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

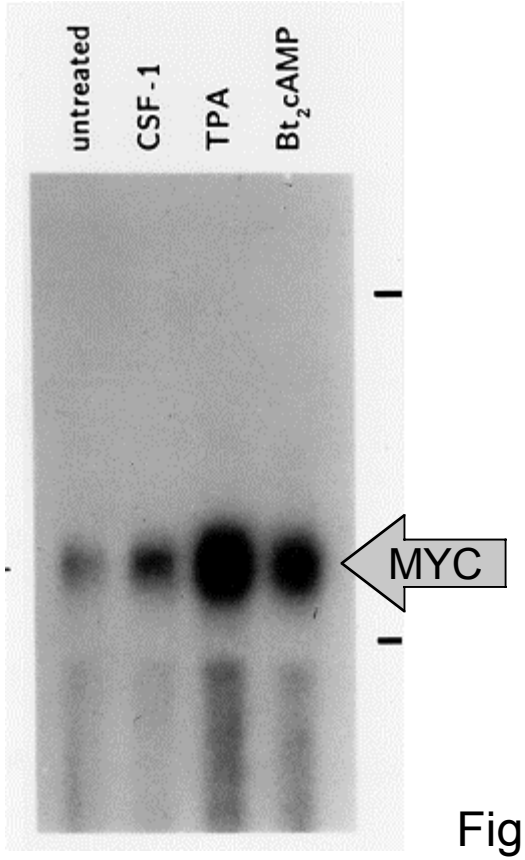


Figure 7

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

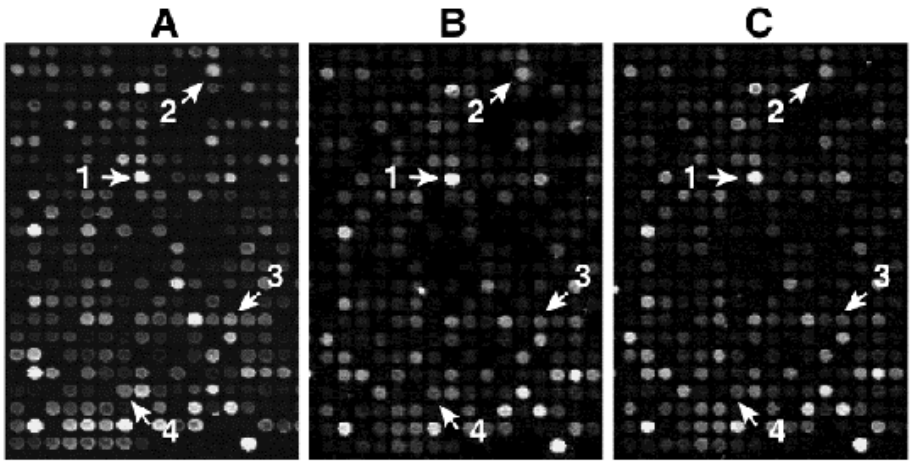


Figure 1

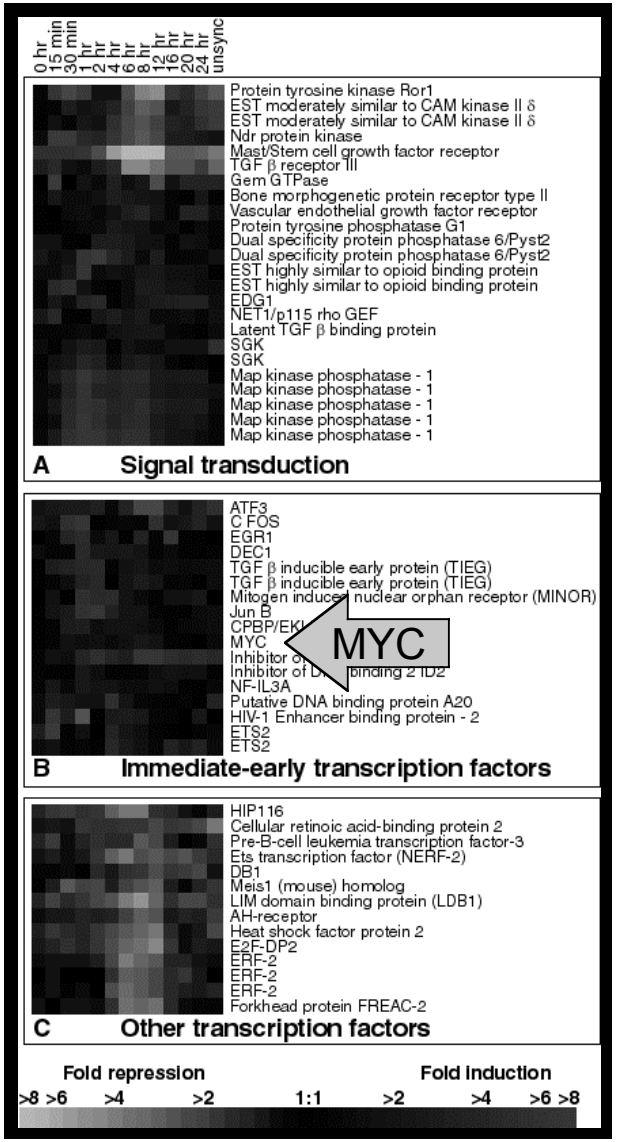
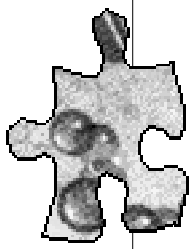
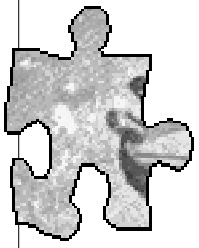
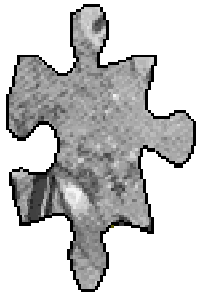


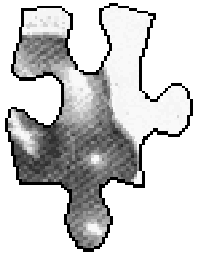
Figure 4

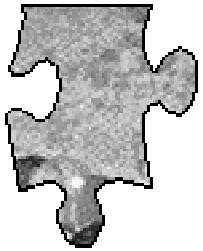
Pieces of the Puzzle

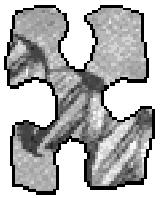












Sequencing the Genome

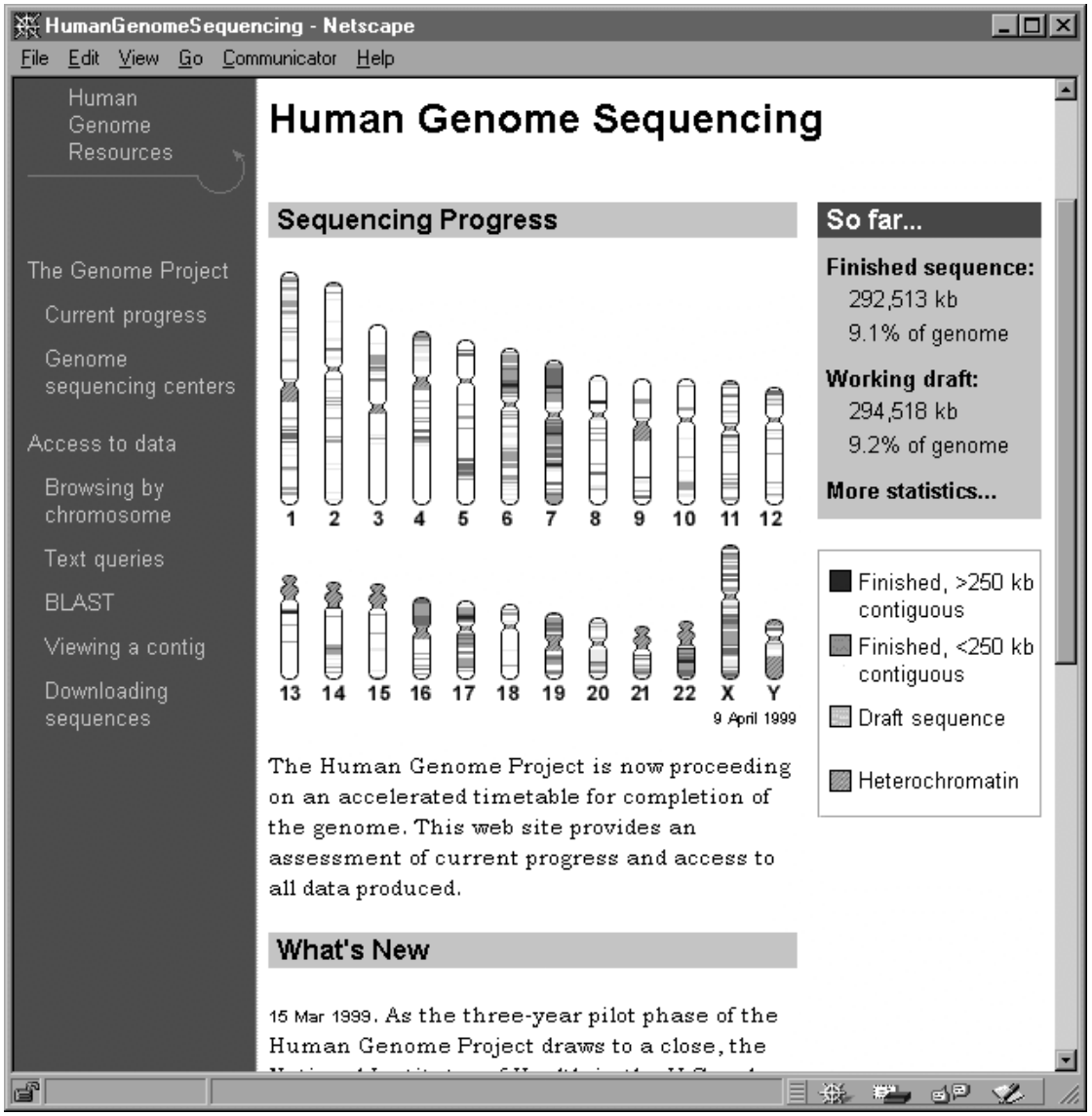
Approx. 3,200,000 kb of DNA, to be finished in 2003

Working draft sequence by 2000

Estimated 65,000 to 80,000 genes

About 100,000 kb of coding DNA (3% of the genome)

97% “junk DNA”



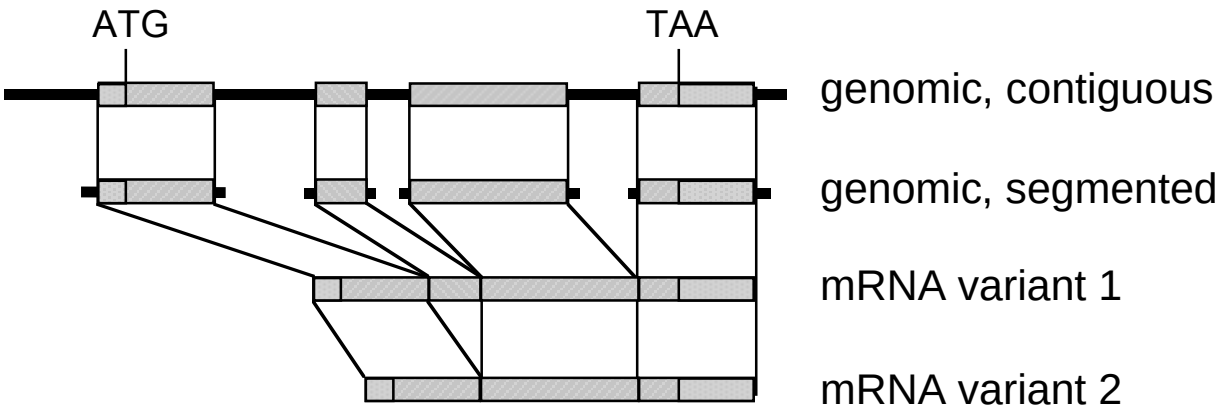
www.nbci.nlm.nih.gov/genome/seq/

One Gene, Many Sequences

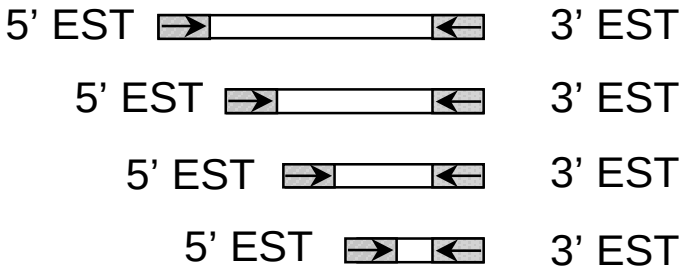
GenBank is an archive of published sequences

May be many representatives of a given gene

Characterized Genes

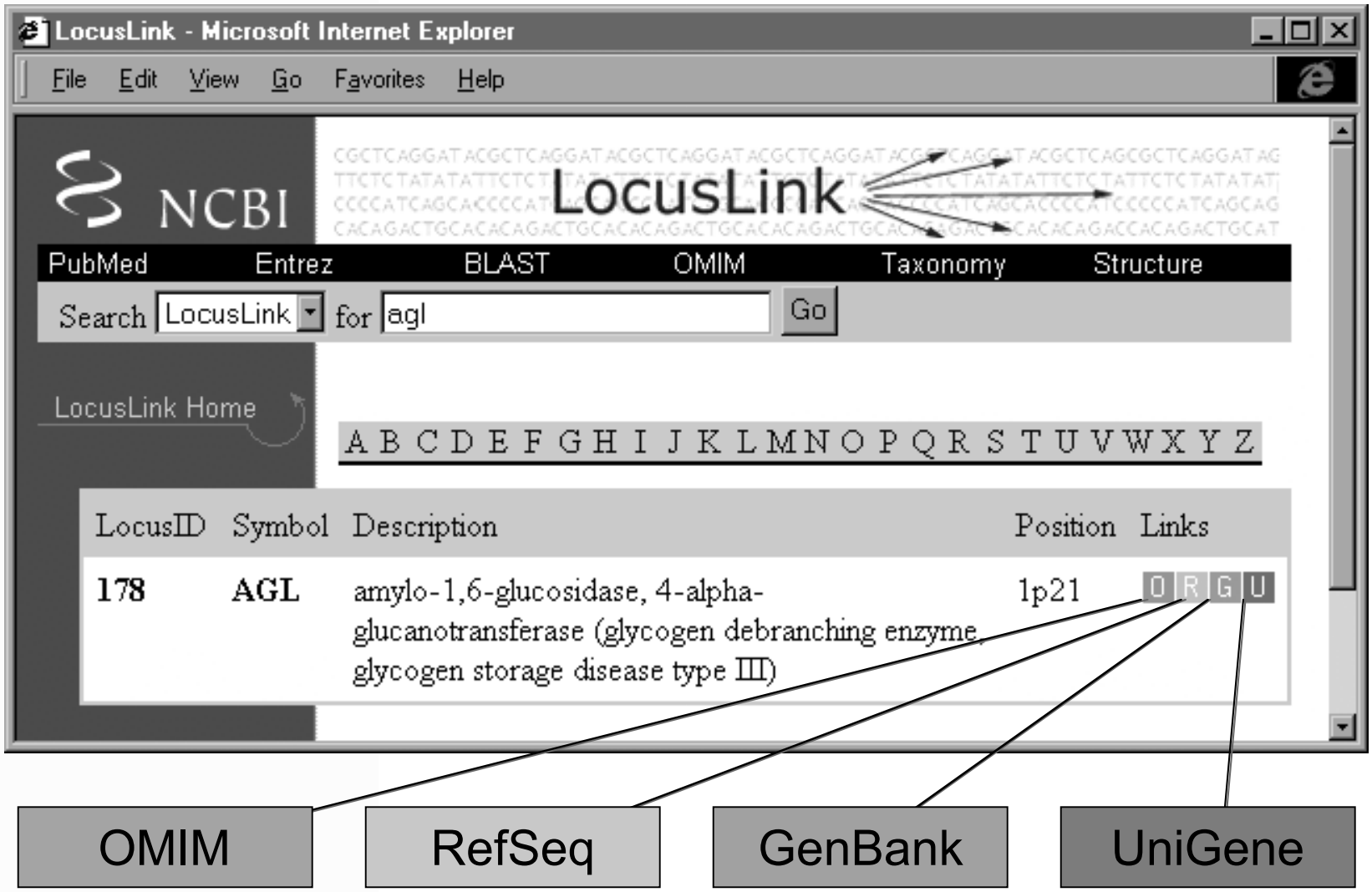


Expressed Sequence Tags



LocusLink

Example: AGL amylo-1,6-glucosidase,4-alpha-glucanotransferase



RefSeq

Homo sapiens

Official Gene Symbol and Name

AGL: **amylo-1,6-glucosidase, 4-alpha-glucanotransferase (glycogen debranching enzyme, glycogen storage disease type III)**

Locus Information

Locus ID: 178
Alternate Symbols: GDE;
Product:
amylo-1,6-glucosidase,
4-alpha-glucanotransferase
Alias: glycogen debranching enzyme;
EC number: 2.4.1.25
EC number: 3.2.1.33
Chromosome: 1
Position: 1p21
OMIM: 232400
UniGene: Hs.904
Phenotype:
Glycogen storage disease IIIa
Glycogen storage disease IIIb

Summary: Glycogen debranching enzyme is involved in glycogen degradation and has two independent catalytic activities: a
4-alpha-glu
(EC 2.4.1.2
amylo-1,6-g
(EC 3.4.1.3
occur at dif
single polyp
Mutations i
glycogen st
wide range
enzymatic v
glycogen d
some of wh
tissue-spec

Homo sapiens AGL Reference sequence

Status: REVIEWED

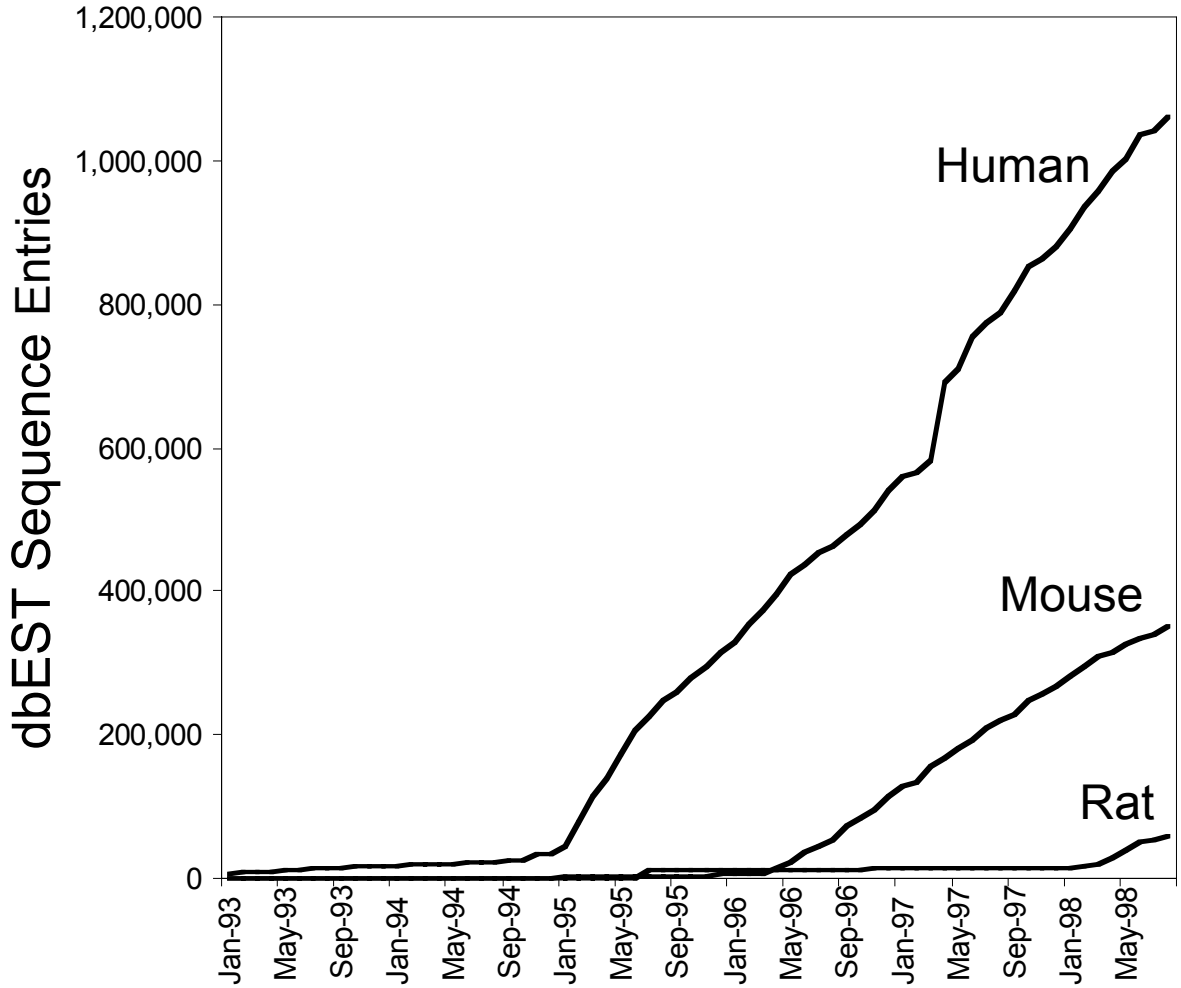
	Source	mRNA	Protein	
AGLa	U84007	NM_000642	NP_000633	amylo-1,6-glucosidase, 4-alpha-glucanotransferase
AGLb	U84008	NM_000644	NP_000635	amylo-1,6-glucosidase, 4-alpha-glucanotransferase
AGLc	U84009	NM_000643	NP_000634	amylo-1,6-glucosidase, 4-alpha-glucanotransferase
AGLd	U84010	NM_000028	NP_000019	amylo-1,6-glucosidase, 4-alpha-glucanotransferase
AGLe	M85168	NM_000645	NP_000636	amylo-1,6-glucosidase, 4-alpha-glucanotransferase
AGLf	U84011	NM_000646	NP_000637	amylo-1,6-glucosidase, 4-alpha-glucanotransferase

Expressed Sequence Tags (ESTs)

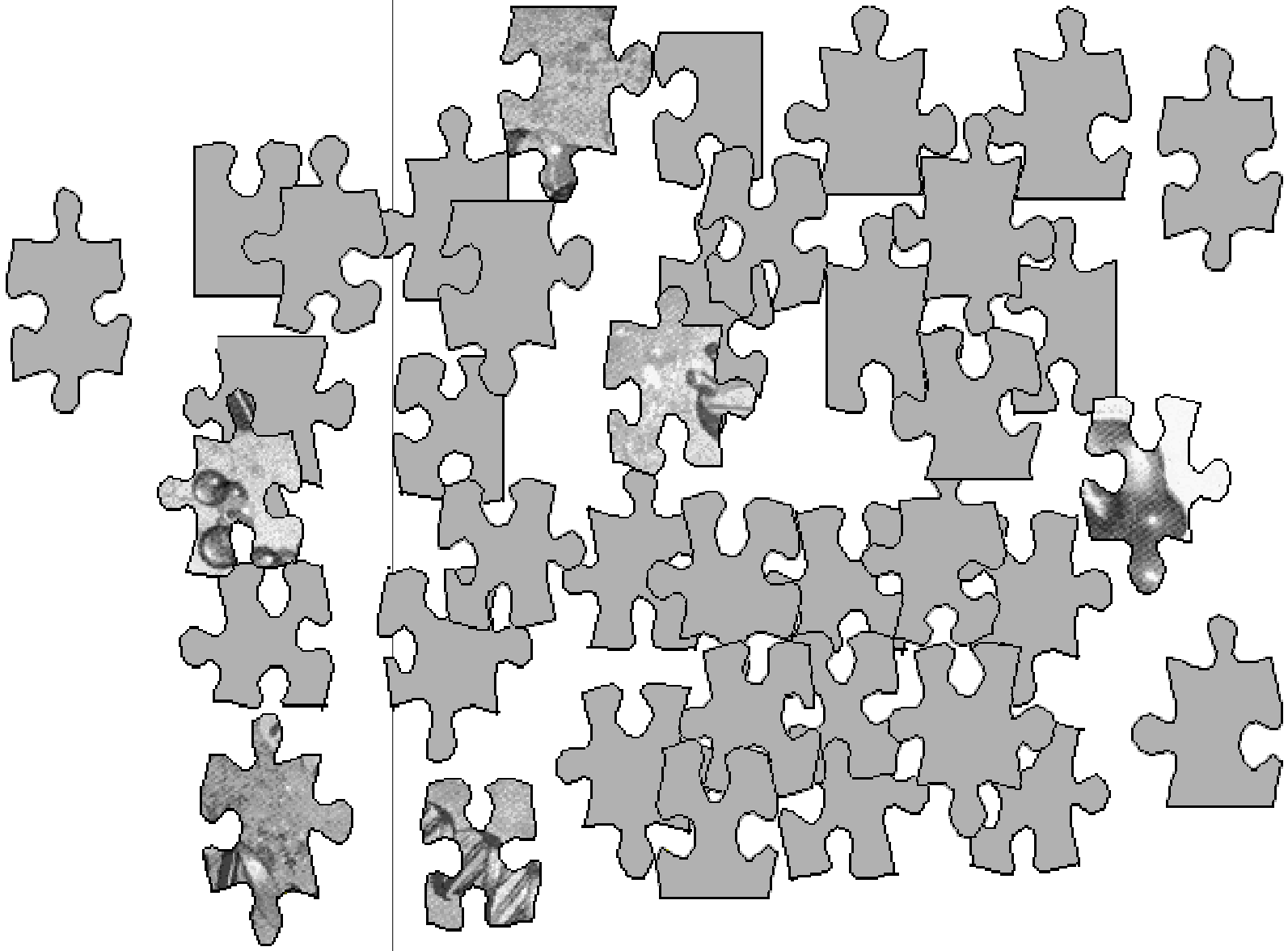
Clones chosen at random from many different cDNA libraries

High-throughput single-pass sequencing

Now over two million ESTs



An Abundance of Puzzle Pieces

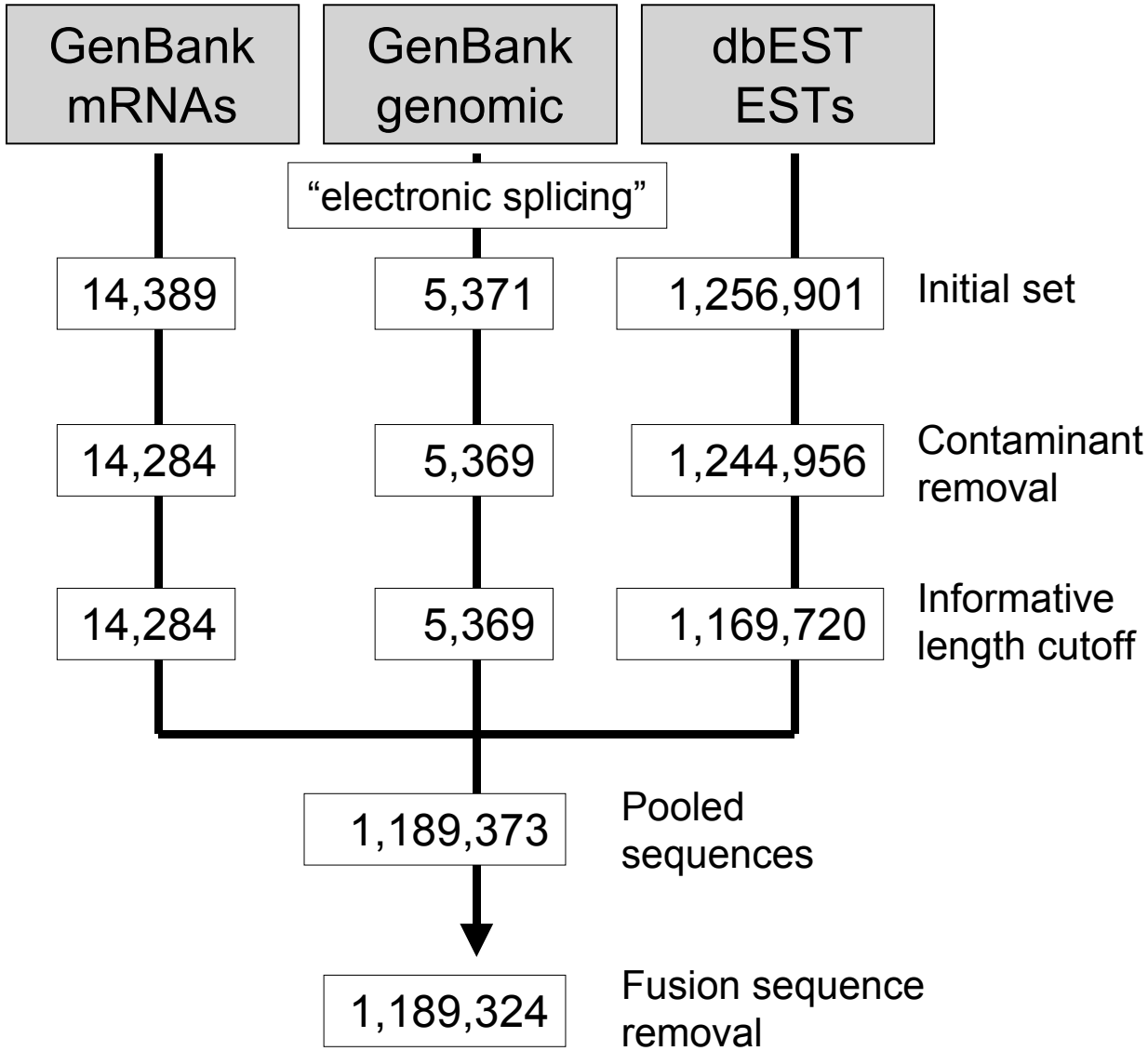


Selection of Sequences for UniGene

Screened for E.coli, yeast, rRNA and mitochondrial

Screened for vector, repetitive elements and simple repeats

Manually maintained exclusion list for fusion sequences

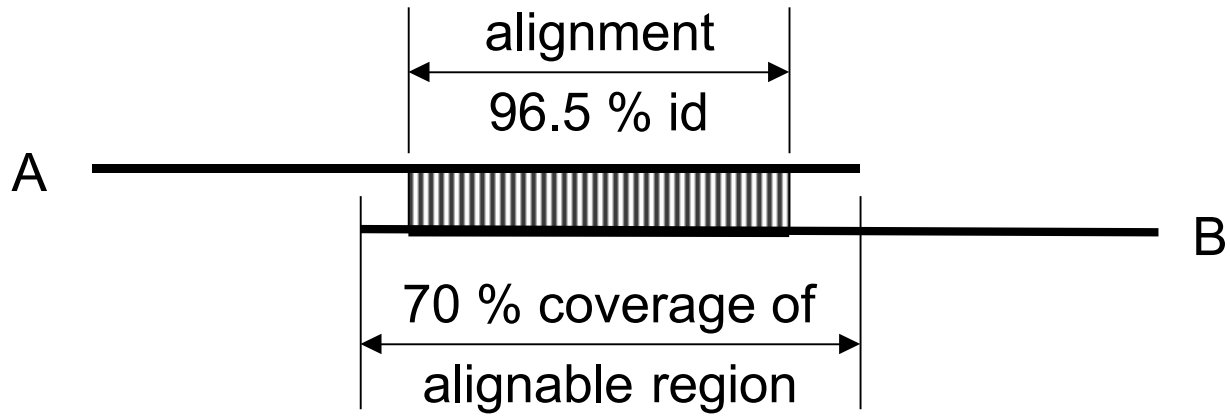


Sequence Similarity Relationships

Sequence alignments using MegaBLAST (Zhang, Schwartz, and Miller, unpublished)

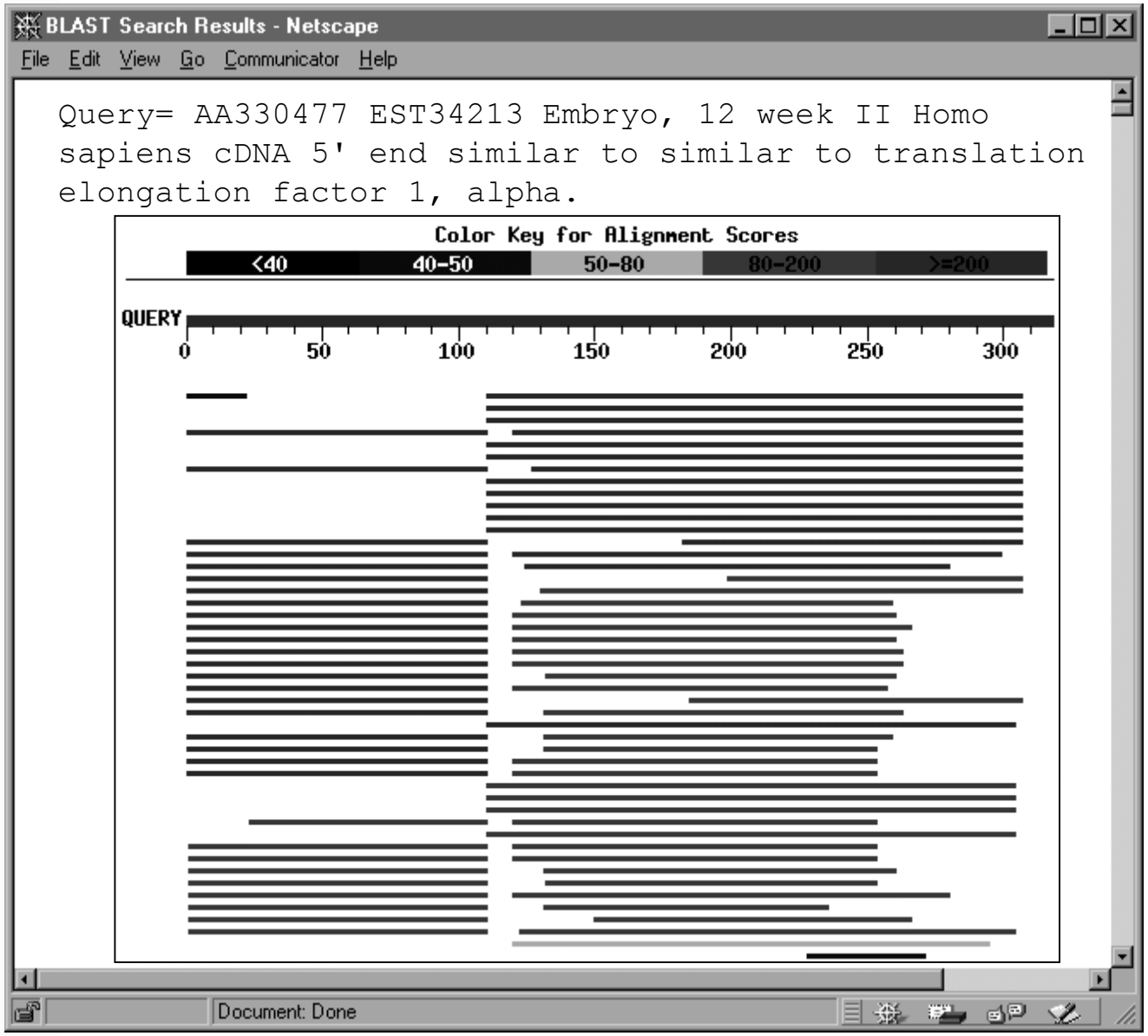
Constraints on alignment quality and coverage

Coverage requirement reduces problems caused by chimeric sequences



Chimeric Sequences

EST that matches both hemoglobin α and elongation factor 1α

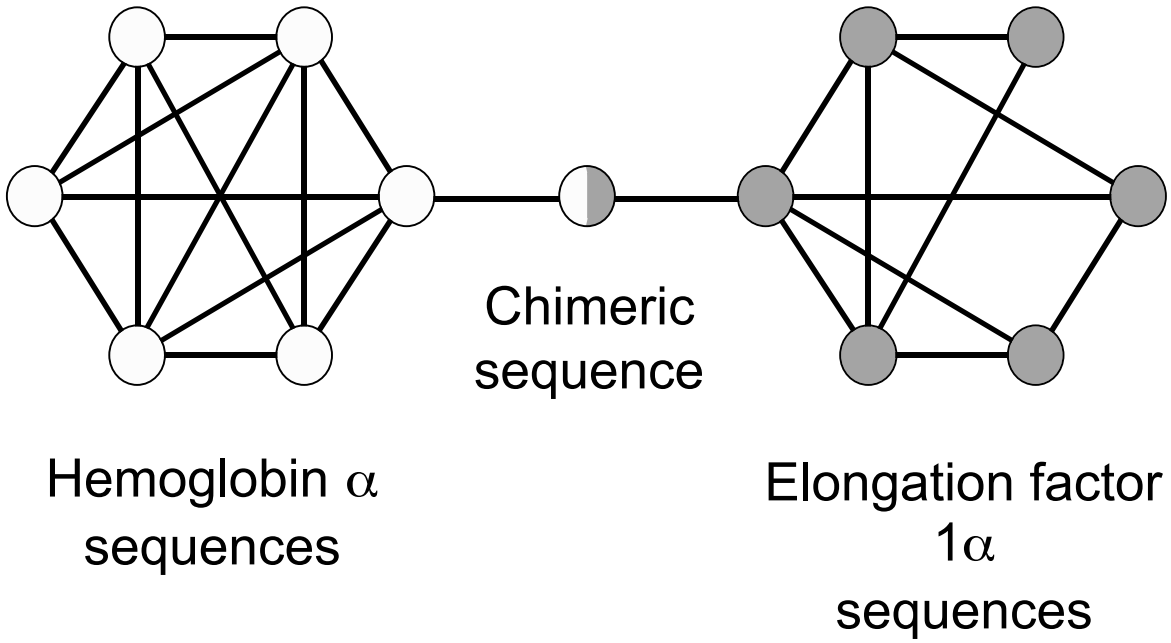


Effect of Chimeric Sequences

One bad sequence can corrupt two good clusters

Some chimeras can be found by graph analysis

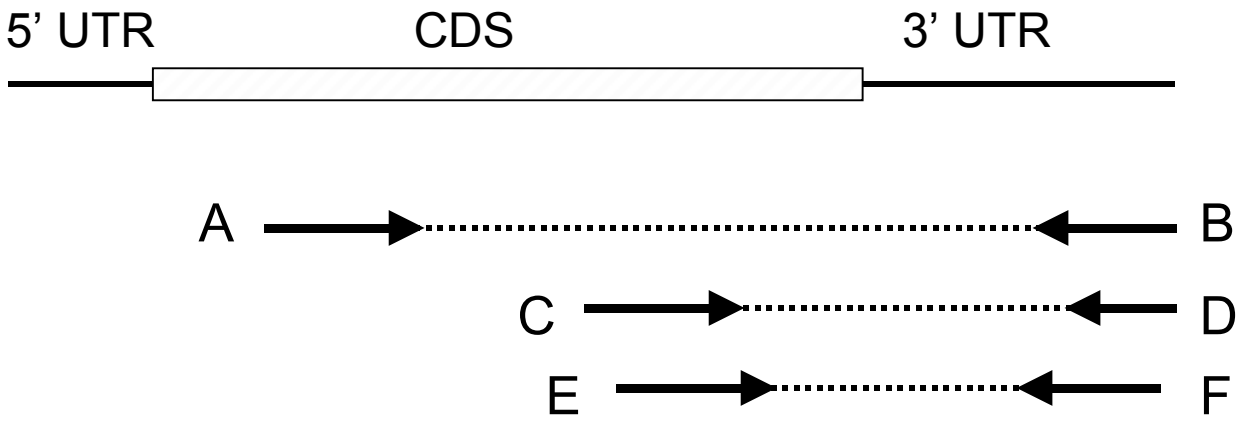
For known genes, use reference sequences to force apart

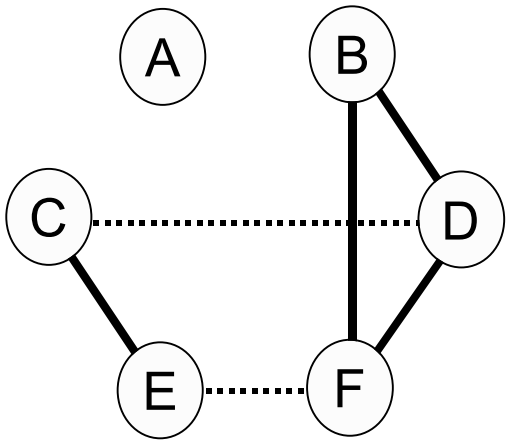


Clone Relationships

ESTs from opposite ends of a clone often do not overlap

Clone IDs can bring ends together only if at least 2 clones agree

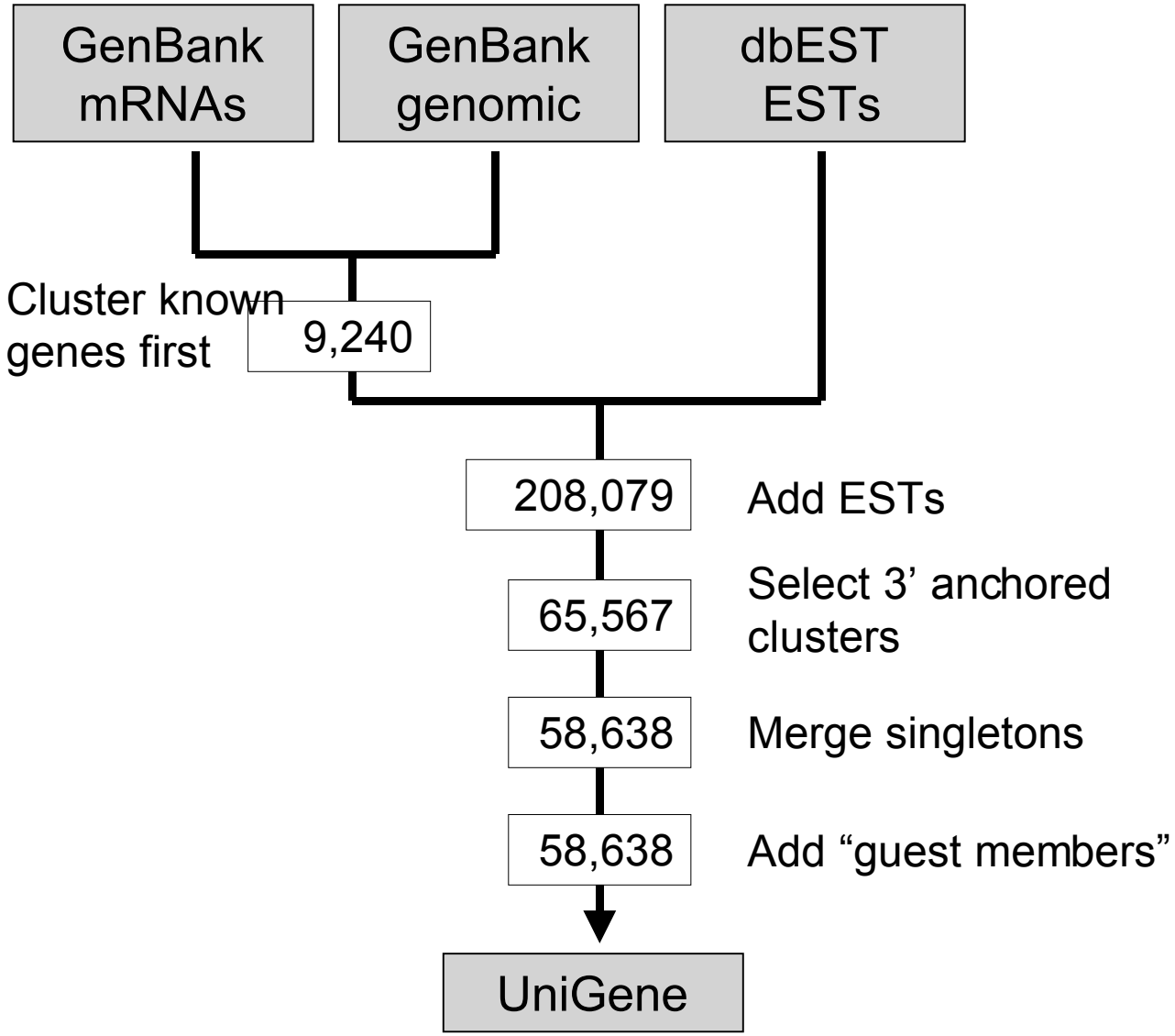




Multi-step Clustering

Multi-step clustering prevents bad ESTs from corrupting good mRNA/gene clusters

3’ ends desired to avoid multiple clusters per gene

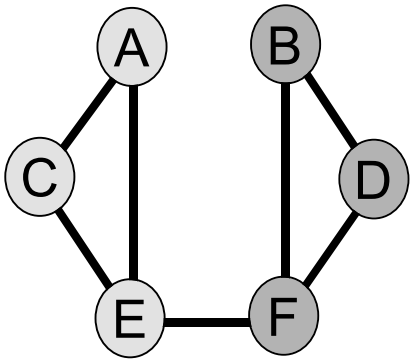


Lumping and Splitting

Estimates of lumping and splitting estimated from partitioning of STSs or locus IDs

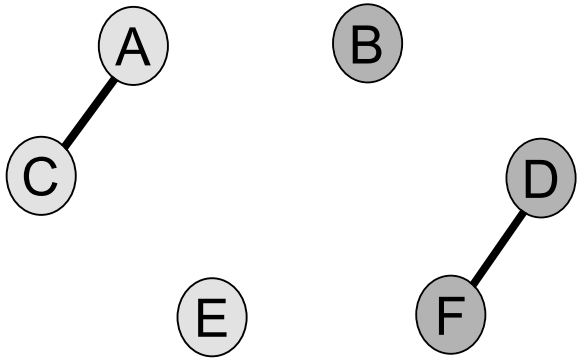
Lumping:
More than one gene in a cluster

2.4% lumping



Splitting:
Multiple clusters for one gene

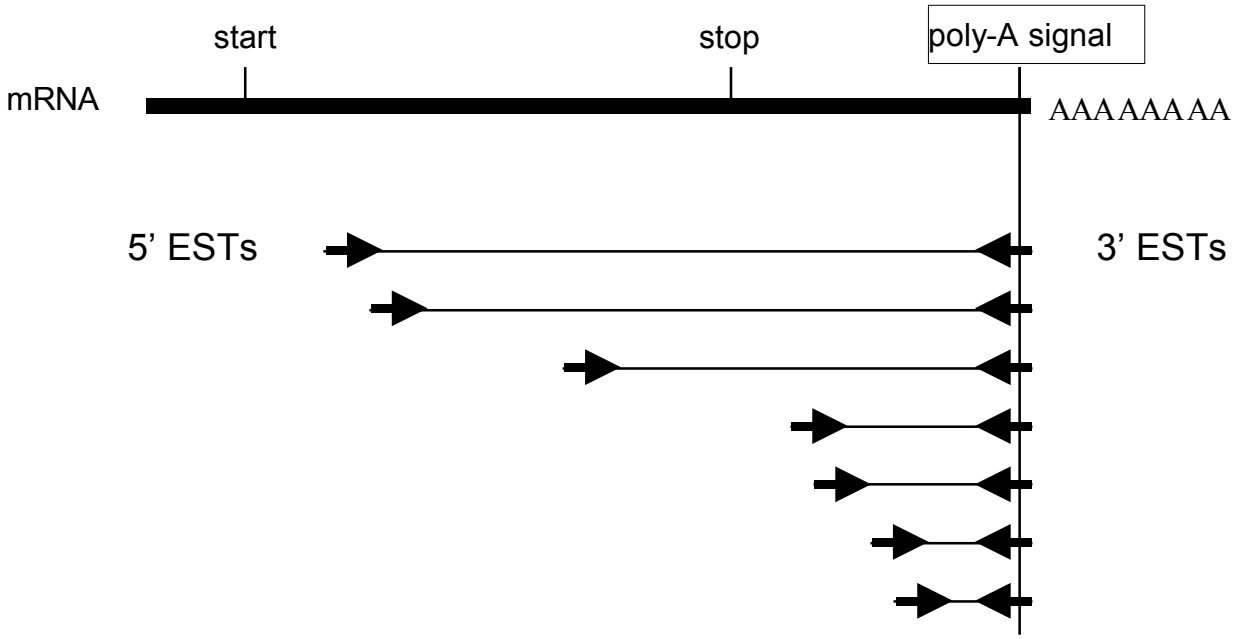
1.2% splitting



3’ Anchoring

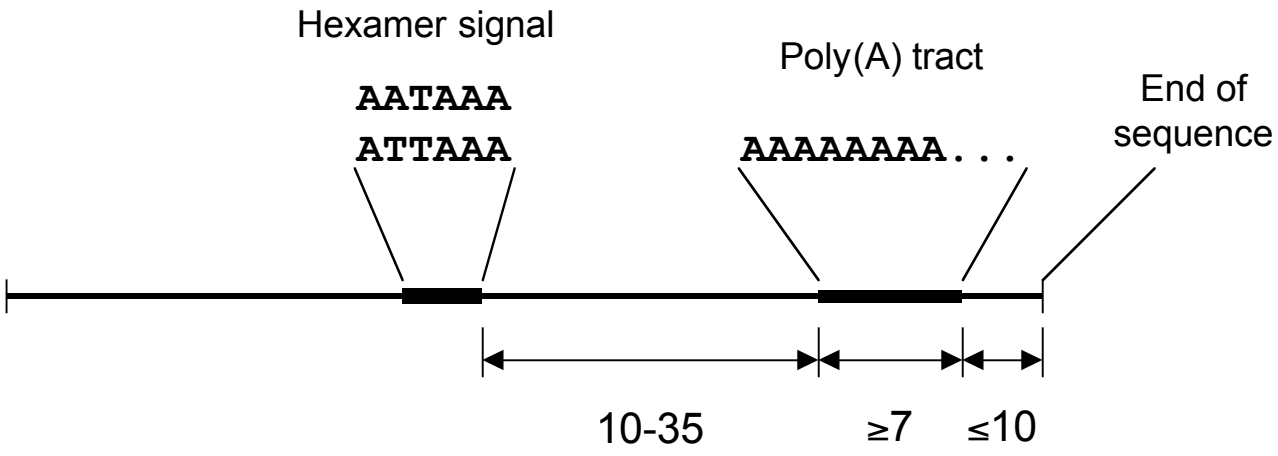
Requiring 3’ end reduces splitting

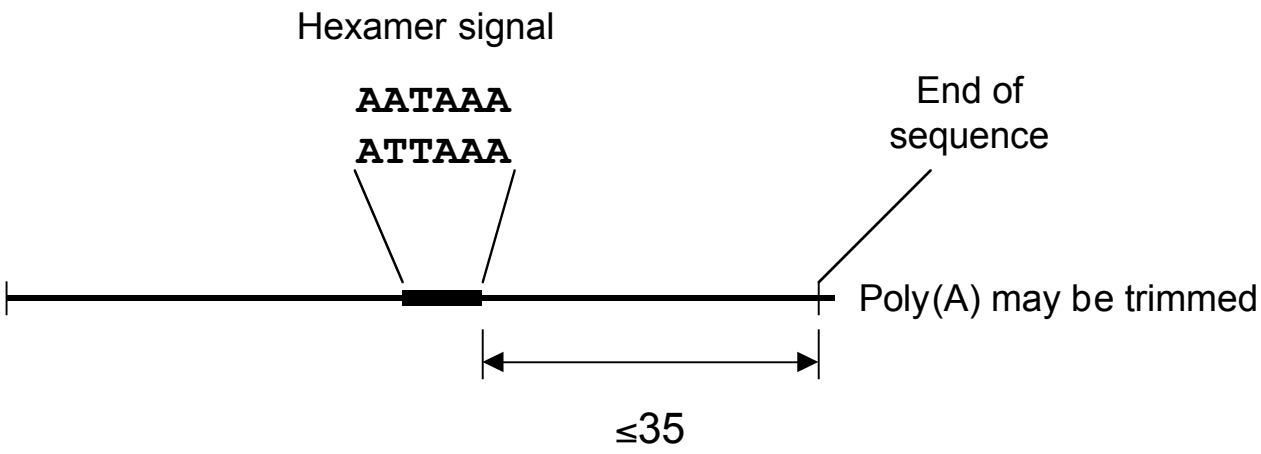
Over 90% of libraries are oligo-dT primed



Poly(A) Signal Detection

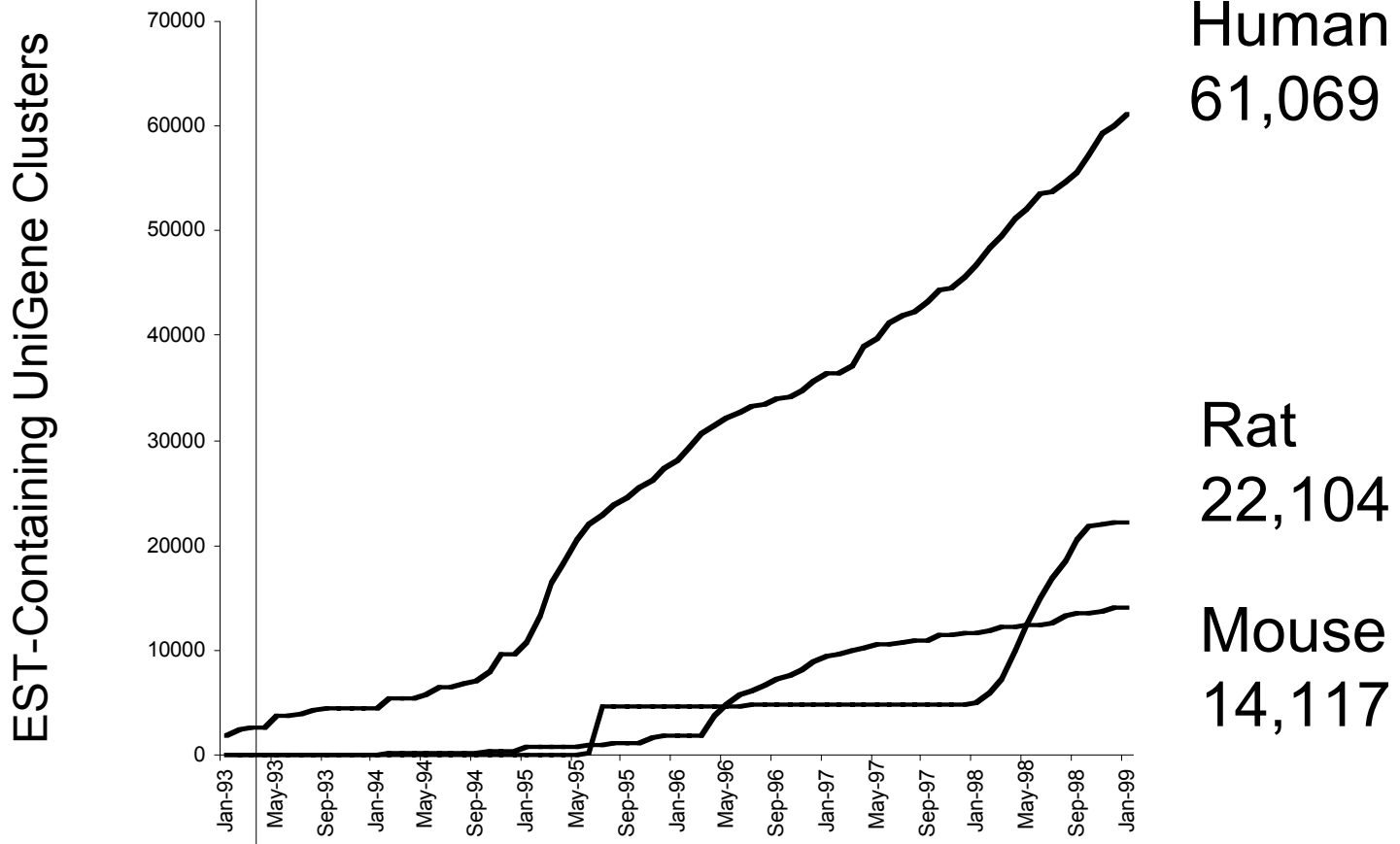
Algorithm is based on published data but tolerates over and under trimming





Wahle & Keller (1992) Annu Rev Biochem 61, 419-440

Gene Discovery Rate



UniGene Query - Netscape

File Edit View Go Communicator Help

Hs UniGene NCBI ?

Search for: prostate specific antigen

4 records satisfy the query

UniGene	Description	Symbol
Hs.1548	Prostate specific antigen	APS
Hs.1915	PROSTATE-SPECIFIC MEMBRANE ANTIGEN	
Hs.74647	T cell receptor alpha-chain	TCRA
Hs.76294	CD63 antigen (melanoma 1 antigen)	CD63

Questions or comments? Write to the NCBI Help Desk

Prostate specific antigen - Netscape

File Edit View Go Communicator Help

Hs UniGene NCBI ?

Search for:

Hs.1548 *Homo sapiens* **APS**

Prostate specific antigen

See also: APS entry in OMIM

BEST SWISS-PROT HIT

Accession:	P07288
P-value:	2.0e-314
Name:	PROSTATE SPECIFIC ANTIGEN PRECURSOR [Homo sapiens]
Function:	PRESUMABLY HYDROLYZE THE HIGH MOLECULAR MASS SEMINAL VESICLE PROTEIN THUS LEADING TO THE LIQUEFACTION OF THE SEMINAL COAGULUM.

MAPPING INFORMATION

Chromosome:	19
MIM Gene Map:	19q13
Gene Map 98:	sts-F18096 , Chr.19, D19S413-D19S221
Gene Map 98:	M21896 , Chr.19, D19S425-D19S418

Document: Done

Prostate specific antigen - Netscape

File Edit View Go Communicator Help

EXPRESSION INFORMATION

Note: Highly represented in many libraries

cDNA sources: Blood, Brain, Breast, Colon, Eye, Germ Cell, Liver, Lung, Ovary, Parathyroid, Prostate, Thymus, Tonsil, Whole embryo

mRNA/GENE SEQUENCES (7)

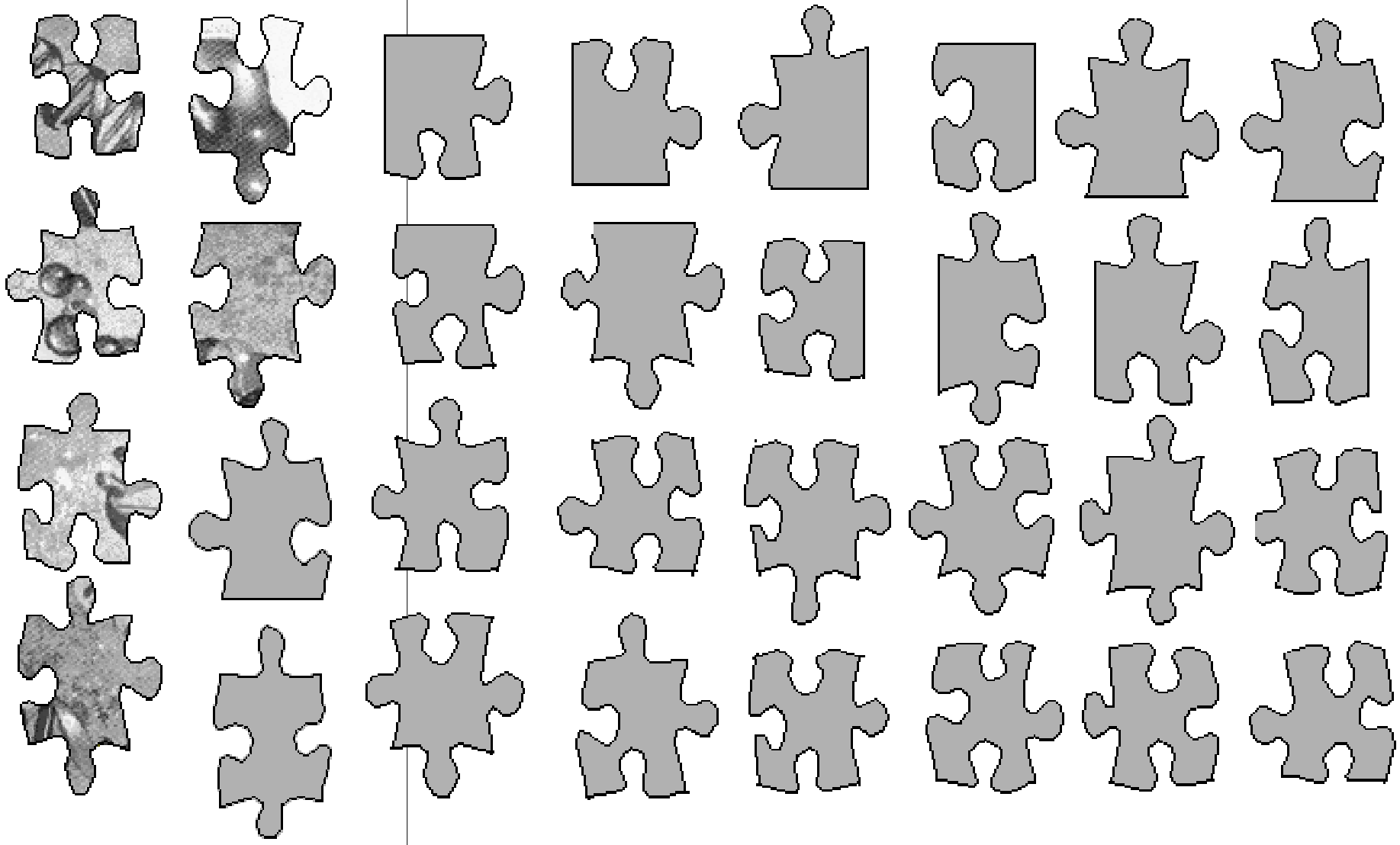
M26663	Homo sapiens prostate-specific antigen mRNA, complete cds	P A
X05332	Human mRNA for prostate specific antigen	P A
U17040	Human prostate specific antigen precursor mRNA, complete cds	P
M27274	Human prostate specific antigen gene, complete cds	P
M24543	Human prostate-specific antigen (PA) gene, complete cds	P A S
X14810	Human DNA for prostate specific antigen (PSA)	P
X07730	Human mRNA for prostate specific antigen	P A

EST SEQUENCES (10 of 438) [Show all ESTs]

AA631839	cDNA clone IMAGE:1132391	Prostate		7.8 kb	C
AA622817	cDNA clone IMAGE:1132329	Prostate		5.5 kb	C
AA230145	cDNA clone IMAGE:1010410	Prostate		1.8 kb	C
AA228273	cDNA clone IMAGE:1010410	Prostate		1.8 kb	P C
AA213735	cDNA clone IMAGE:683228	Tonsil	5' read	1.8 kb	C
AA631696	cDNA clone IMAGE:1132496	Prostate		1.7 kb	C
AA579039	cDNA clone IMAGE:915708	Prostate		1.6 kb	C

Document: Done

A Catalog of Puzzle Pieces

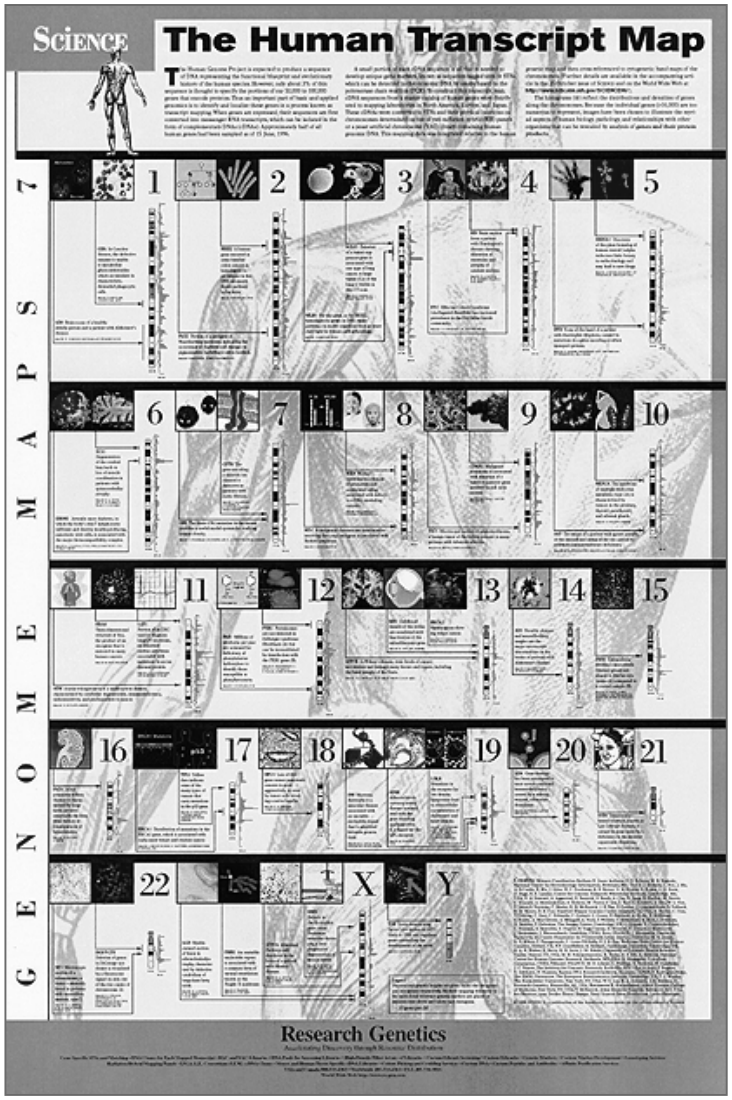


Transcript Mapping

International RH mapping consortium

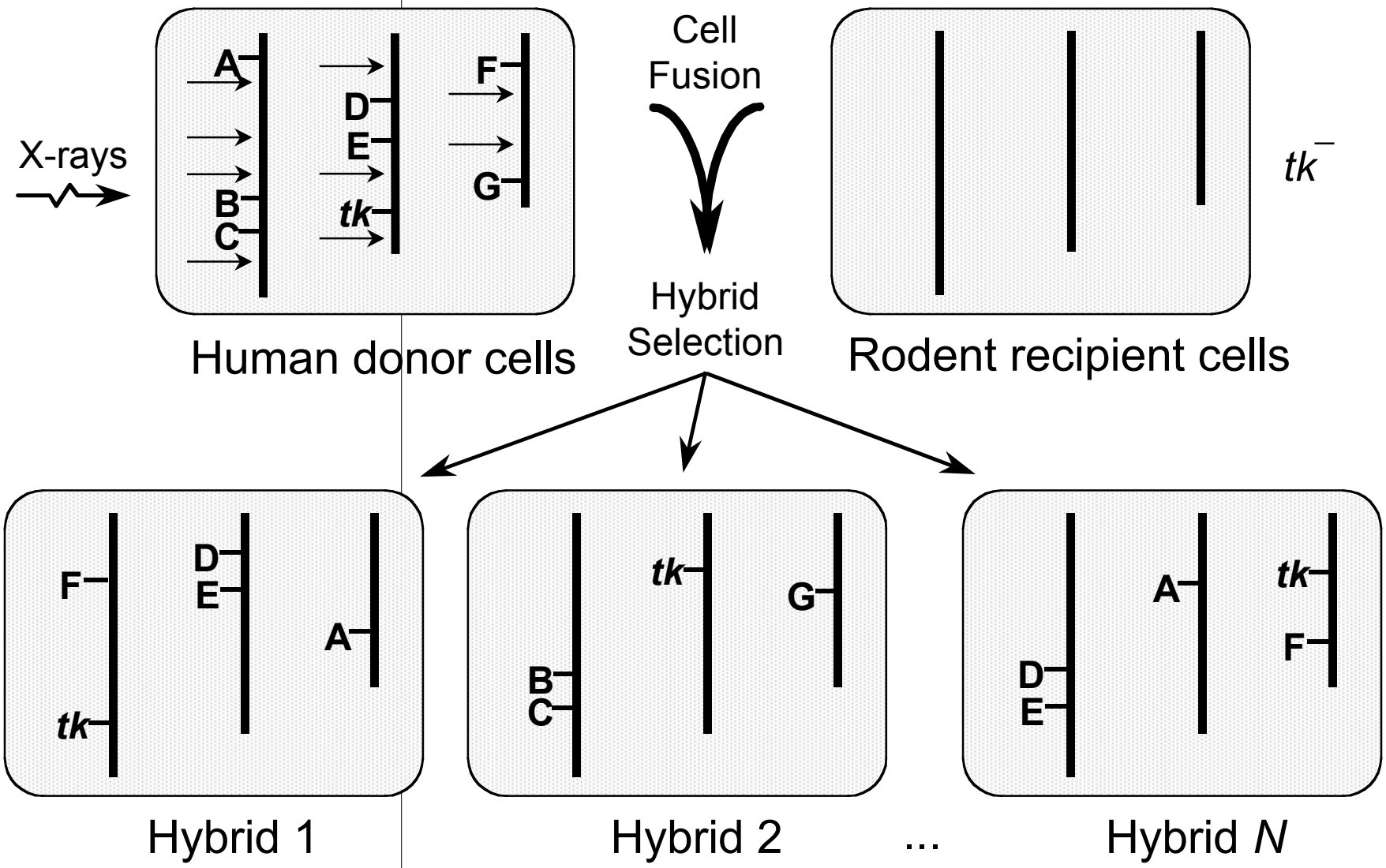
STSs designed from cDNA sequences

Mapped using GeneBridge 4 (GB4) and Stanford G3 panels



Science 1996 Genome Wall Chart

Radiation Hybrid Mapping

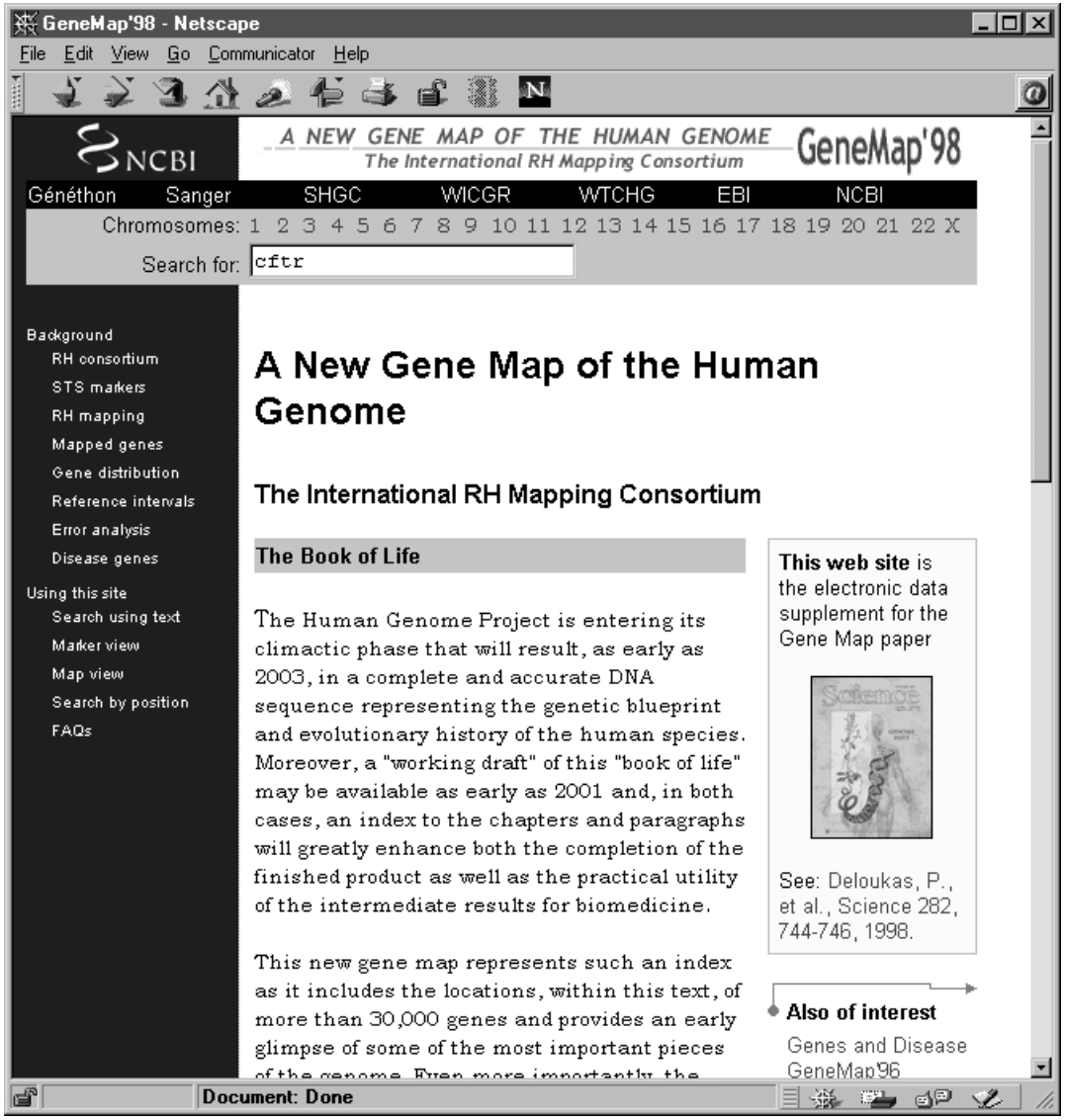


Transcript Mapping

1998 update had over 30,000 unique gene STSs

About 2000 genetic markers used for integration

Accuracy improved approx 3-fold from 1996 release



www.ncbi.nlm.nih.gov/genemap

GM98: Search Results - Netscape

File Edit View Go Communicator Help

NCBI

A NEW GENE MAP OF THE HUMAN GENOME
The International RH Mapping Consortium
GeneMap'98

Généthon Sanger SHGC WICGR WTCHG EBI NCBI

Chromosomes: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X

Search for: cftr

Background
- RH consortium
- STS markers
- RH mapping
- Mapped genes
- Gene distribution
- Reference intervals
- Error analysis
- Disease genes

Using this site
- Search using text
- Marker view
- Map view
- Search by position
- FAQs

Search Results

3 records satisfy the query

sts-M28668, ...	Cystic fibrosis conductance regulator
sts-M96936	Unknown
SHGC-5982	Unknown

Questions or Comments?
Contact the NCBI Service Desk

Document: Done

GM98: Locus details - Netscape

File Edit View Go Communicator Help

NCBI

A NEW GENE MAP OF THE HUMAN GENOME
The International RH Mapping Consortium
GeneMap'98

Généthon Sanger SHGC WICGR WTCHG EBI NCBI

Chromosomes: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X

Search for:

Background
- RH consortium
- STS markers
- RH mapping
- Mapped genes
- Gene distribution
- Reference intervals
- Error analysis
- Disease genes

Using this site
- Search using text
- Marker view
- Map view
- Search by position
- FAQs

Cystic fibrosis conductance regulator

Cross-References

UniGene Hs.663 Cystic fibrosis conductance regulator (CFTR)

RH Mapping Results

SHGC-9783	G3 Map:	Chr.7
	Reference interval:	D7S523-D7S655 (123.9-126.5 cM)
	Physical position:	5752 cR_{10000} (F)
	RH details:	RHdb RH2062
	Typed by:	Stanford (see SHGC-9783)
sts-M28668	GB4 Map:	Chr.7
	Reference interval:	D7S523-D7S655 (123.9-126.5 cM)
	Physical position:	558.49 cR_{3000} (P0.02)
	RH details:	RHdb RH40688
	Typed by:	Genethon

Electronic PCR Results

Genomic (from GenBank PRI division)

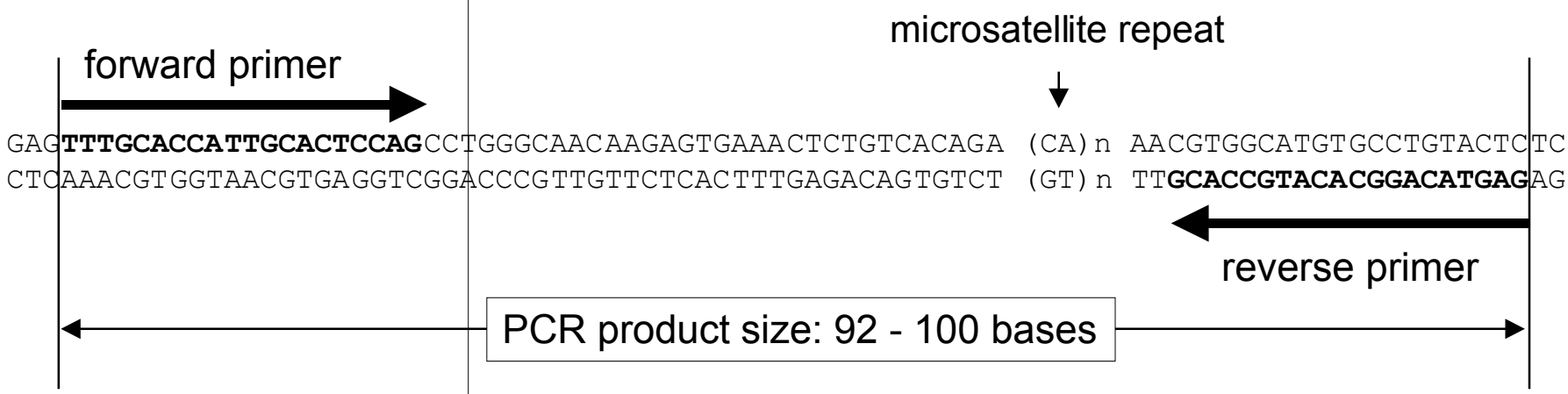
AC000061 Human BAC clone 133K23 from 7q31.2, complete sequence [Homo sapiens]

STS 57464 ... 57729 bp: SHGC-9783

Document: Done

Electronic PCR (E-PCR)

STS marker D6S1606



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

Schuler (1997), *Genome Research* 7, 541-550

GM98: Locus details - Netscape

File Edit View Go Communicator Help

Electronic PCR Results

Genomic (from GenBank PRI division)

AC000061 Human BAC clone 133K23 from 7q31.2, complete sequence [Homo sapiens]

STS 57464 ... 57729 bp: SHGC-9783
STS 58776 ... 58924 bp: sts-M28668

M55131 Human cystic fibrosis transmembrane conductance regulator (CFTR) gene, exon 24

STS 1754 ... 1904 bp: sts-M28668
STS 393 ... 658 bp: SHGC-9783

mRNAs (from GenBank PRI division)

M28668 Human cystic fibrosis mRNA, encoding a presumed transmembrane conductance regulator (CFTR)

STS 4578 ... 4843 bp: SHGC-9783
STS 5890 ... 6039 bp: sts-M28668

ESTs (from GenBank EST division)

AA503064 ne44b04.s1 NCI_CGAP_Co3 Homo sapiens cDNA clone IMAGE:900175 similar to gb:M28668 CYSTIC FIBROSIS TRANSMEMBRANE CONDUCTANCE REGULATOR (HUMAN);

STS 92 ... 240 bp: sts-M28668

AA515982 nf68a02.s1 NCI_CGAP_Co3 Homo sapiens cDNA clone IMAGE:925034 similar to gb:M28668 CYSTIC FIBROSIS TRANSMEMBRANE CONDUCTANCE REGULATOR (HUMAN);

STS 93 ... 241 bp: sts-M28668

GM98: Locus details - Netscape

File Edit View Go Communicator Help

NCBI

A NEW GENE MAP OF THE HUMAN GENOME
The International RH Mapping Consortium
GeneMap'98

Généthon Sanger SHGC WICGR WTCHG EBI NCBI

Chromosomes: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X

Search for:

Background
- RH consortium
- STS markers
- RH mapping
- Mapped genes
- Gene distribution
- Reference intervals
- Error analysis
- Disease genes

Using this site
- Search using text
- Marker view
- Map view
- Search by position
- FAQs

Cystic fibrosis conductance regulator

Cross-References

UniGene Hs.663 Cystic fibrosis conductance regulator (CFTR)

RH Mapping Results

SHGC-9783	G3 Map:	Chr.7
	Reference interval:	D7S523-D7S655 (123.9-126.5 cM)
	Physical position:	5752 cR_{10000} (F)
	RH details:	RHdb RH2062
	Typed by:	Stanford (see SHGC-9783)
sts-M28668	GB4 Map:	Chr.7
	Reference interval:	D7S523-D7S655 (123.9-126.5 cM)
	Physical position:	558.49 cR_{3000} (P0.02)
	RH details:	RHdb RH40688
	Typed by:	Genethon

Electronic PCR Results

Genomic (from GenBank PRI division)

AC000061 Human BAC clone 133K23 from 7q31.2, complete sequence [Homo sapiens]

STS 57464 ... 57729 bp: SHGC-9783

Document: Done

GM98: Chr.7 - Netscape

File Edit View Go Communicator Help

NCBI

A NEW GENE MAP OF THE HUMAN GENOME

The International RH Mapping Consortium

GeneMap'98

Généthon Sanger SHGC WICGR WTCHG EBI NCBI

Chromosomes: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X

Search for:

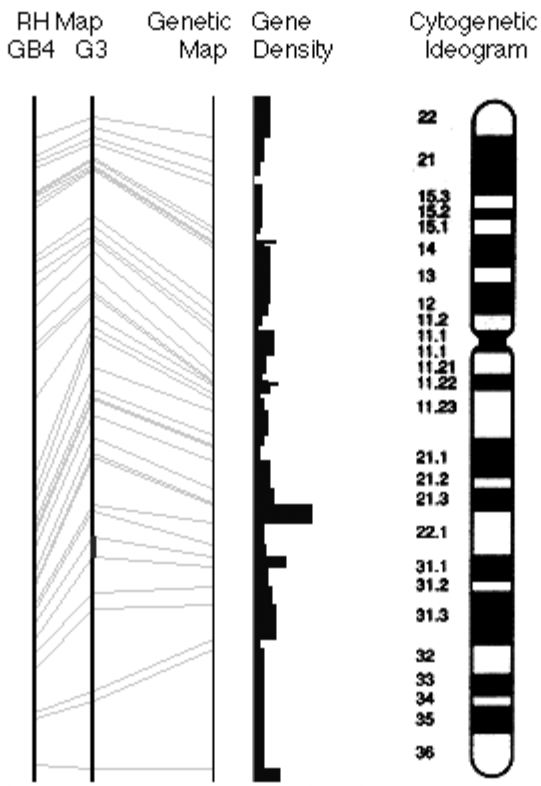
Background
- RH consortium
- STS markers
- RH mapping
- Mapped genes
- Gene distribution
- Reference intervals
- Error analysis
- Disease genes

Using this site
- Search using text
- Marker view
- Map view
- Search by position
- FAQs

Error Flags

*	Minor positional discrepancy
**	Major positional discrepancy
***	Chromosome assignment discrepancy

Chromosome 7: D7S523-D7S655



Document: Done

GM98: Chr.7 - Netscape

File Edit View Go Communicator Help

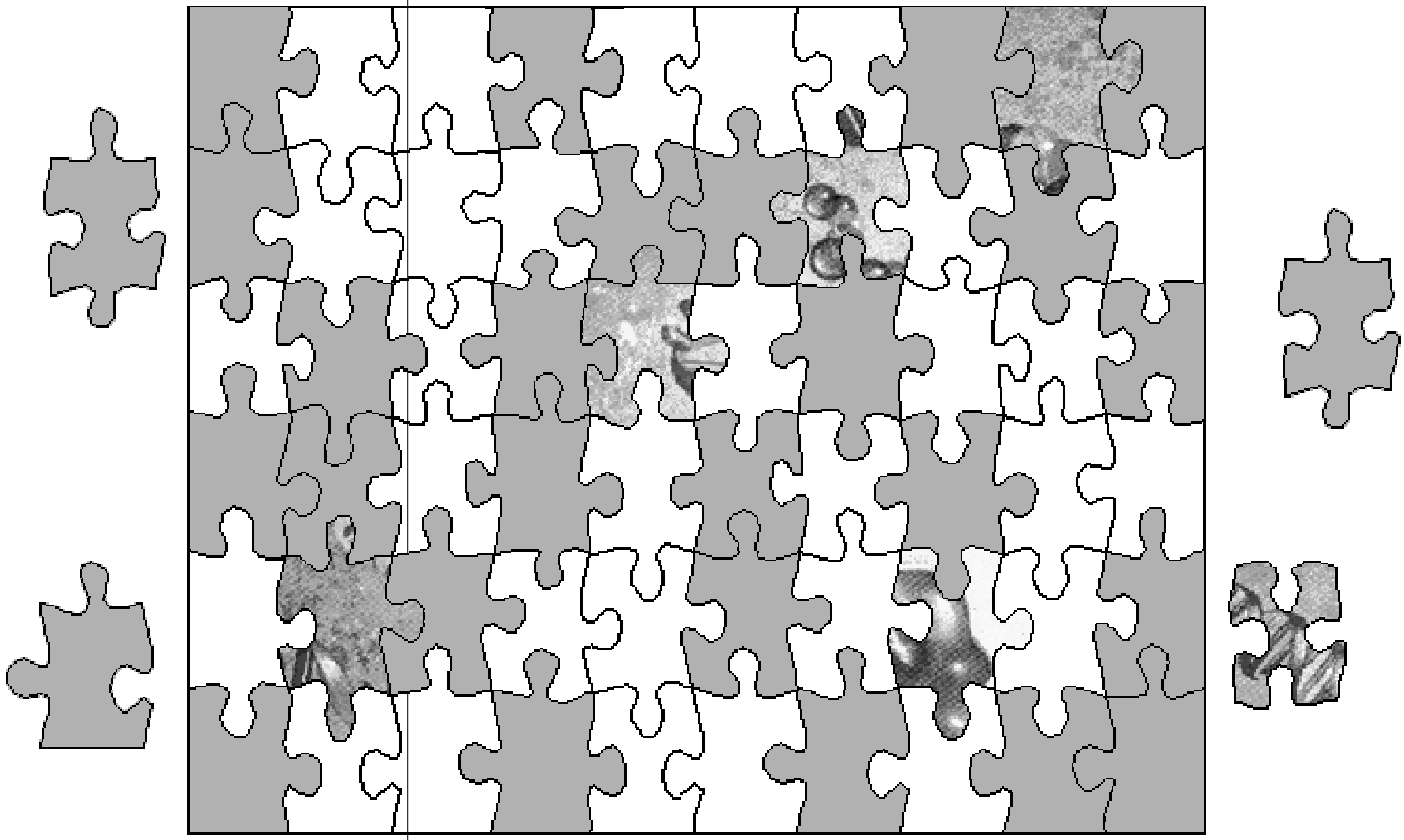
Next interval up

123.9	**5560**	**F**	**AFM242ye3**	**D7S523**	**Microsatellite anchor marker AFM242ye3 (SHGC-1**
	5588	F	SHGC-13594		ESTs
	5613	F	SHGC-8664	PPP1R3	Protein phosphatase 1, regulatory (inhibito..
	5613	F	SHGC-12021	PPP1R3	Protein phosphatase 1, regulatory (inhibito..
123.9	5624	F	AFMa062zd9	D7S2554	Microsatellite marker AFMa062zd9 (SHGC-22..
123.4	5629	F	AFM323yg5	D7S687	Microsatellite marker AFM323yg5 (SHGC-191..
125.0	5662	F	AFMb343xc1	D7S2502	Microsatellite marker AFMb343xc1 (SHGC-22..
125.3	5671	F	AFMa045xb1	D7S2543	Microsatellite marker AFMa045xb1 (SHGC-22..
125.3	5680	F	AFM098xg9	D7S486	Microsatellite marker AFM098xg9 (SHGC-46)
125.1	5695	F	AFM242yc3	D7S522	Microsatellite marker AFM242yc3 (SHGC-999..
	5722	F	SHGC-10385		Homo sapiens clone 24651 mRNA sequence
125.1	5727	F	AFM197xf10	D7S2460	Microsatellite marker AFM197xf10 (SHGC-22..
	5727	F	SHGC-10669		Human activated met oncogene mRNA, partial..
	5744	F	SHGC-5660		ESTs
	5748	F	SHGC-12677	WNT2	Wingless-type MMTV integration site 2, hum..
	5748	F	SHGC-8708		ESTs
	5752	F	SHGC-9783	CFTR	Cystic fibrosis conductance regulator
	5773	F	SHGC-5982		Unknown
	5777	F	SHGC-35613		Homo sapiens cystic fibrosis transmembrane ..
	5791	F	SHGC-10294		ESTs
126.5	**5806**	**F**	**AFM263wg9**	**D7S655**	**Microsatellite anchor marker AFM263wg9 (SHGC-3**

Next interval down

Document: Done

A Framework for the Puzzle



SAGEmap

Presentation of SAGE data generated by CGAP, Greg Riggins (Duke)

Highlights expression differences between different libraries

SAGE tags are mapped to UniGene

CGAP SAGE results - Microsoft Internet Explorer

File Edit View Go Favorites Help

Color = RED if expression of tag in Group A > GroupB

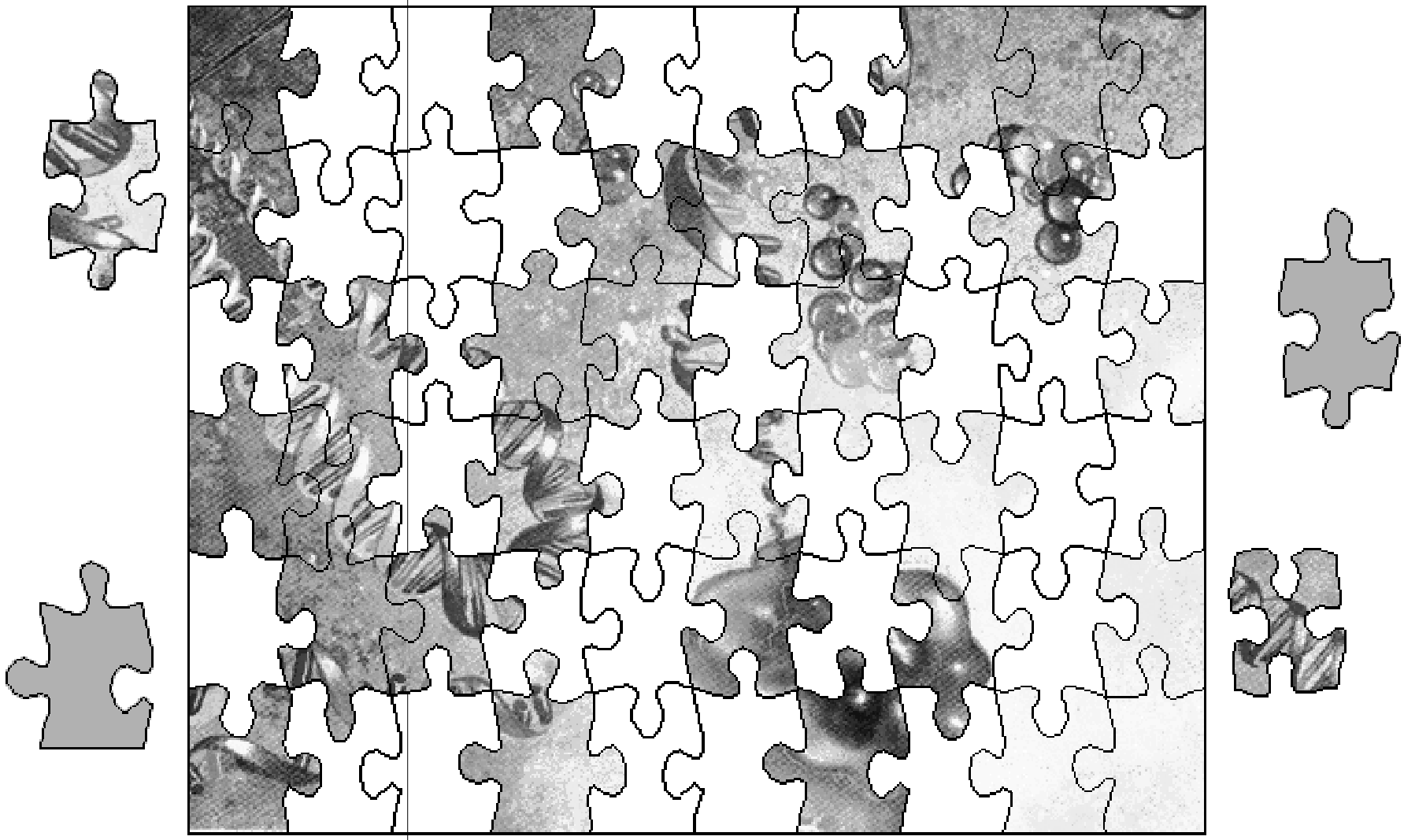
Color = GREEN if expression of tag in Group B > Group A

#	SAGE tag	UniGene id	Gene description	A:B	Grp A	Grp B	A:B >3x
1	AAGCATTAAA	Hs.1519	Protein kinase, cAMP-dependent, regulatory, type I, beta	AB	5	84	100%
2	AATAAAGCTA	Hs.90297	Synuclein, beta	AB	3	73	100%
3	CCAAGGCCCC	Hs.154658	H.sapiens mRNA from TYL gene	AB	1	52	100%
4	GGGGTGCTGT	Hs.126	DYNAMIN-1	AB	7	139	100%
5	GTTGTGGTTA	Hs.75415	BETA-2-MICROGLOBULIN PRECURSOR	AB	290	43	100%
6	TGCTGACTCC	Hs.29076	ESTs	AB	82	7	100%
7	AAATACTGCC	Hs.32916	ESTs	AB	1	46	99%
8	AATAGTTTGA	Hs.18551	Homo sapiens mRNA for AMY, complete cds	AB	1	37	99%
		Hs.76493	Peroxisomal acyl-coenzyme A oxidase [human, liver, mRNA, 3086 nt]				
9	ACAAAAACTA	N/A	**WARNING**: Tag matches mitochondrial DNA	AB	9	189	99%
10	ACAACACTAC	Hs.121515	ESTs	AB	4	80	99%
11	ACGCACATTA	Hs.156007	ZAKI-4 mRNA in human skin fibroblast, complete cds	AB	--	29	99%
12	AGAACCTTCC	Hs.119732	MHC class I protein HLA-A (HLA-A28,-B40, -Cw3)	AB	62	4	99%
13	ATTAAAGTCA	Hs.78748	Human mRNA for KIAA0237 gene, complete cds	AB	--	30	99%
14	CACAGTTTGC	Hs.71346	Human gene for neurofilament subunit M (NF-M)	AB	--	34	99%

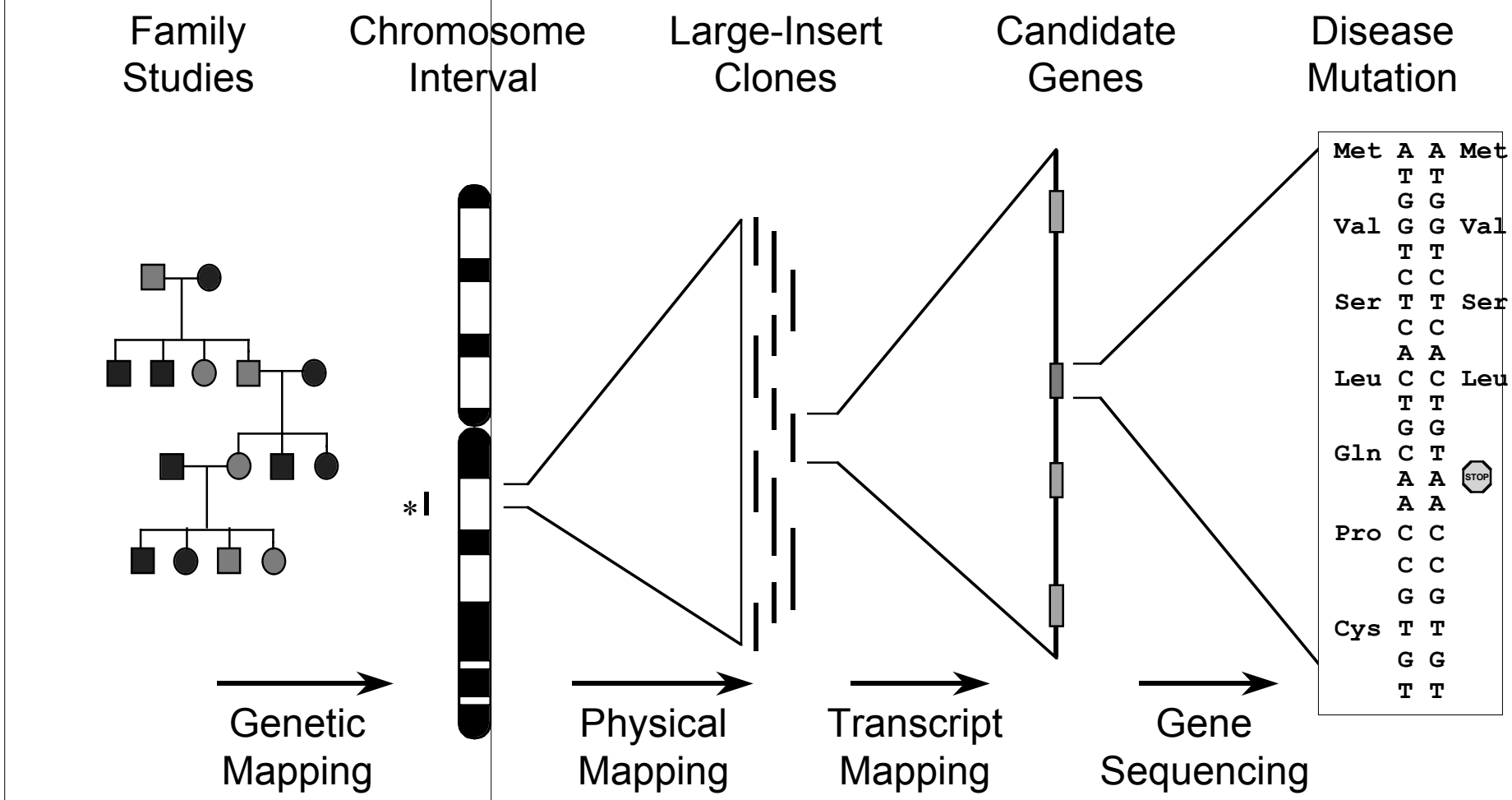
Internet zone

www.ncbi.nlm/nih.gov/SAGE

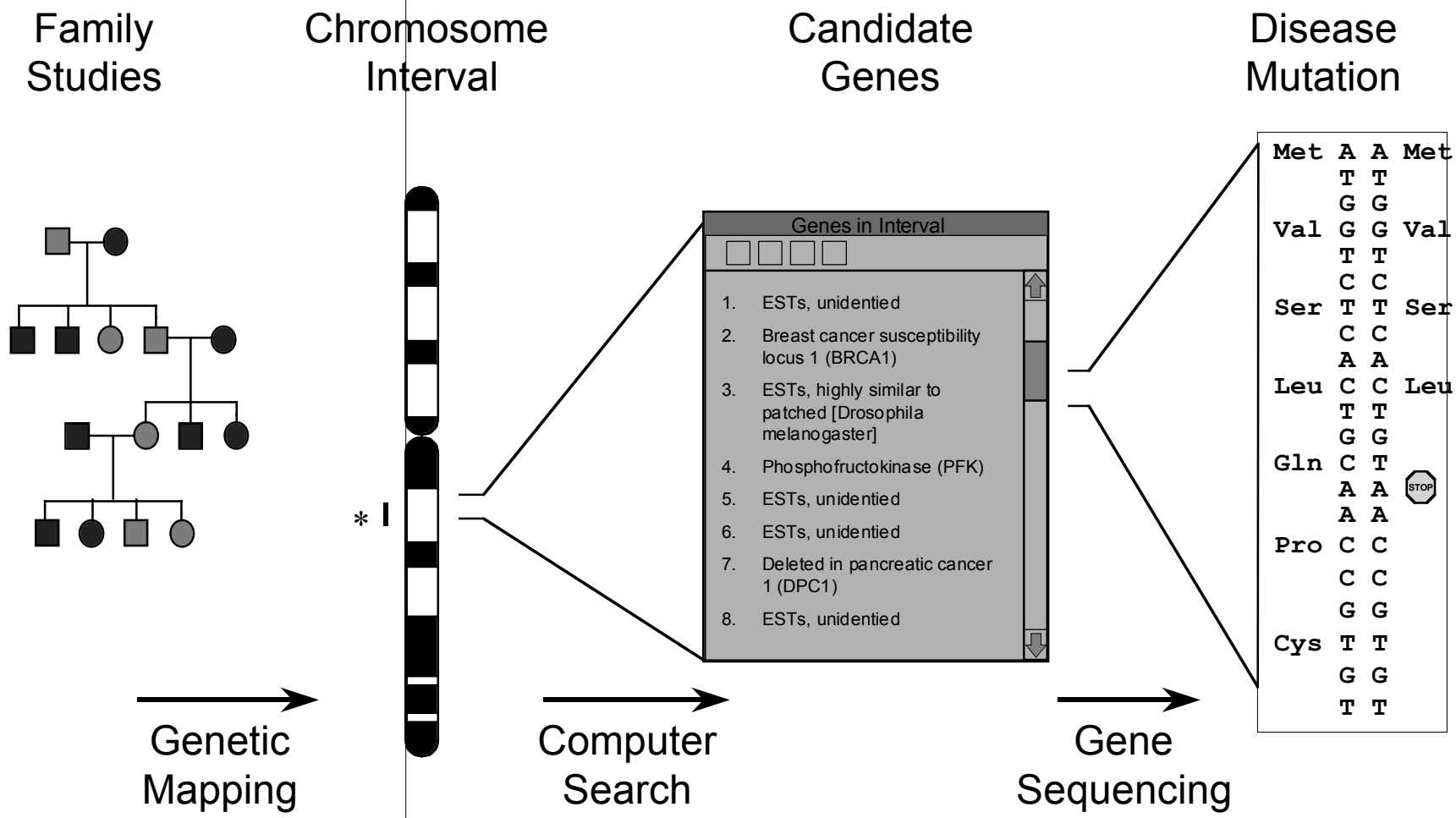
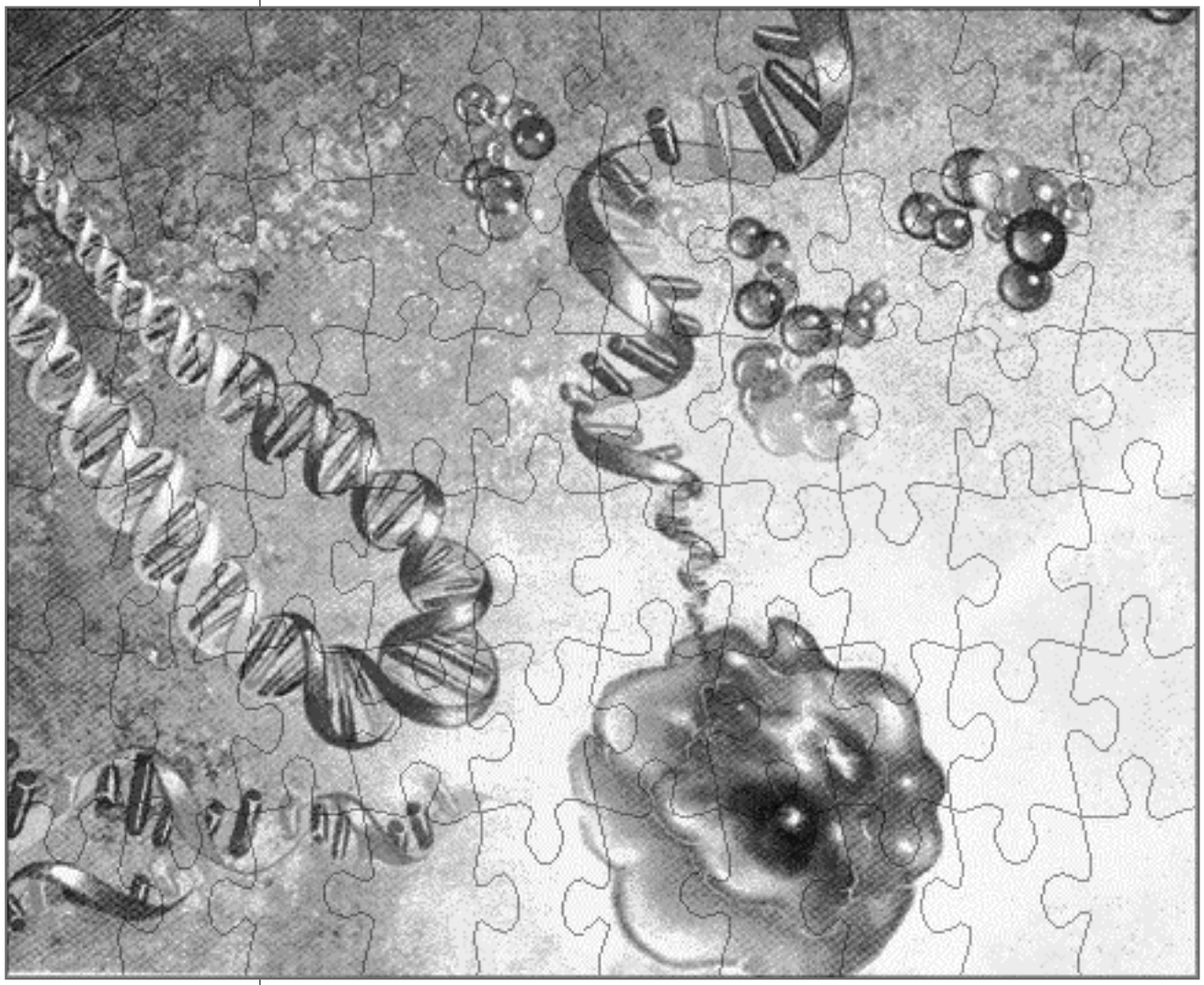
Revealing the Picture



Positional Cloning



Positional Candidate Cloning



Completing the Puzzle