

Grundpraktikum Bioinfo - Week 2

Biological sequence analysis

Carl Herrmann
IPMB - Universität Heidelberg



UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Themen des Grundkurses



• Statistische Datenanalyse - Woche 1

- Beschreibende Statistik, graphische Darstellung von Daten
- Inferenzstatistik
- Datenanalyse
- Umsetzung der Konzepte mit R

*Carl Herrmann
Maiwen Caudron-Herger*

• Sequenzanalyse - Woche 2

- Untersuchung von unbekannten Sequenzen
- Sequenzalignments
- Phylogenetische Rekonstruktion
- Genomische Daten - Pre-Prozessierung von RNA-seq Daten

*Carl Herrmann
Benedikt Brors*

• Analyse von RNA-seq Daten - Woche 3

- Normalisierung
- Differentielle Analysen

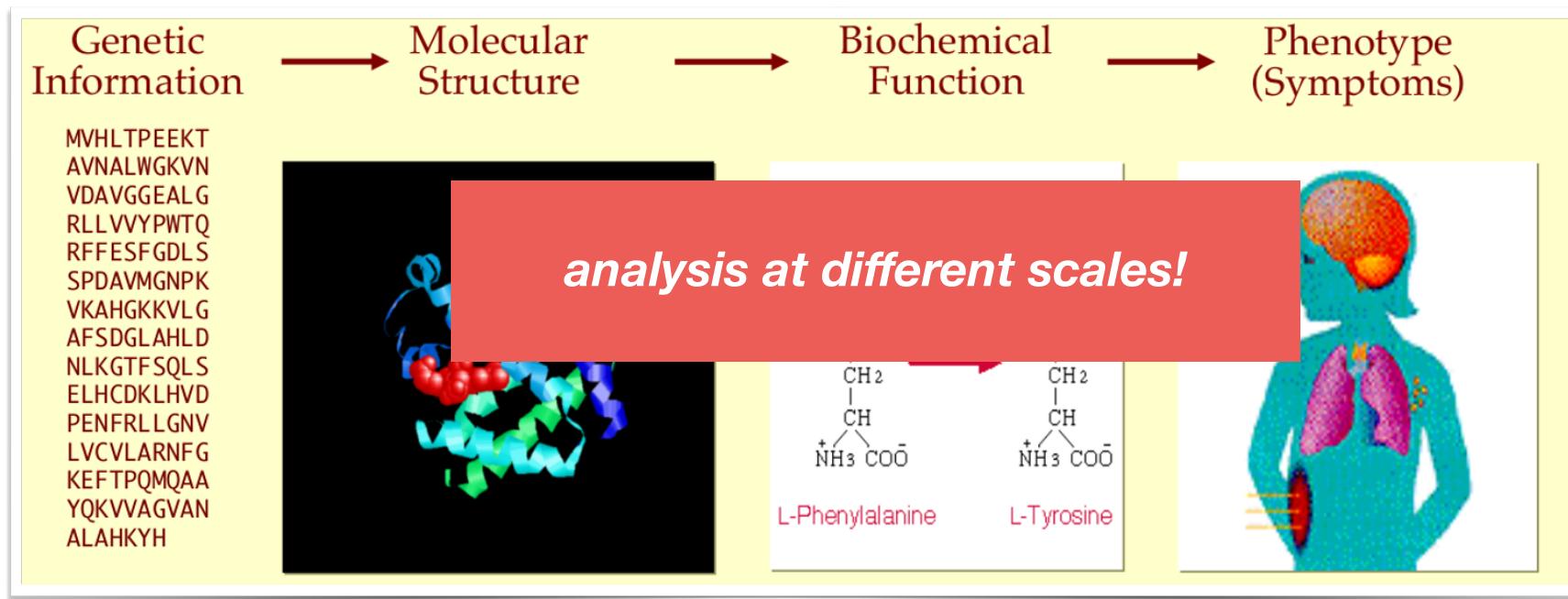
Simon Anders

Woche 2 - Sequenzanalyse

	Vormittags	Nachmittags (Albert Li, Robin Droit)
Montag	Einführung in Sequenzanalyse	
Dienstag	Alignments; BLAST	Annotation! Annotation metagenomischer Sequenzen
Mittwoch	Multiples Sequenzalignment; phylogenetische Rekonstruktion	
Donnerstag	Hochdurchsatzsequenzierung (Benedikt Brors)	Untersuchung von HTS-Daten mit Genome Browser (IGV)
Freitag		pre-prozessierung von RNA- seq Daten

What is Bioinformatics??

Bioinformatics: The science of **information** and **information flow** in biological systems, especially of the use of **computational methods** in genetics and **genomics**. (Oxford English Dictionary)



Why is bioinformatics necessary?

```

>chromosome:GRCh38:12:124911004:124917968:-1
TATGTGTTTACCTTACCAAACTTTAACGACAAAACGACGACACAAACAAAGCCCAA
TAGTTATAACTCTTCTAATTGGCTTAAAGTAGGCCAATTGGAGACTAGCTAAGTCG
GTTTGACCGCTGAGATGAATGGGTTATTAACTGAGCTGGAGCGTTGTGTTCCACCA
CAGTTATACAGCACAACAGGAAACAGGAAGAAGAACACATTCCCCAGGTTTTTTTT
ACTCTCTTGCAGGCTCTAACCTGAGCGCTTGTGACAACCTTTATGGTTATAT
AAGCGATGACAACAGGCAAGGCCAACCTTCCACCCACTCTGTGATTCCACCCCTTTCT
GGGGGAAGCAAGTTGTGACCTCTACTCATCTCTAGAGCTTCATGGGCTGGATT
AACTCACATACAAGGTGTAATTCAAGGGGGACTGAATCCCTAGGGGAGGGAA
AAACGAAGGCCGGAGGAGCAGCTGATCAACCCGCCCTCCCAGCTGAGGCCAGCA
CCCCGGCCCTTGTGACTCTACAGCTCATCACAGCTCCTTCCCCTGCCAACATCCGCC
GTTGCTGGGCTGTGCTCCCTTCATTGGTGCAGCTGAAAGCAATCTTCTGGGC
GCCGCAAGGAGGGATCTGGGTGGAACGAACTGAAAACGGATCCAGTGACTCATCC
GATTCTGCAACATCTAGACATCGCATTCTAGGTGAGCAGGGTTAGGTTAAAGAGCG
TGACTGTATACCTGACTCTGAGCTTCAAGGGACATTAGACTTTGGGAACCCCTTG
AAAACAGAAAAGCAGTTGTGACAACACAAAAATCTTGTAGGACAGGGTTCTTAATT
TCAGCACTTAAATATGTTGGCTGTATAATTCTTGTGTTGGGACTGTGCTCTCATAG
AATGTTCTGGCCATCCCTGGTCTACCAACAGCTGCTAGTATTAGTAGCATCCCT
AACCTCCCTGAGTTGTGACAACCAAAATCTCCAGATATTGCCAAATGTCCCTGAA
AGAGCAAATCACTCCTGTGAGAACCATGTTGAGAGAAACATCAGACACCTGCGTT
TGGGAGGCCATTCTTCTCTGTGACCTCTAGTGACTTGTGAGAAAGGTTATTCTGTT
GCTTCAATAACCTTAAGTCAAGGGAATCTGGGATTTCAATAATTTTCTGCTA
GAGACAACATTGTTCTCTTGTGACCTTAAGCTGGCCATTACCCACCTACCCCTCAT
CCCCCTCTCTCTTTCTCTGTGACCTCTGGGACTCTGAAAGCTATGTCAAAGCTGGT
GGCTGTTTTCTCTGTACCTTTGGACTCTGAAAGCTATGTCAAAGCTGGTAC
TCAGCCTTAAACACATGAATCACAGGCTTGAGCAGAGTAACAGCAGGAAACCA
GGCGTGTGAAAGCAGGCCCTGGTACCTCAGAGAAAGATCTCAGCTTAACTCCCT
AAAAACTCTTCCCGAAATGTAGGGGAAACAAAAACAAAAACTTGTAGGAAATGTG
ATTGGAAATCCTGGGTATCTGCAACTCTGGAAATGACACTTTATTCTGTTT
AATTCTCCAGTAACACATGAATTCTCTTAAATGTTAAAGTGGAGATGTCAA
GTCGTGTTGAAACATGCACAGATCTGGATTCTAAATAACACAGAAACATGGGTAC
CATAGAAAATATAGCAGGGTCTTGTGATGCTATTTAAGTCTGCCGTTAA
ATAATGTAATGTCATAACCTAAAACAGAATGCAGCTTTCTAAATGCCAAAATT
TTCACTGGGTTGAGCTCTTCTTTCTTTTGAGCTTTAAGGGTAGAAAATCAAAAT
GCTCCCTACCTCCAATACATTGTGAAACAGGGCTATTGTAGAAAGACTCCAGAAG
AACAGTGGCTGGGAAATGAGGGCATGAGGGAGACTTTCACTGTACACTTAAAGAC
ACATTAAAATCTACCATGTCGTATTTAAACATTAAAATACTGAAGA
AATGACAGCTGGGATCACAAAATAGGTGTTGTTAAAATTTAAAAGCTGGCTGGT
GCTTCTCTCCACTCTTCAACAAATAGGTGGGTTCTTCTGAGCTTAAACCTGGAC
TGACACCCCTACCTGCACTGAAATAAGGTGGGATTCTATTCTGTCTTTAAATCTC
GGTTAAATGGGCTGATTATGAATGAGCTCTGCTTACAATACTGTAACCCATTGAAAC
CAGAGAAATTTCCATGCCCTCTGGGTCATAGGAGCCTTCAACCTGGGCCCTC
AAAGTACAGGAAGGGTGGAAACAGCTAGAGGGATGTGCTTCGCTCAGCTTGGCT
CCAGCTAAAATAAAATCTGGTGGGTTTCCGCTCTTTTCAAAATTAAACCTGGAC
CCAGCTCTCTGCACTGCTCCCTGGAAAGTCTCGAGCTTCCCGACGCTTGGGCC
CGCCGCCCTGAGATCTGGCAGTCACTGGCTTCCGCTTCAACCTGGGCCCTC
GACGGCTGGGAGGCTAGGGAGGTGAAGGGGGCTGAGCAAAAGGAAGGCCGCTT

```

GCCGCCGGGCGCTCGGGAGGAGGGTGTGGAGACGCCAAGGGCTGTAGCTG
GGTCGGCAGAACGGTCTCCGTAACCTGGGGTGGGGGAGGCAGCAAATGCCGGCT
GTTCCCAGTCTGAATGGAAGACGCTGTGAGGGGGCTGTGAGGCTGTGAAACAAAG
TGGGGGGCATTGGTGGGGCAAGAACCAAGGGTCTGGAGGCTCTGCTAATGCGGAAAC
CTCTTATCGGGTGGAGATGGGCTGGGGCACCATCTGGGACCTGACGTGAAGTTGTC
CTGACTGGAGAACATCGCTTGTGCTCTGGCGGGGGCAGTTAGTGGGGTGCCTG
GGCAGTGACCCGTAACCTTGGGAGCGCCGCCCCCTCGTGTGAGCTCACCCGTT
TGTGGCTTATAATGAGGGTGGGGCACCTGCGGTTAGGTGCTGGTAGCTTCTCC
GTGCAAGCAGCAGGGTCTGGGCTAGGGTAGGCTCTCTGAATGCACAGGGCCGGAC
TCTGGTGGAGGGGGAGGATAAGTGGGGCTCAGTTCTCTGGCTGGTTATGACCTAT
TCTTAAGTAGCTGAAGCTCCGGTTTGAACATCTGGCTGGGGTTGGAGGTGTGTT
GTAGAAGTTTTAGGCACCTTGGAAAATGTAATCATTTGGCTAATATGTAATTTCAGT
GTAGAGTAGTAAATTGTCGGCTAAATTCTGGCGTTTTGGCTTTTTGTTAGACAAT
CAGATCTTCGTAAGAACCTGACTGGTAGAGGACCATCACCCTGAGGGTTGAGCCAGTGAC
ACCATCAGGAATCTCAAGGAAAGATCAGGAAAGGAGCATTCCCTGAGGACAGCA
AGGCTGATTTGCTGGAAAACAGCTGGAGAGATGGGCGCACCTGTCGTACTACAACATC
CAGAAAGAGTCCACCCCTGCACCTGGTGTCCGTCAGAGGTGGGATGCAAATCTCGT
AAGGACACTACTGGCAAGGACCATCACCCTGGAGGTCAGGCCAGTCAGACCCATGAGAAC
GTCAAAAGCAAGATCAGGAAAGGACATTCTCCCTGACCCAGCAGAGGTTGATCTT
GCCGGAAAGCAGCTGGAGATGGGCGCACCTGTCGTACTACAACATCCAGAAAGAGTCT
ACCTGCACTTGTGCTCGTCTAGAGGTGGGATGCACTTCTGTAAGACCCCTGACT
GGTAAGGACATCACCTCTGAGGGTGGGACCCAGTGCACCATCGAGAATGTCAGGAA
ATCCAAGATAAGGAGGCTTCCCTGATCAGCAGAGGTTGATCTTGGCGGAAACAA
CTGGAGAGTGGTGTGACCCCTGTCGACTACAACATCCAGAAAGAGTCCACCTGAC
GTACTCCGCTCTAGAGGTGGGATGCAAATCTCTGTAAGACACTACTGGCAAGACCA
ACCTTGGAGTGGAGGCCAGTGGACACTATCGAGAACGTCAGAAAGATCCAGAACAG
GAAGGGCATCTCTGACCCAGCAGAGGTTGATCTTGGCGGAAAGCAGCTGGAGATGG
CGCACCTGTCGACTACAACATCCAGAAAGAGTCTACCCCTGCACCTGGTGTCCGCTC
AGAGGTGGGATGCAAGATCTTCGTAAGGACCCCTGACTGTGAGAACCCATACTCTGAG
GAGCCGAGTGACACCATCTGAGAAAGATGTCAGGCAAAAGATCCAAGAGGACAGGATCCT
CTGGCACCCAGAGGTTGATCTTGGCCGAAAACAGCTGGAGAGTGGTGTACCCCTGCT
GACTACAACATCCAGAAAGAGTCCACCTGCACTGGTGTCCGTCAGAGGTGGGATG
CAGATCTTCGTAAGAACCCCTGACTGTGAGAACCCATACTCTGAGGTGGAGGCCAGTGAC
ACCATTGAGAATGTCAGGCAAAGATTCAGACAAGGAAGGGCATCCCTGACCCAG
AGGGTGTACTTGTGGGAAACAGCTGGAGAAGATGTCAGGCCACCCCTGTCGACTACAACAT
CAGAAAGAGTCCACCCCTGCACCTGGTGTCCGTCAGAGGTGGGATGCACTTCTG
AAGACCCCTGACTGTGAGAACCCATACTCTGAGAAGGGCAGTGACACCATGAGAAAT
GTCAGGAAAGATCTCAAGGAAAGGACATCCCTGACCCAGCAGAGGTTGATCTT
GCTGGGAAACAGCTGGAGAGTGGACGCCACCTGTCGACTACAACATCCAGAAAGAGTCC
ACCTGCACTTGTGCTCGTCTAGAGGTGGGATGCACTTCTGTAAGACCCCTGACT
GTTAGAACGACCATCTCTGAGAAGGGCAGTGACACCATCTGAGAAAGATGTCAGGAAACAA
ATCCAAGACAAGGAGGCTTCCCTGACCCAGCAGAGGTTGATCTTGTGGGAAACAA
CTGGAGAGTGGAGGCCACCTGTCGACTACAACATCCAGAAAGAGTCCACCCCTGAC
GTGCTCCGCTCTAGAGGTGGGATGCAAATCTCTGTAAGACCCCTGACTGTGAGAACCCAT
ACCTGCACTTGTGCTCGTCTAGAGGTGGGATGCACTTCTGTAAGACCCCTGACT
GAAGGGCATCTCTGACCCAGCAGAGGTTGATCTTGTGGGAAACAGCTGGAGATGG

Why is bioinformatics necessary?



Remember this is **6,965 bp**
out of
3,234,830,000 bp
in the human genome...

Why is bioinformatics necessary?



```

>chromosome:GRCh38:12:124911004:124917968:-1
TATGTGTTTACCAACATAATTAAAGACAAAACCGACGACGACAAACAAAGCCCAA
TAGTTAAATAACTCTTAATGTCCTAAAGGGATTATTAACTGAGCTGGAGCTTGACTGCG
TTTGACGCCCTGAGATGAATGGGTTATTAACTGAGCTGGAGCTTGACTGCG
CAGTTATACAGCACAGAAACAGGAAGAAGAACACATCCCCAGGTTTTTTTTT
ACTCTCTGTGCGAGCCCTTAACCTGCAGCGCTCTTGACAACCTTTATGTTTATAT
AAGCGATGACAAGGAAAGGCAACCCCTCCACCTCTGTGTTCCACCTTTCTTCTT
GGGGGAACGAAGATTGTGACCTCACTCATGCTCTCAGACAGTTCTAGGGGGCTGATT
AAGTCACATACAAGGTGTAATTTCAGAAGGGGAACGTAATCTTAGGGGAGGGAA
AACAGGAAGGGGGAGGAGGACGCCGATCACCCCGCCCGCCAGCTGAGCCAGCA
CCCCGGCCCTTGTGACTCTACCTCATCACAGCTCTTCGCGTGCCTACCCGCT
GTGCTGGCCCTGTGCTCTTCATTGGTGCCAGCTTGAGCAACATTTCTGGGC
GCCTGCAAAAGGAGCGGATCTGTGGTGAAACGAACTGAAAACGGATCCAGTGACTCATCC
GATTCTTGCACTTCTAGCACATGCCATTAGTGAGCTAGGGGTTAGGTTTAAAGAGCC
TGACTGTATACCTGACTCTGTGCAAGTGGACATTAGCTTGGGAAACCTTCTTGG
AAAACAGAAAAGCAGTTGTGCAACACAAAAATCTTGTAGGACAGGGTTCTTAATT
TCAGCACTATAATATGTTGGCTGATAATTCTTGTGTTGGGACTGTCCCTCATAG
AATGTTGCTGGACCTCCCTGGTGTACCAACAGCTGAGCTATTAGTAGTACCTCCC
AACTCCCTCTGAGTTGTGACACCAAGAATATCTCCAGATATTGCAAAATGCTCCC
AGAGCAATCACTCTTGTGAGAACATTGTTAGAGAACATCAGACACCTGCGTT
TGGGAGGCCATTCTCTCCCTGCACTCTAGTGACTTGGAGAACGGTTATTCTGTGTT
GCTTCAATAACATTAACTGAGGGAACTGCGGATGTGGAGTGTTCATAATTTTTCTGCTA
GAGACAAACATTGTTCTCTTGCACTTAACCTTTGGCCATTACCCACCTACCCCTCAT
CCCTCTCTCTCTCTCTCTGCTACCTTTGGACTCTGAAAAGGATATGTCACAAAGCTGGTAC
GCCGCTTTTCTCTGACCTTTGGACTCTGAAAAGGATATGTCACAAAGCTGGTAC
GCCGCTTTTAAACACATGAACTCAACAGGCTTGCGAGGTAACAGCAAAAGCAACCA
GGCGCTGGTAAGGGAGGCCGGCTGGTCACCTCAGAACAGAACATCAGCTTTTAACTCTT
CAAAACTCTCCCGAAATGAGGGAAAAACAAAAACAAAAACTTGCAGAACATGTT
ATTGGAAATCCTGTTGATCTGCAACATTCTGGAAATGACACTTTTATATTGCTTT
ATTCTCCGAAACATGAACTCATTCTTTAAATGTTAAATGGAGATGCTAA
GTCGTGTTGCAACATGCAACATGTCGATTTGATTCATAAAACACAGAACATGGTTAC
CATAGAAAATATATGAGGTTCATTTGCTTGATGCATATTAACTGTCGCGTTAA
ATAATGTAATGTCATAACCTAAACAAAGATCAGGCCCTTCTCAATGCGAAATATT
TTCTGGCTGGTAGCCCTTCTTCTTTTGTGTTGGGTTTAAGGGTAGAAATCGAAATA
GCCCTCTCCACCTTCAACATATTGCTGAGAACACGGGACTATTGAGAAGCTGCCAGAAG
AACAGTGGCTGGGAAATTGAGGGCATGAGGGAGACTTTCACTGTACACTTAAAGAC
ACATTAAAATCTCACCATGTCGATTTGATTTAAAACATTAAAAAAATACATGAAAGA
AATGACAGCTGGGATCACAAAATAGGTTGTTGTAATTTAAAAGCTGCTGGTTAC
GCTTCTCTCCGACCTTCTTACCAAAATAGGTTGAGAATTCCTGCTGAGAAATCACCTTCAATT
TGACACCTCACCTGCACTGAAATAAGGTTGGGATTCTATTCTGCTTTAAATCTCC
GGTTTAAATGGCTCCGATTATGTAATGGGATCTGCCATACATAATCGACCATCTGGAAAC
CAGAGAAATTCTCATGCCCTCTGGCATCAACTGAGGCCATTCTCACCTGCCCTC
AAATGTCAGCAAGGGTGGAAACAGCTGAGGATCTGAGGAGCTGCTGGCTGAGCCCTTGGCTT
CCAGCTAAATAAAACTGTTGGGTTCCGCCCTTTTCCAAATTAACTGGACAC
CCAGCTCCCTCTGAGATCTGCCAGTCTGCCCTGGAAACTTCTCGAGGCCCTCCCGGTTAGGGCCA
CGCCGGCCGGCTGAGATCTGCCAGTCTGGCTCTGGCCCGGCGGCGGCGGCGGCG
GACCCGGCTGGAGGCTGAGGAGCTGGAGGTTGAAGGGGCTGACCAAAAGAACCCCCCTCATTA

```

**2 exons
of a gene
encoding
ubiquitin!**

GGCCGCCGGGCCGCTCGTGGGACGGAGGGCTGTGGAGAGACCGCCAAGGGCTGTAGTCG
GGTCGGCAGAACAGGTTCCCTGAECTGGGGGGGGGGGGCCGCAAAATGGCCGCT
GTTCCCGAGTCTTGAAAGACGCTTGTGAGGGGGGCTGTGAGGTCCTGAAACAAAGG
TGGGGGCCATGTTGGGGCGCAAGAACCAAGGTTCTGAGGGCTCTGCTAATGCCGAAAGG
CTCTTATTGGGTGAGATGGGCTGGGCACCACATCTGGGACCCCTGACCGTAAGATTGTC
CTGACTGGAGAACATTCGCTTGTGCTCTGGCGGGGGCCGCTATTGCGGTTECCGGT
GGCAGTCACCCGCTACCTTGTGGAGGCCGCCCTCGCTGTGCTGAGGTCACCCGGT
TGTGCGCTATAATGCAAGGTTGGGGGCCACTGCGGTAGGTGCTGGTAGGCTTCTCC
GTCCGAGGACGCCAGGGTCTGGGCTTAGGGCTCTCTGAATCGACAGGCCGCCGACC
TCTGGTGGGGGGGGGATAAGTGGAGGCCAGTGTCTGGCTGGTTATGACCTATC
TCTTAAGTAGCTGAACCTGGCTTTGAACTATGCGCTGGGGCTGGGAGTGCTT
GTGAAGTTTTAGGCACCTTTGAAATGTAATCATGGGCTAATATGTAATTTCACT
GTTAGACTAGAAATTGTCGCTAAATTCTGGCGTTTTGGCTTTTGTAGACAACT
CAGATCTTCGTAAGAGCTGACTGGTAAGGACCATCACCTCGAGGGTGGAGCCAGTGAC
ACCATCGAGAAATCTCAAGGCAAAGATCCAAGATAAGGAGGACATCCCTGAGGACAG
AGGCTGATTTGCTGGAAAACAGCTGGAGATGGGCCACCTGTCGACTACAACATC
CAGAAAGAGTCACCCCTGCCACCTGGTGTCCGCTCAAGGGTGGGATGCAAATCTCGT
AAGGACACTCACTGCCAGAACGACCATCACCTTGAGGTGAGGCCAGTGAACCCATCGAGAAC
GTCAACAGGAAAGATCTCAAGGCAAAGGAGGCCATTCTCTGACCGAGGTTGATCTT
GCCGGAAAGCAGCTGGAGATGGGCCACCTGTCGACTACAACATCCAGAAAGAGTCT
ACCCCTGACCTGGTGTCTGGCTTCTAGAGGTTGGGATGCACTCTGTCGAGGACCTGACT
GGTAAGGACATCACCTCTGAGGTGGAGGCCAGTGACACCATCGAGAAATGTCAGGCAAAG
ATCCAAGATAAGGAGGCCATTCTCTGACTCACCGAGGTTGATCTTCTGGGAAAACAG
CTGGAGATGGTCGTACCTGTCGACTACAACATCCAGAAAGAGTTCACCTGCACT
GTAACCTCTCTGAGGGTGGGATGCAAATCTCTGTCGAGACACTCACTGCCAAGGAC
ACCCCTGAGGTGCGAGGCCAGTGACACTATCGAGAAAGCTGCCAAAGGAGATCCAAG
GAAGGGCATCTCTGACCCAGGAGGTTGATCTTGGCGGAAAGCAGCTGGAGATGG
CGCACCTGTCGACTACAACATCCAGAAAGAGTCTACCTGCACTGGTGTCCGCTC
AGAGGTGGGATCTCAGAGTCTTCGTAAGGACCTGACTGTTGAAGACCATCTCTGCAAGT
GAGCCGAGTGACACCATCTGAGAAATGTCAGGCAAAGATGCCAGAAAGGAGCATCC
CTTGACCCAGGAGGTTGATCTTCTGGGAAAACAGCTGGAGATGGTCACCTCTGTC
GACTACAACATCCAGAAAGAGTCCACCTGCACTGGTGTCCGCTCAGAGGTGGGATG
CAGATCTTCGTAAGGACCTGACTGGTAAGGACCATCACTCTGAGGTGGAGGCCAGTGAC
ACCATTGAGAAATGTCAGGCAAAGATCCAAGACAAGGAAGGCAAGGCTACCCCTCTGAC
AGGGTGTACTTTGCTGGGAAACAGCTGGAGAGATGTCAGGCCACCCCTGTCGACTACAACAT
CAGAAAGAGTCCACCCCTGCACTGGTGTCCGTTAGAGGTGGGATGCGAGATCTCGT
AAGACCCCTGACTCTGTAAGGACCATCACTCTGCAAGTGGGCCAGTGACACCCATTGAGAAT
GTCAGGCAAAGATCTCAAGAACAGGAAGGCACTCCCTGACCCAGGAGTTGATCTT
CTGGGAAAACAGCTGGAGATGGACGCCACCTGTCGACTACAACATCCAGAAAGAGTCC
ACCCCTGACCTGGTGTCCGTTAGAGGTGGGATGCGAGATCTCTGTCAGACCCCTGACT
GGTAAGGACCATCACTCTGCAAGTGGGCCAGTGACACCCATTGAGAATGTCAGGCAAAG
ATCCAAGACAAGGAAGGCACTCCCTGACCCAGGAGTTGATCTTCTGGGAAAACAG
CTGGAGATGGAGCGCACCTGTCGACTACAACATCCAGAAAGAGTTCACCCCTGACCT
GTGCTCTGGCTCTCAGAGGTGGGATGCAAATCTCTGTAAGACCCCTGACTGTCAGGACCAT
ACCCCTGACCTGGGAGGTTGAGGACCATCTGAGAAATGTCAGGCAAAGATCCAAGGAA
GAAGGGCATCCCTCTGATCTGACAGGAGGTTGATCTTCTGTCAGGCAAAGCTGGAGATGG

Genome annotation

Find german
words for animals

Y	A	Z	E	C	W	F	I	U	H	A	S	T	G	W
G	I	M	T	X	N	Z	G	S	K	O	B	M	Y	R
F	J	B	H	U	N	D	J	M	C	R	J	A	N	P
R	K	Z	P	U	O	Y	H	W	X	T	R	U	J	Z
H	A	R	M	B	S	R	W	O	G	A	N	S	F	G
I	F	G	C	P	F	P	R	L	I	H	F	O	J	B
T	F	S	G	A	J	F	M	F	R	T	Z	N	W	K
C	E	U	F	S	X	E	H	I	K	A	T	Z	E	T
H	T	X	O	G	K	R	M	W	G	U	A	R	G	P
I	R	G	H	W	N	D	Y	M	T	F	K	X	J	C
K	Z	N	U	H	J	B	U	S	A	O	T	R	B	G
W	J	N	H	R	Y	O	T	I	G	E	R	S	P	F
A	G	R	N	J	I	C	H	M	Z	S	G	G	K	X

Genome annotation



Find klingon words for whatever...

Without bioinformatics...



Genome wide association studies

EDITORIALS

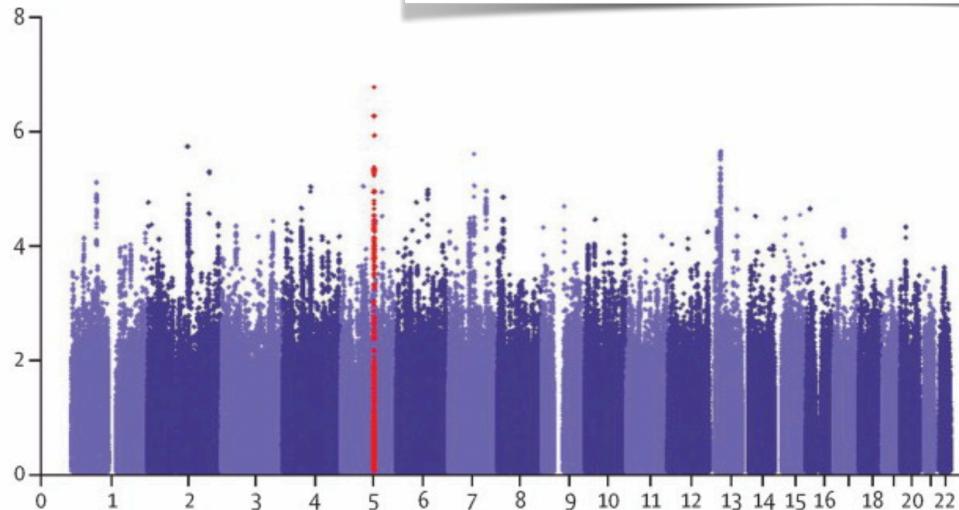
Finding a Needle in the Haystack

Leveraging Bioinformatics to Identify a Functional Genetic Risk Factor for Sepsis Death*

Meyer, Nuala J. MD, MS

[Author Information](#) 

Critical Care Medicine 43(1):p 242-243, January 2015. | DOI: 10.1097/CCM.0000000000000664



- Single-nucleotide variants (SNPs)
- ~ 4-5 million per genome
- 600 million different SNPs identified so far...
- **Which ones are associated to disease phenotype?**

[Rautanen et al., Lancet Resp. Med (2015)]

Biological databases

Bioinformatics

Sequence alignments

Genome annotation

ML & Artificial Intelligence

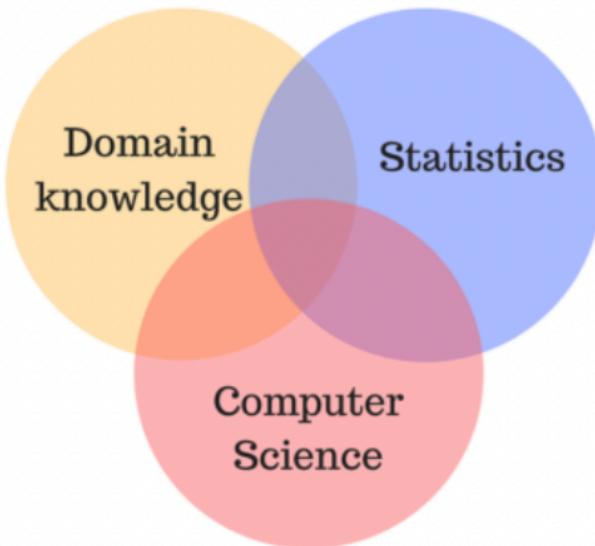
Computational Biology

Data Science

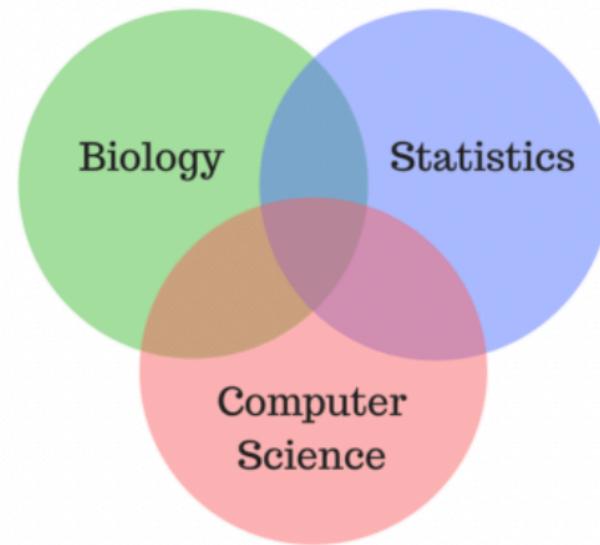
Systems Biology

Bioinformatics = data analysis?

Data Science



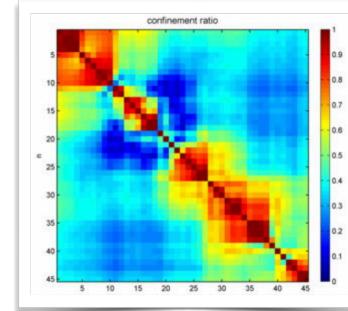
Bioinformatics



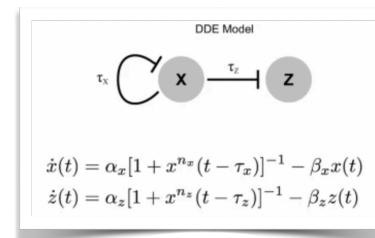
[OMGenomics]

Aspects of bioinformatics/computational biology

- 1. Data analysis:** Going from raw data, to clean data, to statistical and visual interpretations of the results
- 2. Bioinformatics software development:** Developing software to do bioinformatics analysis, big enough software products to publish as independent methods papers and to be used by other scientists.
- 3. Modeling:** Performing simulations and writing equations to represent biological systems ("systems biology")



```
from __future__ import print_function
from multiprocessing import Process
device = '/dev/ttys0'
def read():
    print('READING')
    f = open(device, 'rb')
    while True:
        out = f.read(1)
        if out != '':
            print(out, end='')
p = Process(target=read)
p.start()
f = open(device, 'w')
```



[OMGenomics]

Aspects of bioinformatics/computational biology

1. **Data analysis:** Going from raw data, to clean data, to statistical and visual interpretations of the results

2. **Bioinformatics software development:** Developing software to do bioinformatics analysis, big enough software products to publish as independent methods papers and to be used by other scientists.

3. **Modeling:** Performing simulations and writing equations to represent biological systems.

***what we are doing
in these 3 weeks !***

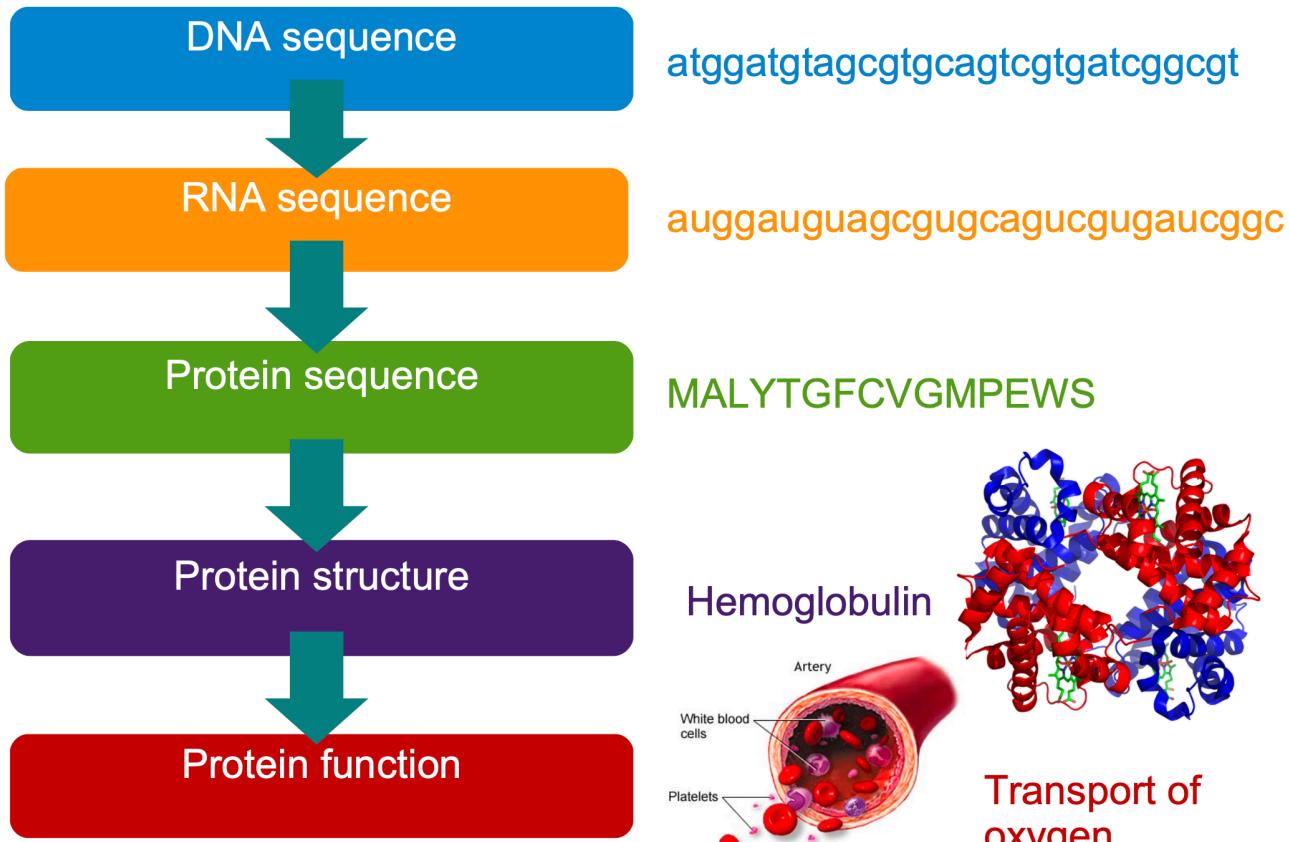
data =

- ***numerical data***
- ***sequence data***
- ***genomics data***

[OMGenomics]

Sequence analysis

What sequences?



What sequences?

DNA / RNA sequences
4 bases

5' A G T A C G 3'
3' T C A T G C 5'

Protein sequences
20 amino acids

N_{term} M H P G C V C_{term}

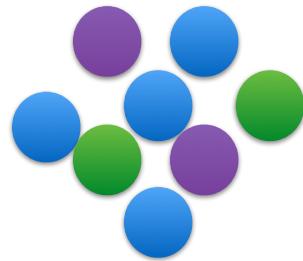
FASTA formated sequence:

```
>free text annotation
AGTACGGGTAATGCGTAGCATG
TACGTATGCTATCTGTACGA
```

FASTA formated sequence:

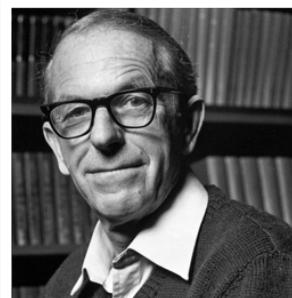
```
>free text annotation
MHPGCVGTRFSPMNWVAFGRDS
```

Initial protein models...

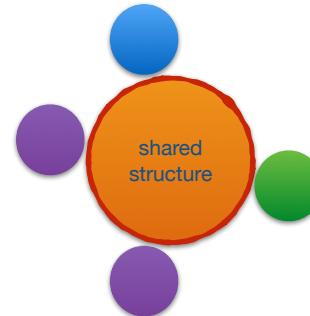


"soup" of amino-acids?

Fred Sanger,
sequencing of
insulin protein
1951/1952



"Proteins have a unique linear sequence"



Shared structure + micro-heterogeneity?

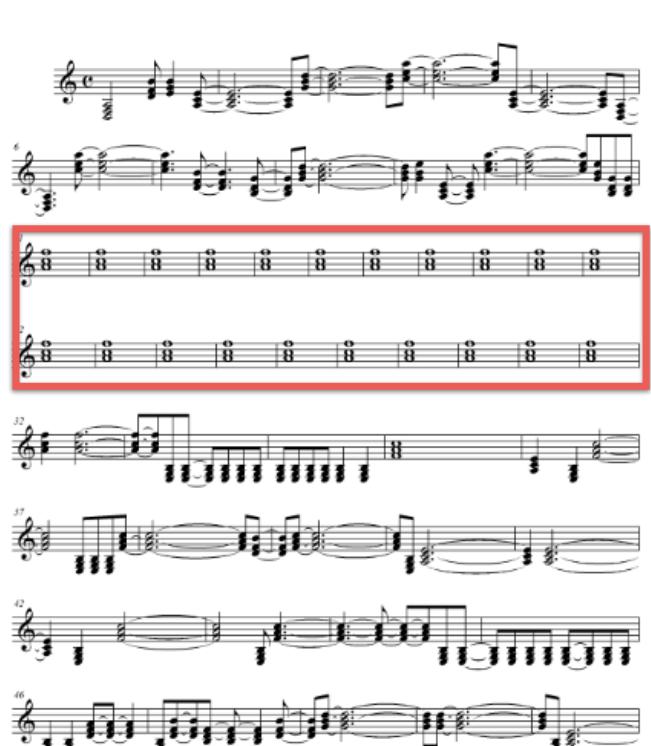


Nonetheless, the correctness of the concept that all molecules of a very carefully prepared protein have the same unique structure and configuration is still controversial, as shown by the divergent conclusions in two recent reviews (124,

[Colvin, Smith, Cook (1954)]

Initial protein models...

Huntingtin



The musical score consists of six staves of music. The first staff starts with a treble clef, common time, and a key signature of one sharp. It features a continuous eighth-note pattern. The second staff begins with a treble clef, common time, and a key signature of one sharp, also showing an eighth-note pattern. The third staff starts with a treble clef, common time, and a key signature of one sharp, continuing the eighth-note pattern. The fourth staff starts with a treble clef, common time, and a key signature of one sharp, continuing the eighth-note pattern. The fifth staff starts with a treble clef, common time, and a key signature of one sharp, continuing the eighth-note pattern. The sixth staff starts with a treble clef, common time, and a key signature of one sharp, continuing the eighth-note pattern. A red box highlights the first four staves.

32 37 42 46

gene2music 2007

Bio-music

[Takahashi & Miller (2007)]

>sp|P42858|HD_HUMAN_Huntingtin
MATLEKLMKAESLKS PQQQQQQQQQQQQQQQQQQQPPPPPPPPPQLPQPPPQAQP
LLPQPQPPPPPPPPPPPUVVAEEPLIURDKKEIGATKKDIVNHCLTICENIVAQSVRNSPE
FQKLLGIAMELFILLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG
NFANDNEIKVLLKAFIANLKSSPTIRRTAAGSAVSICQHSRRTQYFYSWLLNVLLGLLV
PVEDEHSTLLILGVLLTRYLVPLLQQVKDTSLKGSGFGVTRKEMEVSPSAEQLVQVYEL
TLHHTQHQDHNVVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAAKEESGGRSRSGSI
VELIAGGGSSCSPVLSRKQKGKVLLGEAALEDDSESRSVDSSALTASVKDEISGELAA
SSGVSTPGSAGHDIIITEQPRSQHTLQADSDVLDASCDLTSSATDGDEEDILSHSSSQVSAV
PSDPAMDNLNDGTQASSPISDSSQTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD
EDEEATGILPDEASEAFRNSSMALQOAHLLKNMSHCRQPSDSSVDFVLRDEATEPGDQE
NKPCRIKGDIGQSTDDDSAPLVHCVRLLASFLTGKNVLVPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLICSL

Bio-literature

Decode biological sequences...

VNNBM SSDII XZOWR TIEHI FQKNN
WKCSS DGLHG GXEMJ IWKBA YZGWJ
QTAHK AUSVR SCJTR OQ

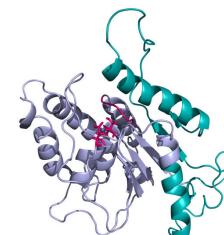
MTQFDKQYNSIIKDIXINNGISDEEFDVRTKWDSDGTPAHTLSVISQMRFDNSEVPILTT
KKVAWKTAIKELLWIWQLKSNDVLNMMGVHIWDQWKQEDGTIGHAYGFQLGKKNRSLN
GEKVDQVDYLLHQLKNNPSSRRHITMLWNPDLELDAMALTPCVYETQWYVKHGKLHLEVRA
RSNDMALGNPFNVFQYNVLQRMIAQVTGYELGEYIFNIGDCHVYTRHIDNLKIQMEREQF
EAPELWINPEVKDFYDFTIDDFKLINYKHGDKLLFEVAV



*Mignonne allons voir si la rose
Qui ce matin avait éclosé
Sa robe pourpre au soleil*

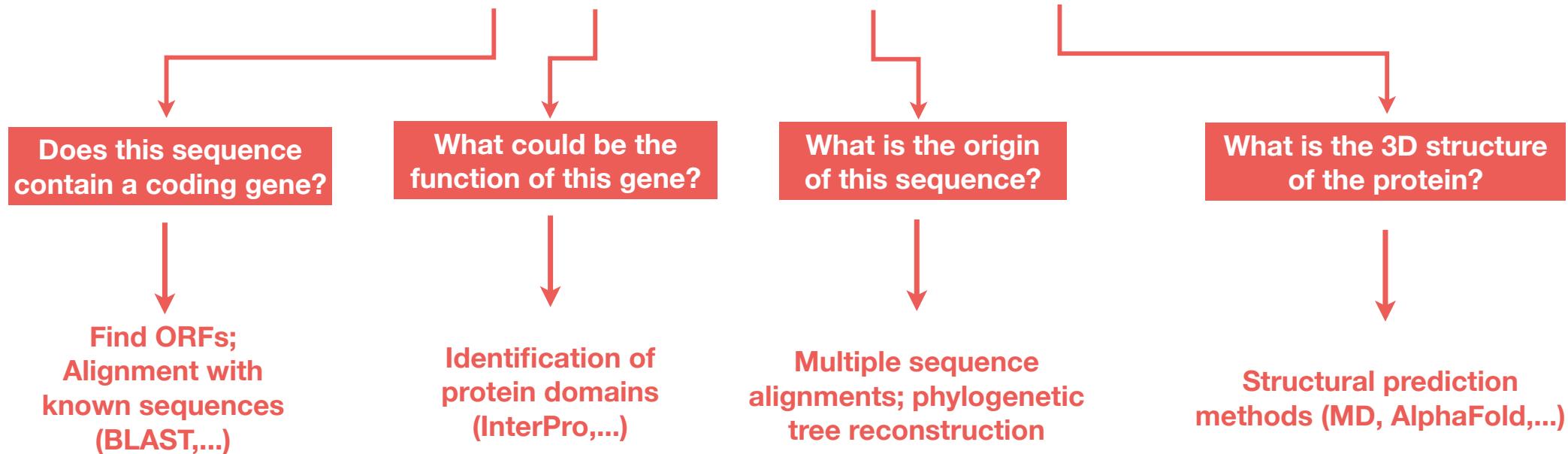


Tyrosine kinase /
response to osmotic stress



Which questions can we answer?

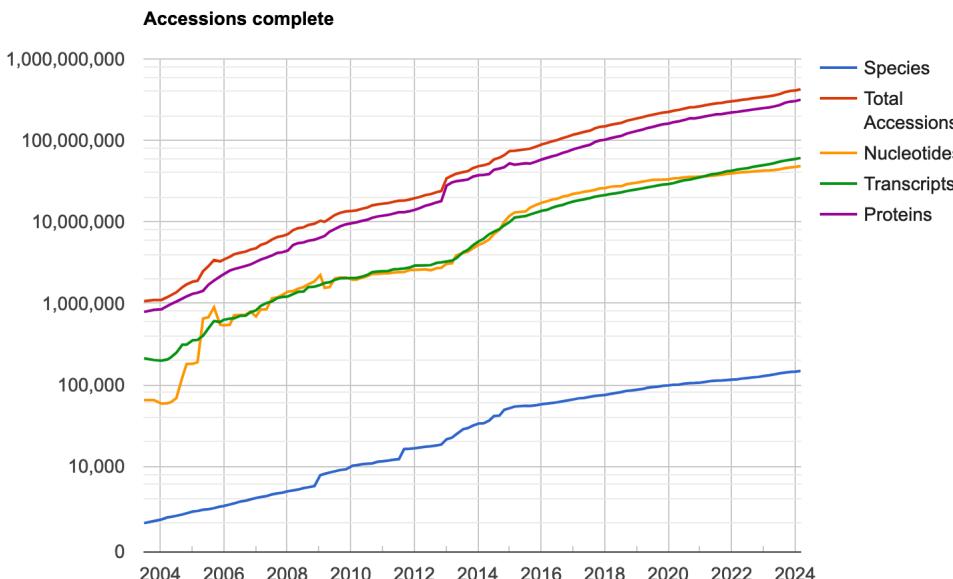
GGAACACTCGCGCTCGTTGAGGTAGACGGCGTTGCTGTGCGAGTCGCCAGCACCTGGATCTCGATGTGGCGGGCTCCTCCACGAACCTCTCCACCAGC
ATGGCGTCGCGCAAAGGAGGCAGCCGCGCTCCCTGGATGGAGAGGTGGTAGCCCTCGCGCCTCGGTGTCGTTGAGGCCACGCCATGCCCTGCGC
CGCCGCCGGCGAGGCCCTGATCATGACCGGGTAGCCGATCTCCGCGCATCTCGAGCATGTCGCCCTCGTCGCGCACCTCGCCCCAACAAAGCCGGCAC
CACGTTGATGCCCGCCTCGGTGCGATGCGCTTGCGATCTGGCGCCATCACCTCGAGCGCGTGTGCTGCTTGGGCCCCACGAAGGTGACGCCGTGC
TTCTCGAGCAGGTGCGACAATTGGTGTCTCCGAGAGGAACCCGTAGCCGGGGTGGACCGCGTCGGCGCCGTCGACGACCCCTCGAGGATGCGGT
CCATGGCCAGGTAGGACTCGGAGGACGCGGGCCGGCCCACGTGCGCTCGACTCGTCCGCCATCTGAACGTGCTTGGACGACGGCGGGCGACGAGTACACC
GCCACCGTGGGATCCCCATCTGGCGCGGGTGCACAAACACGGGACCGCATCTGCCCGGTTGGCACCAAGGAGCTTGGTCACTCCGGTGGAGGCTG
CTGCTGCGGCTGCGCGGAGAGCCTGCGCGAGGGCTCCCTCACCAACCGCCACGTTCTGGACCGTGTGCGCGGATCA



Sequence databases

NCBI GeneBank / RefSeq

comprehensive, integrated, non-redundant,
well-annotated set
of sequences, including genomic DNA,
transcripts, and proteins



Homo sapiens oligodendrocyte transcription factor 2 (OLIG2), RefSeqGene on chromosome 21

NCBI Reference Sequence: NG_011834.1

[FASTA](#) [Graphics](#)

[Go to:](#)

LOCUS NG_011834 10262 bp DNA linear PRI 15-JUN-2020
DEFINITION Homo sapiens oligodendrocyte transcription factor 2 (OLIG2), RefSeqGene on chromosome 21.
ACCESSION NG_011834
VERSION NG_011834.1
KEYWORDS RefSeq; RefSeqGene.
SOURCE Homo sapiens (human)
ORGANISM [Homo sapiens](#)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

COMMENT REVIEWED [REFSEQ](#): This record has been curated by NCBI staff. The reference sequence was derived from [AP000286.1](#) and [AP000287.1](#). This sequence is a reference standard in the RefSeqGene project.

Summary: This gene encodes a basic helix-loop-helix transcription factor which is expressed in oligodendroglial tumors of the brain. The protein is an essential regulator of ventral neuroectodermal progenitor cell fate. The gene is involved in a chromosomal translocation t(14;21)(q11.2;q22) associated with T-cell acute lymphoblastic leukemia. Its chromosomal location is within a region of chromosome 21 which has been suggested to play a role in learning deficits associated with Down syndrome. [provided by RefSeq, Jul 2008].

PRIMARY REFSEQ_SPAN PRIMARY_IDENTIFIER PRIMARY_SPAN COMP
1-4501 AP000286.1 18016-22516
4502-10262 AP000287.1 1-5761

FEATURES source Location/Qualifiers
gene 1..10262
/organism="Homo sapiens"
/mol_type="genomic DNA"
/db_xref="taxon:9606"
/chromosome="21"
/map="21q22.11"
/gene="OLIG2"
/locus="NG_011834.1"
/protein="OLIG2_Protein"
/refseq="RefSeqGene"
/species="Homo sapiens"
/taxon_id="9606"

[Link to the sequence]

Sequence databases

UniProt/Swissprot

reference protein databases containing manually curated entries and automatic translations



Total number of entries in this release of UniProtKB

Section	Number of entries in total	Number of entries with an annotation update
UniProtKB	248,805,733	117,146,151
Reviewed (Swiss-Prot)	571,282	436,850
Unreviewed (TrEMBL)	248,234,451	116,709,301

Q13516 · OLIG2_HUMAN

Proteinⁱ Oligodendrocyte transcription factor 2
Geneⁱ OLIG2
Statusⁱ UniProtKB reviewed (Swiss-Prot)
Organismⁱ Homo sapiens (Human)

Amino acids 323 (go to sequence)
Protein existenceⁱ Evidence at protein level
Annotation scoreⁱ 5/5

Entry Variant viewer 337 Feature viewer Genomic coordinates Publications External links History

BLAST Download Add Add a publication Entry feedback

Functionⁱ

Required for oligodendrocyte and motor neuron specification in the spinal cord, as well as for the development of somatic motor neurons in the hindbrain. Functions together with ZNF488 to promote oligodendrocyte differentiation. Cooperates with OLIG1 to establish the pMN domain of the embryonic neural tube. Antagonist of V2 interneuron and of NKX2-2-induced V3 interneuron development. [By Similarity](#)

GO annotationsⁱ

Access the complete set of GO annotations on QuickGO [↗](#)

[Link to the sequence]

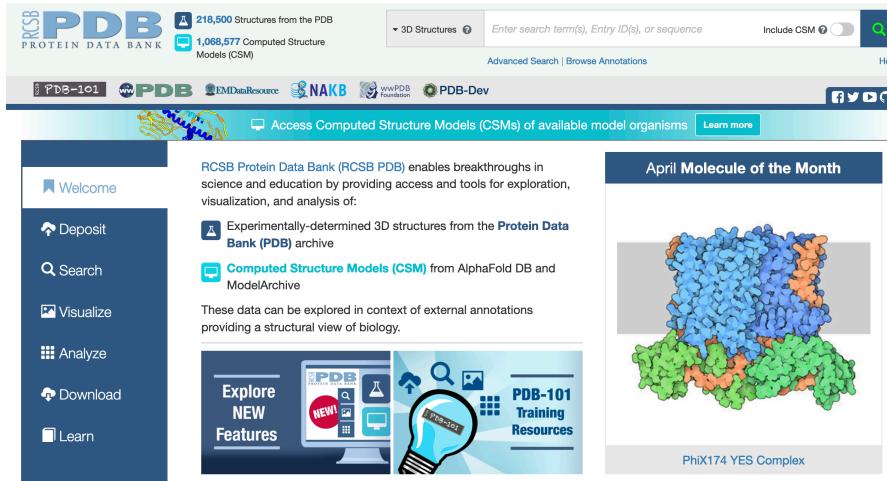


Sequence databases

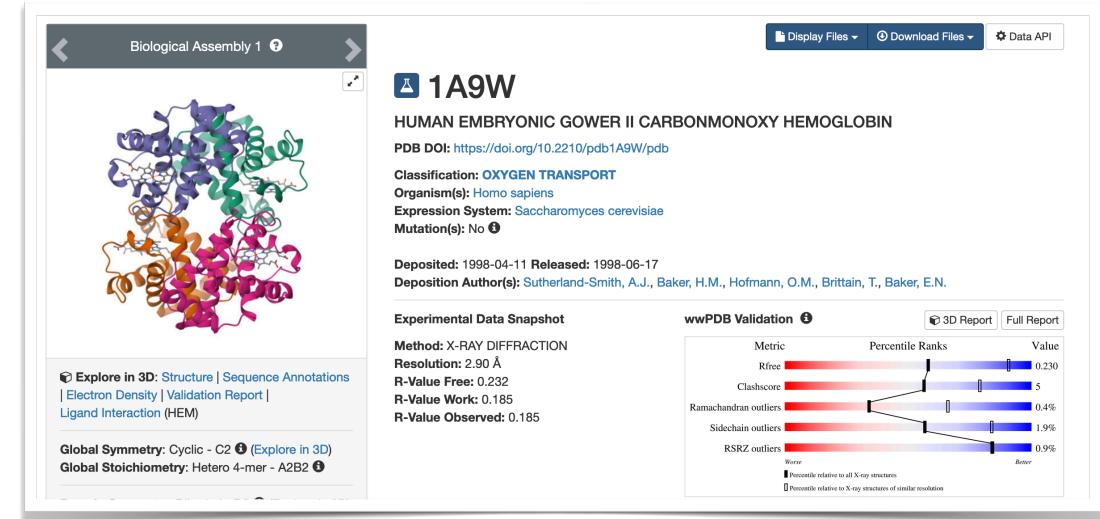


Protein Databank (PDB)

database of protein sequences including
structural 3D conformation



The screenshot shows the homepage of the RCSB Protein Data Bank (PDB). It features a search bar at the top with options for "3D Structures", "Enter search term(s), Entry ID(s), or sequence", and "Include CSM". Below the search bar are links for "Advanced Search" and "Browse Annotations". The main content area includes a "Welcome" sidebar with links for "Deposit", "Search", "Visualize", "Analyze", "Download", and "Learn". A central panel displays the "April Molecule of the Month" for the PhiX174 YES Complex, showing a 3D ribbon model of the protein structure.



The screenshot shows the detailed entry page for the protein structure 1A9W. The page title is "1A9W HUMAN EMBRYONIC GOWER II CARBONMONOXY HEMOGLOBIN". It provides the PDB DOI (<https://doi.org/10.2210/pdb1A9W/pdb>), classification (OXYGEN TRANSPORT), organism (Homo sapiens), expression system (Saccharomyces cerevisiae), and deposition details (1998-04-11, Sutherland-Smith, A.J., Baker, H.M., Hofmann, O.M., Brittan, T., Baker, E.N.). The page features a large 3D ribbon model of the protein. On the right, there is a "wwPDB Validation" section with a table comparing validation metrics against all X-ray structures and structures of similar resolution. The validation table includes:

Metric	Percentile Ranks	Value
Rfree	5	0.230
Clashscore	0.4%	5
Ramachandran outliers	1.9%	0.185
Sidechain outliers	0.9%	0.185
RSRZ outliers	0.9%	0.185

[link to entry]

Genomic databases

Gene Expression Omnibus (GEO)
database of gene expression / high-throughput sequencing data (RNA-seq / ChIP-seq / ATAC-seq / ...)

NCBI

GEO
Gene Expression Omnibus

GEO Publications | FAQ | MIAME | Email GEO | Login

NCBI » GEO » Info » GEO Overview

GEO Overview

- General overview
- Data organization
- Query and analysis

General overview

GEO is an international public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community.

(→ Thursday lab!)

GEO DataSets skin RNA-seq | Create alert Advanced

Summary ▾ 20 per page ▾ Sort by Default order ▾ Send to: ▾

Search results

Items: 1 to 20 of 6150

<< First < Prev Page 1 of 308 Next > Last >>

1. Neutral evolution of snoRNA Host Gene long non-coding RNA affects cell fate control
(Submitter supplied) A fundamental challenge in molecular biology is to understand how evolving genomes can acquire new functions. Several recent studies have underscored how non-conserved sequences can contribute to organismal diversification in the primate lineage. Actively transcribed, non-coding parts of the genome provide a potential platform for the development of new functional sequences, but their biological and evolutionary roles remain largely unexplored. more...
Organism: Homo sapiens
Type: Expression profiling by high throughput sequencing
Platform: GPL24676 8 Samples
Download data: CSV
Series Accession: GSE263910 ID: 200263910

2. [BD in fibroblasts in vitro model of circadian genetic and genomic studies: A temporal analysis \(RNA-Seq\)](#)
(Submitter supplied) Bipolar disorder (BD) is a heritable disorder characterized by shifts in mood that manifest in manic or depressive episodes. Clinical studies have identified abnormalities of the circadian system in BD patients as a hallmark of underlying pathophysiology. Fibroblasts are a well-established in vitro model for measuring circadian patterns. We set out to examine the underlying genetic architecture of circadian rhythm in fibroblasts, with the goal to assess its contribution to the polygenic nature of BD disease risk. more...
Organism: Homo sapiens
Type: Expression profiling by high throughput sequencing
Platform: GPL20301 78 Samples
Download data: TSV
Series Accession: GSE263713 ID: 200263713

3. [BD in fibroblasts in vitro model of circadian genetic and genomic studies: A temporal analysis \(ATAC-Seq\)](#)
(Submitter supplied) Bipolar disorder (BD) is a heritable disorder characterized by shifts in mood that manifest in manic or depressive episodes. Clinical studies have identified abnormalities of the circadian system in BD

Genomic databases

Gene Expression Omnibus (GEO)
database of gene expression / high-throughput sequencing data (RNA-seq / ChIP-seq / ATAC-seq / ...)

The screenshot shows the NCBI GEO Overview page. At the top, there's a sidebar with links to 'General overview', 'Data organization', and 'Query and analysis'. Below this, the main content area has a heading 'GEO Overview' and a sub-section 'General overview' which states: 'GEO is an international public repository for all forms of high-throughput functional genomic data'. A red box highlights this text. To the right, there's a table for 'GSE263713' with columns for 'Supplementary file', 'Size', 'Download', and 'File type/resource'. A red box highlights the 'Supplementary file' column, and another red box highlights the 'Download' column for the file 'GSE263713_raw_counts.tsv.gz'. A red arrow points from the 'Download' link on the GEO page to the 'Download' link on the right-hand search results page.

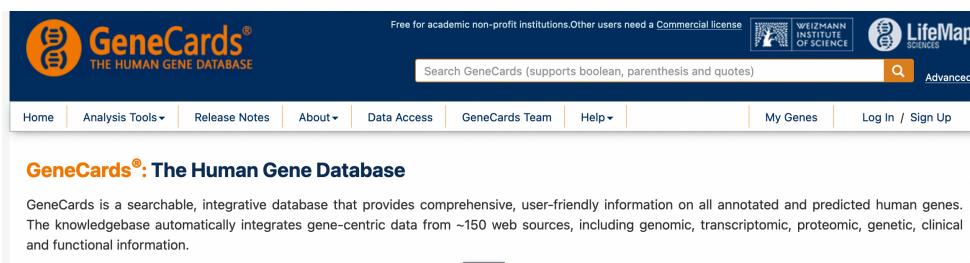
The screenshot shows the GEO DataSets search results page for 'skin RNA-seq'. The search bar at the top contains 'skin RNA-seq'. The results list includes three entries, each with a checkbox, a title, a detailed description, and metadata like 'Organism', 'Type', 'Platform', and 'Download data'. The third entry is highlighted with a red box around its title and description. A red arrow points from the 'Download' link in the GEO Overview table on the left to the 'Download data: TSV' link for this entry on the right.

Supplementary file	Size	Download	File type/resource
GSE263713_raw_counts.tsv.gz	4.7 Mb	(ftp)(http)	TSV

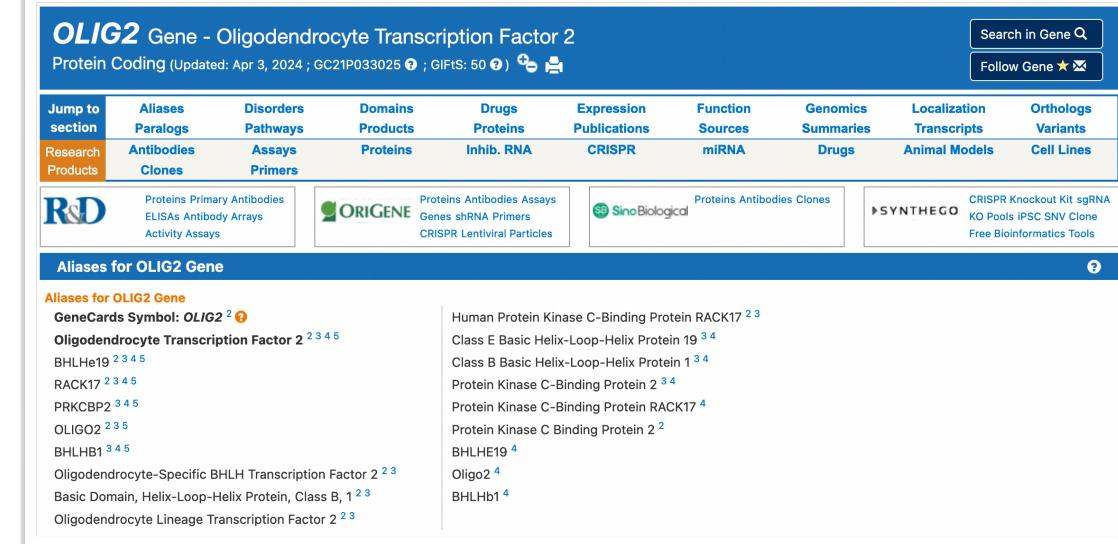
Functional database

GeneCards

Gene centric database summarizing all information about human genes (diseases, functions, proteins, tissues, ...)



The screenshot shows the GeneCards homepage. At the top, there's a banner for the Weizmann Institute of Science and LifeMap Sciences. Below the banner, there's a search bar with placeholder text "Search GeneCards (supports boolean, parenthesis and quotes)" and an "Advanced" link. The main menu includes "Home", "Analysis Tools", "Release Notes", "About", "Data Access", "GeneCards Team", "Help", "My Genes", and "Log In / Sign Up". A large orange arrow points downwards from this section towards the bottom of the page.



The screenshot shows the GeneCards entry for the **OLIG2** gene. The header includes the gene name, its function as "Oligodendrocyte Transcription Factor 2", and protein coding information. It also features a "Search in Gene" and "Follow Gene" button. Below the header is a navigation menu with tabs like "Aliases", "Disorders", "Domains", etc., and a "Jump to section" dropdown. The main content area displays various resources for the gene, such as R&D services, ORIGENE, SinoBiological, and SYNTHEGO. A section titled "Aliases for OLIG2 Gene" lists numerous aliases with their counts: GeneCards Symbol: **OLIG2** (2), Oligodendrocyte Transcription Factor 2 (2, 3, 4, 5), BHLHe19 (2, 3, 4, 5), RACK17 (2, 3, 4, 5), PRKCBP2 (2, 3, 5), OLIGO2 (2, 3, 5), BHLHB1 (2, 3, 4, 5), Oligodendrocyte-Specific BHLH Transcription Factor 2 (2, 3), Basic Domain, Helix-Loop-Helix Protein, Class B, 1 (2, 3), and Oligodendrocyte Lineage Transcription Factor 2 (2, 3).

Version 5.20: 10 April 2024

- › **466,349 Entries**, 43,839 HGNC approved, 21,601 Protein coding, 291,492 RNA genes including 130,365 lncRNAs, 111,811 piRNAs, and 49,316 other ncRNAs.
- › **Malacards**: Revamped user interface, with AI-generated summaries in a new Overview section, a new navigation pane, searchable and downloadable tables, and more. Visit www.malacards.org - we welcome your feedback.

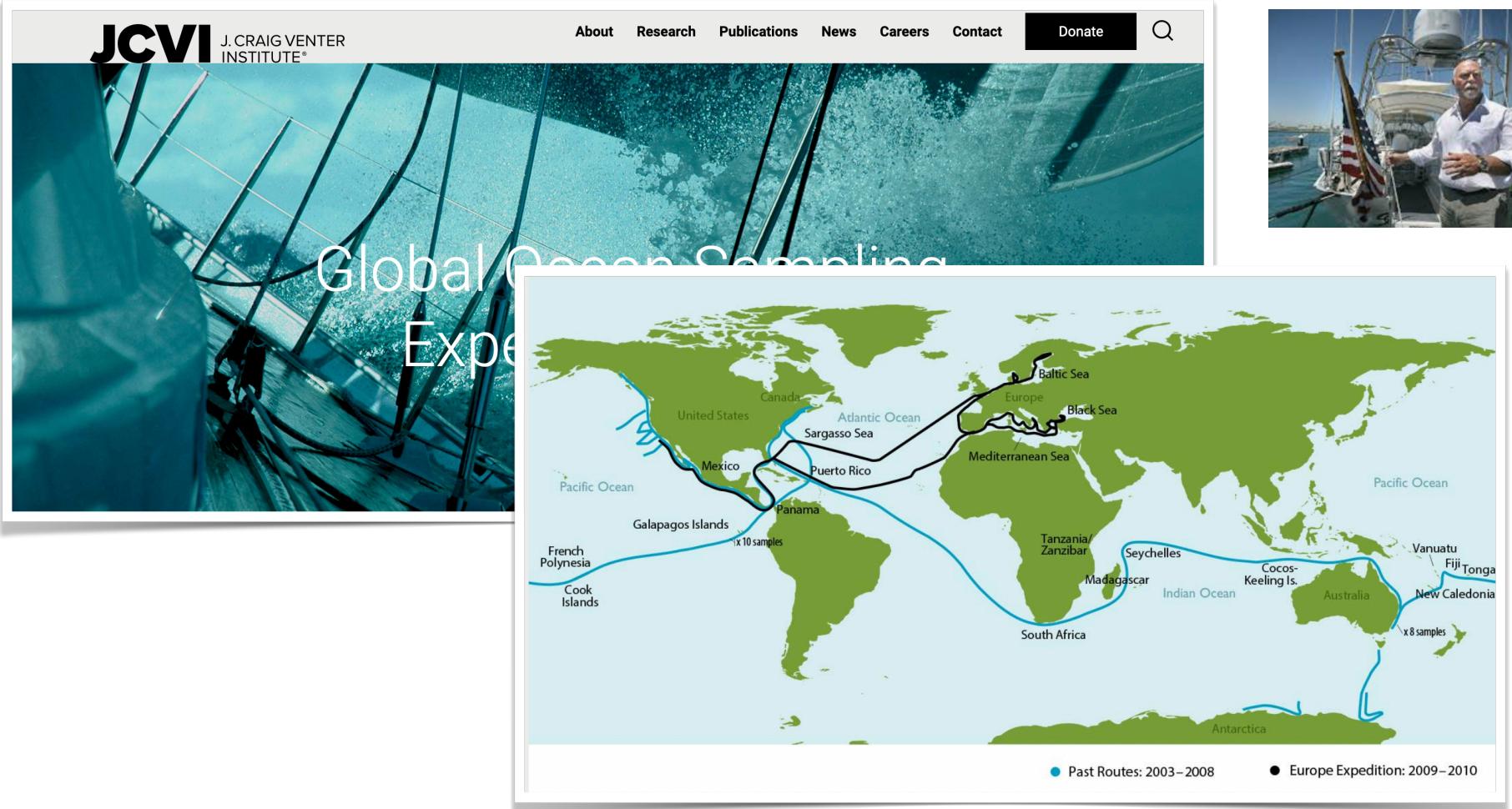
[link to entry]



Analysis of metagenomic sequences

Ocean metagenomics





The screenshot shows the JCVI website's homepage. At the top left is the JCVI logo: "JCVI J. CRAIG VENTER INSTITUTE®". At the top right are navigation links: About, Research, Publications, News, Careers, Contact, Donate, and a search icon. Below the header is a large image of a sailboat's hull and rigging in the water, with the text "Global Ocean Sampling" overlaid. To the right of the main content area is a smaller photo of a man standing on a boat deck, wearing a light blue shirt and tan pants, with an American flag visible in the background.

Global Ocean Sampling

Expe


A world map illustrating the Global Ocean Sampling Expeditions. The map shows the major oceans and seas with a network of blue lines representing sampling routes. Labels indicate specific locations and routes, such as the Pacific Ocean, French Polynesia, Cook Islands, Galapagos Islands, Mexico, United States, Canada, Sargasso Sea, Puerto Rico, Panama, South Africa, Tanzania/Zanzibar, Madagascar, Seychelles, Indian Ocean, Australia, New Caledonia, Vanuatu, Fiji, Tonga, and Antarctica. A legend at the bottom indicates that blue dots represent "Past Routes: 2003–2008" and black dots represent the "Europe Expedition: 2009–2010".

- Past Routes: 2003–2008
- Europe Expedition: 2009–2010



ID	Sample Location	Country	Date, mm/dd/yy	Time	Location	Sample Depth, m	Water Depth, m	T (°C) ^a	S ^b (ppt)	Size Fraction (μm)	Habitat Type	Chl a Sample Month (Annual ± SE) mg/m ⁻³	Good Sequences
GS00a	Sargasso Stations 13 and 11	Bermuda (UK)	02/26/03	3:00	31°32'6" n; 63°35'42" w	5.0	>4,200	20.0 20.5	36.6	0.1-0.8	Open ocean	0.17 (0.09 ± 0.02)	644,551
GS00b	Sargasso Stations 13 and 11	Bermuda (UK)	02/26/03	10:10	31°10'50" n; 64°19'27" w	5.0	>4,200	20.0 20.5	36.6	0.22-0.8	Open ocean	0.17 (0.09 ± 0.02)	317,180
GS00c	Sargasso Stations 3	Bermuda (UK)	02/25/03	13:00	32°09'30" n; 64°00'36" w	5.0	>4,200	19.8	36.7	0.22-0.8	Open ocean	0.17 (0.09 ± 0.02)	368,835
GS00d	Sargasso Stations 13	Bermuda (UK)	02/25/03	17:00	31°32'6" n; 63°35'42" w	5.0	>4,200	20.0	36.6	0.22-0.8	Open ocean	0.17 (0.09 ± 0.02)	332,240
GS01a	Hydrostation S	Bermuda (UK)	05/15/03	11:40	32°10'00" n; 64°30'00" w	5.0	>4,200	22.9	36.7	3.0-20.0	Open ocean	0.10 (0.10 ± 0.01)	142,352
GS01b	Hydrostation S	Bermuda (UK)	05/15/03	11:40	32°10'00" n; 64°30'00" w	5.0	>4,201	22.9	36.7	0.8-3.0	Open ocean	0.10 (0.10 ± 0.01)	90,905
GS01c	Hydrostation S	Bermuda (UK)	05/15/03	11:40	32°10'00" n; 64°30'00" w	5.0	>4,202	22.9	36.7	0.1-0.8	Open ocean	0.1 (0.1 ± 0.01)	92,351
GS02	Gulf of Maine	USA	08/21/03	6:32	42°30'11" n; 67°14'24" w	1.0	106	18.2	29.2	0.1-0.8	Coastal	1.4 (1.12 ± 0.19)	121,590
GS03	Brown's Bank, Gulf of Maine	Canada	08/21/03	11:50	42°51'10" n; 66°13'2" w	1.0	119	11.7	29.9	0.1-0.8	Coastal	1.4 (1.12 ± 0.19)	61,605
GS04	Outside Halifax, Nova Scotia	Canada	08/22/03	5:25	44°8'14" n; 63°38'40" w	2.0	142	17.3	28.3	0.1-0.8	Coastal	0.4 (0.78 ± 0.17)	52,959
GS05	Bedford Basin, Nova Scotia	Canada	08/22/03	16:21	44°41'25" n; 63°38'14" w	1.0	64	15.0	30.2	0.1-0.8	Embayment	6 (6.76 ± 0.98)	61,131
GS06	Bay of Fundy, Nova Scotia	Canada	08/23/03	10:47	45°6'42" n; 64°56'48" w	1.0	11	11.2	0.1-0.8	Estuary	2.8 (1.87 ± 0.18)	59,679	
GS07	Northern Gulf of Maine	Canada	08/25/03	8:25	43°37'56" n; 66°50'50" w	1.0	139	17.9	31.7 ^c	0.1-0.8	Coastal	1.4 (1.12 ± 0.19)	50,980
GS08	Newport Harbor, RI	USA	11/16/03	16:45	41°29'9" n; 71°21'4" w	1.0	12	9.4	26.5 ^c	0.1-0.8	Coastal	2.2 (1.59 ± 0.17)	129,655
GS09	Block Island, NY	USA	11/17/03	10:30	41°5'28" n; 71°36'8" w	1.0	32	11.0	31.0 ^c	0.1-0.8	Coastal	4.0 (2.72 ± 0.24)	79,303
GS10	Cape May, NJ	USA	11/18/03	4:30	38°56'24" n; 74°41'6" w	1.0	10	12.0	31.0 ^c	0.1-0.8	Coastal	2.0 (2.75 ± 0.33)	78,304
GS11	Delaware Bay, NJ	USA	11/18/03	11:30	39°25'4" n; 75°30'15" w	1.0	8	11.0	0.1-0.8	Estuary	4.8 (9.23 ± 1.02)	124,435	
GS12	Chesapeake Bay, MD	USA	12/18/03	11:32	38°56'49" n; 76°25'2" w	1.0	25	3.2	3.47 ^c	0.1-0.8	Estuary	21.0 (15.0 ± 1.01)	126,162
GS13	Off Nags Head, NC	USA	12/19/03	6:28	36°0'14" n; 75°23'41" w	1.0	20	9.3	0.1-0.8	Coastal	3.0 (2.24 ± 0.25)	138,033	
GS14	South of Charleston, SC	USA	12/20/03	17:12	32°30'25" n; 79°15'50" w	1.0	31	18.6	0.1-0.8	Coastal	1.70 (1.92 ± 0.25)	128,885	
GS15	Off Key West, FL	USA	01/08/04	6:25	24°29'19" n; 83°4'12" w	2.0	47	25.3	36.0	0.1-0.8	Coastal	0.2 (0.27 ± 0.09)	127,362
GS16	Gulf of Mexico	USA	01/08/04	14:15	24°10'29" n; 84°20'40" w	2.0	3,333	26.4	35.8	0.1-0.8	Coastal sea	0.16 (0.11 ± 0.01)	127,122
GS17	Yucatan Channel	Mexico	01/09/04	13:47	20°31'21" n; 85°24'49" w	2.0	4,513	27.0	35.8	0.1-0.8	Open ocean	0.13 (0.09 ± 0.01)	257,581
GS18	Rosario Bank	Honduras	01/10/04	8:12	18°2'12" n; 83°47'5" w	2.0	4,470	27.4	35.4	0.1-0.8	Open ocean	0.14 (0.09 ± 0.01)	142,743
GS19	Northeast of Colón	Panama	01/12/04	9:03	10°42'59" n; 80°15'16" w	2.0	3,336	27.7	35.4	0.1-0.8	Coastal	0.23 (0.15 ± 0.02)	135,325
GS20	Lake Gatún	Panama	01/15/04	10:24	9°5'52" n; 79°50'10" w	2.0	4	28.5	0.06	0.1-0.8	Fresh water	296,355	
GS21	Gulf of Panama	Panama	01/19/04	16:48	8°7'45" n; 79°41'28" w	2.0	76	27.6	30.7	0.1-0.8	Coastal	0.50 (0.73 ± 0.22)	131,798
GS22	250 miles from Panama City	Panama	01/20/04	16:39	6°29'34" n; 82°54'14" w	2.0	2,431	29.3	32.3	0.1-0.8	Open ocean	0.33 (0.28 ± 0.02)	121,662
GS23	30 miles from Cocos Island	Costa Rica	01/21/04	15:00	5°38'24" n; 86°33'55" w	2.0	1,139	28.7	32.6	0.1-0.8	Open ocean	0.07 (0.19 ± 0.02)	133,051
GS25	Dirty Rock, Cocos Island	Costa Rica	01/28/04	10:51	5°33'10" n; 87°5'16" w	1.1	30	28.3	31.4	0.8-3.0	Fringing reef	0.11 (0.19 ± 0.01)	120,671
GS26	134 miles NE of Galapagos	Ecuador	02/01/04	16:16	1°15'51" s; 90°17'42" w	2.0	2,376	27.8	32.6	0.1-0.8	Open ocean	0.22 (0.28 ± 0.02)	102,708
GS27	Devil's Crown, Floreana	Ecuador	02/04/04	11:41	1°12'58" s; 90°25'22" w	2.0	2.3	25.5	34.9	0.1-0.8	Coastal	0.40 (0.38 ± 0.03)	222,080
GS28	Coastal Floreana	Ecuador	02/04/04	15:47	1°13'1" s; 90°19'11" w	2.0	156	25.0 ^c	0.1-0.8	Coastal	0.35 (0.35 ± 0.02)	189,052	
GS29	North James Bay, Santigo	Ecuador	02/08/04	18:03	0°1'2" s; 90°50'7" w	2.0	12	26.2	34.5	0.1-0.8	Coastal	0.40 (0.39 ± 0.03)	131,529
GS30	Warm seep, Roca Redonda	Ecuador	02/09/04	11:42	0°16'20" n; 91°38'0" w	19.0	19	26.9	0.1-0.8	Warm seep		359,152	
GS31	Upwelling, Fernandina	Ecuador	02/10/04	14:43	0°18'4" s; 91°39'6" w	12.0	19	18.6	0.1-0.8	Coastal upwelling	0.35 (0.39 ± 0.03)	436,401	
GS32	Mangrove, Isabella	Ecuador	02/11/04	11:30	0°35'38" s; 91°4'10" w	0.3	0.67	25.4	0.1-0.8	Mangrove		148,018	
GS33	Punta Cormorant Lagoon, Floreana	Ecuador	02/19/04	13:35	1°13'42" s; 90°25'45" w	0.2	0.33	37.6	46 ^c	0.1-0.8	Hypersaline		692,255
GS34	North Seamount	Ecuador	02/19/04	17:06	0°2'59" s; 90°16'47" w	2.0	35	27.5	0.1-0.8	Coastal	0.36 (0.35 ± 0.02)	134,347	
GS35	Wolf Island	Ecuador	03/01/04	16:44	1°23'21" n; 91°49'1" w	2.0	71	21.8	34.5	0.1-0.8	Coastal	0.28 (0.31 ± 0.02)	140,814
GS36	Cabo Marshall, Isabella	Ecuador	03/02/04	12:52	0°1'15" s; 91°11'52" w	2.0	67	25.8	34.6	0.1-0.8	Coastal	0.65 (0.45 ± 0.05)	77,538
GS37	Equatorial Pacific TAO Buoy	International	03/17/04	16:38	1°58'26" s; 95°0'53" w	2.0	3,334	28.8	0.1-0.8	Open ocean	0.21 (0.24 ± 0.02)	65,670	
GS47	201 miles from French Polynesia	International	03/28/04	15:25	10°7'53" s; 135°26'58" w	30.0	2,400	28.6	37.3	0.1-0.8	Open ocean		66,023
GS51	Rangiroa Atoll	French Polynesia	05/22/04	7:04	15°8'37" s; 147°26'6" w	1.0	10	27.3	34.2	0.1-0.8	Coral reef atoll		128,982
Total													7,697,926

^aTemperature.

^bSalinity.

^cMeasurements were acquired from nearby vessels and/or research stations.

doi:10.1371/journal.pbio.0050077.t001

[Rusch DB, Halpern AL, Sutton G,
Heidelberg KB, Williamson S,
Yoosheph S, et al. (2007)]

Goals

Perform a full bioinformatics annotation of an unknown metagenomic sequence

- Pick a random sequence from the GOS dataset Use bioinformatics online tools to ask:
 - does the DNA code for a **protein**?
 - does the DNA have **homologs** in GENBANK?
 - can we predict the protein's **function**?
 - can we predict from what **life form** the DNA came from?

Annotathon

C-Herrmann @ GKBioinfo Disconnect français
d-42

Home Rule Book Cart Forum Explore Messages Evaluations Grades Activity Logs Accounts Team Help

Total annotations received: 5 k
Annotations saved as version: 3

Modify annotations of GOS_11632010.3

Save your annotations regularly!
After 4 hours of inactivity (no clics), your session will automatically be disconnected and all modifications since last save could be lost.
Having entered new annotations, do not forget to clic on [Save your annotations](#)

General comments

Genomic Sequence  **DNA sequence in FASTA format**

```
>GOS_11632010 [ Global Ocean Sampling expedition : Open Ocean : Sargasso Sea : Bermuda (UK) : Sargasso Sea, Station 11 ]  
GGAAACACTCGCGCTGTTGAGGTAGACGGCGTGTGCGAGTCGGCAGACCTGGATCTGGATGTGGCGAGCCACGAACTTCTCCACCAAC  
ATGGCGTGTGCGCAAAGGGCGCGGCCCTCTGGATGGAGAGGTGGTAGGCCCTCGCGGCCCTCGTGTGTTGAGCCACGGCATGCCCTGGCG  
CGCCCGCCGGCGAGGCCCTGATCATGAGCGGTAGCCGATCTCGCGCGATCTCGAGCATGTGCGCTCTGCGCCACCTGCCACAAAGCGGCAC  
CACGTGATGCCGGCTGGCTGGCGATGCCGCTTGCGGATGCCGATCTGGCGCCCATACCTCGAGCGGTGCTGCTGGGCCACGAAGGTGACGGCGTGC  
TTCTCGAGCAGGTGCAAGAACCTGGTCTCGAGAGAACCGTAGCCGGGACCGCTGCGCTGACTCGCCGATCTGAACTGCTTGAGACGAGCGCGAGATCGGT  
CCATGGCCAGGTAGGACTGGAGGACGCCGGCGCCACGCTGCGCTGACTCGCCGATCTGAACTGCTTGAGACGAGCGCGAGATCAC  
GCCACCGTCGGGATCCCCTGCGCGCTGCAAACAGCGGACCGGAGCTGGCCACCGGAGCTGGGACCCAGGAGCTTGGTCACTCGGTGGAGGCTG  
CTGCTGGCTGGCGAGACGGCTGGCGAGACGGCTCTCACCACGGCACGGTTGGACCGTGTGCGGGATCA
```

OPEN  ACCESS Freely available online

PLOS BIOLOGY

Community Page

Metagenome Annotation Using a Distributed Grid of Undergraduate Students

Pascal Hingamp*, Céline Brochier, Emmanuel Talla, Daniel Gautheret, Denis Thieffry, Carl Herrmann

[Hingamp, ... Herrmann, PLOS Biology (2008)]



Workflow

Sample random metagenomic sequence

Step 1
identify open reading frames

Step 2
identify protein domains / protein families

Step 3
identify homologous sequences using BLAST

Step 4
Perform multiple sequence alignment

Step 5
Reconstruction phylogenetic tree

Step 6
Summarize your findings

- Is the DNA fragment **coding**?
- What are the protein **domains**?
- What are **homologous** sequences?
- What could be the biological **function** of the gene?
- What is the **evolutionary origin** of this sequence?

Let's dive...