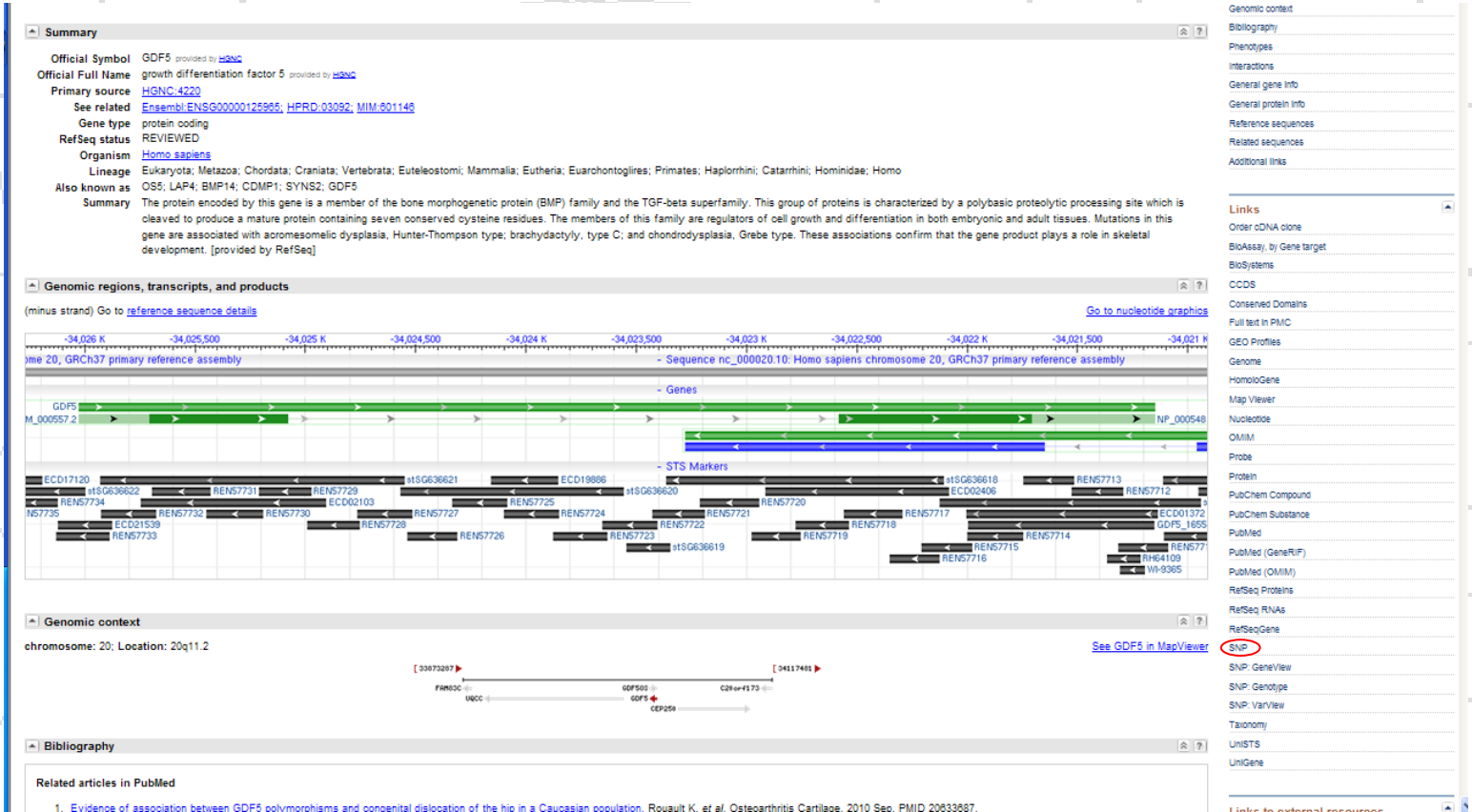
SNP: IUPAC ambiguity codes

SNP bases are frequently identified by these codes

R = A or G
K = G or T
S = G or C
Y = C or T
M = A or C
W = A or T
B = not A (C, G or T)
H = not G (A, C or T)
D = not C (A, G or T)
V = not T (A, C or G)
N = any nucleotide

Finding a SNP

Find a Gene of interest and find the SNP link.



Summary

Official Symbol GDF5 provided by HGNC
Official Full Name growth differentiation factor 5 provided by HGNC
Primary source HGNC:4220
See related Ensembl:ENSG00000125965; HPRD:03092; MIM:601146
Gene type protein coding
RefSeq status REVIEWED
Organism Homo sapiens
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as OS5; LAP4; BMP14; CDMP1; SYNS2; GDF5
Summary The protein encoded by this gene is a member of the bone morphogenetic protein (BMP) family and the TGF-beta superfamily. This group of proteins is characterized by a polybasic proteolytic processing site which is cleaved to produce a mature protein containing seven conserved cysteine residues. The members of this family are regulators of cell growth and differentiation in both embryonic and adult tissues. Mutations in this gene are associated with acromesomelic dysplasia, Hunter-Thompson type; brachydactyly, type C; and chondrodysplasia, Grebe type. These associations confirm that the gene product plays a role in skeletal development. [provided by RefSeq]

Genomic regions, transcripts, and products

(minus strand) Go to reference sequence details — Go to nucleotide graphics

Genomic context

chromosome: 20; Location: 20q11.2 — See GDF5 in MapViewer

Bibliography

Related articles in PubMed

1. Evidence of association between GDF5 polymorphisms and congenital dislocation of the hip in a Caucasian population. Rouault K, et al. Osteoarthritis Cartilage. 2010 Sep. PMID 20633687.

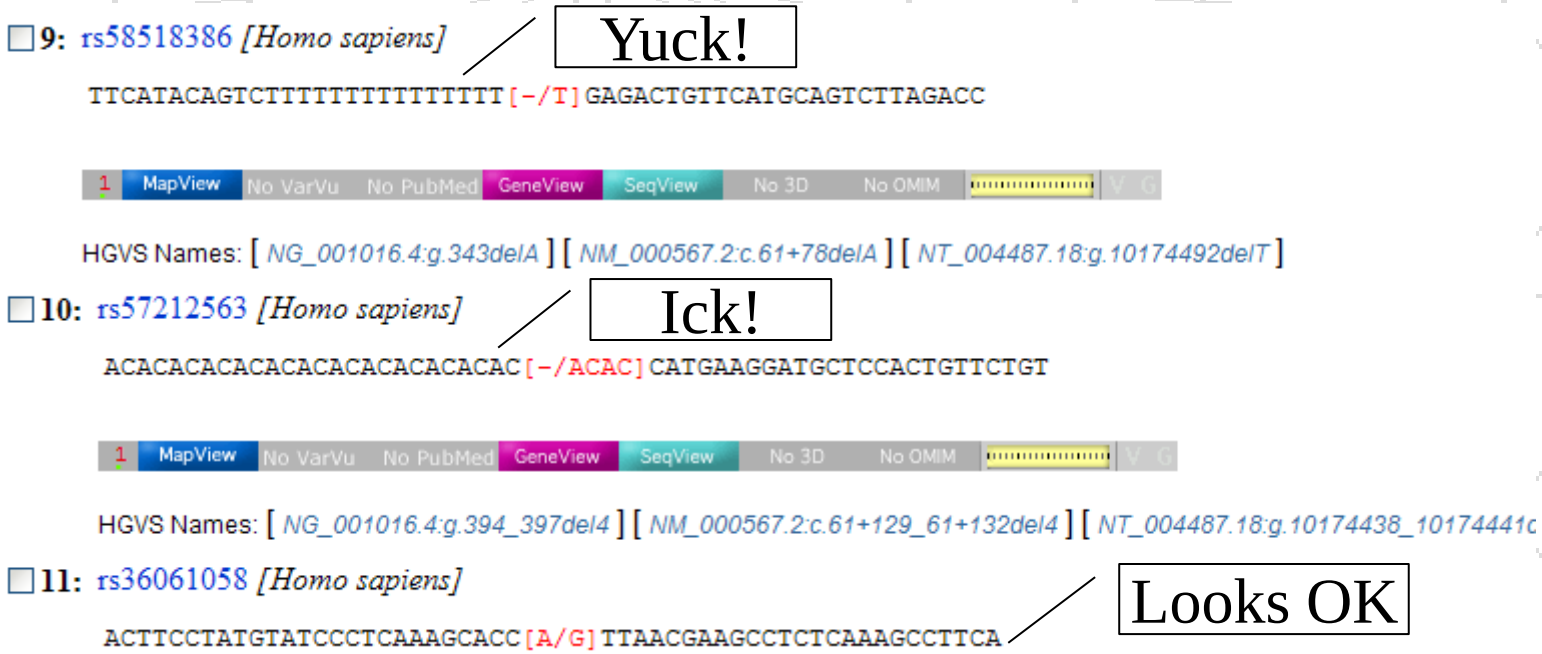
Genomic context
Bibliography
Phenotypes
Interactions
General gene info
General protein info
Reference sequences
Related sequences
Additional links

Links
Order cDNA clone
BioAssay, by Gene target
BioSystems
CCDS
Conserved Domains
Full text in PMC
GEO Profiles
Genome
HomoloGene
Map Viewer
Nucleotide
OMIM
Probe
Protein
PubChem Compound
PubChem Substance
PubMed
PubMed (GeneRIF)
PubMed (OMIM)
RefSeq Proteins
RefSeq RNAs
RefSeqGene
SNP
SNP: GeneView
SNP: Genotype
SNP: VarView
Taxonomy
UniSTS
UniGene

Finding a SNP

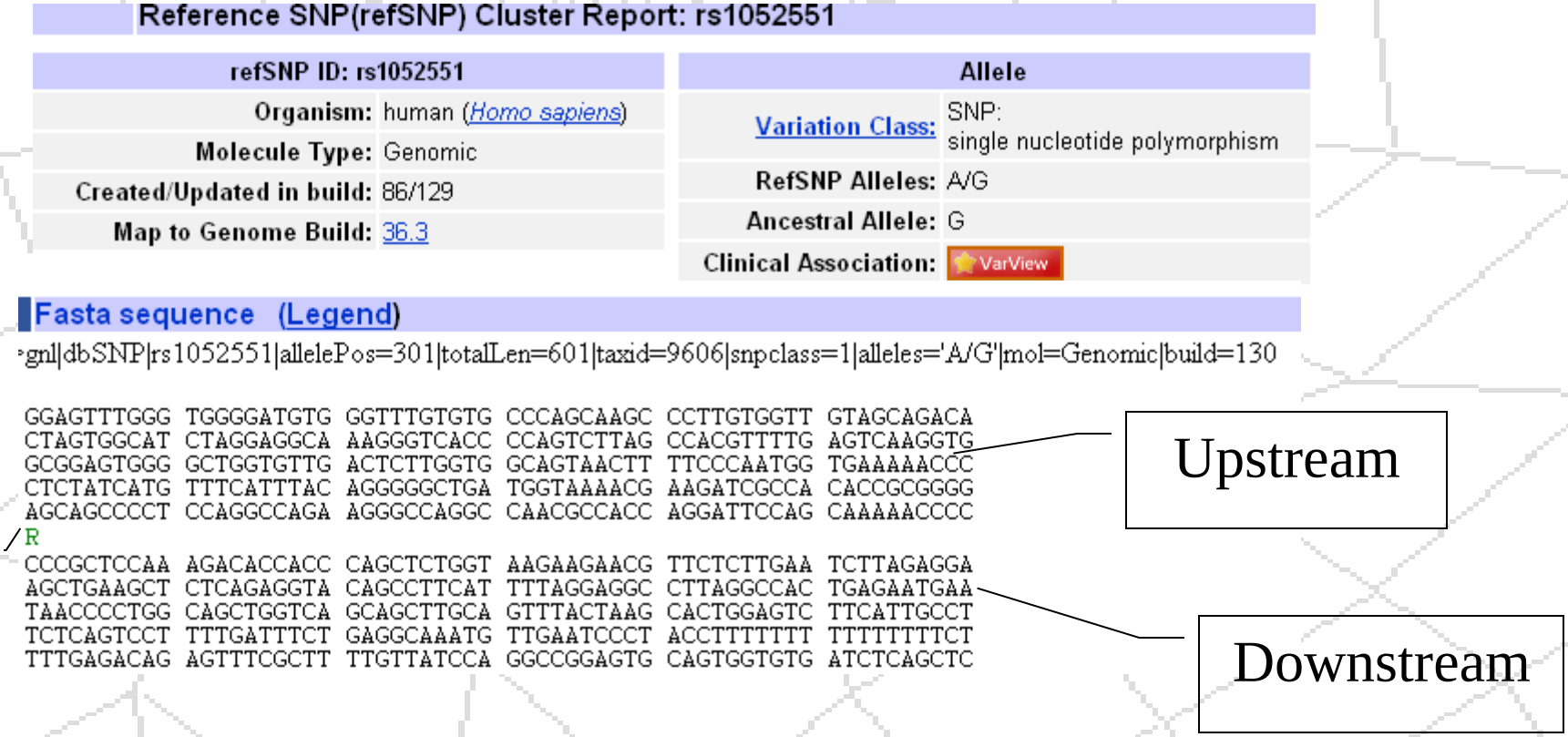
Choose one of the SNPs from the Gene link

Avoid obviously repetitive sequences.



SNP

Single Nucleotide Polymorphisms



Reference SNP(refSNP) Cluster Report: rs1052551

refSNP ID: rs1052551	
Organism:	human (*Homo sapiens*)
Molecule Type:	Genomic
Created/Updated in build:	86/129
Map to Genome Build:	36.3

Allele	
Variation Class:	SNP: single nucleotide polymorphism
RefSNP Alleles:	A/G
Ancestral Allele:	G
Clinical Association:	VarView

Fasta sequence (Legend)

>gnl|dbSNP|rs1052551|allelePos=301|totalLen=601|taxid=9606|snpclass=1|alleles='A/G'|mol=Genomic|build=130

GGAGTTTGGG TGGGGATGTG GGTTTGTGTG CCCAGCAAGC CCTTGTGGTT GTAGCAGACA
CTAGTGGCAT CTAGGAGGCA AAGGGTCACC CCAGTCTTAG CCACGTTTTG AGTCAAGGTG
GCGGAGTGGG GCTGGTGTTG ACTCTTGGTG GCAGTAACTT TTCCCAATGG TGAAAAACCC
CTCTATCATG TTTCATTTAC AGGGGGCTGA TGGTAAAACG AAGATCGCCA CACCGCGGGG
AGCAGCCCCT CCAGGCCAGA AGGGCCAGGC CAACGCCACC AGGATTCCAG CAAAAACCCC
R
CCCGCTCCAA AGACACCACC CAGCTCTGGT AAGAAGAACG TTCTCTTGAA TCTTAGAGGA
AGCTGAAGCT CTCAGAGGTA CAGCCTTCAT TTTAGGAGGC CTTAGGCCAC TGAGAATGAA
TAACCCCTGG CAGCTGGTCA GCAGCTTGCA GTTTACTAAG CACTGGAGTC TTCATTGCCT
TCTCAGTCCT TTTGATTTCT GAGGCAAATG TTGAATCCCT ACCTTTTTTT TTTTTTTTCT
TTTGAGACAG AGTTTCGCTT TTGTTATCCA GGCCGGAGTG CAGTGGTGTG ATCTCAGCTC

SNP

Upstream

Downstream

Hands On

Find a Gene of interest

See if there is a link to “SNP” under Links

Open the SNP link and choose a SNP from among those linked

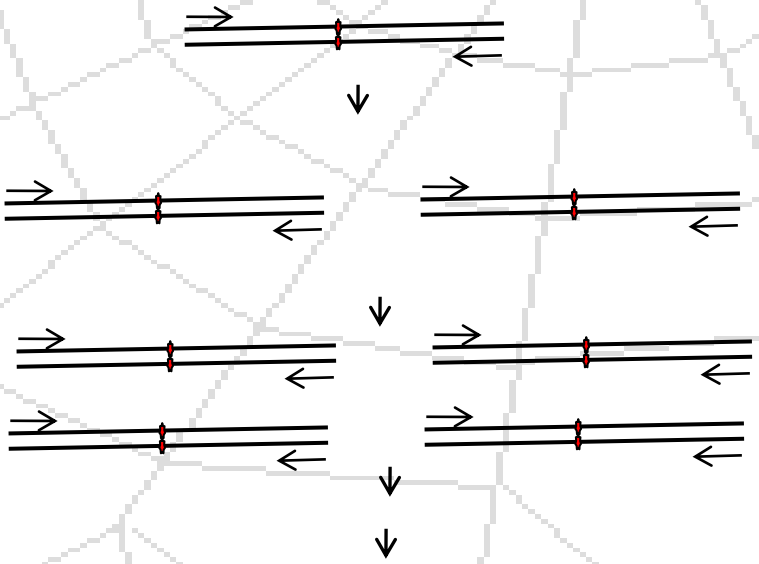
Copy the sequence information and SNP sequence to a text document and save it.

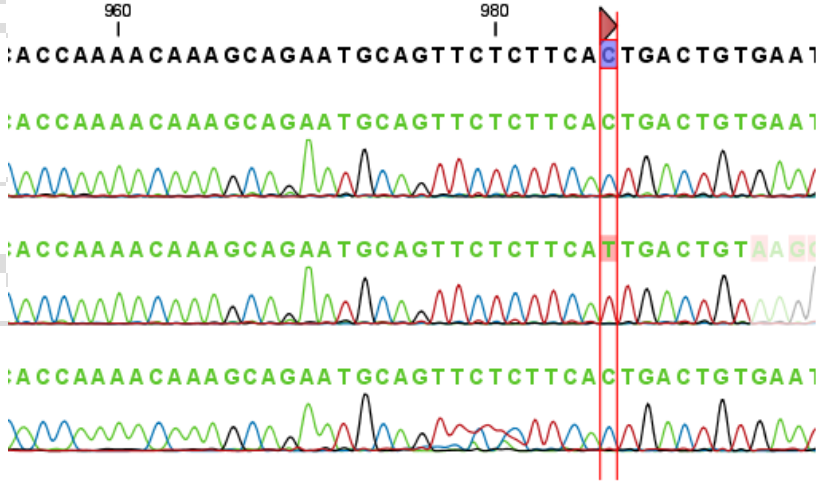
SNP detection

SNP detection:

There are a number of specialized technologies.

PCR followed by sequencing is one of the simplest.





Primer Design

There are many tools which will design primers.

Ensuring that primers are specific is difficult.

BLAST is commonly used for this purpose.

BLAST

- Basic Local Alignment Search Tool
- Compares a query sequences against all sequences in a database.
- Very powerful for finding biologically significant relationships and full gene sequences in the database when you have a fragment etc.
- Different types:
  - Nucleic acid – Nucleic Acid
  - Protein- Protein
  - Nucleic Acid Translation – Protein
  - Protein – Nucleic Acid Translation
  - Translation - Translation

BLAST: primer analysis

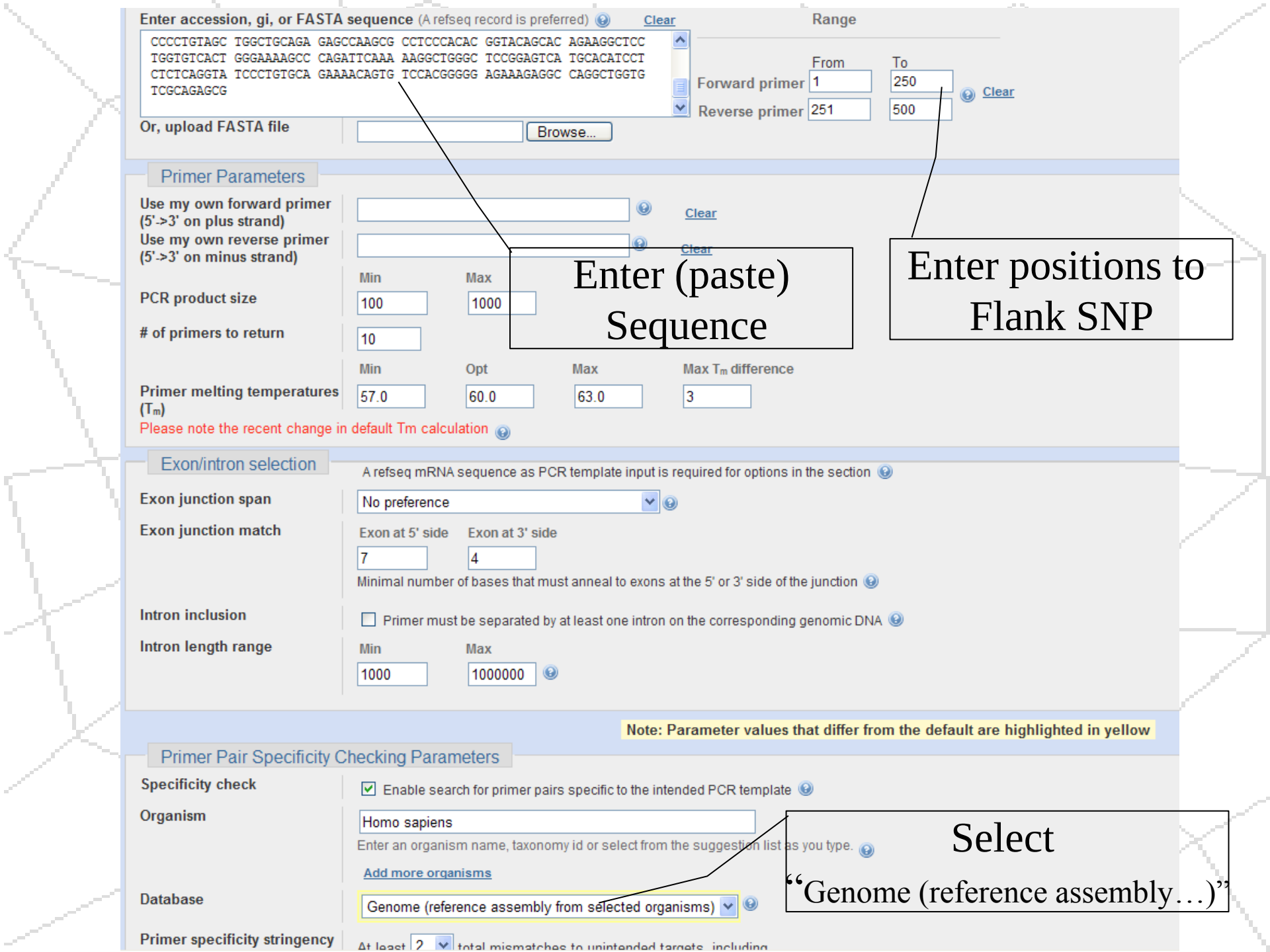
Blast results are time consuming and difficult to interpret.

Blast Output example.htm

Easier primer design and specificity check

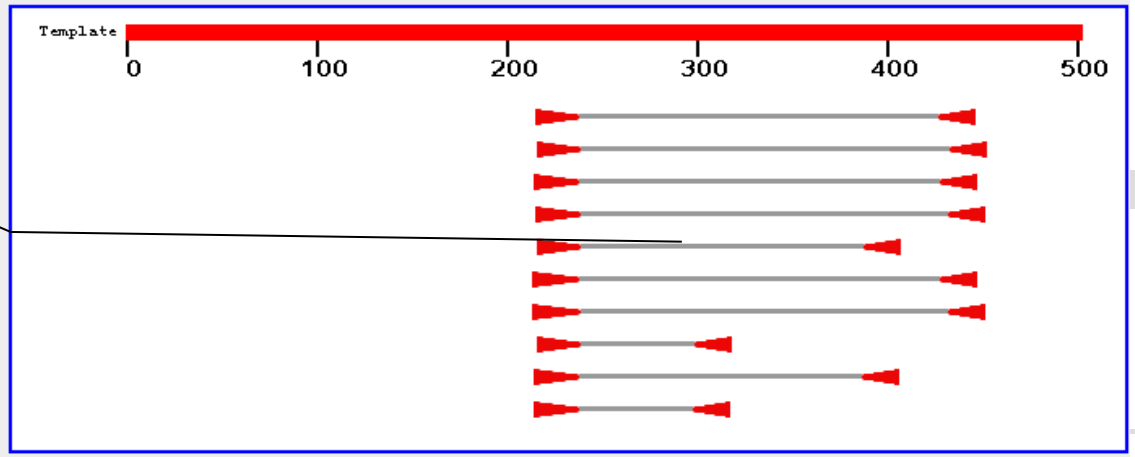
Primer BLAST

Performs primer design and BLAST at the same time.


Enter accession, gi, or FASTA sequence (A refseq record is preferred) Clear
CCCCTGTAGC TGGCTGCAGA GAGCCAAGCG CCTCCCACAC GGTACAGCAC AGAAGGCTCC
TGGTGTCACT GGGAAAAGCC CAGATTCAAA AAGGCTGGGC TCCGGAGTCA TGCACATCCT
CTCTCAGGTA TCCCTGTGCA GAAAACAGTG TCCACGGGGG AGAAAGAGGC CAGGCTGGTG
TCGCAGAGCG
Or, upload FASTA file
Browse...
Range
From
To
Forward primer 1 250
Reverse primer 251 500
Clear
Primer Parameters
Use my own forward primer (5'->3' on plus strand)
Use my own reverse primer (5'->3' on minus strand)
Clear
Clear
Min
Max
PCR product size 100 1000
# of primers to return 10
Min
Opt
Max
Max Tm difference
Primer melting temperatures (Tm) 57.0 60.0 63.0 3
Please note the recent change in default Tm calculation
Enter (paste) Sequence
Enter positions to Flank SNP
Exon/intron selection
A refseq mRNA sequence as PCR template input is required for options in the section
Exon junction span No preference
Exon junction match
Exon at 5' side
Exon at 3' side
7
4
Minimal number of bases that must anneal to exons at the 5' or 3' side of the junction
Intron inclusion
Primer must be separated by at least one intron on the corresponding genomic DNA
Intron length range
Min
Max
1000
1000000
Note: Parameter values that differ from the default are highlighted in yellow
Primer Pair Specificity Checking Parameters
Specificity check
Enable search for primer pairs specific to the intended PCR template
Organism
Homo sapiens
Enter an organism name, taxonomy id or select from the suggestion list as you type.
Add more organisms
Database
Genome (reference assembly from selected organisms)
Primer specificity stringency
At least 2 total mismatches to unintended targets, including
Select
"Genome (reference assembly…)"

Primer BLAST results

mmary of primer pairs



Positions of Primers pairs

tailed primer reports

Primer pair 1

	Sequence (5'->3')	Strand on template	Length	Start	Stop	Tm	GC%
Forward primer	TGACTCACCAGAGAGCCTTCAGC	Plus	23	215	237	58.09	56.52%
Reverse primer	CCTCCGGGTGGCCCATGGAA	Minus	20	444	425	60.54	70.00%
Product length	230						

Products on potentially unintended templates

>NT_167187.1 Homo sapiens chromosome 8 genomic contig, GRCh37 reference primary assembly

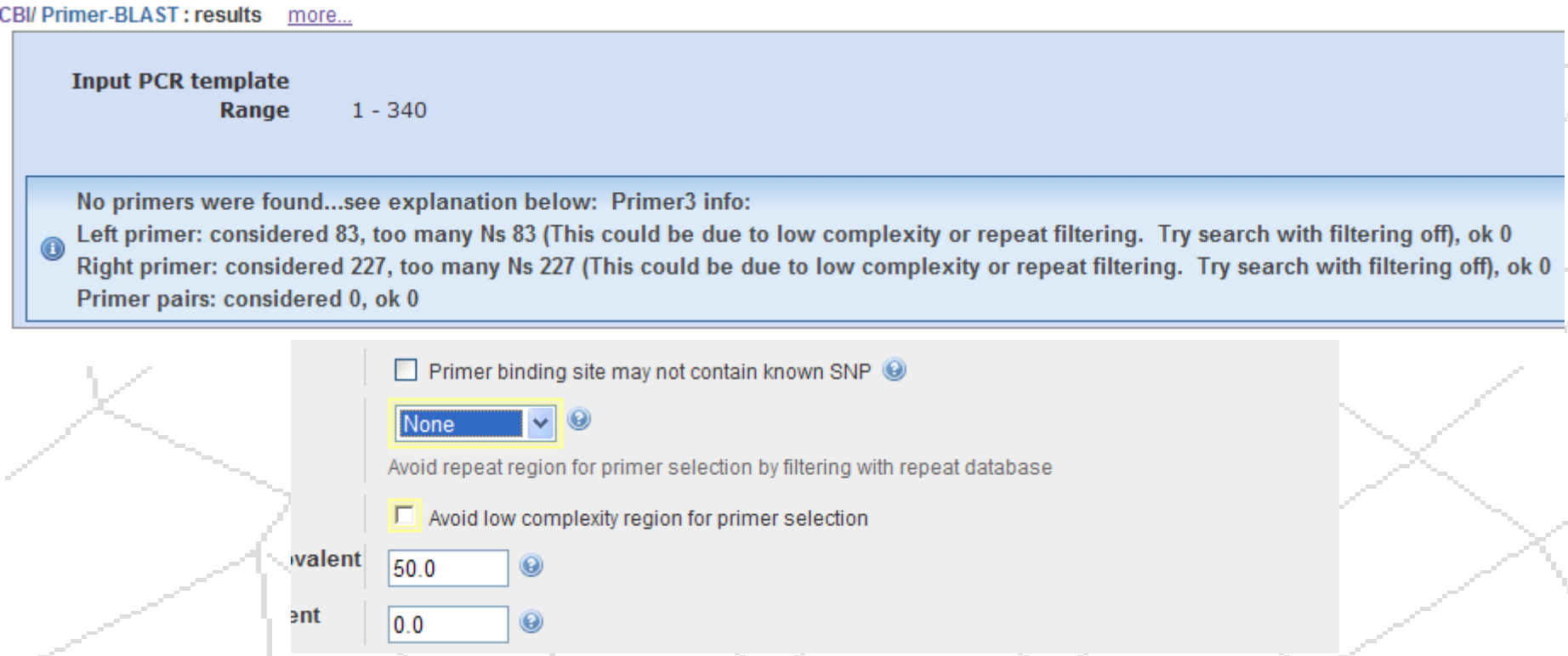
```
product length = 230
Features associated with this product:
   ankyrin 1 isoform 9

   ankyrin 1 isoform 1
```

General Primer information

Specificity

Possible problems with Primer BLAST

You may get a “No Primers were found message”. Read the details.



General Utilities

- http://www.bioinformatics.org/sms/
  - Translation
  - Restriction Digestion
  - Reformatting (alternately FASTA Formatter)
  - Complement/Reverse
  - Etc.
- http://www.promega.com/biomath/calc11.htm
  - Melting Temperature of an oligo.

Using Microsoft Word

- Clean a sequence:
  - Replace all numbers (^#)
  - Replace all spaces (“ “)
  - Replace all returns (^p)
- Count bases
  - Review: Word count
- Find sequences with Find feature

  This will find primers in the forward orientation only. You must enter the Reverse Complement of the other.

Hands on

Design primers for the SNP sequence you saved.

Mark the positions of the primers on your saved SNP sequence.

Homework

Due October 29

- Identify a SNP in a gene of interest
  - Report the Gene ID and Official Full Name of the Gene in which the SNP exists
  - Report the sequence around the SNP being sure to make the location and identity of the actual polymorphism obvious.
  - Cite at least one paper which describes the sequence.
    Note that not all genes have publications associated with them. You will need to find a gene which does have an associated paper.
  - Give a very brief (1-2 sentences) summary of why the gene is interesting or what it does.
  - Design primers that will amplify the polymorphism for that SNP.
  - Display a RefSeq mRNA sequence with the locations of the primers clearly indicated. The sequence must be in a monospaced font such as Courier.
    I suggest highlighting the primers, but any method that clearly shows where they are in the sequence is OK.