# Analysis of One Gene

The role of the c-myc oncogene and hemopoietic growth factors in tumorigenesis

Gregory D. Schuler

Ph.D. Thesis, Princeton University, 1988

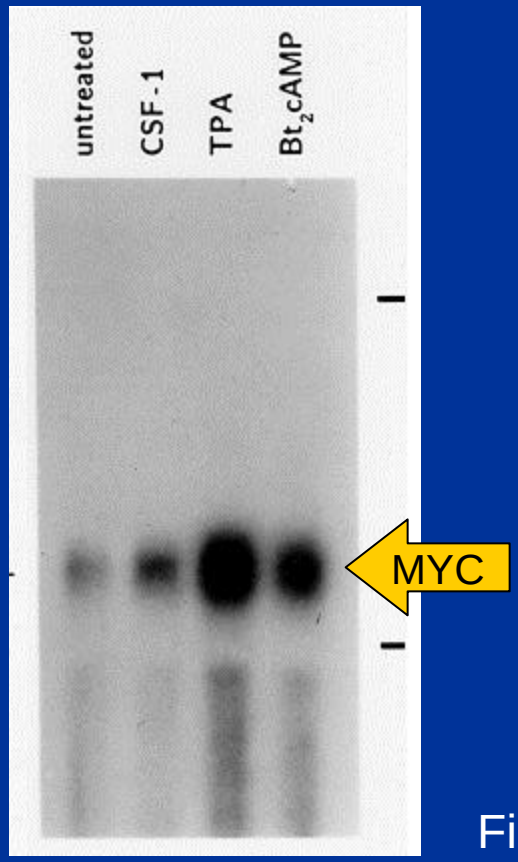



Figure 7

# Analysis of Many Genes

Science Jan 1 1999: 83-87

The Transcriptional Program in the Response of Human Fibroblasts to Serum

Vishwanath R. Iyer, Michael B. Eisen, Douglas T. Ross, Greg Schuler, Troy Moore, Jeffrey C. F. Lee, Jeffrey M. Trent, Louis M. Staudt, James Hudson Jr., Mark S. Boguski, Deval Lashkari, Dari Shalon, David Botstein, Patrick O. Brown

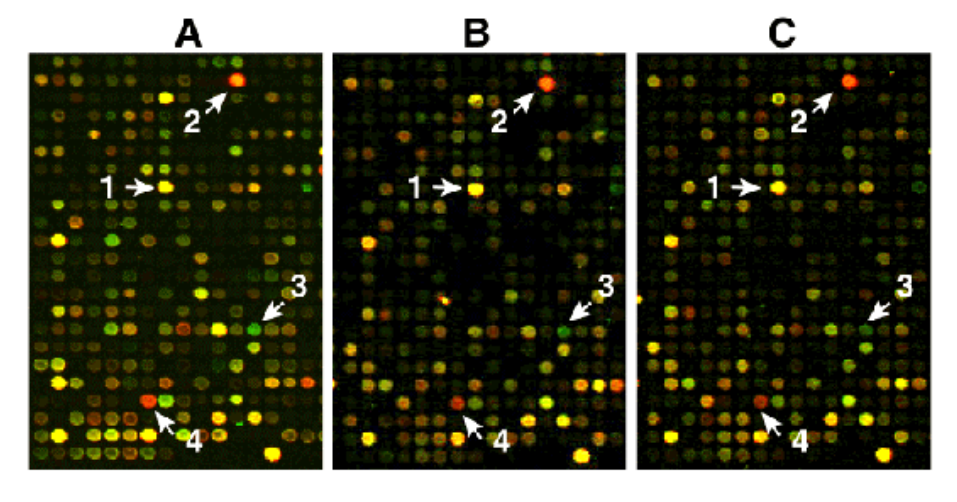



Figure 1

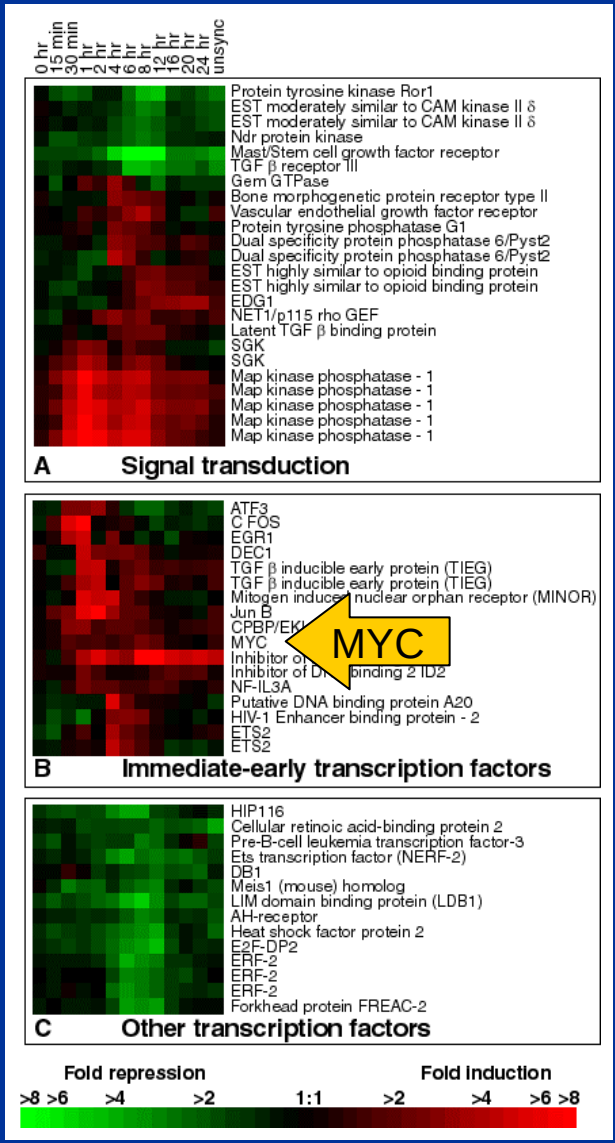
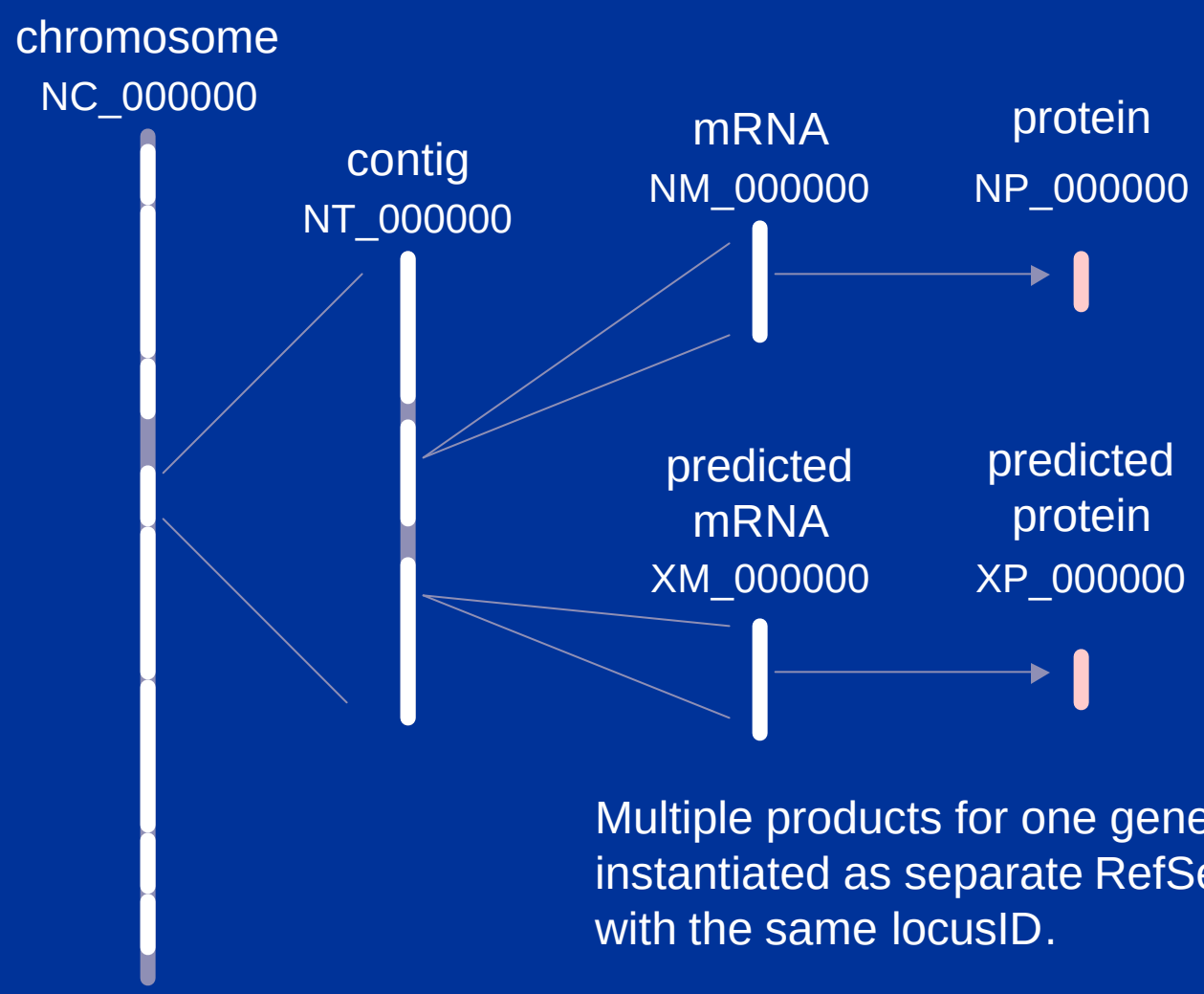



Figure 4

# *Reference sequences*

Goal: One sequence entry for each naturally occuring DNA and protein molecule




Multiple products for one gene are instantiated as separate RefSeqs with the same locusID.

NCBI

Kim Pruitt

Donna Maglott

Ken Katz

Hugues Sicotte

Jim Ostell

Collaborators:

Human gene nomenclature committee

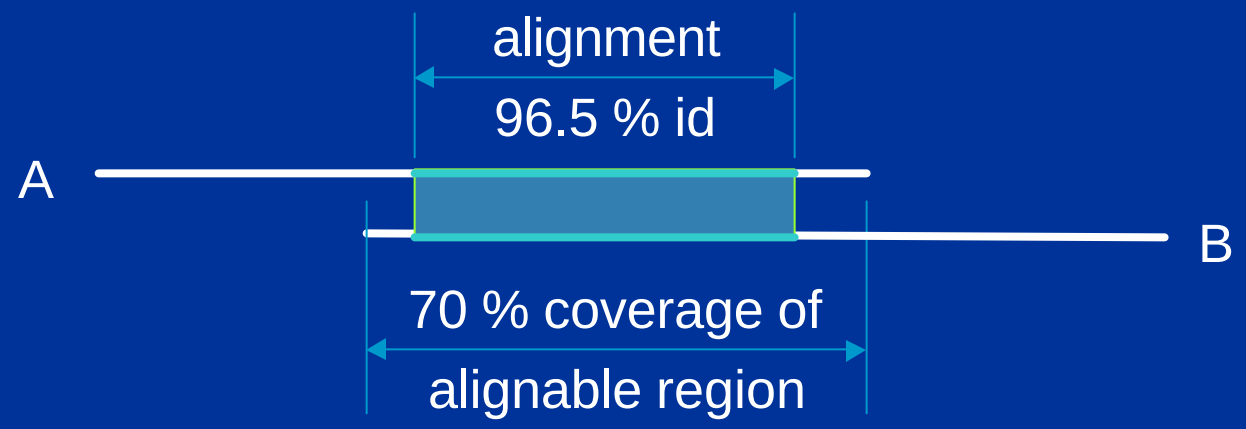
Online Mendelian Inheritence in Man

## *It's the Genes, Stupid!*

"If something like 98% of the genome is junk, then the best strategy would be to find the important 2%, and sequence it first"

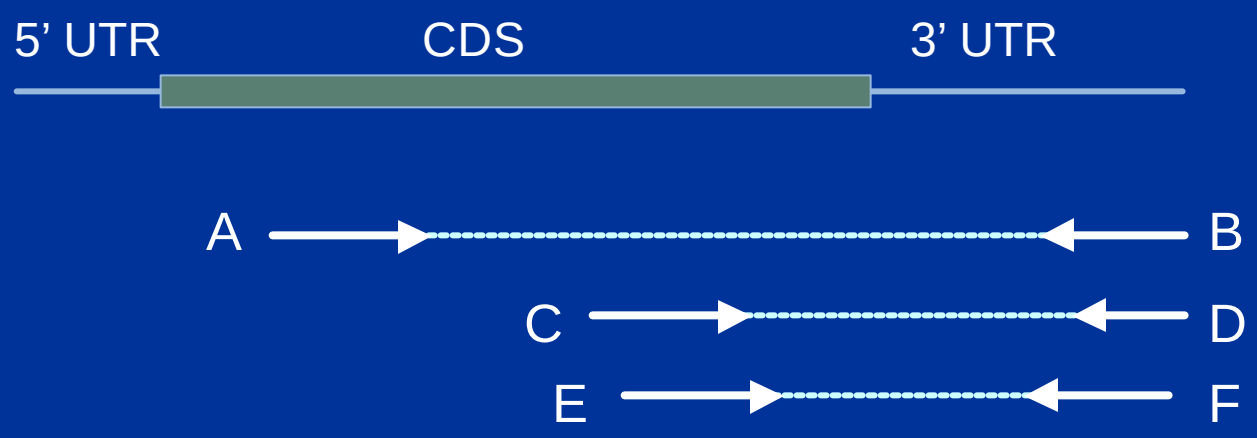
– Sydney Brenner, 1990

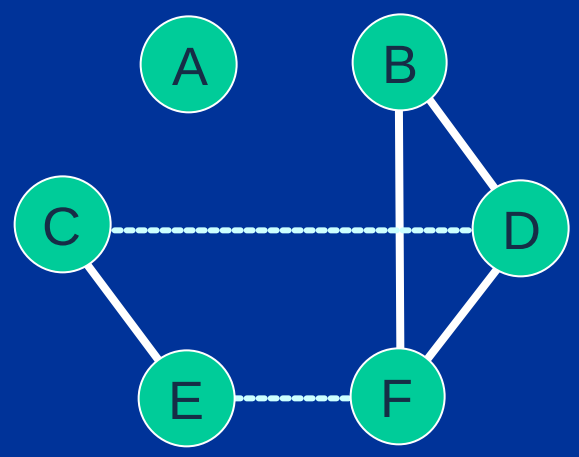
# UniGene: Sequence Matching




NCBI
Lukas Wagner
Joan Pontius
Greg Schuler

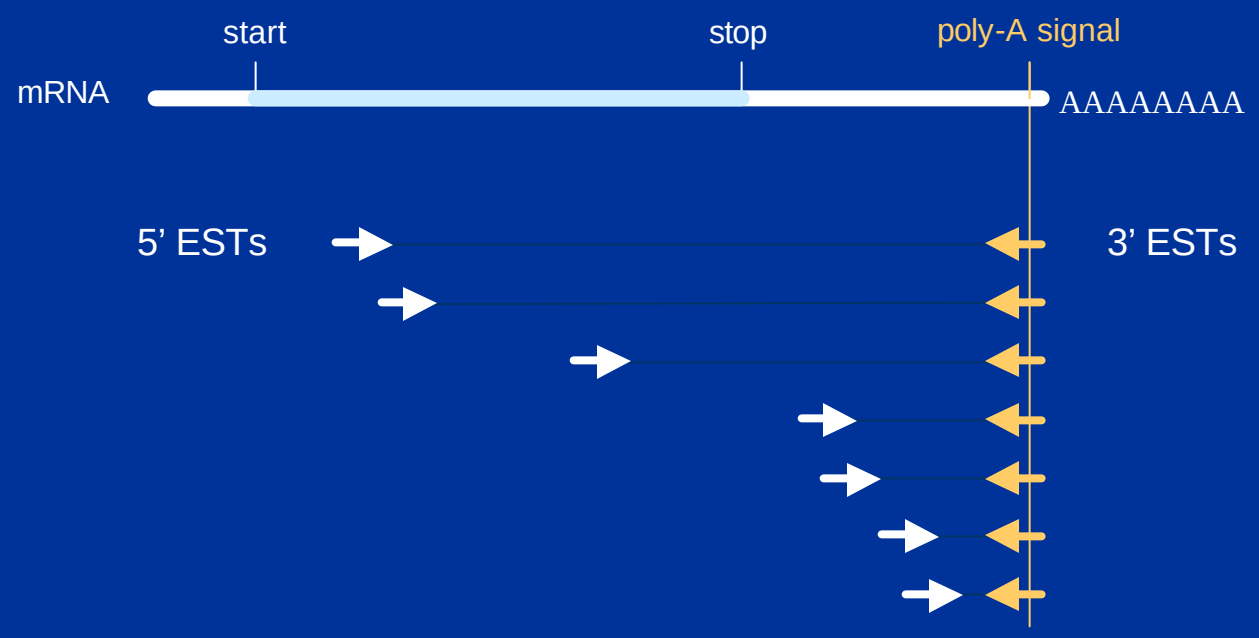
# UniGene: Clone Pairs




NCBI
Lukas Wagner
Joan Pontius
Greg Schuler

# UniGene: 3’ Anchoring



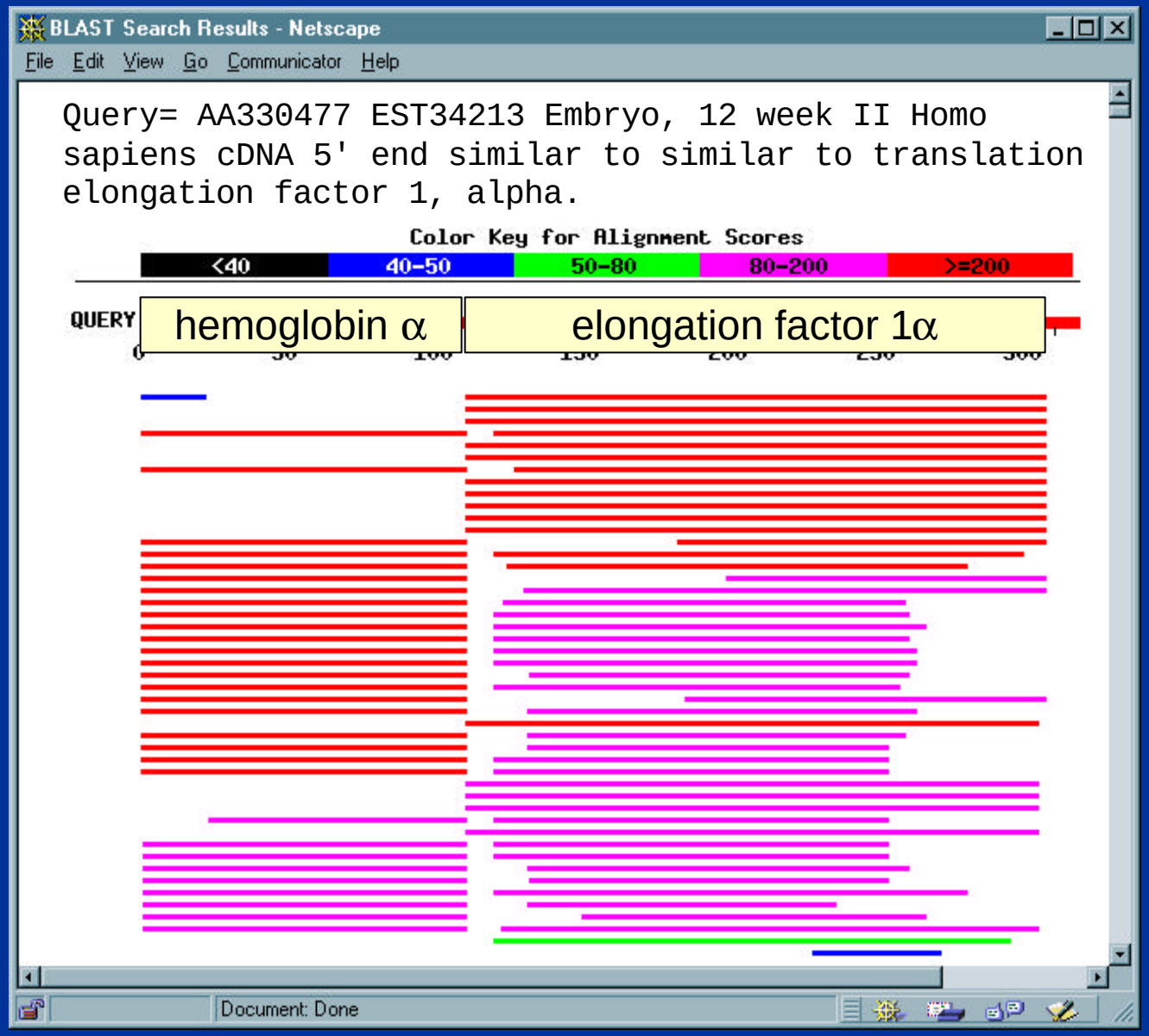

NCBI

Lukas Wagner

Joan Pontius

Greg Schuler

# Chimeric Sequences





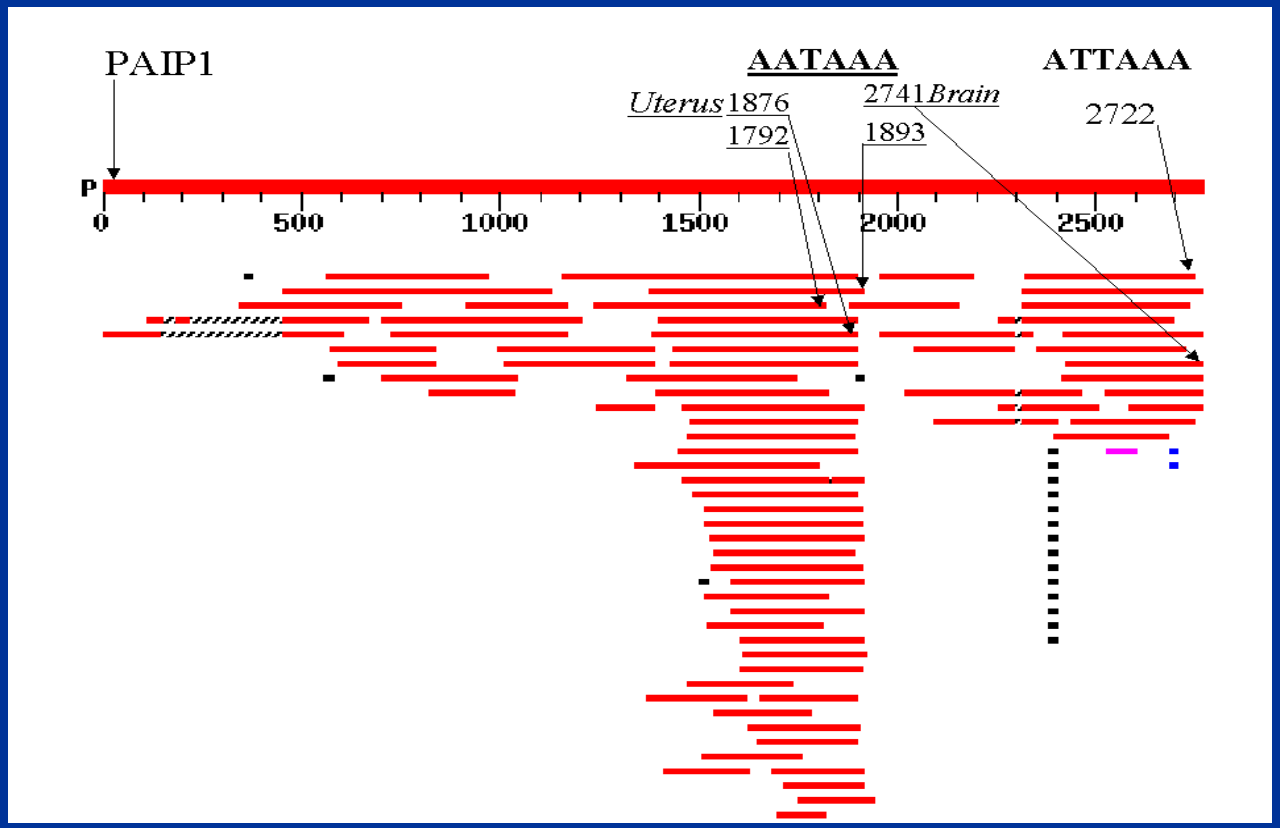
NCBI

Lukas Wagner

Joan Pontius

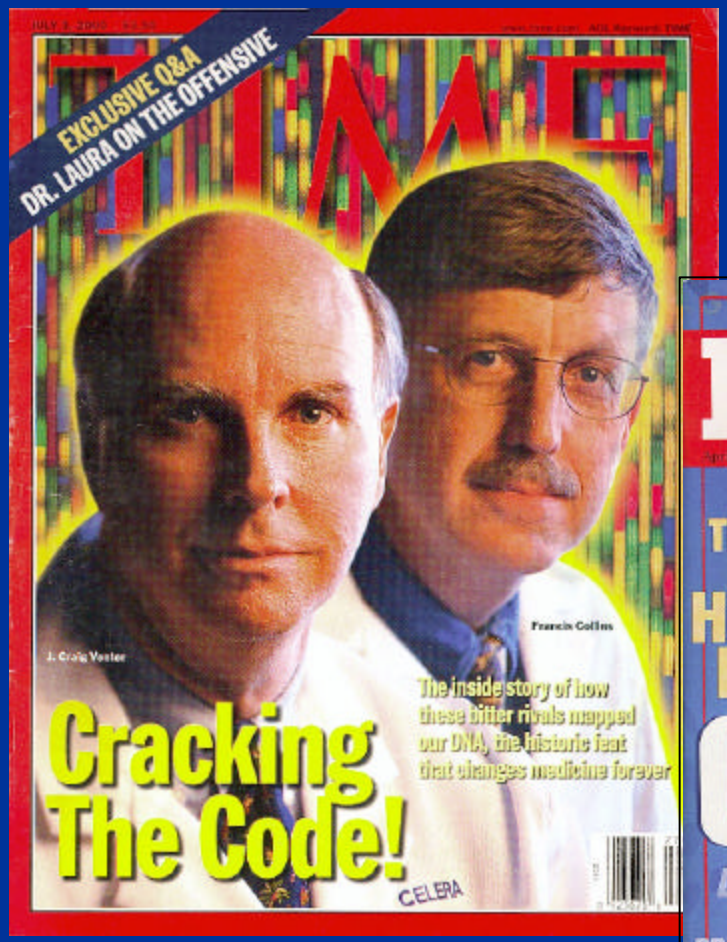
Greg Schuler

# Alternative 3' ends

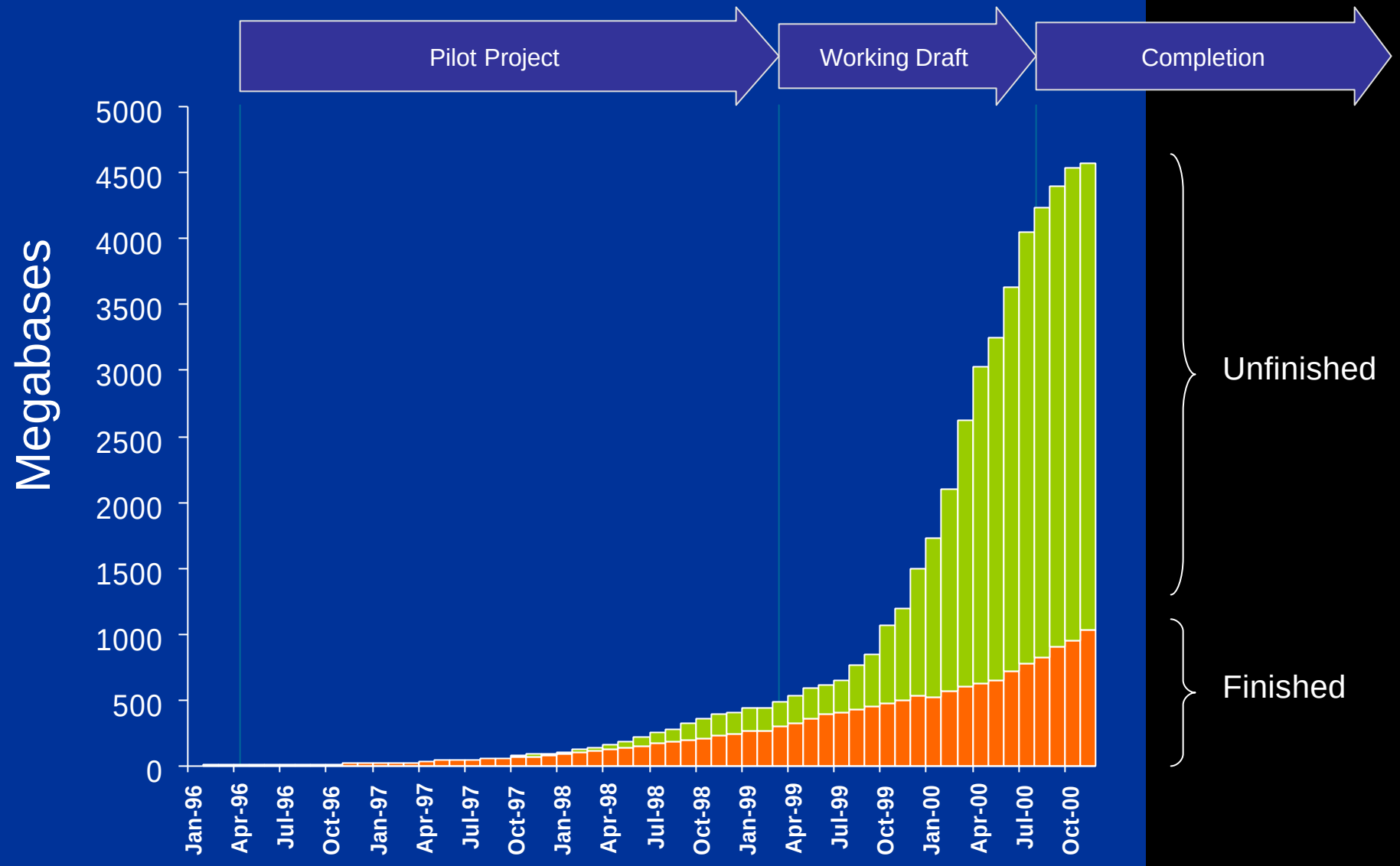



NCBI
Lukas Wagner
Joan Pontius
Greg Schuler

# Human Genome In the News



# Human Genome Sequence in GenBank



# Draft Entries in Genbank

NCBI Sequence Viewer - Netscape

```
LOCUS       AC079399    174446 bp    DNA              HTG       30-SEP-2000
DEFINITION  Homo sapiens chromosome 4 clone RP11-414P19, WORKING DRAFT
            SEQUENCE, 9 unordered pieces.
```

```
*          1      1089: contig of 1089 bp in length
*       1090      1189: gap of unknown length
*       1190      6890: contig of 5701 bp in length
*       6891      6990: gap of unknown length
*       6991     16392: contig of 9402 bp in length
*      16393     16492: gap of unknown length
*      16493     31956: contig of 15464 bp in length
*      31957     32056: gap of unknown length
*      32057     44673: contig of 12617 bp in length
*      44674     44773: gap of unknown length
*      44774     58419: contig of 13646 bp in length
*      58420     58519: gap of unknown length
*      58520     91391: contig of 32872 bp in length
*      91392     91491: gap of unknown length
*      91492    118691: contig of 27200 bp in length
*     118692    118791: gap of unknown length
*     118792    174446: contig of 55655 bp in length.
```

```
Insert size: 173646; sum-of-contigs
Quality coverage: 5.30 in Q20 bases; agarose-fp
Quality coverage: 5.11 in Q20 bases; sum-of-contigs
-----------------------------------------------------.
```

# Draft Entries in Genbank

NCBI Sequence Viewer - Netscape

File Edit View Go Communicator Help

NCBI Entrez Nucleotide

PubMed Nucleotide Protein Genome Structure Popset

Search Nucleotide for Go Clear

Limits Index History Clipboard

Display Default View as HTML Save Add to Clipboard Hide Brief and LinkBar

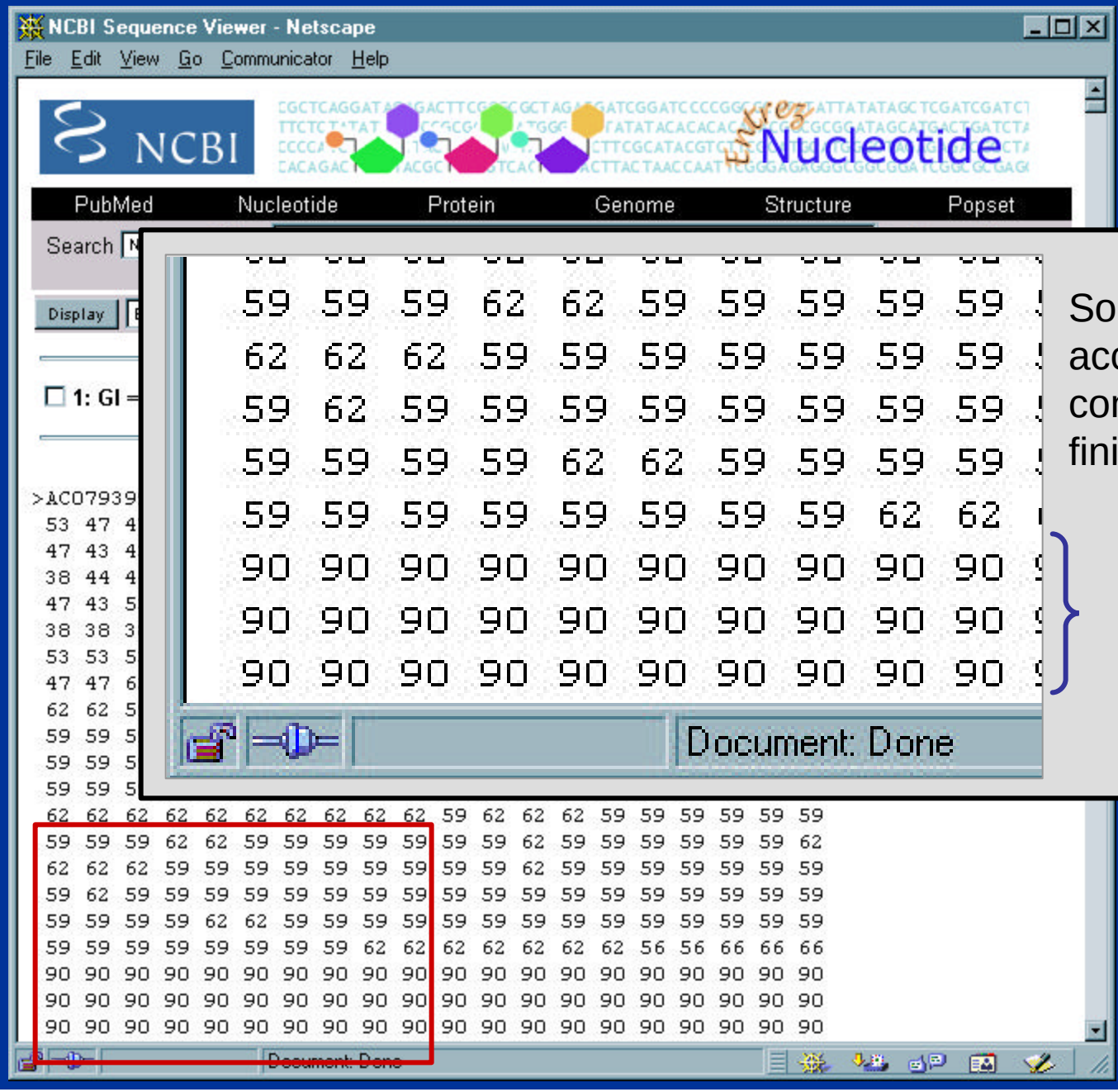
ASN.1
FASTA
GenBank
Graphics
Base Quality
XML
Default

1: GI [GenBank] Taxonomy

Display option “Base Quality”

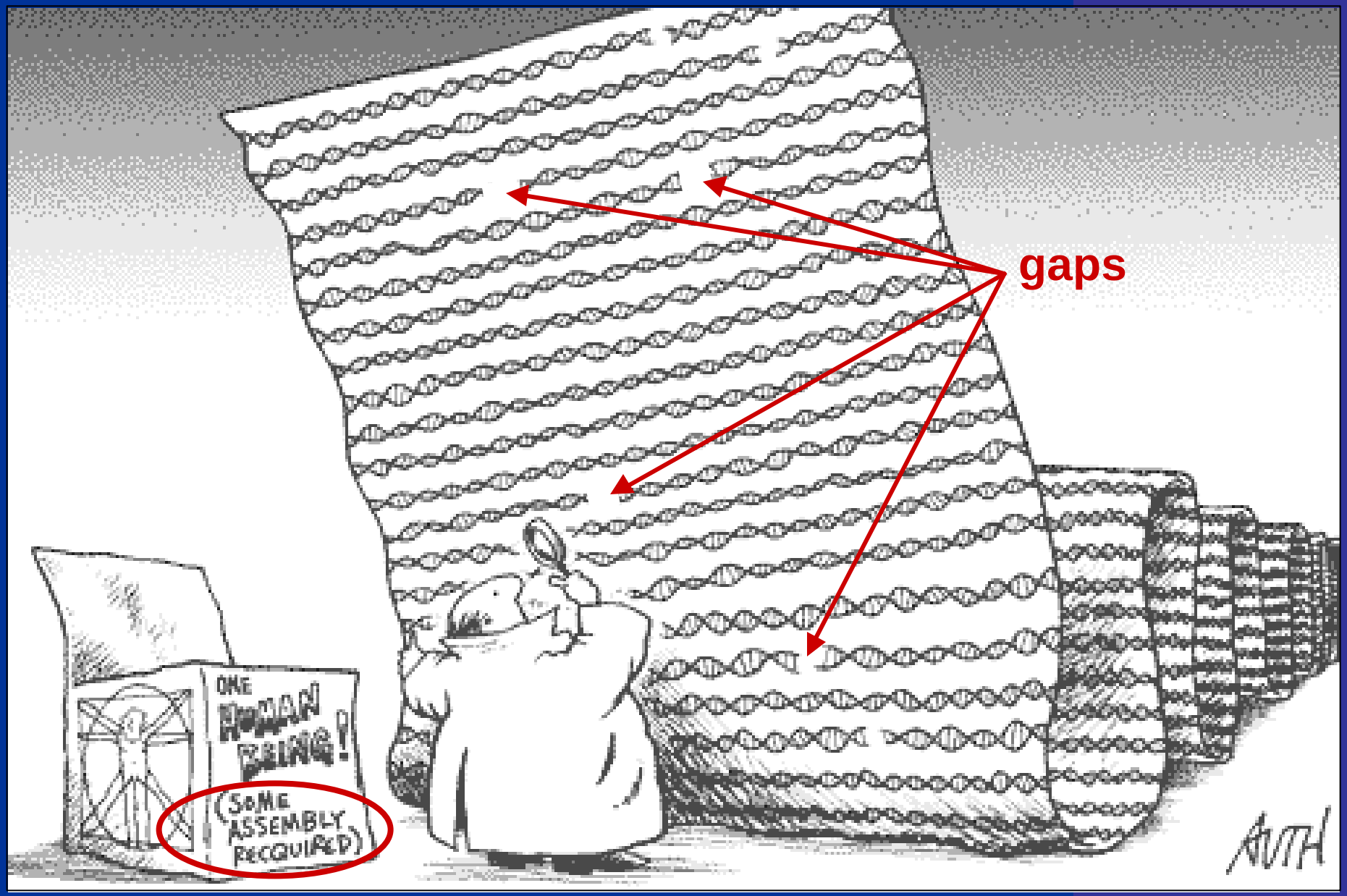
```
LOCUS       AC079399    174446 bp    DNA             HTG       30-SEP-2000
DEFINITION  Homo sapiens chromosome 4 clone RP11-414P19, WORKING DRAFT
            SEQUENCE, 9 unordered pieces.
ACCESSION   AC079399
VERSION     AC079399.2  GI:10440754
KEYWORDS    HTG; HTGS_PHASE1; HTGS_DRAFT.
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 174446)
  AUTHORS   Waterston,R.H.
  TITLE     The sequence of Homo sapiens clone
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 174446)
  AUTHORS   Waterston,R.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (31-AUG-2000) Genome Sequencing Center, Washington
            University School of Medicine, 4444 Forest Park Parkway, St. Louis,
            MO 63108, USA
COMMENT     On Sep 30, 2000 this sequence version replaced gi:9954846.
```

# Base quality values



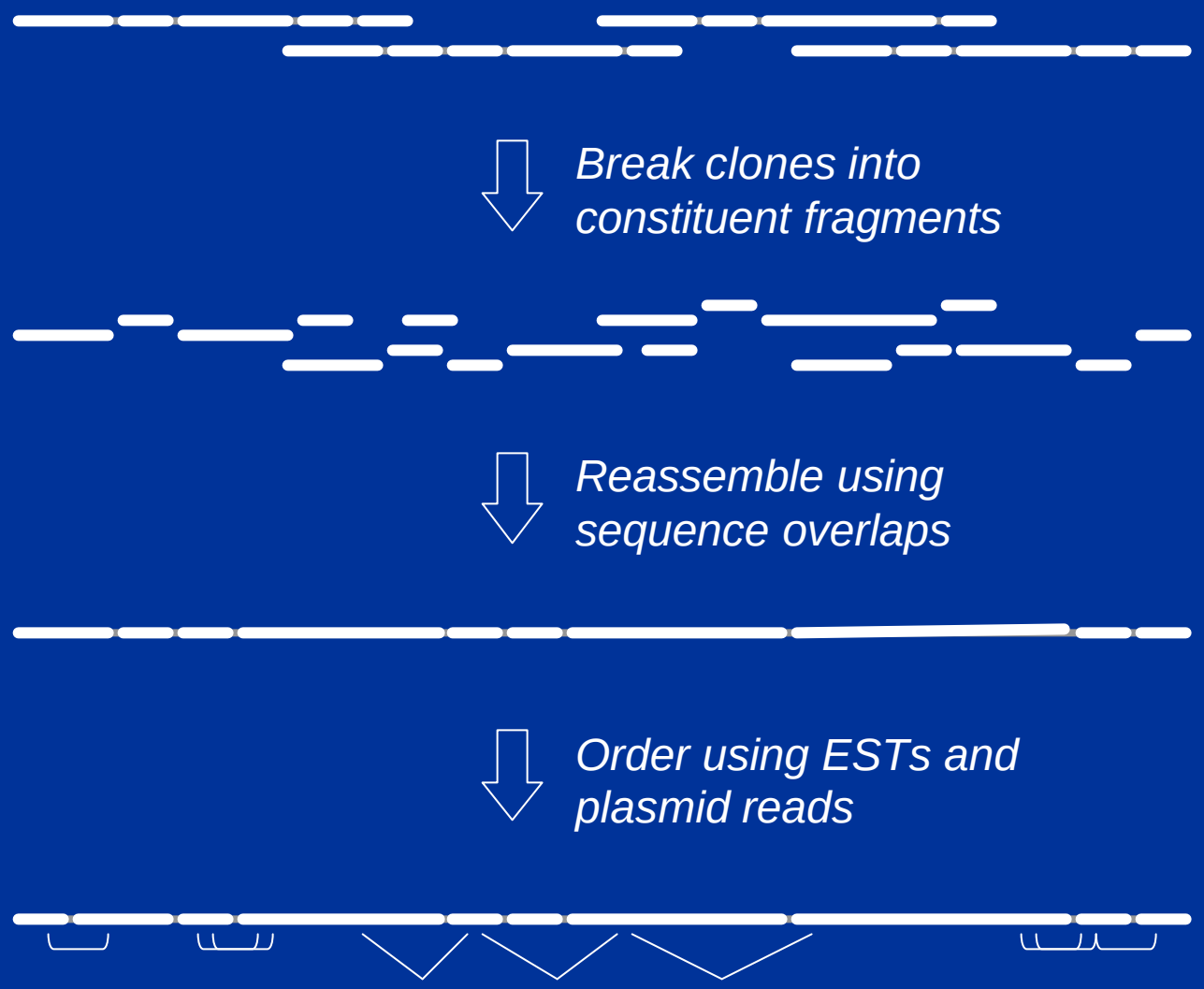

Some regions have accuracy comparable to finished sequence

# *Working Draft Sequence*



# *Some assembly required*

Overlaping draft clones




NCBI

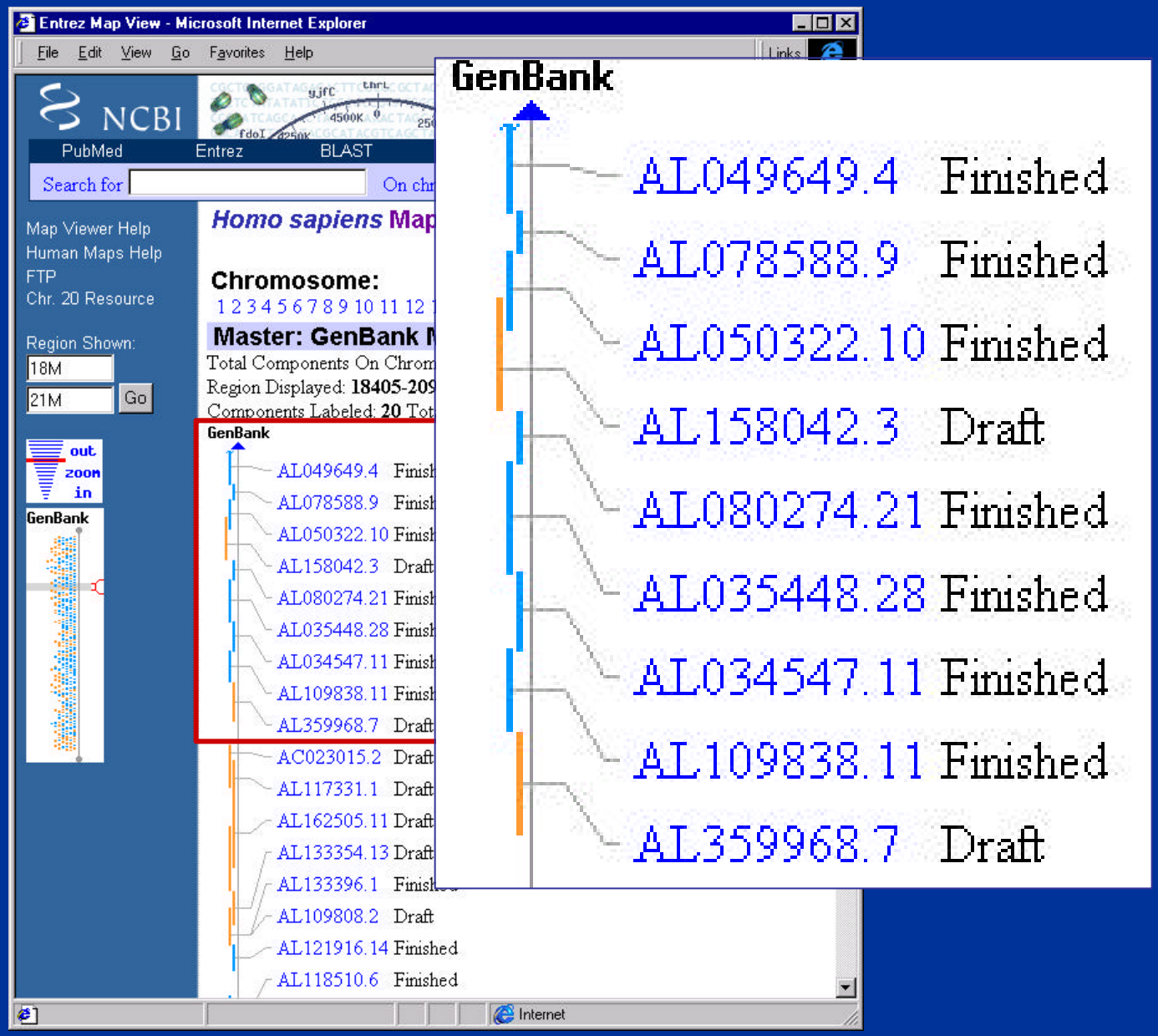
Richa Agarwala

David Lipman

Paul Kitts

Greg Schuler

# Genome Assembly: Map Viewer





NCBI

Slava Chetvernin

Deanna Church

Wonhee Jang

Jonathan Kans

Donna Maglott

Jim Ostell

Kim Pruitt

Sergei Resenchuk

Steve Sherry

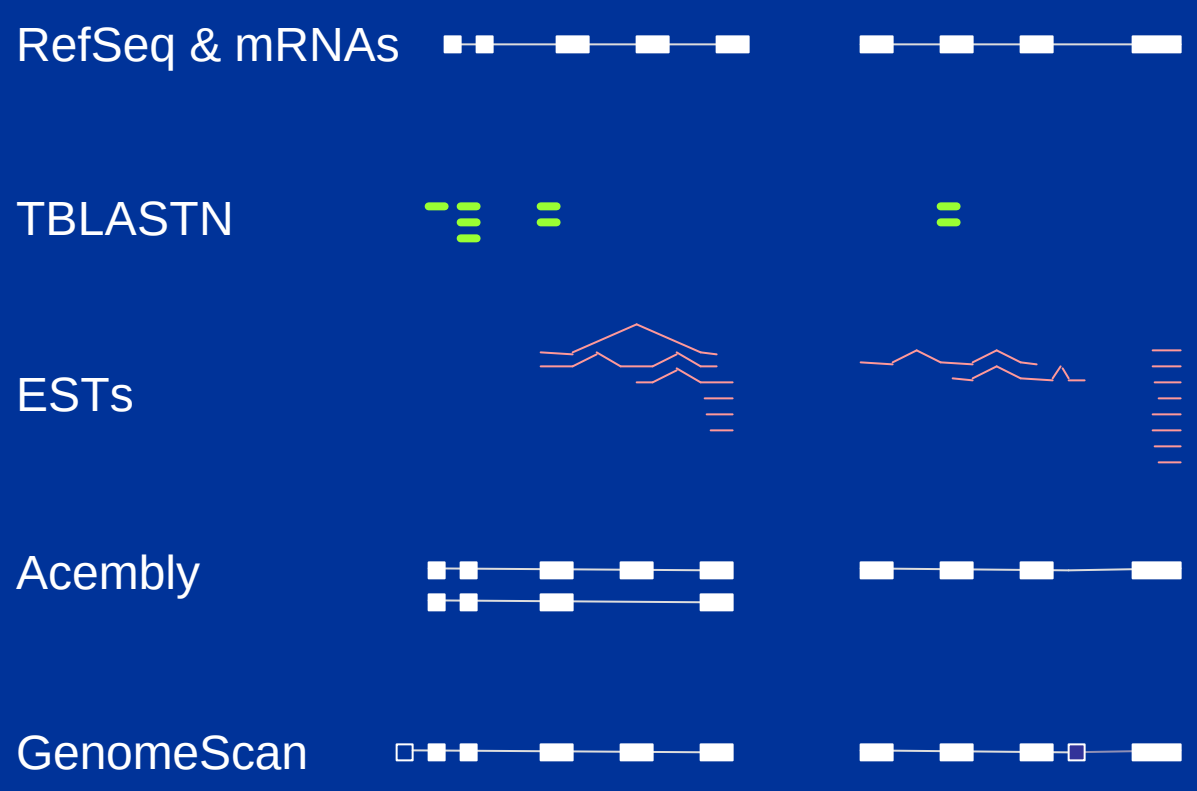
Tatiana Tatusov

Tyra Wolfsberg

# Draft Contig Genbank Entry

```
misc_feature    join(937343..937570,939593..939601)
                /gene="Hs20_11485_18:7_7"
                /note="rve; Score: 123, E_value: 2.782970e-08, bit_score:
                52.323100; Region: Integrase core domain"
STS             941009..941130
                /gene="Hs20_11485_18:7_7"
                /db_xref="dbSTS:37004"
mRNA            join(956106..956175,963889..964061,965530..965646,
                966237..966314)
                /gene="Hs20_11485_18:7_7"
                /product="TR00007142_m"
CDS             join(956106..956175,963889..964061,965530..965646,
                966237..966314)
                /gene="Hs20_11485_18:7_7"
                /codon_start=1
                /exception="No explanation supplied"
                /product="TR00007142_p"
mRNA            complement(join(1014364..1014543,1015872..1015990,
                1019203..1019266,1019414..1019541,1020138..1020264,
                1020563..1020675,1021740..1021832,1026612..1026749,
                1027991..1028138,1029542..1029694,1030068..1030243,
                1030313..1030450,1037361..1037560,1048006..1048182,
                1054663..1054783,1060514..1060586,1063766..1064067,
                1064371..1064704))
                /gene="Hs20_11485_18:7_10"
                /product="TR00007143_m"
gene            complement(1014364..1064704)
                /gene="Hs20_11485_18:7_10"
CDS             complement(join(1014364..1014543,1015872..1015990,
                1019203..1019266,1019414..1019541,1020138..1020264,
                1020563..1020675,1021740..1021832,1026612..1026749,
                1027991..1028138,1029542..1029694,1030068..1030243,
                1030313..1030450,1037361..1037560,1048006..1048182,
                1054663..1054783,1060514..1060586,1063766..1064067,
                1064371..1064704))
```

# Gene prediction




NCBI

Jean Thierry-Mieg

Danielle Thierry-Mieg

Adam Lowe

David Lipman

Greg Schuler

Wonhee Jang

Whitehead / MIT

Chris Burge

# Finding the Genes

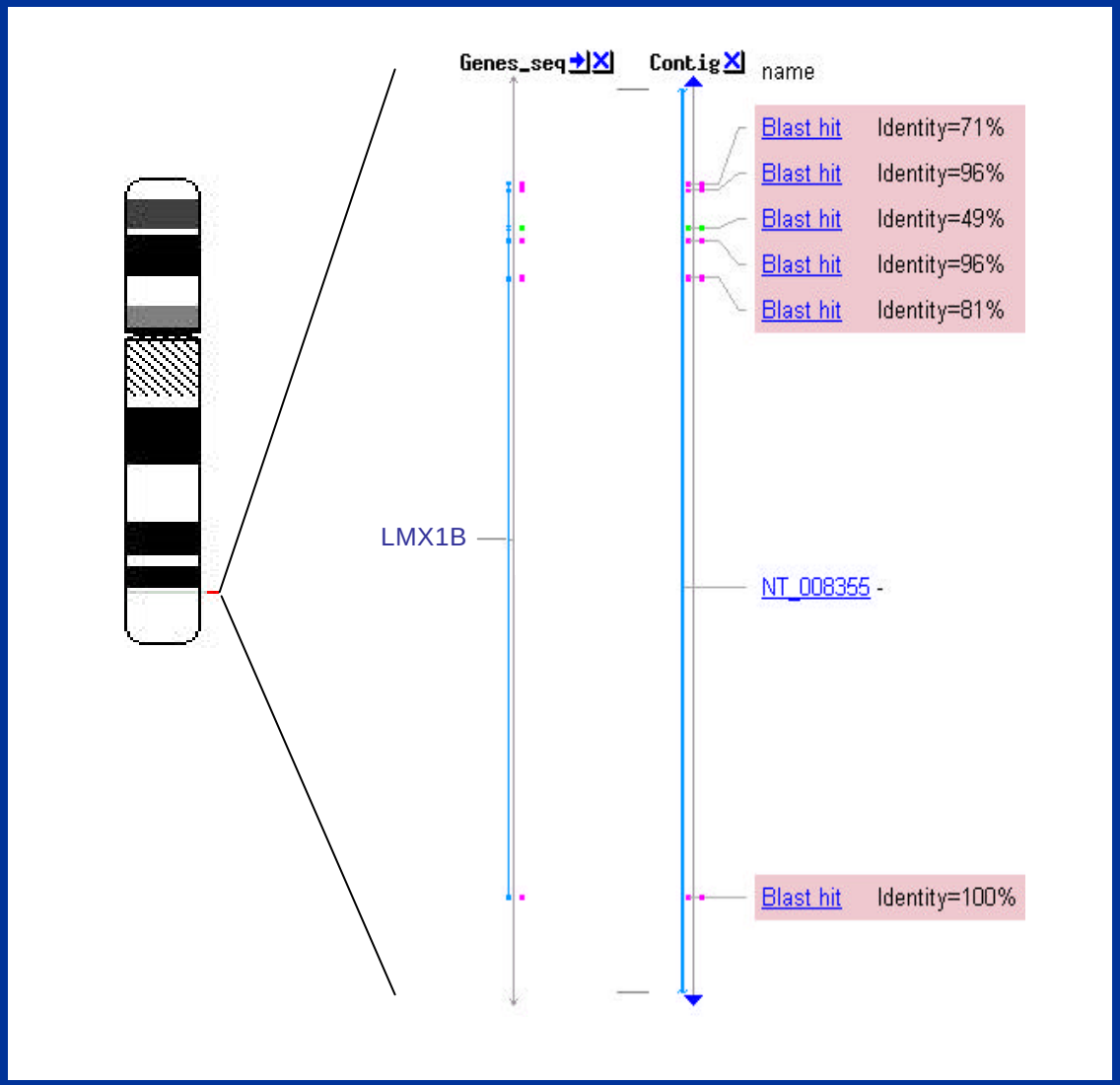
| Quality of match | Min length of mRNA that aligned | %base identity over aligned region | Number of known genes successfully aligned from 10272 RefSeqs (%) |
|---|---|---|---|
| Excellent | 98% | 99% | 5364 (52%) |
| Very Good | 90% | 97% | 7041 (69%) |
| Good | 80% | 97% | 7663 (75%) |
| OK | 50% | 97% | 8691 (85%) |

NCBI

Jean Thierry-Mieg

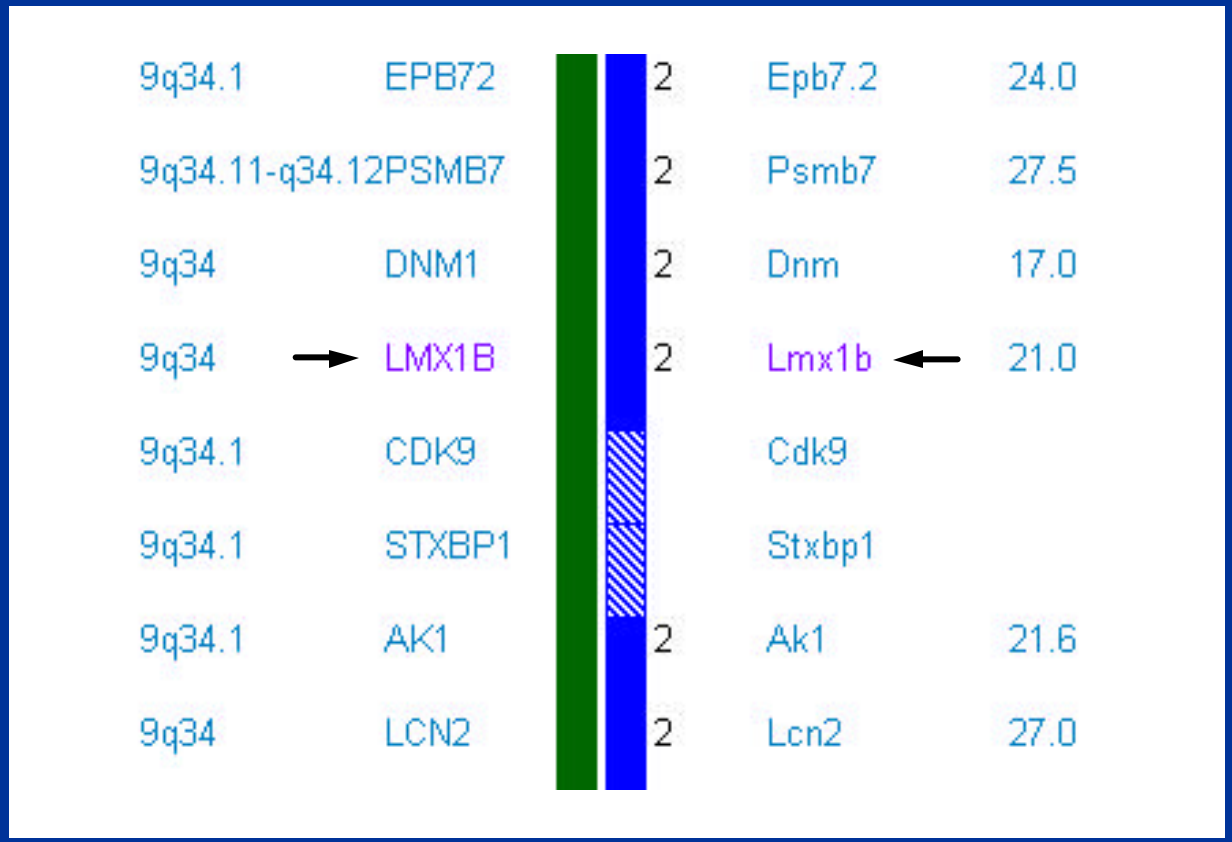
Danielle Thierry-Mieg

# Finding an Ortholog



# Finding an Ortholog

NCBI
Deanna Church
Donna Maglott

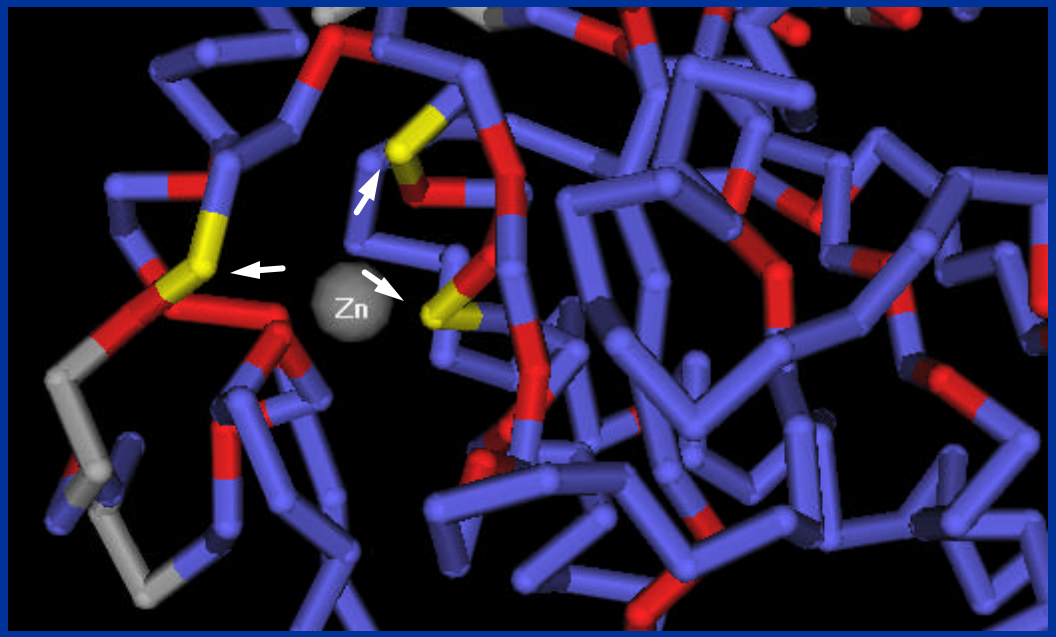


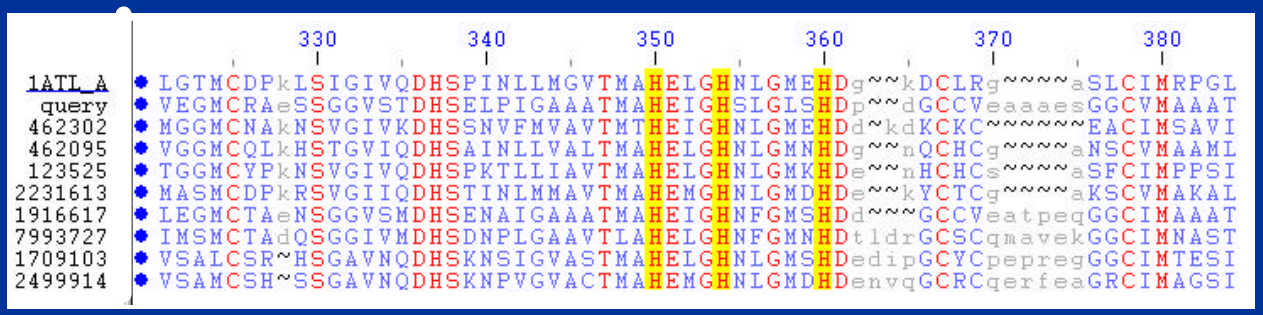
# Finding a Paralog

| Sequences producing significant alignments: | | Score (bits) | E value |
|---|---|---|---|
| gnl\|Pfam\|pfam01421 | Reprolysin, Reprolysin (M12B) family zinc metalloprotease | 244 | 9e-66 |
| gnl\|Smart\|DISIN | Homologues of snake disintegrins; Snake disintegrins inhibit t... | 74.1 | 2e-14 |
| gnl\|Pfam\|pfam00200 | disintegrin, Disintegrin | 65.6 | 8e-12 |

NCBI
Tyra Wolfsberg
Jo McEntyre

# Finding a Paralog



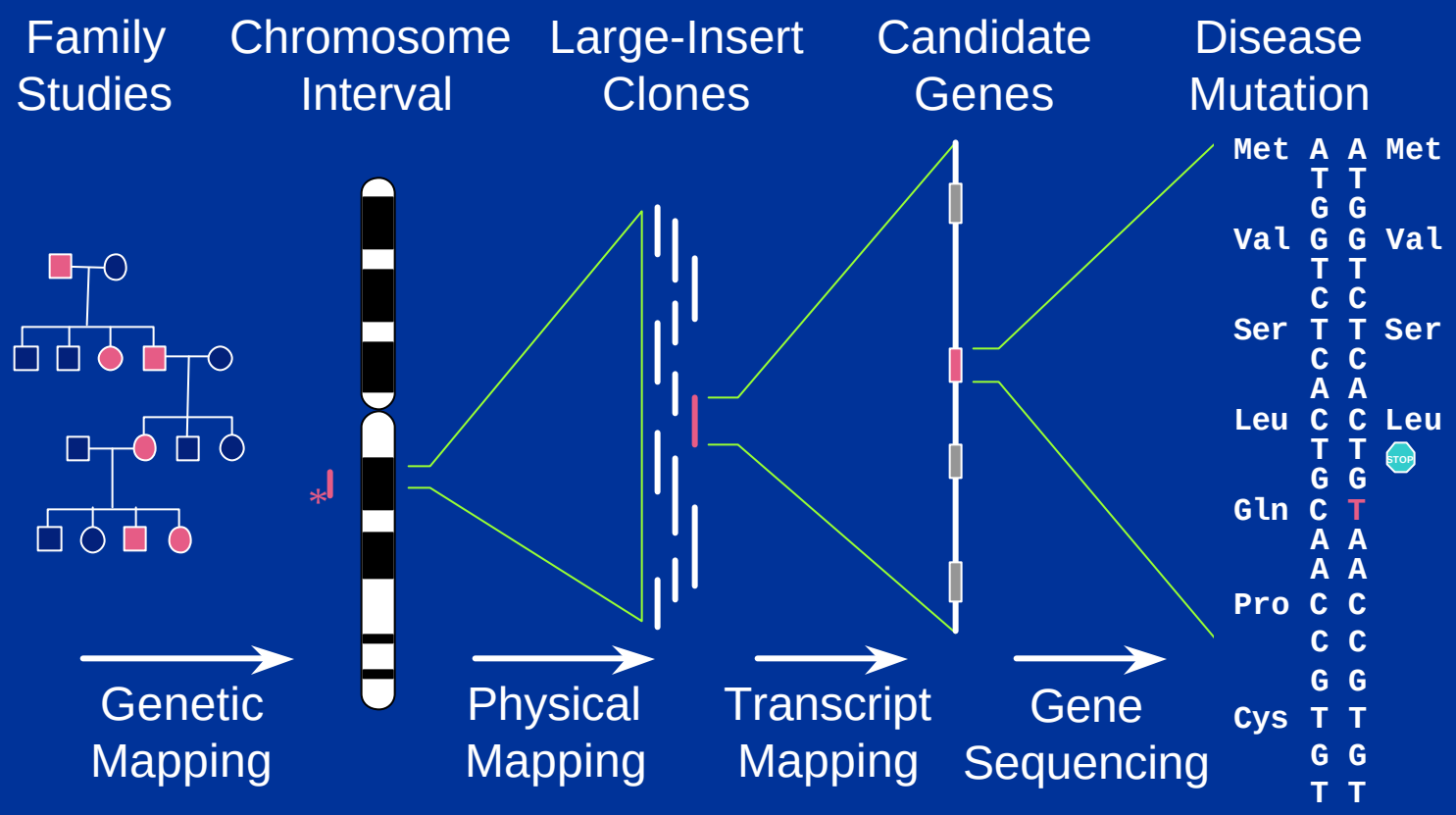




NCBI

Tyra Wolfsberg

Jo McEntyre

# Positional Cloning



# Disease Gene Hunting

© 1998 Nature America Inc. • http://genetics.nature.com



## Evidence for a prostate cancer susceptibility locus on the X chromosome

Jianfeng Xu[1*], Deborah Meyers[1], Diha Freije[2], Sarah Isaacs[2], Kathy Wiley[2], Deborah Nusskern[2], Charles Ewing[2], Eric Wilkens[2], Piroska Bujnovszky[2], G. Steven Bova[2,3], Patrick Walsh[2] & William Isaacs[2,4] (*Group 1*)
Johanna Schleutker[5*], Mika Matikainen[5*], Teuvo Tammela[5], Tapio Visakorpi[5] & Olli-P. Kallioniemi[12] (*Group 2*)
Rebecca Berry[6*], Daniel Schaid[7*], Amy French[6], Shannon McDonnell[7], Jennifer Schroeder[6], Michael Blute[8] & Stephen Thibodeau[6] (*Group 3*)
Henrik Grönberg[9*], Monika Emanuelsson[9], Jan-Erik Damber[10], Anders Bergh[11] & Björn-Anders Jonsson[11] (*Group 4*)
Jeffrey Smith[12*], Joan Bailey-Wilson[12], John Carpten[12], Dietrich Stephan[12], Elizabeth Gillanders[12], Isaac Amundson[12], Tommi Kainu[12], Diana Freas-Lutz[12], Agnes Baffoe-Bonnie[13], Anne Van Aucken[12], Raman Sood[12], Francis Collins[12], Michael Brownstein[12] & Jeffrey Trent[12] (*Group 5*)
*These authors contributed equally to this work.*

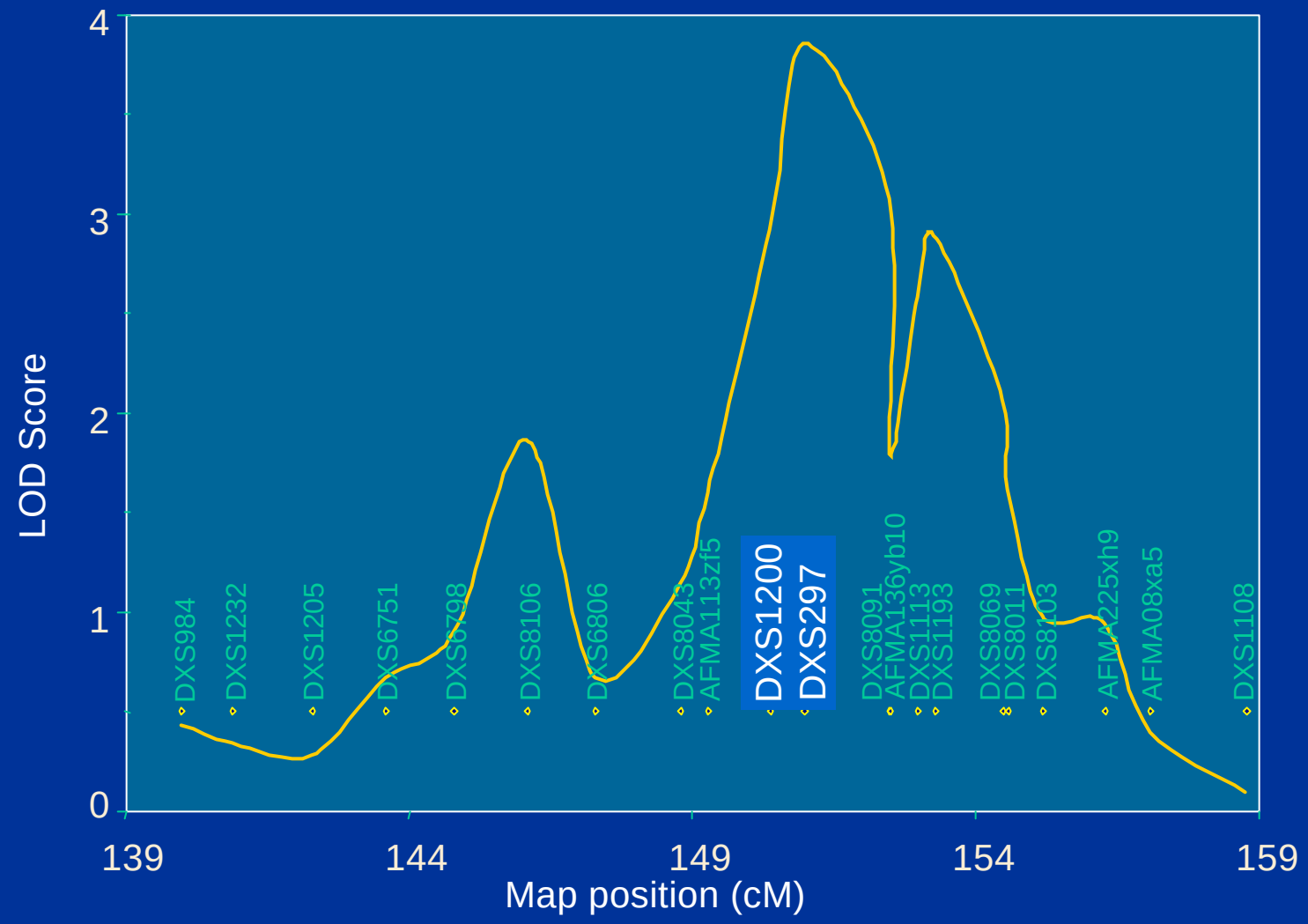
**Over 200,000 new prostate cancer cases are diagnosed in the United States each year, accounting for more than 35% of all cancer cases affecting men, and resulting in 40,000 deaths annually[1]. Attempts to characterize genes predisposing to prostate cancer have been hampered by a high phenocopy rate, the late age of onset of the disease and, in the absence of distinguishing clinical features, the inability to stratify patients into subgroups relative to suspected genetic locus heterogeneity. We previously performed a genome-wide search for hereditary prostate cancer (HPC) genes, finding evidence of a prostate cancer susceptibility locus on chromosome 1 (termed *HPC1*; ref. 2). Here we present evidence for the location of a second prostate cancer susceptibility gene, which by heterogeneity estimates accounts for approximately 16% of HPC cases. This HPC locus resides on the X chromosome (Xq27–28), a finding consistent with results of previous population-based studies suggesting an X-linked mode of HPC inheritance. Linkage to Xq27–28 was observed in a combined study population of 360 prostate cancer families collected**

**Genetic mapping of the locus represents an important initial step in the identification of an X-linked gene implicated in the aetiology of HPC.**

Despite the medical significance of prostate cancer in terms of morbidity, mortality and health-care costs, our understanding of the molecular determinants of prostate cancer susceptibility remains rudimentary. Epidemiological studies supporting the existence of hereditary forms of prostate cancer have led to the initiation of genome-wide searches for loci contributing to hereditary prostate cancer. A previous scan for linkage resulted in suggestive evidence (lod>1.0) for prostate cancer susceptibility loci on several chromosomes, including 1q, 4q, 5p, 7p, 13q and Xq (ref. 2). Statistically significant evidence was achieved only for the locus 1q24–25 (*HPC1*). Subsequent stratification of pedigrees showed that families linked to *HPC1* tended to have an early mean age of diagnosis (under 65 years) and a large number of affected members (>4). Even in this subset, this locus accounts for only approximately one-half of the families[3].

1998 Nature America Inc. • http://genetics.nature.com

# Parametric Multipoint Analysis



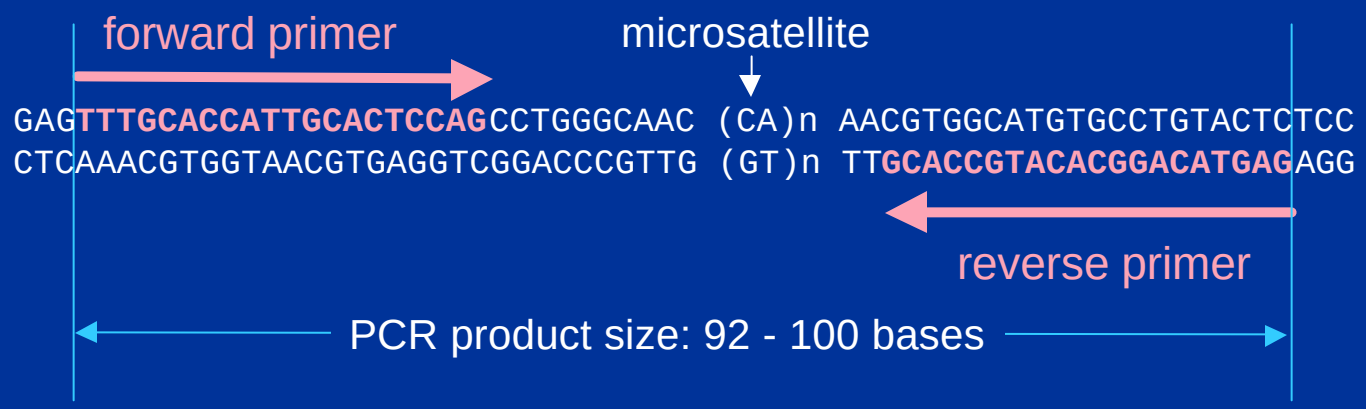

Reference

Xu, J. et al. Nature Genetics 20, 195-197 (1998).

# *Electronic PCR (e-PCR)*

Search DNA sequences for exact matches to both primers in correct order, orientation, and spacing to be consistent with known PCR product size.

STS marker D6S1606




Reference

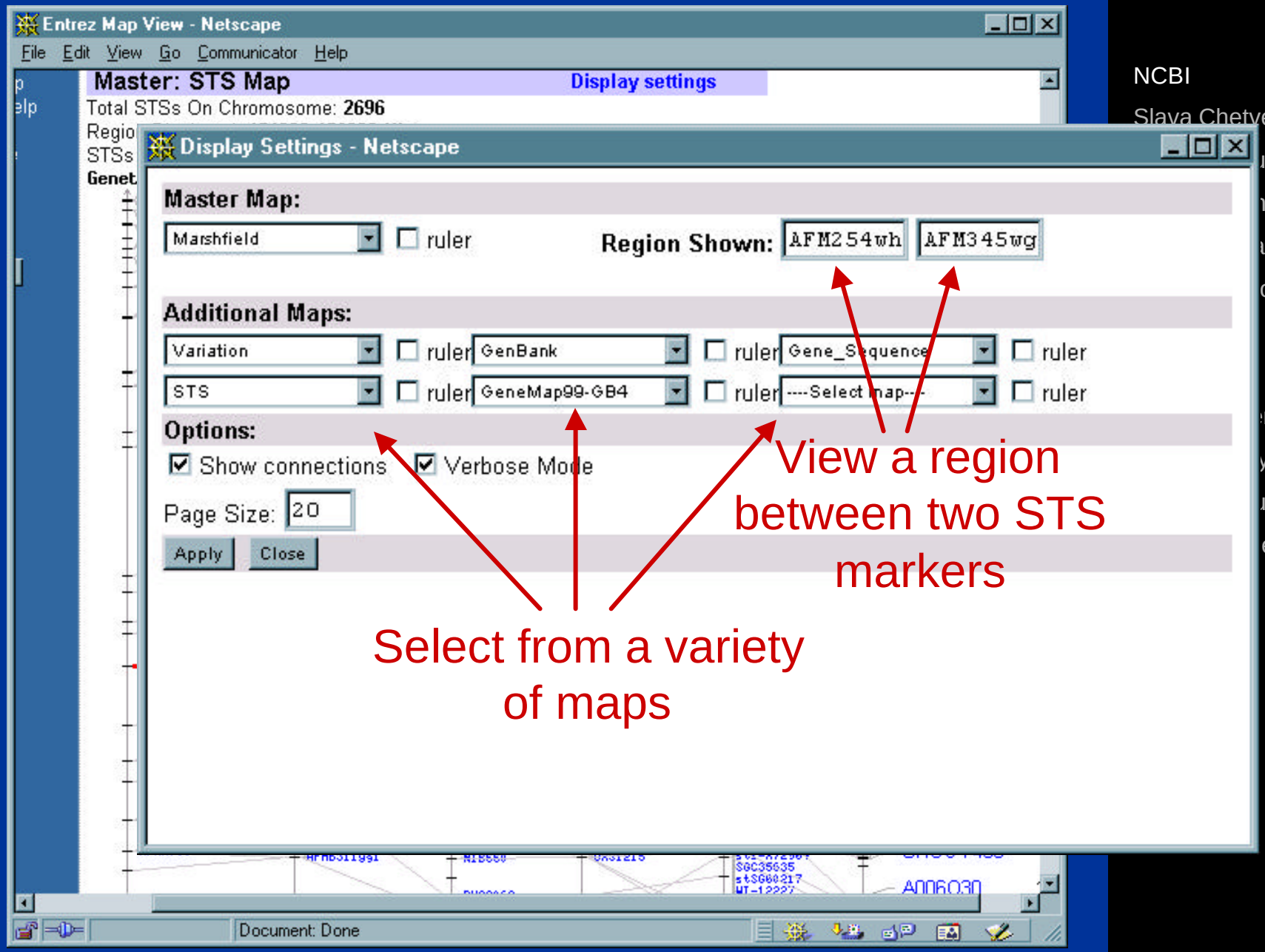
Schuler, Genome Res 7, 541-550 (1997).

# Electronic PCR (e-PCR)

NCBI
Wonhee Jang
Greg Schuler

| | | Hit sequence | | Repetitive | |
|---|---|---|---|---|---|
| *Map* | *STSs* | *STSs* | *(%)* | *STSs* | *(%)* |
| Généthon | 5264 | 4309 | 81.9% | 186 | 3.5% |
| Marshfield | 7667 | 6011 | 78.4% | 278 | 3.6% |
| Whitehead | 10451 | 8379 | 80.2% | 446 | 4.3% |
| GeneMap99 | 44328 | 34204 | 77.2% | 1997 | 4.5% |
| Stanford G3 | 7154 | 5851 | 81.8% | 141 | 2.0% |
| Overall | 60405 | 46831 | 77.5% | 2668 | 4.4% |

# Search by position




NCBI

Slava Chetvernin

urch

ng

ans

ott

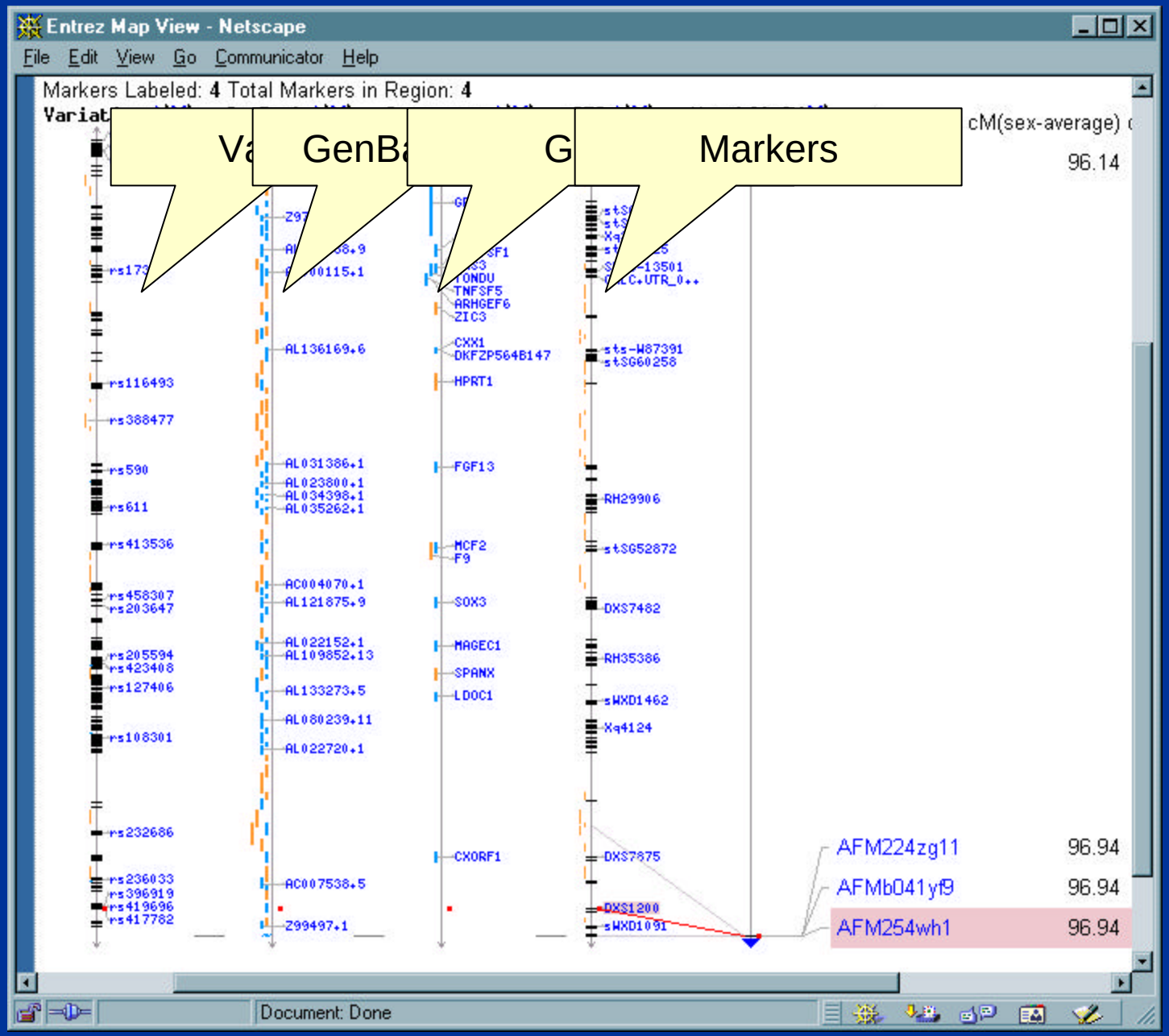
enchuk

y

usov

erg

# Search by position




NCBI

Slava Chetvernin

Deanna Church

Wonhee Jang

Jonathan Kans

Donna Maglott

Jim Ostell

Kim Pruitt

Sergei Resenchuk

Steve Sherry

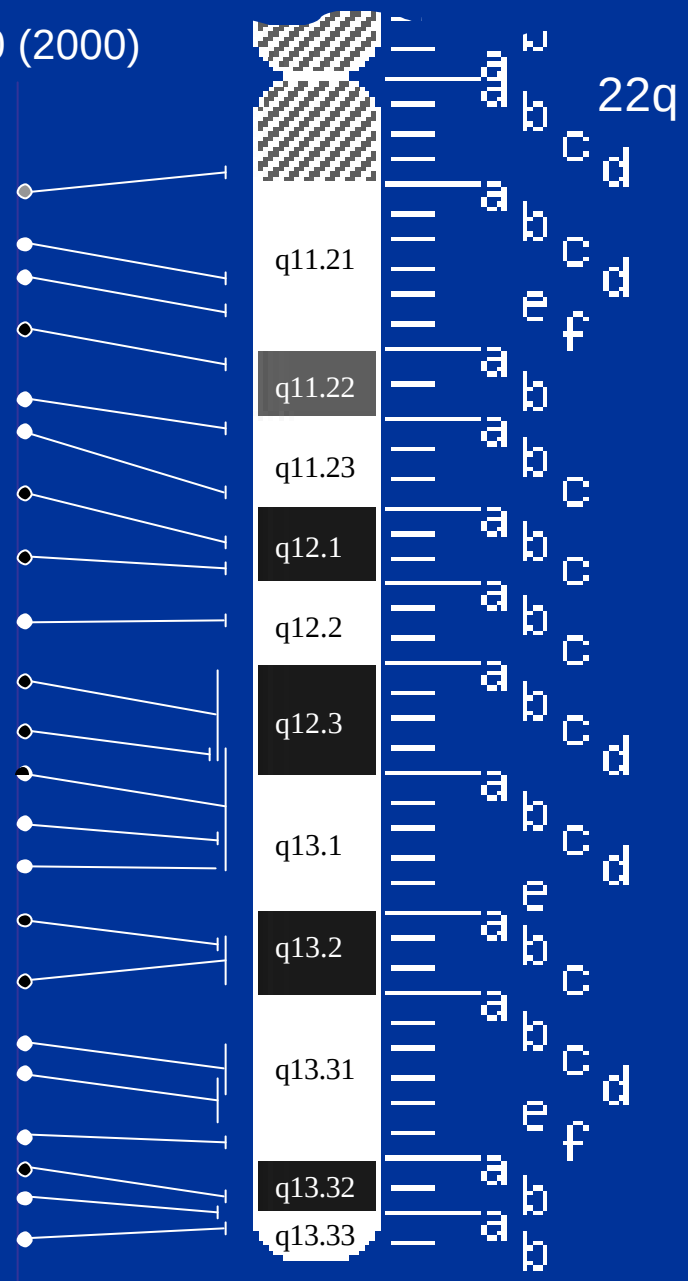
Tatiana Tatusov

Tyra Wolfsberg

# Sequence to cytogenetics

Kirsch, et al., Nat Genet. 24, 339-340 (2000)

| FISH | BAC |
|---|---|
| 22q11.1(d) | CTA-115F6 |
| 22q11.21(d) | CTA-154H4 |
| 22q11.21(e) | CTA-433F6 |
| 22q11.22(a) | CTA-526G4 |
| 22q11.23(a) | CTA-322B1 |
| 22q11.23(c) | CTA-221G9 |
| 22q12.1(b) | CTA-922D9 |
| 22q12.1(c) | CTA-57G9 |
| 22q12.2(b) | CTA-99F11 |
| 22q12.3(a-d) | CTA-514G2 |
| 22q12.3(d) | CTA-221H1 |
| 22q12.3(d)-13.1(d) | CTA-212A2 |
| 22q13.1(c) | CTA-390B3 |
| 22q13.1(d) | CTA-150C2 |
| 22q13.2(b) | CTA-229A8 |
| 22q13.2(b-c) | CTA-250D10 |
| 22q13.31(c-d) | CTA-397C4 |
| 22q13.31(d-e) | CTA-268H5 |
| 22q13.31(f) | CTB-1109B5 |
| 22q13.32(b) | CTA-299D3 |
| 22q13.33(a) | CTA-722E9 |
| 22q13.33(b) | CTA-799F10 |




NCBI

Wonhee Jang

Alex Lash

Karl Sirotkin

Greg Schuler

NCI

Raluca Yonescu

Thomas Ried

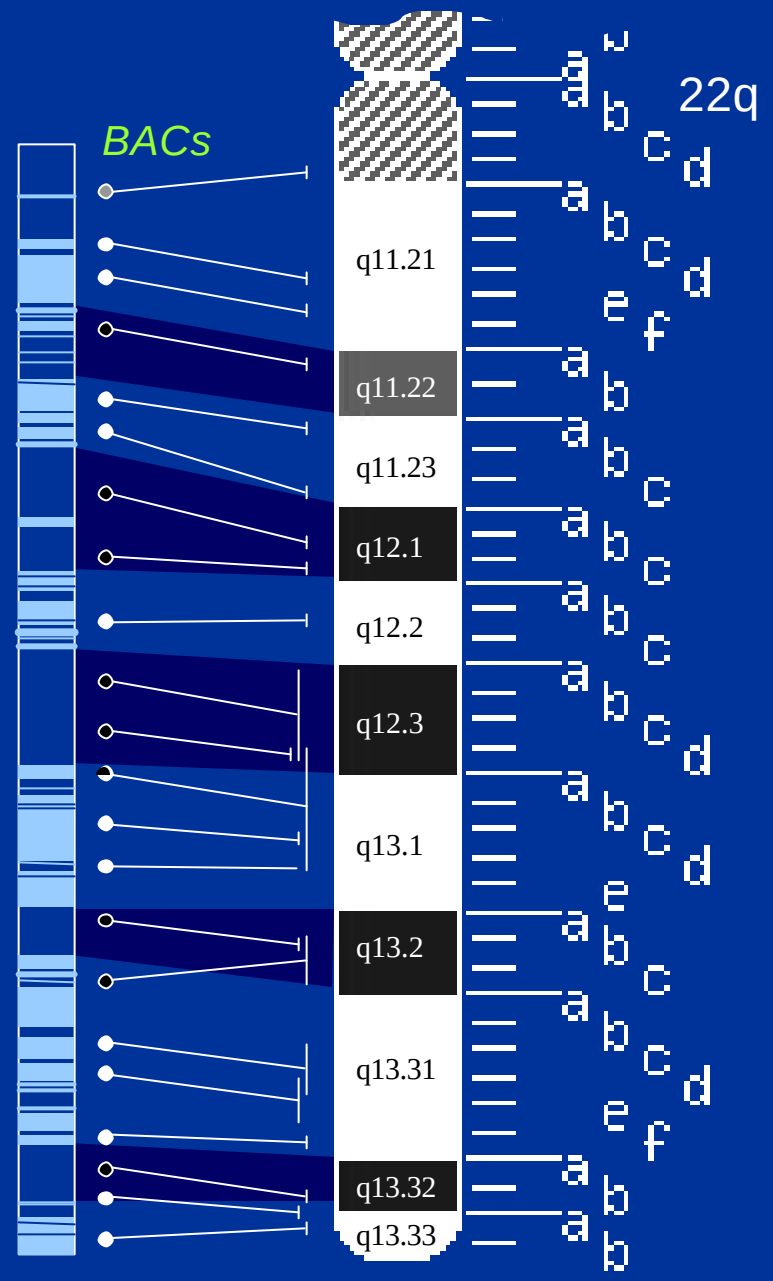
Lonny Kirsch

# Sequence to cytogenetics

Finished sequence of chromosome 22 shaded by G+C content

G+C content

- >45%
- <45%




NCBI

Wonhee Jang

Alex Lash

Karl Sirotkin

Greg Schuler

NCI

Raluca Yonescu

Thomas Ried

Lonny Kirsch

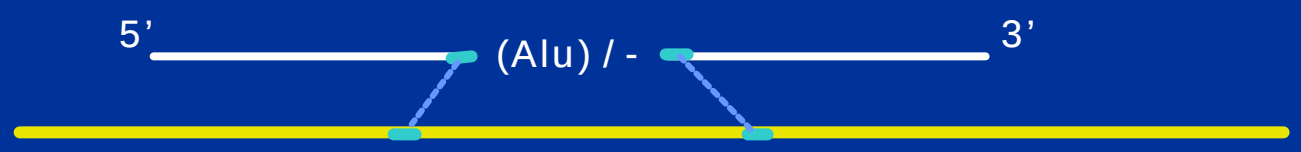
# *Mapping SNPs to Contigs*

Heuristic Rules:

1. HI: 95% flanking sequence, 0-6 mismatch

2. LO: 75% flanking sequence, 3% mismatch

3. Immediately flanking 5’ & 3’ bases must be returned in the alignment




4. Variations with multiple sequence locations are reblasted at higher stringency

NCBI
Steve Sherry
Minghong Ward
Jon Baker

Entrez Map View - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address .00%5D-r%2Cloc%5B13000000.00%3A20000000.00%5D%2Csnp%5B13M%3A20M%5D&BEG=16M&END=18M Go

HELP
FTP
Display settings

Region Shown: 16M 18M Go

out zoom in

SNP

Total SNPs On Chromosome: **685**
Region Displayed: **16000-18000 Kbp**
SNPs Shown: **20** Total SNPs in Region: **35**

STS: RH11351, stSG4778, WI-326, RH80191, stSG29584, RH118685, D22S943, sts-T79756, WI-15274, RH27791, SHGC-14358, stSG60090, stSG40938, A008R36, D22S1036, RH27805, stSG46882, D22S1138, RH27825, WI-178

Genes: GSCL, HIRA, HIRA, CDC45L, TBX1, ARVCF, BZRP, BZRP, GALR3, PVALB, RFPL3, SF3A1, KIAA0330, PK1+3, KIAA0330, KDELR3, SREBF2, KIAA0330, KIAA0330, KIAA0330

SNP: rs11744, rs10314, rs13164, rs1878, rs6268, rs6270, rs4633, rs12814, rs9265, rs23572, rs8443, rs4020, rs9997, rs22399, rs22300, rs5905, rs5906, rs5907, rs5904, rs10710

Map Gene Het (0 – 100%) Validation (>80 >90 >95%) Genotypes Avail Linkout Avail

Map Counts: CH: 1, NT: 1, TOT: 1

Local intranet

# *Functional Analysis of SNPs*

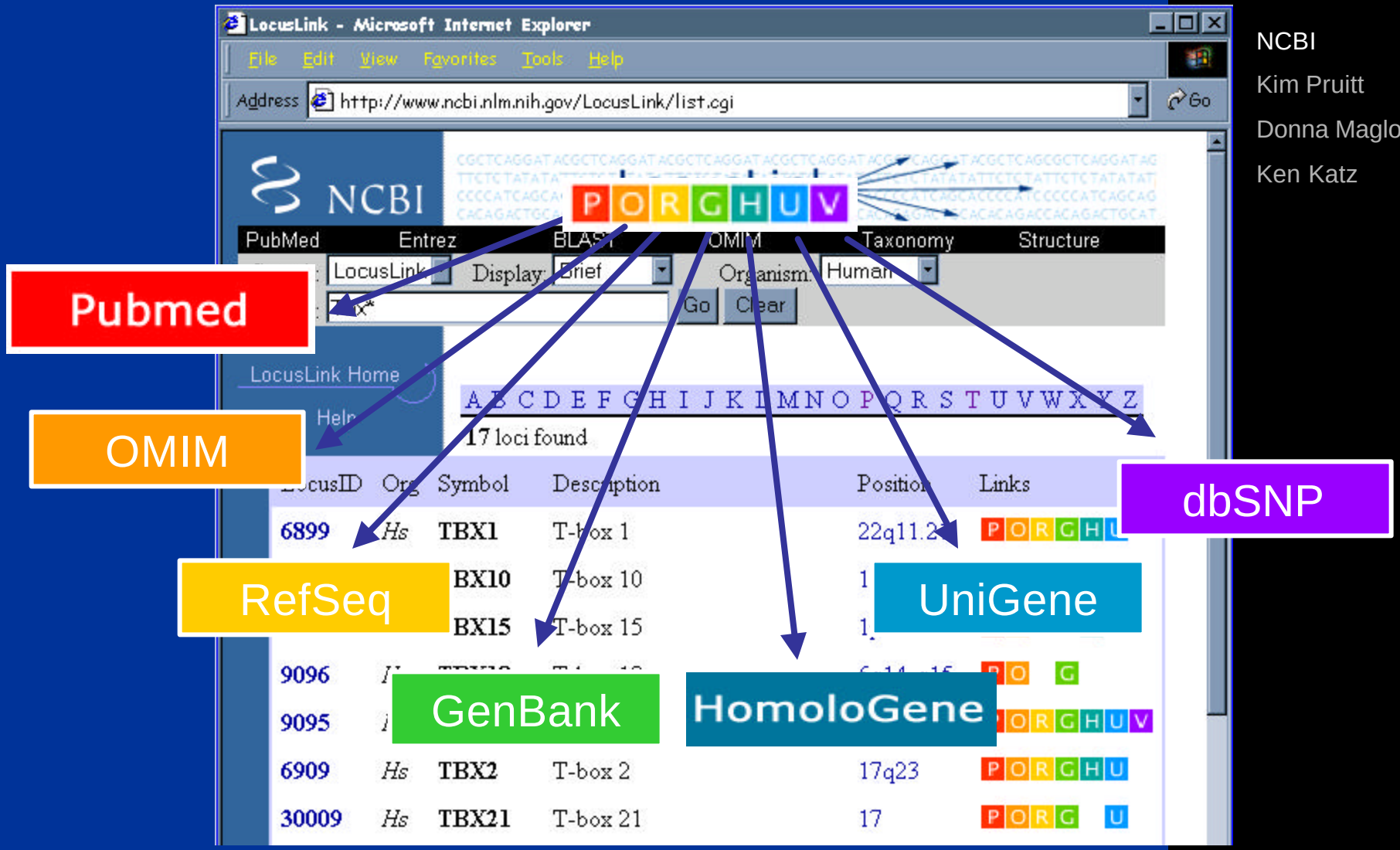
Based on alignment of 7,000 RefSeq mRNAs to human genome sequence

| Gene feature | # SNPs / 10 kb sequence |
|---|---|
| Overall | 6.67 |
| Intron | 6.75 |
| 3’ UTR | 5.00 |
| 5’ UTR | 4.95 |
| exon, overall | 4.20 |
| exon, coding* | 3.77 |

*56% synonymous
44% non-synonymous
relative rates by codon position: 1.2 : 1 : 3

NCBI
Steve Sherry
Gabor Marth
Deanna Church

# LocusLink





NCBI

Kim Pruitt

Donna Maglott

Ken Katz

# LocusLink

LocusLink Report - Microsoft Internet Explorer

File Edit View Favorites Tools Help

NCBI LocusLink

PubMed Entrez BLAST OMIM Taxonomy Structure

Search LocusLink Display Brief Organism: All

Query: Go Clear

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

LocusLink Home
- Collaborators
- Download
- FAQ
- Help
- Statistics

NCBI Genome Guides
- Fruit Fly
- Human
- Mouse
- Rat
- Zebrafish

RefSeq
- About
- Download
- FAQ
- Statistics

Related Resources
- BDGP
- CGAP

PUB UG VAR HOMOL GDB HGMD

***Homo sapiens* Official Gene Symbol and Name (HGNC)**

**A2M: alpha-2-macroglobulin**

**Locus Information** **Submit GeneRIF for A2M**

| | |
|---|---|
| **LocusID:** | 2 |
| **Type:** | gene with protein product, function known or inferred |
| **Product:** | alpha 2 macroglobulin precursor |
| **UniGene:** | Hs.74561 |
| **OMIM:** | 103950 |
| **Phenotype:** | Alzheimer disease, susceptibility to<br>Emphysema due to alpha-2-macroglobulin deficiency |

**Summary:** Alpha-2-macroglobulin is a protease inhibitor and cytokine transporter. It inhibits many proteases, including trypsin, thrombin and collagenase. A2M is implicated in Alzheimer disease (AD) due to its ability to mediate the clearance and degradation of A-beta, the major component of beta-amyloid deposits.

from Nomenclature committee

Locus Information, Phenotype, UniGene

# LocusLink

FAQ
Statistics

Related Resources
BDGP
CGAP
FlyBase
GDB
GeneMap99
HomoloGene
Map Viewer
--fruit fly
--human
--mouse
MGD
Nomenclature
OMIM
RATMAP
RGD
UniGene
ZFIN

**Summary:** Alpha-2-macroglobulin is a protease inhibitor and cytokine transporter. It inhibits many proteases, including trypsin, thrombin and collagenase. A2M is implicated in Alzheimer disease (AD) due to its ability to mediate the clearance and degradation of A-beta, the major component of beta-amyloid deposits.

## Map Information

NEW **Entrez Map Viewer: View Genomic Context**

| | | | |
|---|---|---|---|
| **Chromosome:** | 12 | | |
| **Cytogenetic:** | 12p13.3-p12.3 | HUGO | |
| **Mouse Homology Map:** | 6 62.0 cM | A2m | Hs Mm |
| **Markers(ePCR):** | Chr. 12 | stSG1293 | |
| | Chr. 12 | SGC31674 | |
| | Chr. 12 | stSG1290R | |
| | Chr. 12 | WIAF-1430-STS | |

**Key:** Link to Map Viewer

## *Homo sapiens* A2M Reference Sequence (RefSeq)

**Category:** REVIEWED

| | | |
|---|---|---|
| **1. Nucleotide:** | NM_000014 | |
| **Protein:** | NP_000005 | alpha 2 macroglobulin precursor BL |
| **Domains:** | Alpha-2-macroglobulin family | 2375 |
| | Alpha-2-macroglobulin family N-terminal region | 1754 |
| **GenBank Source:** | M11313 | |

RefSeq review

MapViewer

Comparative Map

ePCR results

RefSeq Info

BLink

# *Acknowledgments*

Richa Agarwala
Hsiu-Chuan Chen
Slava Chetvernin
Deanna Church
Renata Geer
Wonhee Jang
Jonathan Kans
Ken Katz
Paul Kitts
David Lipman
Adam Lowe
Donna Maglott
Danielle Thierry-Mieg
Jean Thierry-Mieg
Jim Ostell
Kim Pruitt
Greg Schuler
Hugues Sicotte
Steve Sherry
Andrei Shkeda
Serge Resenchuk
Tatiana Tatusova
Lukas Wagner
Sarah Wheelan
Tyra Wolfsbefg

RefSeq Staff
BLAST Team
Entrez Team
Omim Staff
The Jackson Laboratory
Rat Genome Database
Nomenclature committee

**http://www.ncbi.nlm.nih.gov/genome/guide/**