Gene Annotation

Michael Schatz

October 12, 2022

Lecture 13. Applied Comparative Genomics

Assignment 4: Variant Analysis and bedtools
Due Monday Oct 17 by 11:59pm

github.com/schatzlab/appliedgenomics2022/tree/main/assignments/assignment4#readme

README.md

Assignment 4: Variant Analysis and Bedtools

Assignment Date: Monday, October 10, 2022
Due Date: Monday, Oct 17, 2022 @ 11:59pm

Assignment Overview

In this assignment you will consider the fundamental properties for variant calling, and get some experience with small variant analysis.

In this assignment you will explore WDLs as a workflow language to orchestrate a variety of read mapping tasks. You can execute your WDLs using miniwdl after running `pip install miniwdl`. You may also find learn-wdl to be very helpful.

If you are using a Mac, make sure to disable gRPC FUSE for file sharing. There is a currently a very weird bug in Docker desktop 4.12.0 (docker engine 20.10.17) on Apple M1 where if you try to disable "Use gRPC FUSE for file sharing" via the GUI it will turn itself back on when you try to activate it. Docker is aware of the problem and are working on a fix. In the meantime there is a workaround available: docker/for-mac#6467

The bioinformatics tools you will need for the assignment (bowtie, samtools, etc) are bundled into a Docker container so you wont need to install other software packages. This will get you ready to run your code in the cloud for future assignments. Instructions for running the container are available here: https://github.com/mschatz/wga-essentials. This is known to work on new macs (M1) and older macs (intel chip). It should also run on Linux and Windows but let us know if you have any issues.

As a reminder, any questions about the assignment should be posted to Piazza.

Question 1. Binomial Distribution [10 pts]

- 1a. For coverage n = 10 to 200, calculate the maximum number of minor allele reads (round down) that would make your one-sided binomial test reject the null hypothesis p=0.5 at 0.05 significance. Plot coverage on the x-axis and (number of reads)/(coverage) on the y-axis. Note this is the minimum number of reads that are necessary to believe we might have a heterozygous variant on the second haplotype rather than just mere sequencing error.
- 1b. What asymptote does the plot seem to approach? Why is this?

Question 2. Hello World on AnVIL [5 pts]

https://github.com/schatzlab/appliedgenomics2022/tree/main/assignments/assignment4
Check Piazza for questions!

Agenda

9:30 a.m.
Elana Fertig, PhD, FAIMBE
Professor
Director of the Division and Research Program in Quantitative Sciences
co-Director Convergence Institute
Sidney Kimmel Comprehensive Cancer Center
The Johns Hopkins University School of Medicine
Welcome

9:35 a.m.
Nikolaus Schultz, PhD
Head of Knowledge Systems,
Marie-Josée & Henry R. Kravis Center for Molecular Oncology;
Attending Computational Oncologist,
Department of Epidemiology & Biostatistics
Memorial Sloan Kettering Cancer Center
TBD

10:35 a.m.
Won Jin Ho, MD
Assistant Professor
Cancer Immunology/GI Oncology
Sidney Kimmel Comprehensive Cancer Center
The Johns Hopkins University
Navigating the Multi-Omic Landscape to Unlock Insights into Cancer Immunotherapy

10:55 a.m.
10 minute break

11:05 a.m.
Kellie Smith, PhD
Assistant Professor of Oncology
Director of the FEST and TCR Immunogenomics Core at the Bloomberg~Kimmel Institute for Cancer Immunotherapy at Johns Hopkins
Immunogenomic profiling of tumor-reactive TIL for novel IO target discovery

11:40 a.m.
Benjamin Orsburn, PhD
Instructor of Pharmacology and Molecular Sciences
Johns Hopkins University School of Medicine
Applying global proteomics techniques to single human cells to solve riddles in human pharmacology

Noon
Break for lunch

1:20 p.m.
Mindy Kim Graham, PhD
Research Associate
Radiation Oncology and Molecular Radiation Sciences
Sidney Kimmel Comprehensive Cancer Center at
The Johns Hopkins University
From Atlas to insights: probing the microenvironmental changes in prostate cancer at single cell resolution

1:50 p.m.
Atul Deshpande, PhD
Postdoctoral Associate
Fertig Lab
Sidney Kimmel Comprehensive Cancer Center
The Johns Hopkins University School of Medicine
Identifying molecular changes from spatially interacting latent features in the tumor microenvironment

2:20 p.m.
10 minute break

2:30 p.m.
Michael Schatz, PhD
Bloomberg Distinguished Professor
Department of Computer Science
The Johns Hopkins University
The next 100 years of genome sequencing

16th Annual Symposium on Genomics and Bioinformatics

Thursday, October 13th, 2022
9:30AM to 3:30PM

Please click the link below to join the webinar:
https://jhjhm.zoom.us/j/98939305072?pwd=THEzdEtLSWF1b3BxdkJnOVVMWGZZQT09
Passcode: 606295

Annotation

Goal: Genome Annotations

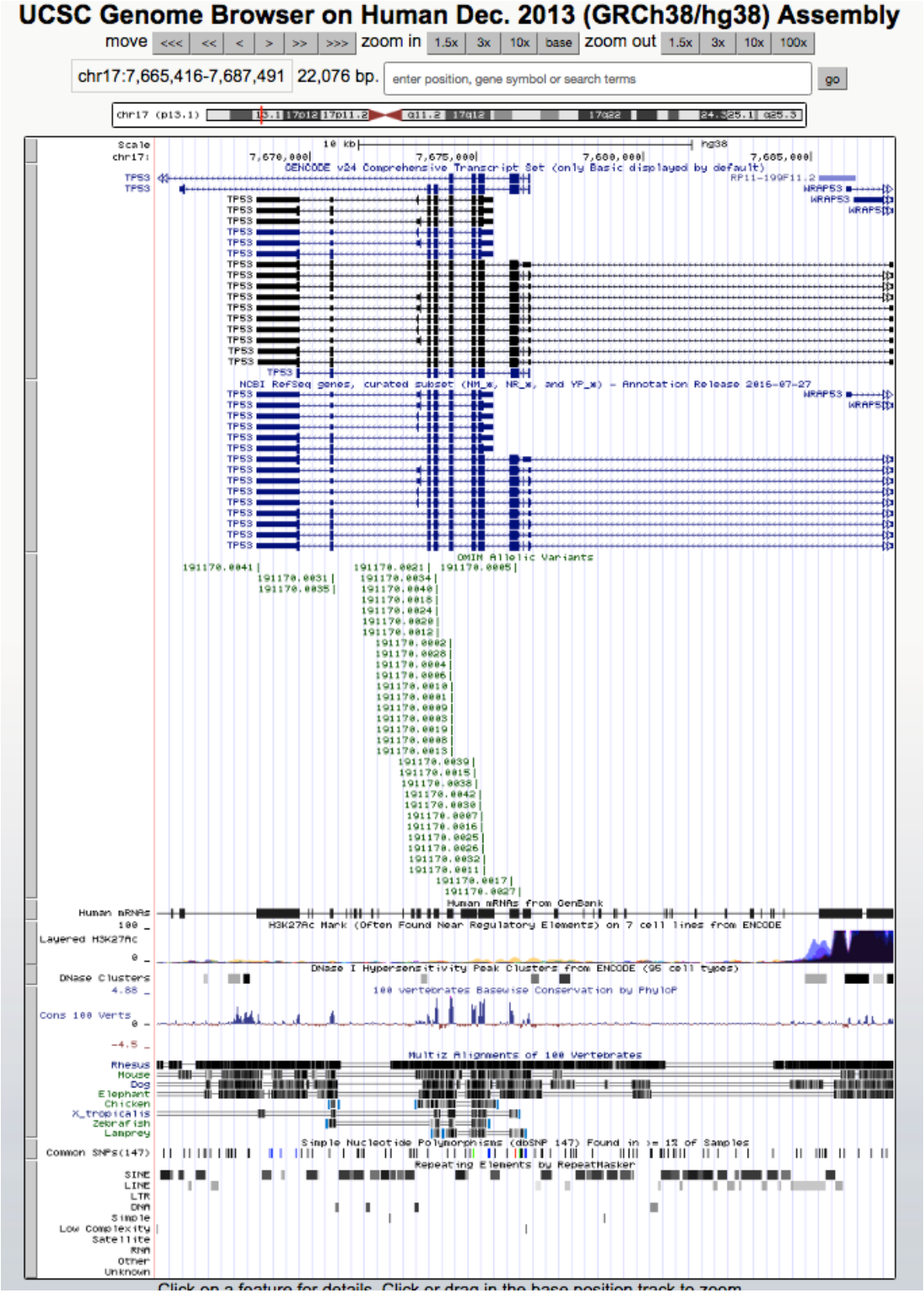
aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaat
gcatgcggctatgcaagctgggatccgatgactatgctaagctgggatccgatgacaatgcatgcggctatgct
aatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaa
tggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcg
gctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcg
gctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcct
gcggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggct
atgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggaatgcatgcg
gctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccg
atgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcatgcggctatgctaagctgggaat
gcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctg
ggatccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcggctatgctaatgaatg
gtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgg
gatttaccttggaatatgctaatgcatgcggctatgctaagctgggaatgcatgcggctatgctaagctgggatc
cgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcg
gctatgctaatgcatgcggctatgctaagctcatgcgg

Goal: Genome Annotations

aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaat
gcatgcggctatgcaagctgggatccgatgactatgctaagctgggatccgatgacaatgcatgcggctatgct
aatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaa
tggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcg
gctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcg
gctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcct
gcggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggct
atgctaatgaatggtcttgggatt **Gene!** gctatgctaagctgggaatgcatgcg
gctatgctaagctgggatccgat atgcggctatgcaagctgggatccg
atgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcatgcggctatgctaagctgggaat
gcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctg
ggatccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcggctatgctaatgaatg
gtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgg
gatttaccttggaatatgctaatgcatgcggctatgctaagctgggaatgcatgcggctatgctaagctgggatc
cgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcg
gctatgctaatgcatgcggctatgctaagctcatgcgg

What are genome intervals?

- Genetic variation:
  - SNPs: 1bp
  - Indels: 1-50bp
  - SVs: >50bp
- Genes:
  - exons, introns, UTRs, promoters
- Conservation
- Transposons
- Origins of replication
- TF binding sites
- CpG islands
- Segmental duplications
- Sequence alignments
- Chromatin annotations
- Gene expression data
- …
- ***Your own observations and data: put them into context!***



Plane Sweep

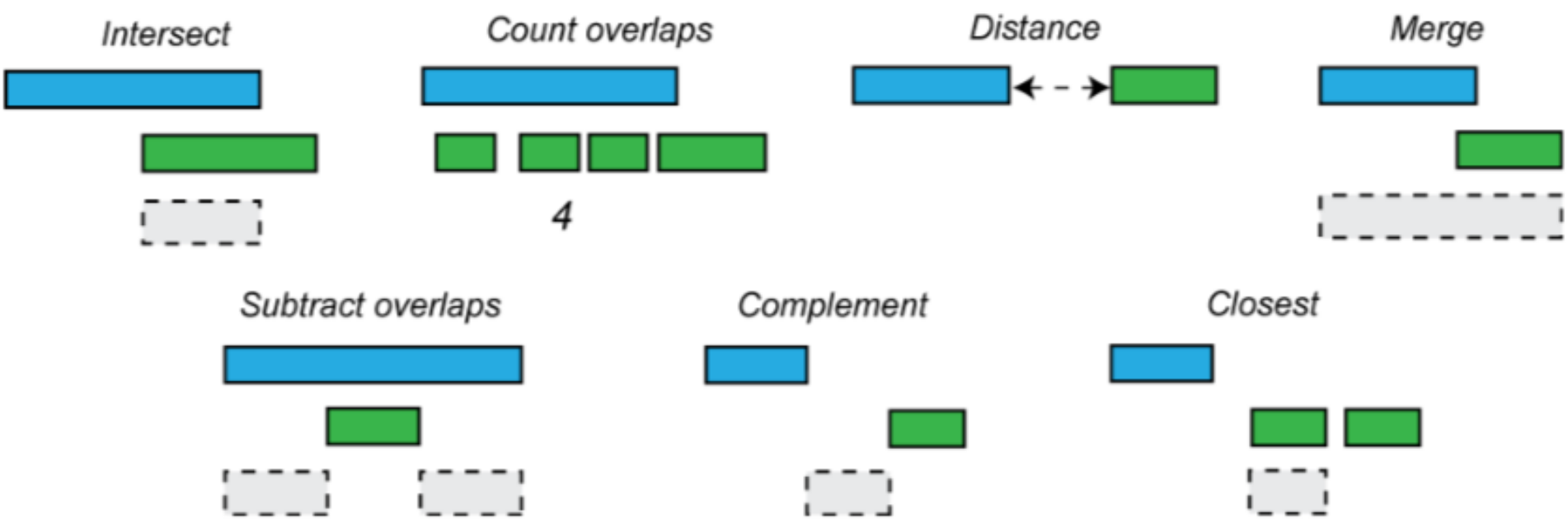
```
pos:   1    5    10   15   20   25   30   35   40   45   50   55
pos:   |    |    |    |    |    |    |    |    |    |    |    |
r1:    [=============================]
r2:        [====================]
r3:             [=================================]
r4:                  [====================]
r5:                                  [===================]

       1    2    3         4    3
```

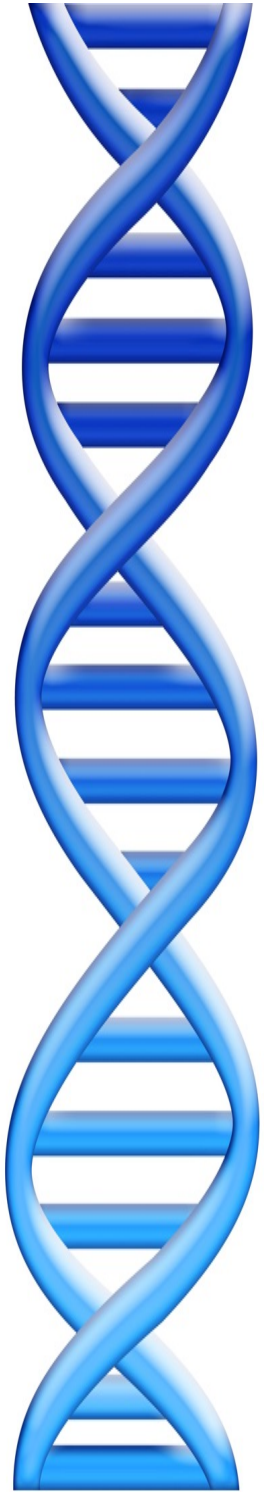
arrive at r5[35,55]:

35 > 25: step down at 25; active set: 30, 40, 45

output (25, 3)

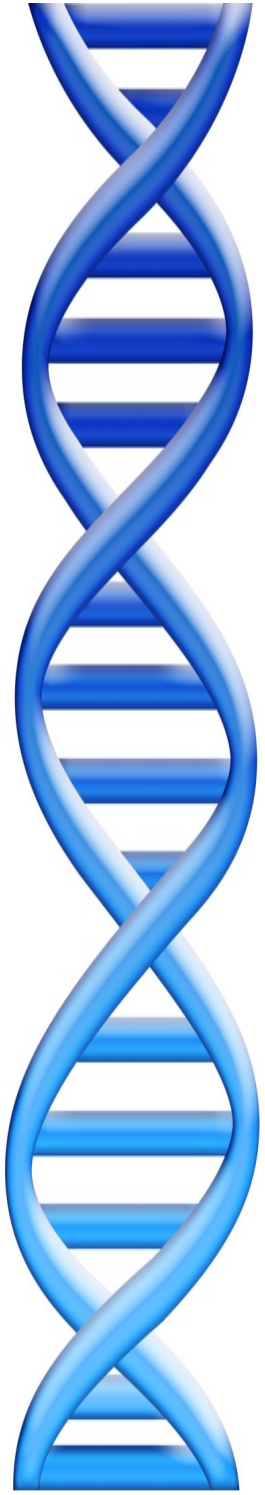
BEDTools to the rescue!





Outline

1. Alignment to other genomes
2. Prediction aka "Gene Finding"
3. Experimental & Functional Assays



Outline

1. Alignment to other genomes
2. Prediction aka "Gene Finding"
3. Experimental & Functional Assays

Basic Local Alignment Search Tool

- Rapidly compare a sequence Q to a database to find all sequences in the database with an score above some cutoff S.
  - Which protein is most similar to a newly sequenced one?
  - Where does this sequence of DNA originate?

- Speed achieved by using a procedure that typically finds "most" matches with scores > S.
  - Tradeoff between sensitivity and specificity/speed
    - Sensitivity – ability to find all related sequences
    - Specificity – ability to reject unrelated sequences

(Altschul et al. 1990)

Seed and Extend

```
FAKDFLAGGVAAAISKTAVAPIERVKLLLQVQHASKQITADKQYKGIIDCVVRIPKEQGV

FLIDLASGGTAAAVSKTAVAPIERVKLLLQVQDASKAIAVDKRYKGIMDVLIRVPKEQGV
```

- Homologous sequences are likely to contain a short high scoring word pair, a seed.
  - Smaller seed sizes make the sense more sensitive, but also (much) slower
  - Typically do a fast search for prototypes, but then most sensitive for final result

- BLAST then tries to extend high scoring word pairs to compute high scoring segment pairs (HSPs).
  - Significance of the alignment reported via an e-value

Seed and Extend

```
FAKDFLAGGVAAAISKTAVAPIERVKLLLQVQHASKQITADKQYKGIIDCVVRIPKEQGV
|  |    | ||| |||||||||||||||||| ||| |  |  |||| |   | ||||||
FLIDLASGGTAAAVSKTAVAPIERVKLLLQVQDASKAIAVDKRYKGIMDVLIRVPKEQGV
```

- Homologous sequences are likely to contain a short high scoring word pair, a seed.
  - Smaller seed sizes make the sense more sensitive, but also (much) slower
  - Typically do a fast search for prototypes, but then most sensitive for final result

- BLAST then tries to extend high scoring word pairs to compute high scoring segment pairs (HSPs).
  - Significance of the alignment reported via an e-value

BLAST E-values

E-value = the number of HSPs having alignment score S (or higher) expected to occur by chance.

→ Smaller E-value, more significant in statistics

→ Bigger E-value, less significant

→ Over 1 means expect this totally by chance

(not significant at all!)

The expected number of HSPs with the score at least S is :

$$E = K*n*m*e^{-\lambda S}$$

K, λ are constant depending on model

n, m are the length of query and sequence

E-values quickly drop off for better alignment bits scores

Very Similar Sequences

```
Query: HBA_HUMAN Hemoglobin alpha subunit
Sbjct: HBB_HUMAN Hemoglobin beta subunit

Score =  114 bits (285),  Expect = 1e-26
Identities = 61/145 (42%), Positives = 86/145 (59%), Gaps = 8/145 (5%)

Query  2   LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGSAQV 55
           L+P +K+ V A WGKV  +  E G EAL R+ + +P T+ +F  F      D   G+ +V
Sbjct  3   LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV 60

Query  56  KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA 115
           K HGKKV  A ++ +AH+D++    + LS+LH  KL VDP NF+LL + L+  LA H
Sbjct  61  KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK 120

Query  116 EFTPAVHASLDKFLASVSTVLTSKY 140
           EFTP V A+  K +A V+  L  KY
Sbjct  121 EFTPPVQAAYQKVVAGVANALAHKY 145
```

Quite Similar Sequences

```
Query: HBA_HUMAN Hemoglobin alpha subunit
Sbjct: MYG_HUMAN Myoglobin

Score = 51.2 bits (121), Expect = 1e-07,
Identities = 38/146 (26%), Positives = 58/146 (39%), Gaps = 6/146 (4%)

Query  2   LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGSAQV  55
           LS  +   V   WGKV A    +G E L R+F   P T   F  F      D    S  +
Sbjct  3   LSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDL  62

Query  56  KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA  115
           K HG  V  AL   +         +  L+  HA K ++     + +S C++  L +  P
Sbjct  63  KKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPG  122

Query  116 EFTPAVHASLDKFLASVSTVLTSKYR  141
           +F      +++K L      + S Y+
Sbjct  123 DFGADAQGAMNKALELFRKDMASNYK  148
```

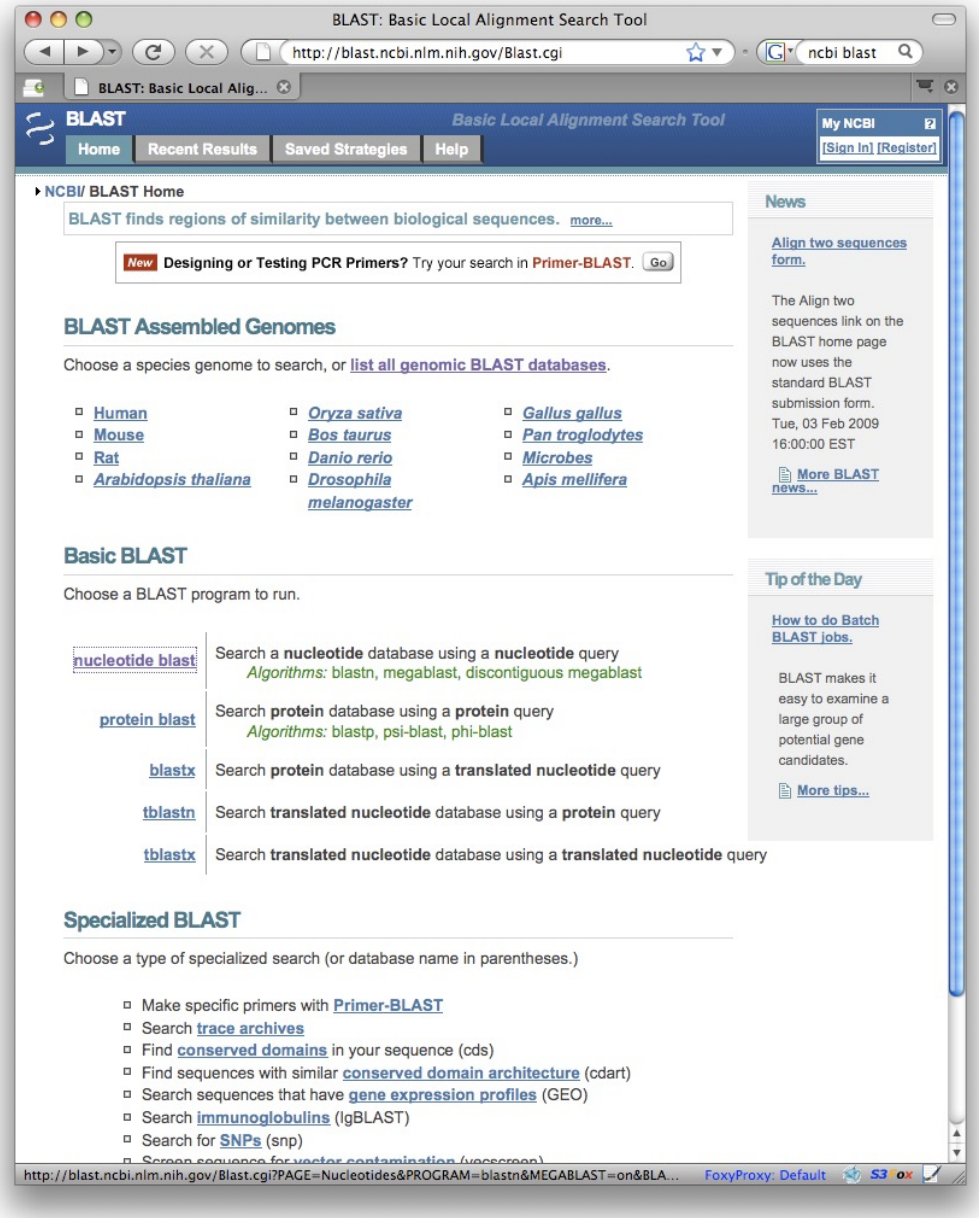
Not similar sequences

```
Query: HBA_HUMAN Hemoglobin alpha subunit
Sbjct: SPAC869.02c [Schizosaccharomyces pombe]

 Score = 33.1 bits (74),  Expect = 0.24
 Identities = 27/95 (28%), Positives = 50/95 (52%), Gaps = 10/95 (10%)

Query  30  ERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAH  89
           ++M  ++P      P+F+ +H  +      + +A AL N   ++DD+  +LSA  D
Sbjct  59  QKMLGNYPEV---LPYFNKAHQISL--SQPRILAFALLNYAKNIDDL-TSLSAFMDQIVV 112

Query  90  K---LRVDPVNFKLLSHCLLVTLAAHLPAEF-TPA  120
           K   L++   ++ ++ HCLL T+   LP++  TPA
Sbjct  113 KHVGLQIKAEHYPIVGHCLLSTMQELLPSDVATPA  147
```

Blast Versions

Program	Database	Query
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Protein	Nucleotide translated into protein
TBLASTN	Nucleotide translated into protein	Protein
TBLASTX	Nucleotide translated into protein	Nucleotide translated into protein

NCBI Blast



- Nucleotide Databases
  - nr: All Genbank
  - refseq: Reference organisms
  - wgs: All reads
- Protein Databases
  - nr: All non-redundant sequences
  - Refseq: Reference proteins