

# Gene Identification and Genome Annotation

Jutta Marzillier, PhD  
Lehigh University  
Department of Biological Sciences  
March 15<sup>th</sup>, 2011

# Outline

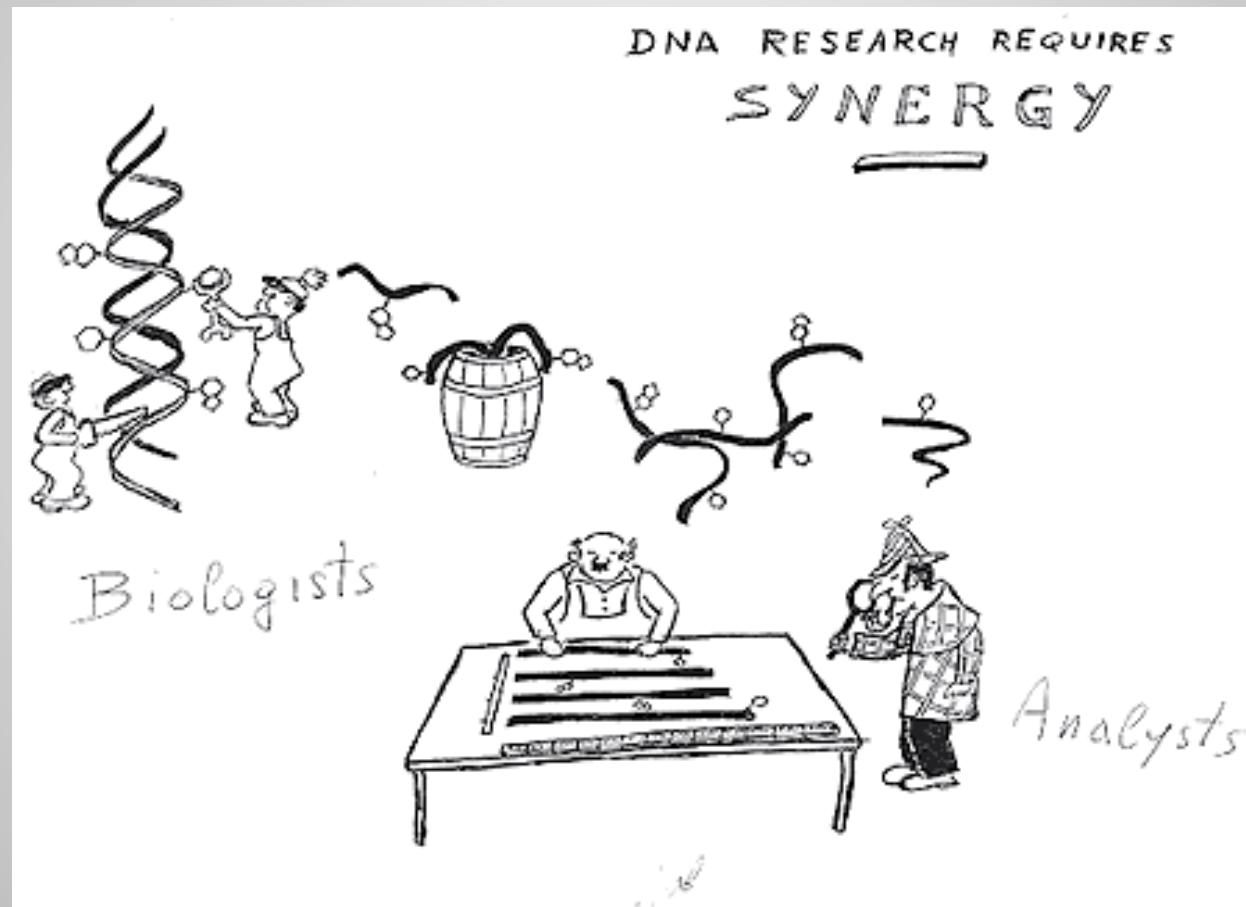
High Throughput Next Generation Sequencing

Data Collection, Organization and Interpretation

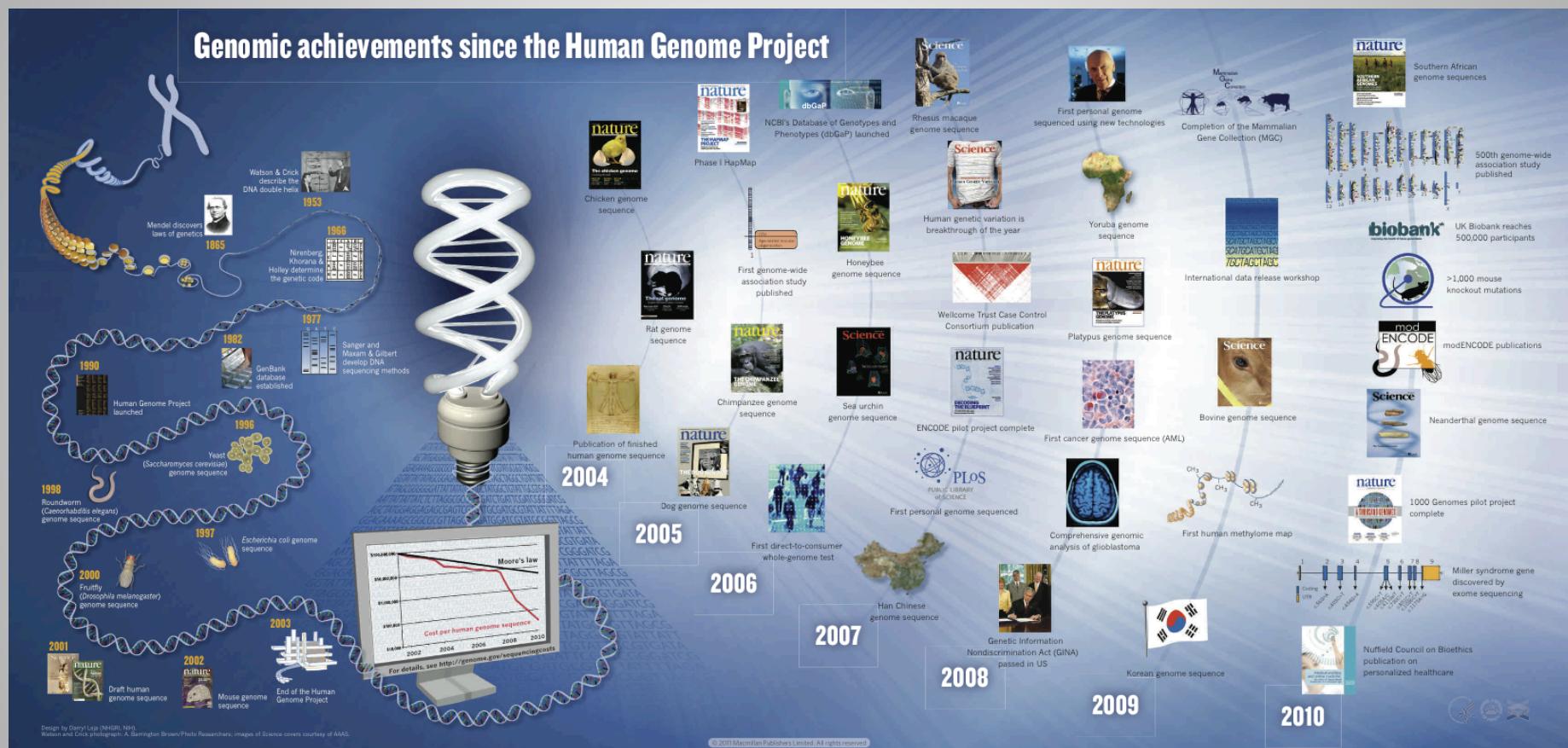
Gene Calling – Definition of a Gene

Gene Calling Software: Glimmer and Genemark

# Life Sciences Thrive Through their Co-laboration with Engineering and BioInformatics



# Genomic Achievements since the Human Genome Project

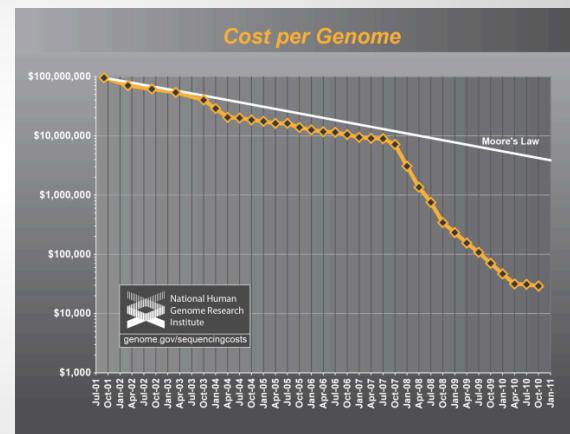
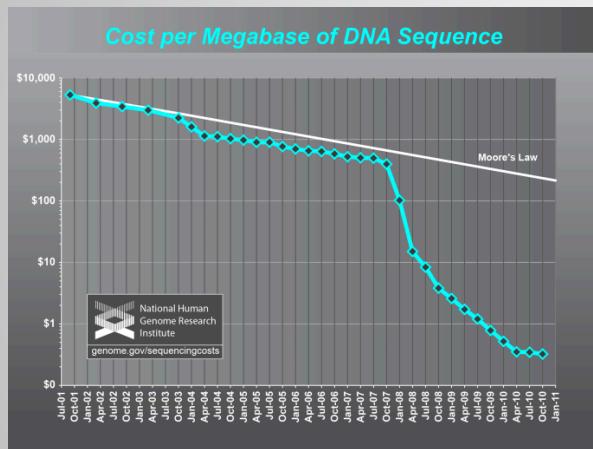


ED Green et al. *Nature* 470, 204-213 (2011) doi:10.1038/nature09764

**nature**

# Sudden and profound out-pacing of Moore's Law

Transition from Sanger Sequencing to Next-Generation Sequencing technologies



Moore's Law, which describes a long-term trend in the computer hardware industry that involves the doubling of 'compute power' every two years

[www.genome.gov/sequencingcosts](http://www.genome.gov/sequencingcosts).

## Dealing with Data



Science, Vol 331, 11 Feb 2011

Science and technology are essential to improving public health and welfare

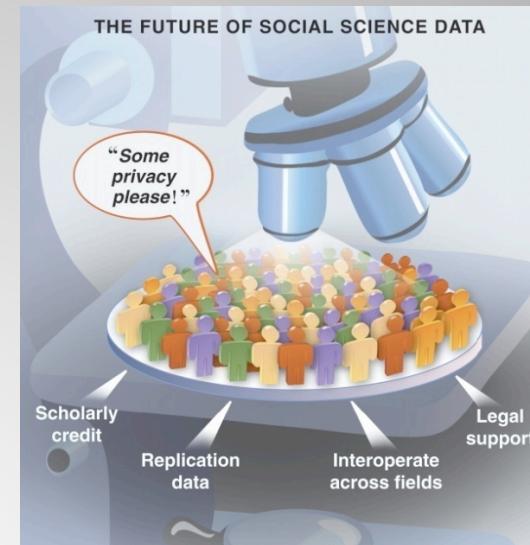
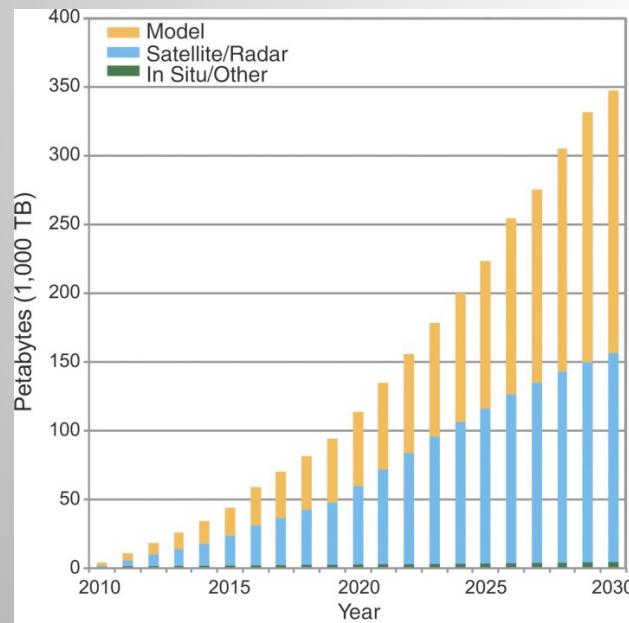
There is a huge influx of research data

Data collection, curation and access are central to look at data in a meaningful way

# Data Collection Examples

Social Sciences

Climate change



Global Health

Science, Vol 331, 11 Feb 2011

# Finishing Draft Sequences using Assembly Software

Primary sequencing reads:

```
cagacgtgtcagtgcactcgatatactgagctagtcact  
tagctagccggatagtattaccagacgtgtcagtgcactcgata  
ccagatcgatcgattcgcgatagctagccggatagtattaccagacgt
```

Align sequences for homologies

The diagram illustrates the alignment of three sequencing reads for homology. The top sequence is cagacgtgtcagtgcactcgatatactgagctagtcact. The middle sequence is tagctagccggatagtattaccagacgtgtcagtgcactcgata. The bottom sequence is ccagatcgatcgattcgcgatagctagccggatagtattaccagacgt. The alignment shows overlapping regions highlighted in yellow and green, indicating coverage levels.

Green: 3-fold coverage

Yellow: 2-fold coverage

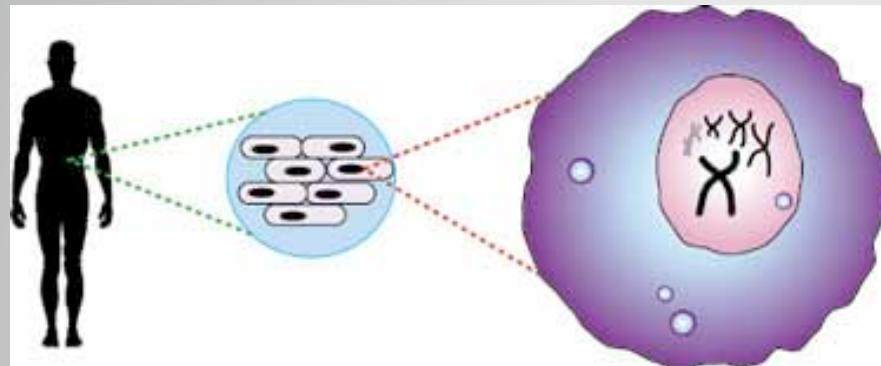
Fragment assembly to contig

The diagram illustrates the assembly of fragments into a contig. The sequence is ccagatcgatcgattcgcgatagctagccggatagtattaccagacgtgtcagtgcactcgatatactgagctagtcact. The assembly shows overlapping regions highlighted in yellow and green.

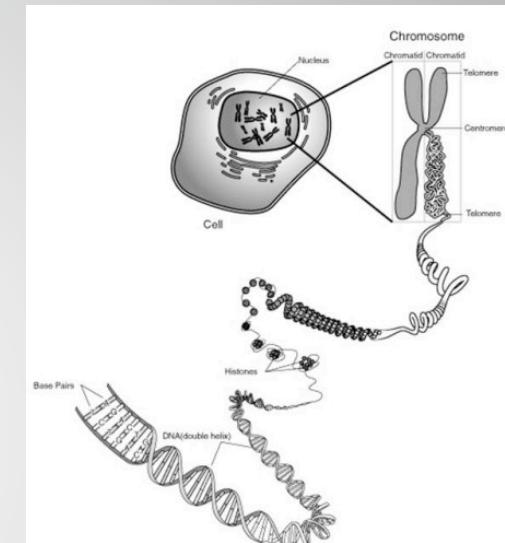
Goal for Genome Sequencing:  
Sanger: 8-fold coverage  
454: 30 fold coverage

TGCGGCTGCCAGATTTGTACGGGTTGGAAGTCGACGGAGAGAACAGCG  
CGGGCCTAGAAGGCCCGTAATGCCCCCTGAGAGCCCCGTAGACGGACG  
AACGGTGC GGATCGATAGATGGCACCGGAGACAAGCGAAGACGGCCGCA  
GAGCCGTCGCCGGCTGACGCCCGCGTAGGAAGATATTGTGAAGTGC  
GTCACATTCTACGGGTGAAACCGCAAAGTGGAAAGGTTCTTACCTATGGAG  
GGGTAAAGGGAGCGAGCTCCAGCGAGCGACCGCACCCGACATAGGTTCT  
TGTCGGGGTAGTCGAACGGAGAGAGACTACCCCTTTAGCGACCTCCGGT  
CGCCCAGGTAGGTACCGAACGATGAGAGAGAGTACCTAGACCGTACAGGCC  
GGGGTTATCCCCCGGCCGATAACAGCATGGTCATTTGGTAGGTACGTTA  
CGTAAGCATCACTCACCAACAGAACCAACAGTGGTTACGTAACCAGGTACGTTA  
CGTACGACTAGATACTACGTAACAGAACCAACTAACCGTGGCCGCCGAAGGC  
GGCCCGCAGCGGGTACGTTGTCAAGGTATGTCACTAGGGAGGGTGAAC  
ATGAATCTCAAGGAGCATCAGAACGGAGCTCTCGTCTACTGCTCGTGTCT  
GAGTCGATGACGTATGAGCGTGCCTACTCGAACGGTGGAGACCATCCTGCA  
CCTCATCCAGTCCTCGACGCCGGAGAACTCGATAAGTGAGCTGGAGTCA  
TCTGACCGGCGCGATCGTCTGCCGACCGACTGGCCTCGCATCCGCCGCG  
AGGTTCTGCGGGCGGCTGGCCACCGCTGCCAGATCCGCTACGCCGGACAT  
CTGCACAGGGATGGCTACCGAGGTTGATCACGTCCGCTACCGCGACGAG  
GCGTCACCTCTGCAGGTGTCGTGCAGACCGTGCCTACGCCGGAGTCCG  
CGATGGAAGGC GTTGCTCAGCGTGCAGACTGCGCGATGAAGAAGCG  
GCCGCCGCCGCCACCCGGCGTAGAACGAACTAGGAGGGACCAGG  
CGTCCCCGAGCCCAGGAGGCGTCATGCCGGTCCAGTGCCCAAGCGATC  
GGACGAACCGCGTCCGGCGCAAATCGCATACGAGTGAGCGCGAGGCTAGC

# Why is the Knowledge about the Human Genome interesting?



[http://www.genomenewsnetwork.org/articles/06\\_00/sequence\\_primer.shtml](http://www.genomenewsnetwork.org/articles/06_00/sequence_primer.shtml)



[http://www.pharmainfo.net/files/images/stories/article\\_images/](http://www.pharmainfo.net/files/images/stories/article_images/)

The human body has about 100 trillion cells.  
Each cell harbors the same genetic information in its nucleus in form of DNA containing chromosomes.  
Depending on cellular, developmental, and functional stage of a cell only a subset of genes is expressed.

## How do we Identify a Gene?

DNA sequence of the insulin gene:

```
AGCCCTCCAGGACAGGCTGCATCAGAAGAGGCCATCAAGCAGATCACTGTCCCTGCCATGGCCCTG  
TGGATGCGCCTCCTGCCCTGCTGGCGCTGCTGGCCCTCTGGGGACCTGACCCAGCCGCAGCCTTG  
TGAACCAACACCTGTGCGGCTCACACCTGGTGGAAAGCTCTACCTAGTGTGCGGGGAACGAGGCTTC  
TTCTACACACCCAAGACCCGCCGGGAGGCAGAGGACCTGCAGGTGGGGCAGGTGGAGCTGGCGGG  
GGCCCTGGTGCAGGCAGCCTGCAGCCCTGGCCCTGGAGGGTCCCTGCAGAACGCTGGCATTGTG  
GAACAATGCTGTACCAGCATCTGCTCCCTCTACCAGCTGGAGAACTACTGCAACTAGACGCAGCCCGCA  
GGCAGCCCCACACCCGCCCTCCTGCACCGAGAGAGATGGAATAAGCCCTGAACCAGCAAAA
```

Unraveling the sequence of the bases was the ‘easy’ part with a defined endpoint

Next comes to figure out the meaning of these sequences of A’s, G’s, T’s, and C’s

Examples:

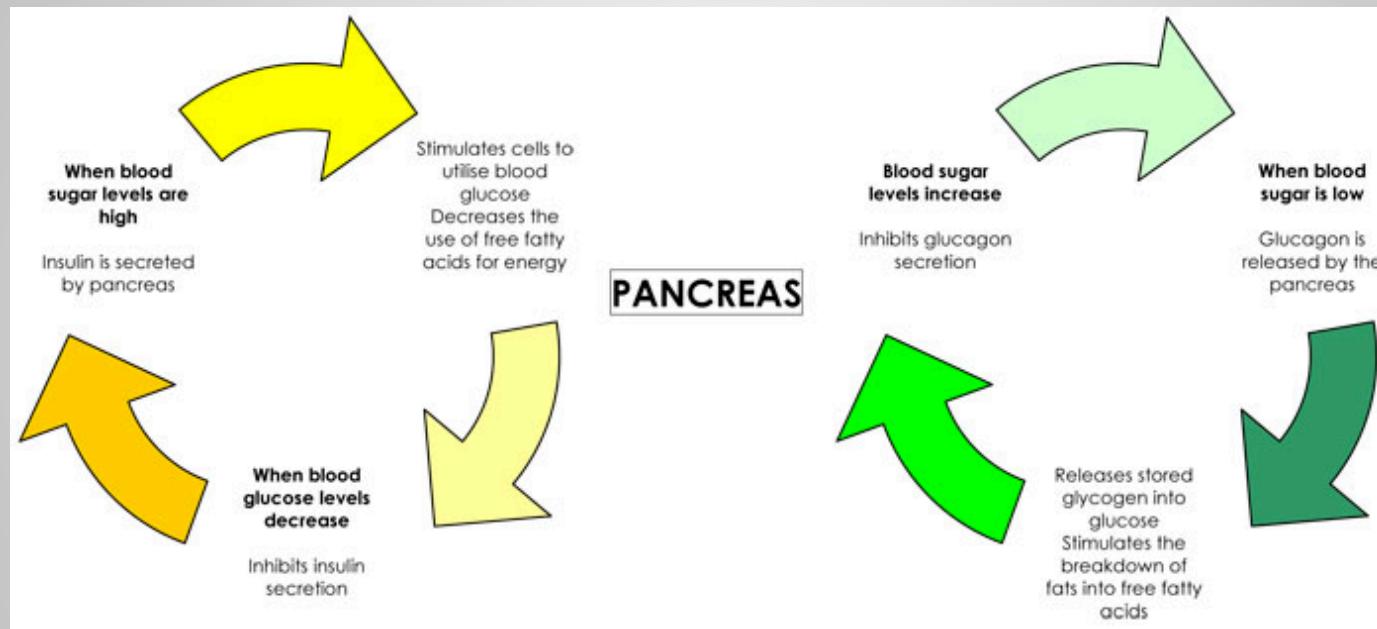
Which stretches of DNA correspond to genes?

When and in what tissues are these genes expressed?

What kind of proteins do they code for?

How do all these proteins interact with each other and function?

## Example: Regulation of Insulin Secretion



[http://www.google.com/imgres?imgurl=http://pilatespower.tv/Portals/0/Articles/LooseWeight-001-](http://www.google.com/imgres?imgurl=http://pilatespower.tv/Portals/0/Articles/LooseWeight-001-.)

# The Biological 'Dogma'

DNA: storage of genetic information

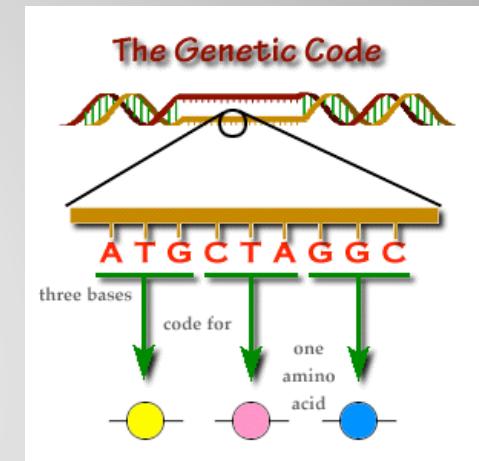
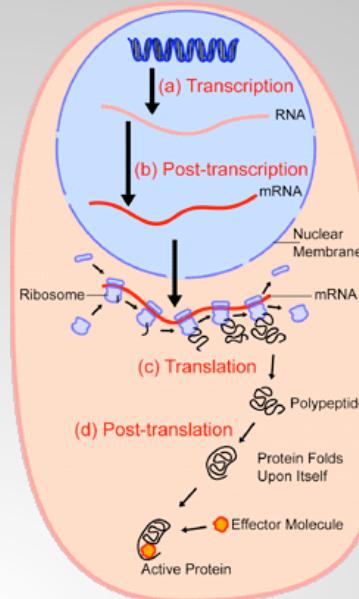


RNA: vehicle to transport genetic information into the cytoplasm where it is translated into proteins



Protein: gene products made of amino acids with very specific functions, such as enzymes, hormones, motility proteins, transport, structural, etc.

Genetic code:  $4^3$  combinations



[www.brooklyn.cuny.edu/.../GP.GeneticCode.GIF](http://www.brooklyn.cuny.edu/.../GP.GeneticCode.GIF)

		Second Letter					
		T	C	A	G		
T		TTT } Phe TTC TTA } Leu TTG	TCT } Ser TCC TCA TCG	TAT } Tyr TAC TAA } Stop TAG	TGT } Cys TGC TGA } Stop TGG Trp	T C A G	
C		CTT } Leu CTC CTA CTG	CCT } Pro CCC CCA CCG	CAT } His CAC CAA } Gln CAG	CGT } Arg CGC CGA CGG	T C A G	
A		ATT } Ile ATC ATA ATG Met	ACT } Thr ACC ACA ACG	AAT } Asn AAC AAA } Lys AAG	AGT } Ser AGC AGA } Arg AGG	T C A G	
G		GTT } Val GTC GTA GTG	GCT } Ala GCC GCA GCG	GAT } Asp GAC GAA } Glu GAG	GGT } Gly GGC GGA GGG	T C A G	

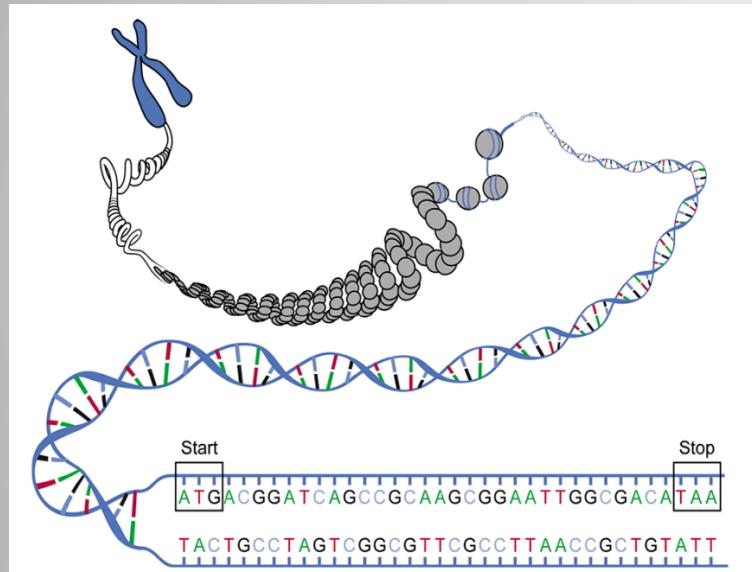
# Start and Stop Codons in the Insulin Gene

Start – ATG

Stop – TAA, TAG, TGA

AGCCCTCCAGGACAGGCTGCATCAGAAGAGGCCATCAAGCAGA  
TCACTGTCTTCTGCCATGGCCCTGTGGATGCGCCTCCTGCC  
CTGCTGGCGCTGCTGCCCTCTGGGGACCTGAGCCAGCGCA  
GCCTTGATGACCAACACCTGTGCGGCTCACACCTGGTGGAAAG  
CTCTCTACCTAGTGTGCAGGGAACGAGGCTTCTTACACACCC  
AAGACCCGCCGGGAGGCAGAGGACCTGCAGGTGGGCAGGT  
GGAGCTGGCGGGGGCCCTGGTGCAGGCAGCCTGCAGCCCT  
GGCCCTGGAGGGTCCCTGCAGAACGCTGGCATTGTGGAACA  
ATGCTGTACCAGCATCTGCTCCCTCTACCAAGCTGGAGAACTACT  
GCAACTAGACGCAGCCGCAGGCAGCCCCACACCCGCC  
CCTGCACCGAGAGAGATGGAATAAGCCCTTGAACCAGCAAAA

# The Genetic Code translates Nucleotide triplets into Amino Acids



Second Letter				
T	C	A	G	
T	TTT TTC TTA TTG } Phe	TCT TCC TCA TCG } Ser	TAT TAC } Tyr	TGT TGC } Cys
C	CTT CTC CTA CTG } Leu	CCT CCC CCA CCG } Pro	CAT CAC } His	CGT CGC CGA CGG } Arg
A	ATT ATC ATA ATG } Ile Met	ACT ACC ACA ACG } Thr	AAT AAC } Asn	AGT AGC } Ser
G	GTT GTC GTA GTG } Val	GCT GCC GCA GCG } Ala	GAT GAC } Asp	GGT GGC GGA GGG } Gly
			GAA GAG } Glu	TCA TCA TCA TCA } Stop

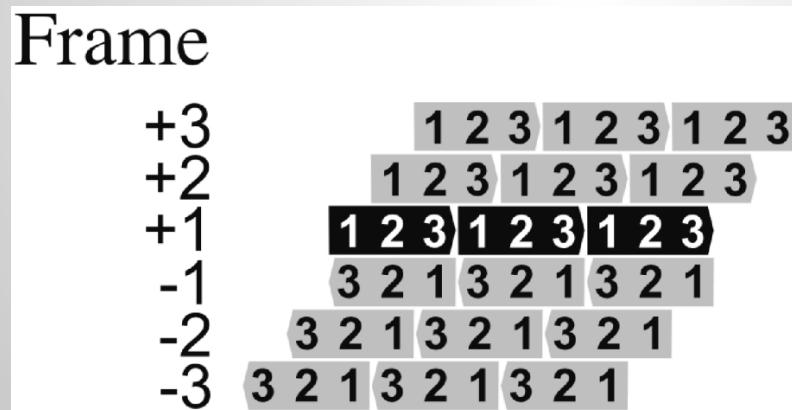
## Frame

+3	1	2	3	1	2	3	1	2	3
+2	1	2	3	1	2	3	1	2	3
+1	1	2	3	1	2	3	1	2	3
-1	3	2	1	3	2	1	3	2	1
-2	3	2	1	3	2	1	3	2	1
-3	3	2	1	3	2	1	3	2	1

# To Identify Genes in a Genome One Needs to Look for Open Reading Frames

A random sequence of nucleotides will by chance encode for a stop signal once every 20 codons (3 Stop codons out of 64).

Open Reading Frames are continuous nucleotide stretches (>100 codons) that lack stop codons.



## Criteria of a Gene

Open Reading Frame

Promoter Consensus Sequence

Ribosomal Binding Site

> 35 amino acids

# The ExPasy Proteomics Server translates proteins in all 6 Reading Frames

Translate Tool - Results of translation - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://expasy.org/cgi-bin/dna\_aa

Most Visited Getting Started Latest Headlines

Translate Tool - Results of translation

SIB Swiss Institute of Bioinformatics ExPASy

ExPASy Proteomics Server

Search ExPASy web site for Go Clear

Databases Tools Services Mirrors About Contact

You are here: ExPASy CH > Tools > DNA -> Protein > Translate

**Translate Tool - Results of translation**

Please select one of the following frames:

**5'3' Frame 1**  
SPPGQAASEEAIKQITVLLPWPCGCASCPCWRWCWPSGDLTQPQPL Stop TNTCAAHTWWKLST Stop CAGNEASSTHPRPAGRQRTCRWGRWSWAGALVQAACSPWPWRG  
PCRSVALWNNAVPASAPSTS WRTTATRRSPQAAPHPPPAPRE Met E Stop SP Stop TSK

**5'3' Frame 2**  
ALQDRLHQKRPSRSLSFCHGPVDAPPAPAGAAGPLGT Stop PSRSLCEPTPVRLTPGGSSLPSVRGTRLLLHTQDPGGRGPAAGGAGRGPCRQPAALGPGVPAE  
AWHCGT Met LYQHLLPLPAGELLQLDAARRQPHTRLLHRERWNKALEPAK

**5'3' Frame 3**  
PSRTGCIRRHQADHCPSA Met ALW Met RLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFYTPKTRREAEDLQVGQVELGGPGAGSLQPLALEGSLQK  
RGIVEQCCTSICSLYQLENYCN Stop TQPAGSPTPAASCTERDGIKPLNQQ

**3'5' Frame 1**  
FCWFKGIFPSLSVQEAGVGLPAGCV Stop LQ Stop FSSW Stop REQ Met LVQH C ST Met PRFCRDPSRAKG C RLPAPGPPPSSTCPTCRSSASRRVLGV Stop KKPRSPHTR Stop  
RASTRCEPHRCWFTKAAAGSGPQRASSASRGRRRIHRA Met AEGQ Stop SA Stop WPLL Met QPVLEG

**3'5' Frame 2**  
FAGSRALFHLSCRRLRVWGCLRAASSCSSSPAGRGSRCWYSIVPQCHASAGTPPGPRAAGCLHQGPRPAPPAPPAGPLPPGGSWCRRSLVPRTLGRELPPGSRTG  
VGSQRLRLGQVPRGPAAPAGAGGASTGPWKQDKSDLLDGLF Stop CSLSWRA

**3'5' Frame 3**  
LLVQGLYSISLGAGGGCGAACGLRLVAVVLQLVEGADAGTALFHNTL LQGPLQGQGLQAAC TRAPAQLHLPHLQVLCLPAGLGCVEEASFPAH Stop VESFHQV Stop AAQV  
LVHKGCGWVRSPEGQQRQQGQEAHPQGHGRRTVICL Met ASSDAACP G

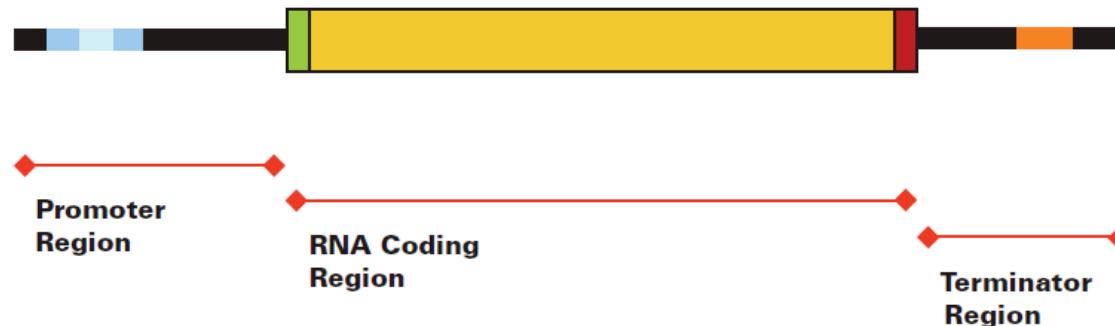
Done

11:00 PM

# Basic Structure of a Gene

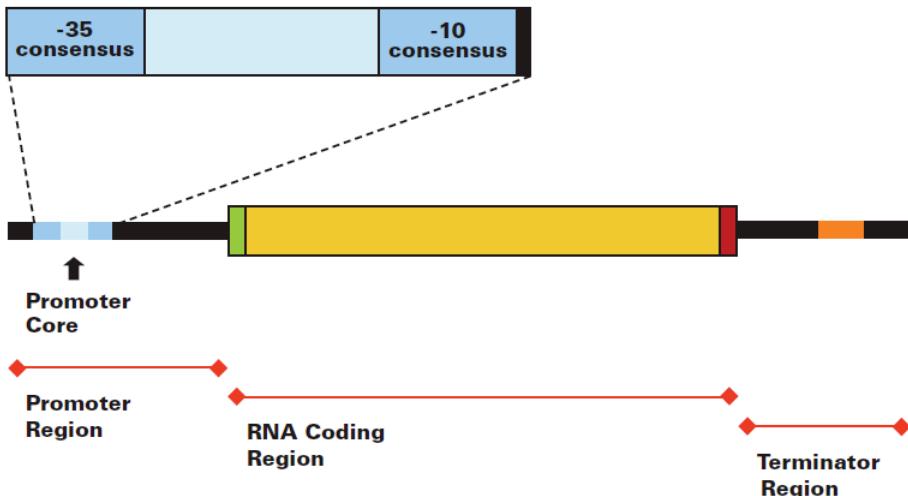
The basic structure of a gene includes the promoter region, the RNA coding region, and the terminator region, indicated in *Analyze fig. 37*.

Analyze fig. 37  
Basic gene structure.

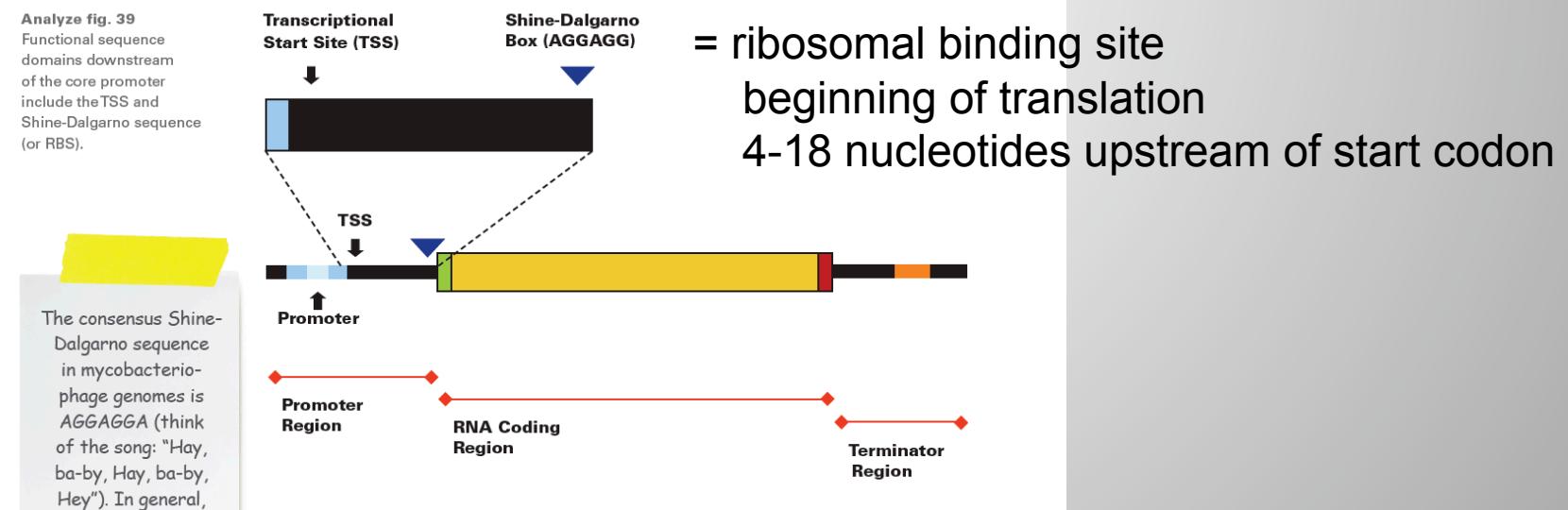


Promoter: RNA polymerase binding site, upstream of the ORF

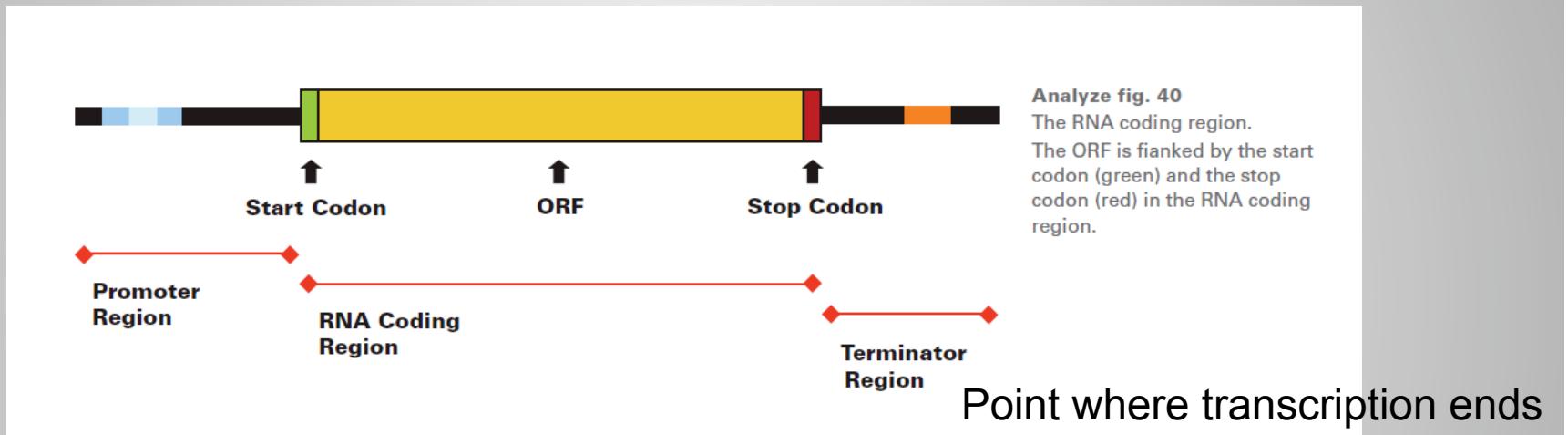
- 10 consensus sequence
- 35 consensus sequence
- the transcriptional start site
- Shine-Dalgarno sequence



Analyze fig. 38  
The core of the promoter (in blue) contains the -10 and -35 consensus sequences.



# The RNA coding region



Mycobacteria have three possible start codons:

ATG – 45%

GTG – 45%

TTG – 10%

# Human Insulin Gene Organization

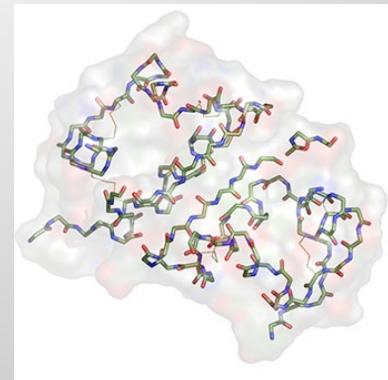
DNA/RNA sequence of insulin gene

AGCCCTCCAGGACAGGCTGCATCAGAAGAGGCCATCAAGCAGATCACTGTCCCTTGCC **ATGGCCCTGTGGATGCGCCCTGC**CCCCTGCTGGCGCTGCTGGCCCTCTGGGACCTGACCCAGCCGCAGCCTTGTGAACCAACACCTGTGCGGCTCACACCTGGTGGAAAGCTCTACCTAGTGTGCGGGGAACGAGGGCTTCTTACACACCCAAGACCCGCCGGAGGCAGAGGACCTGCAGGTGGGGCAGGTGGAGCTGGCGGGGGGCCCTGGTGCAGGCAGCCTGCAGCCCTGGCCCTGGAGGGTCCCTGCAGAACGCTGGCATTGTGGAACAATGCTGTACCATCTGCTCCCTCTACCAGCTGGAGAACTACTGCAACT**TAGACGCAGCCGCAGGCAGCCCCACACCCGCCCTGCACCGAGAGAGATGGAATAAAGCCCTGAACCAGCAAA**

Linear Translated Open Reading Frame

MetAlaLeuTrpMetArgLeuLeuProLeuLeuAlaLeuLeuAlaLeuTrpGlyProAspProAlaAlaAlaPheValAsnGlnHisLeuCysGlySerHisLeuValGluAlaLeuTyrLeuValCysGlyGluArgGlyPhePheTyrThrProLysThrArgArgGluAlaGluAspLeuGlnValGlyGlnValGluLeuGlyGlyProGlyAlaGlySerLeuGlnProLeuAlaLeuGluGlySerLeuGlnLysArgGlyIleValGluGlnCysCysThrSerIleCysSerLeuTyrGlnLeuGluAsnTyrCysAsnEnd

However, no information about  
3D structure, function.....



## Four Major Methods are used to define ‘Genes’

1. Search for ‘Open Reading Frames’ (ORFs)
2. Presence of a consensus sequence for ribosome binding in the immediate vicinity of the start codon
3. Expressed Sequence Tags (EST) analysis
4. Homology of the putative gene to other known genes using BLAST

# Genome Comparisons using ‘BLAST’

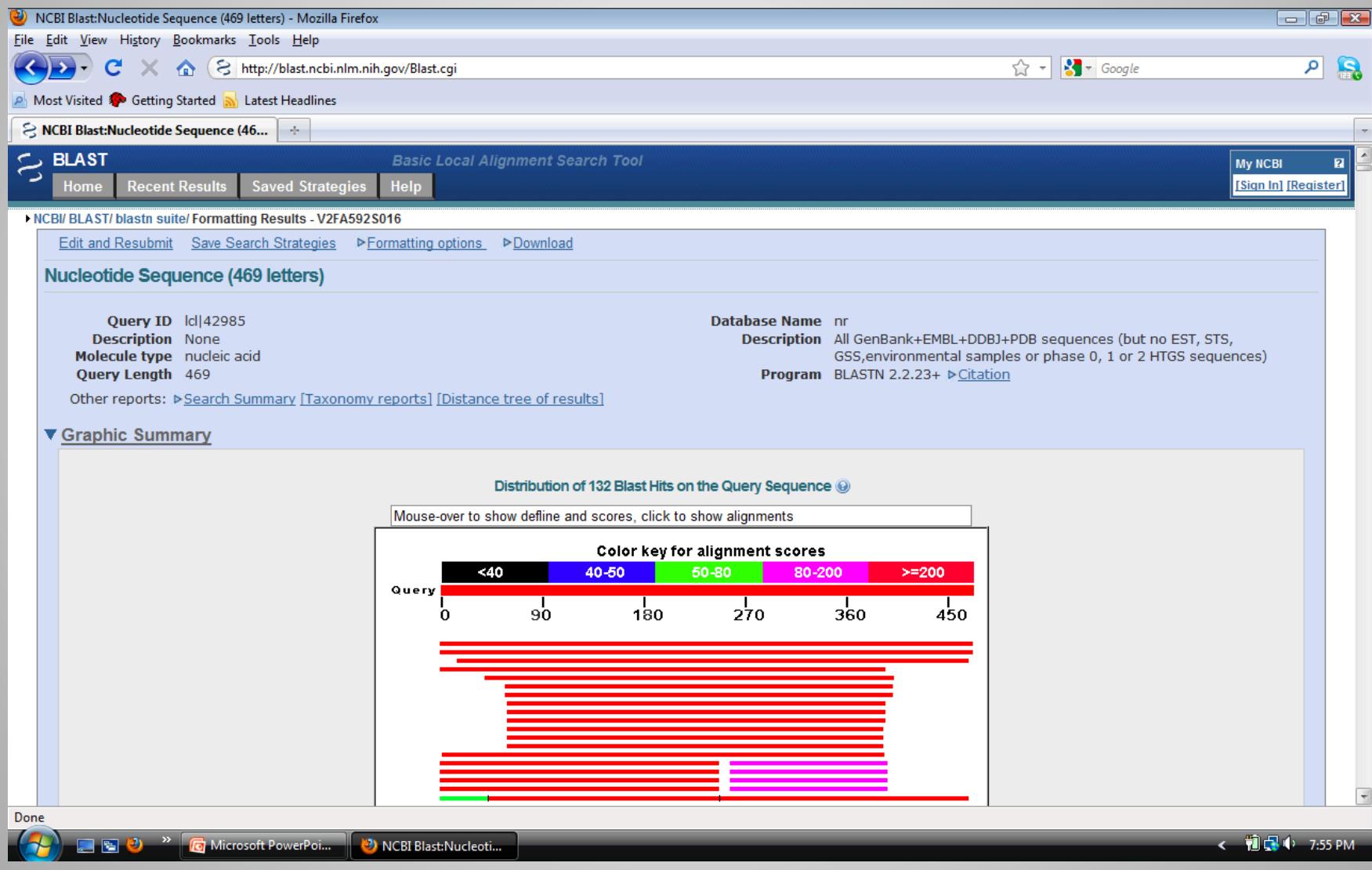
Basic Local Alignment Search Tool

An algorithm for comparing primary biological sequence information, such as the nucleotides of DNA sequences or the amino acid sequences of different proteins

The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches.

<http://blast.ncbi.nlm.nih.gov/Blast.cgi>

# Blast search: Human Insulin



# Blast Search: Human Insulin Continued

NCBI Blast:Nucleotide Sequence (469 letters) - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://blast.ncbi.nlm.nih.gov/Blast.cgi

Most Visited Getting Started Latest Headlines

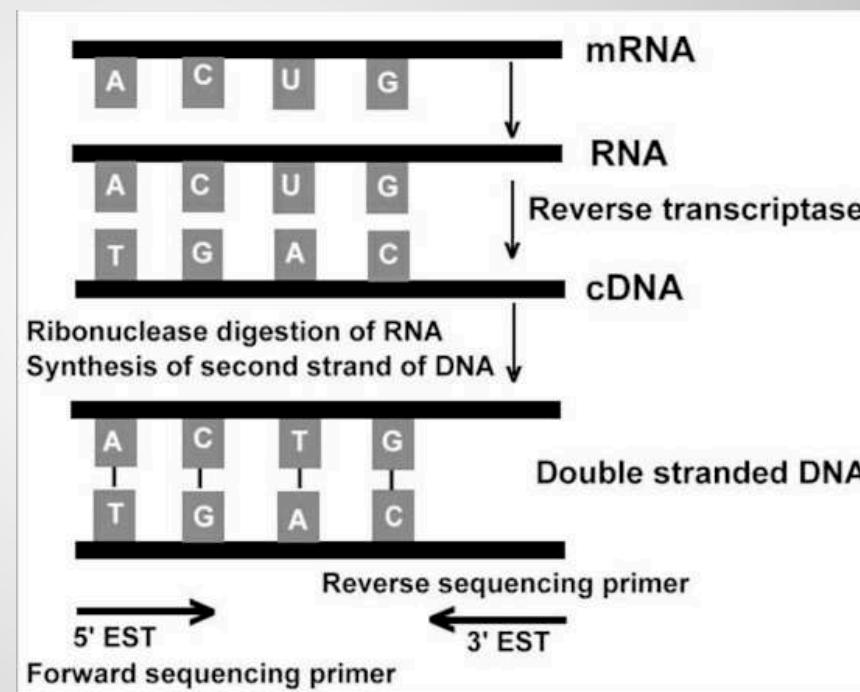
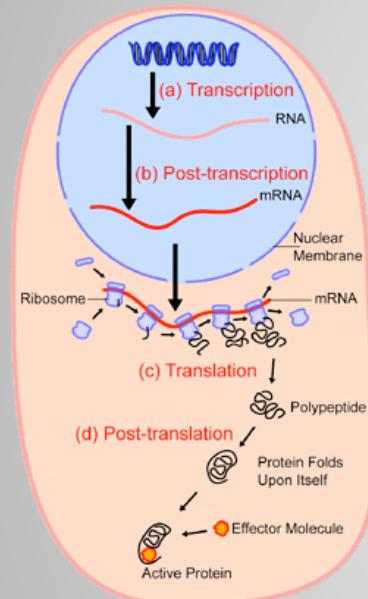
NCBI Blast:Nucleotide Sequence (469 letters)

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:							
Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NM_000207.2	Homo sapiens insulin (INS), mRNA	867	867	100%	0.0	100%	UEG
BC005255.1	Homo sapiens insulin, mRNA (cDNA clone MGC:12292 IMAGE:3950204)	856	856	100%	0.0	99%	UEG
X70508.1	Homo sapiens mRNA for insulinoma pre-proinsulin	821	821	95%	0.0	99%	UEG
NM_001008996.1	Pan troglodytes insulin (INS), mRNA	691	691	83%	0.0	98%	G
AY89304.1	Homo sapiens proinsulin mRNA, complete cds, alternatively spliced	667	667	76%	0.0	100%	UG
DQ893040.2	Synthetic construct clone IMAGE:100005670; FLH192922.01X; RZPD08	621	621	72%	1e-174	99%	G
DQ896283.2	Synthetic construct Homo sapiens clone IMAGE:100010743; FLH192918	616	616	72%	5e-173	99%	GY
BT006808.1	Homo sapiens insulin mRNA, complete cds	616	616	71%	5e-173	100%	UG
AY890084.1	Synthetic construct Homo sapiens clone FLH013503.01X insulin (INS) n	616	616	71%	5e-173	100%	UG
AY890083.1	Synthetic construct Homo sapiens clone FLH013502.01X insulin (INS) n	616	616	71%	5e-173	100%	UG
BT007778.1	Synthetic construct Homo sapiens insulin mRNA, partial cds	612	612	70%	6e-172	100%	UG
AY892564.1	Synthetic construct Homo sapiens clone FLH013499.01L insulin (INS) m	612	612	70%	6e-172	100%	UG
AY892563.1	Synthetic construct Homo sapiens clone FLH013498.01L insulin (INS) m	612	612	70%	6e-172	100%	UG
J00336.1	Monkey (M. fascicularis) preproinsulin mRNA, complete cds	597	597	82%	2e-167	94%	U
NR_003512.1	Homo sapiens INS-IGF2 readthrough transcript (INS-IGF2), transcript v	455	455	52%	1e-124	100%	UEG
NM_001042376.1	Homo sapiens INS-IGF2 readthrough transcript (INS-IGF2), transcript v	455	455	52%	1e-124	100%	UEG
DO104205.1	Homo sapiens INSIGF long transcript variant mRNA, complete cds, alte	455	455	52%	1e-124	100%	G
DO104204.1	Homo sapiens INSIGF short transcript variant mRNA, complete cds, alte	455	455	52%	1e-124	100%	G
NG_007114.1	Homo sapiens insulin (INS) on chromosome 11	409	871	99%	9e-111	100%	G
AC132217.15	Homo sapiens chromosome 11, clone RP11-889I17, complete sequenc	409	871	99%	9e-111	100%	G
M10039.1	Human alpha-type insulin gene and 5' flanking polymorphic region	409	871	99%	9e-111	100%	G
AY138590.1	Homo sapiens insulin (INS) gene, exons 1, 2, 3, and complete cds; and	403	865	99%	4e-109	100%	G
AC130303.8	Homo sapiens chromosome 11, clone RP4-539G11, complete sequence	403	865	99%	4e-109	100%	G
J00265.1	Human insulin gene, complete cds	403	865	99%	4e-109	100%	G
AJ009655.1	Homo sapiens ins gene, partial	403	787	90%	4e-109	100%	G
L15440.1	Homo sapiens tyrosine hydroxylase (TH) gene, 3' end; insulin (INS) ge	403	865	99%	4e-109	100%	EG
V00565.1	Human gene for preproinsulin, from chromosome 11. Includes a highly	403	865	99%	4e-109	100%	EG
NM_001130093.1	Canis lupus familiaris insulin (INS), mRNA	377	377	71%	2e-101	87%	UG
AY127500.1	Canis lupus familiaris insulin (INS), mRNA	275	275	0%	0e+001	100%	UG

Done Microsoft PowerPoint NCBI Blast:Nucleoti... 7:57 PM

# ESTs (Expressed Sequence Tags) are Used to Determine Which Protein Coding Genes are Expressed in a Particular Cell Type or Tissue



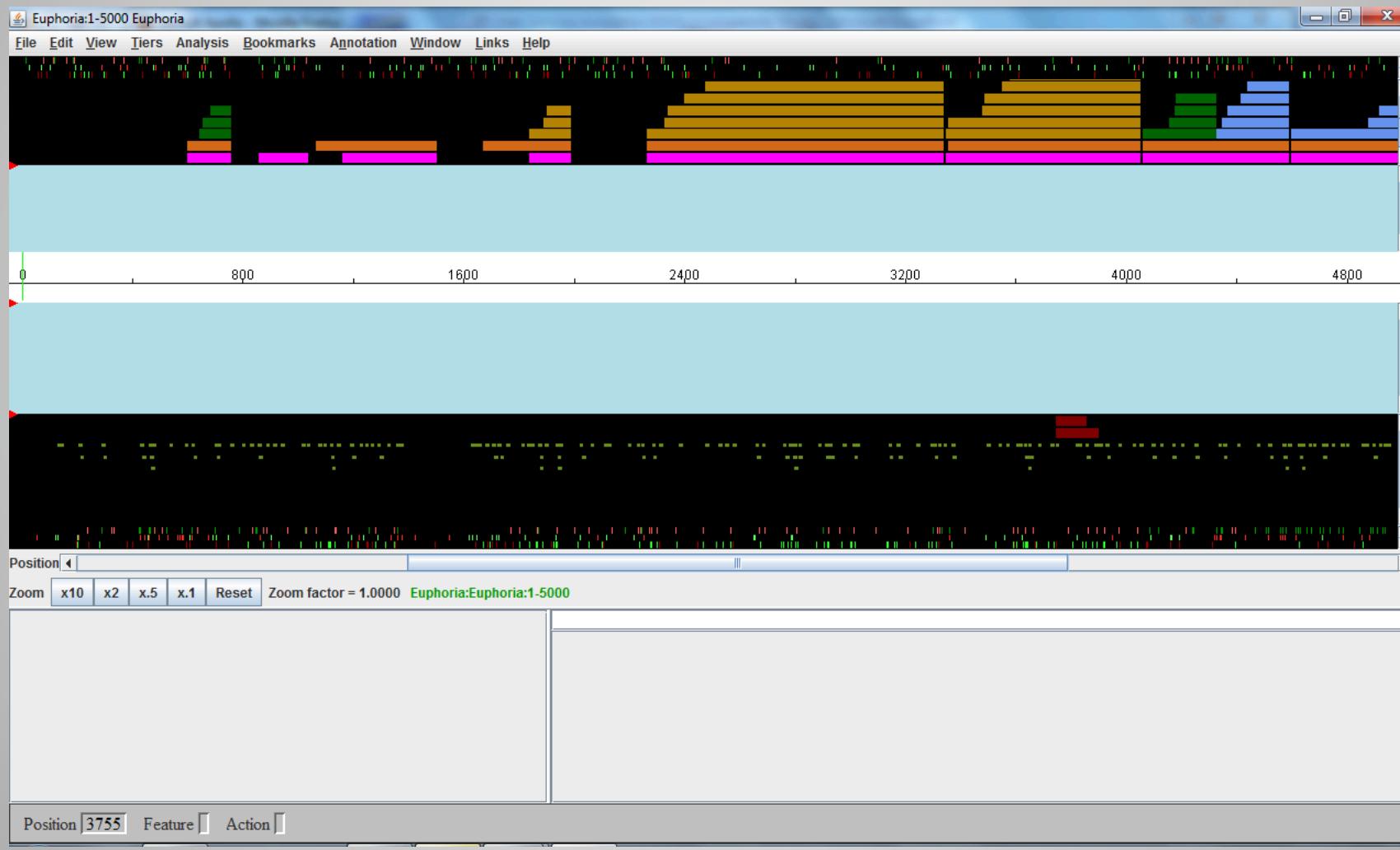
Approximately 200 different cell types  
Only a subset of genes is being expressed on cellular level under certain conditions  
Map expressed genes on cellular level to whole genome

## Gene Calling Softwares

GLIMMER: uses Interpolated Markov models (IMMs)

GeneMark: determine the protein-coding potential of a DNA sequence (within a sliding window) by using species specific parameters of the Markov models of coding and non-coding regions

# Genome Annotation of Bacteriophage EUPHORIA using GeneMark and GLIMMER algorithms displayed in ‘Apollo’

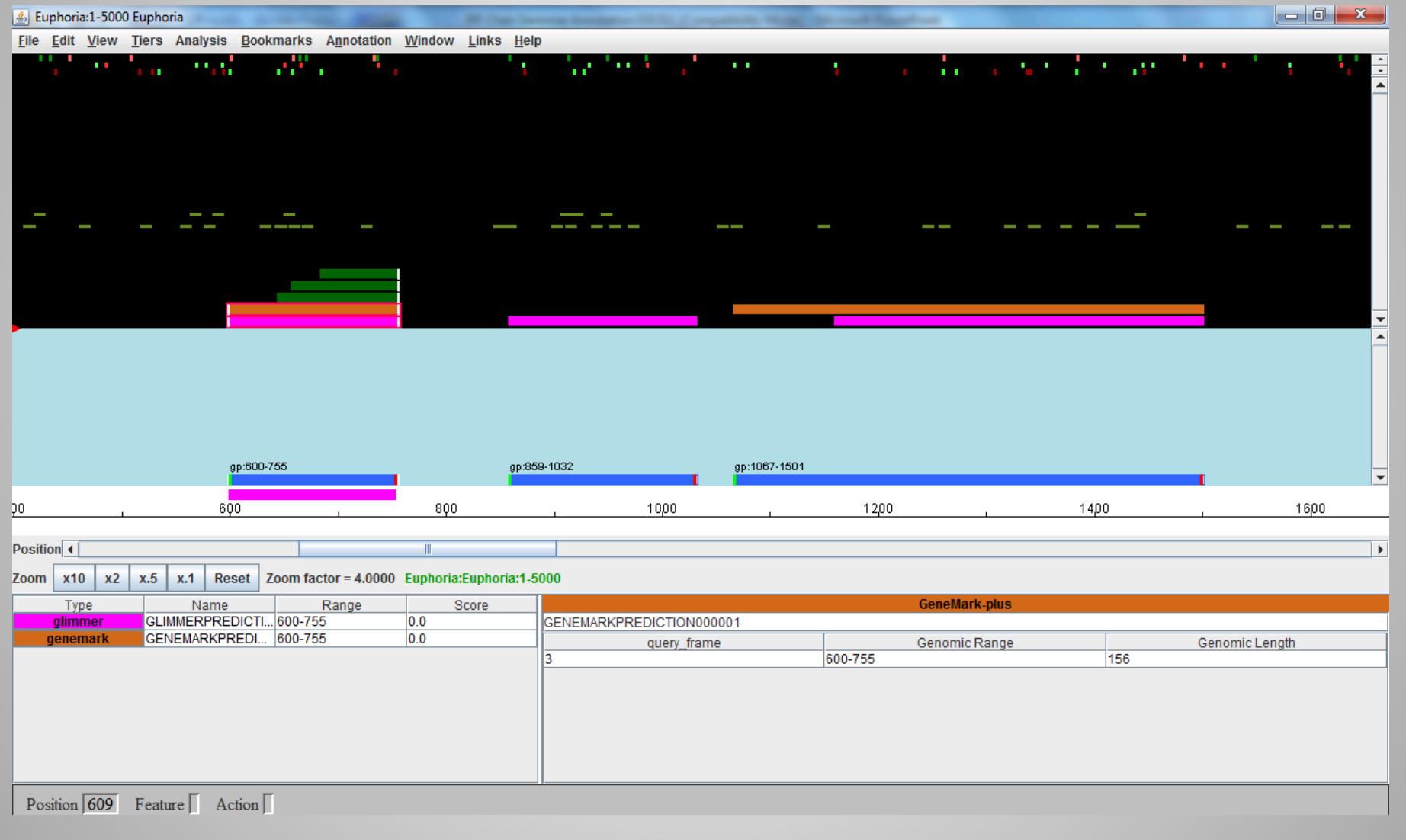


purple: GLIMMER

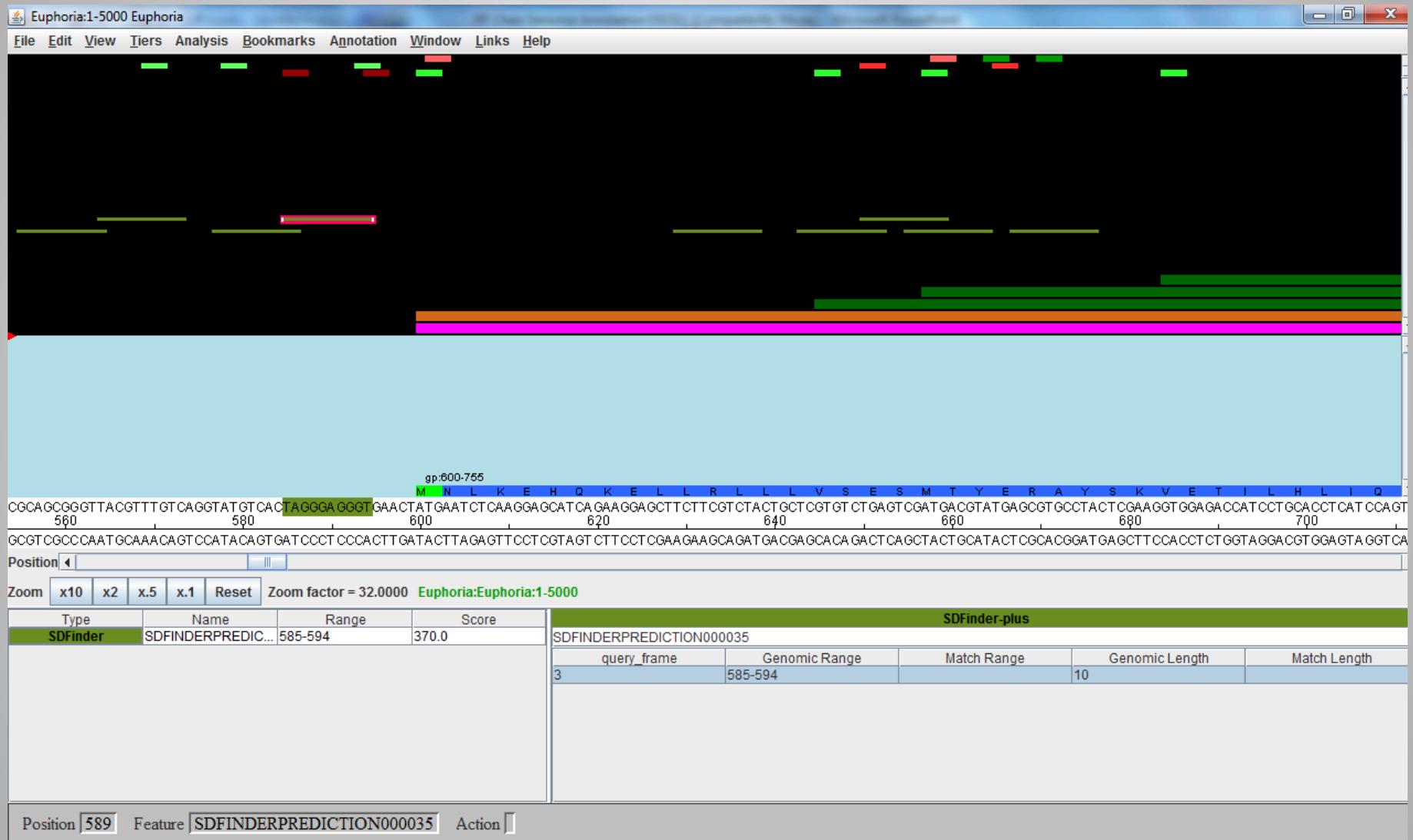
dark brown: GeneMark

light brown: GeneMark-TB

# First three Gene Calls of EUPHORIA genome using GLIMMER and GeneMark Gene Prediction



# Apollo: Zoom into nucleotide resolution



Bright green: Start Codon

Red: Stop Codon

Light green: ribosomal binding site

## Summary

Genome sequencing is followed by gene identification and genome annotation

Genome annotation is the critical process by which functions are assigned to predicted proteins.

Gene annotation performed computationally should always be viewed as generating a hypothesis that needs to be experimentally tested.