

ABS04

- 2004 Applied Bayesian Statistics School -

**STATISTICS & GENE EXPRESSION GENOMICS:
METHODS AND COMPUTATIONS**

Mike West
Duke University

Centro Congressi Panorama, Trento, Italy
15th-19th June 2004

CNR-IAMI/University of Pavia ~ Inaugural Applied Bayesian Statistics School



ABS04



Some Bio basics -The Genome -

**Joseph Nevins
Holly Dressman
Mike West**

Duke University

DNA the molecule of life

Trillions of cells

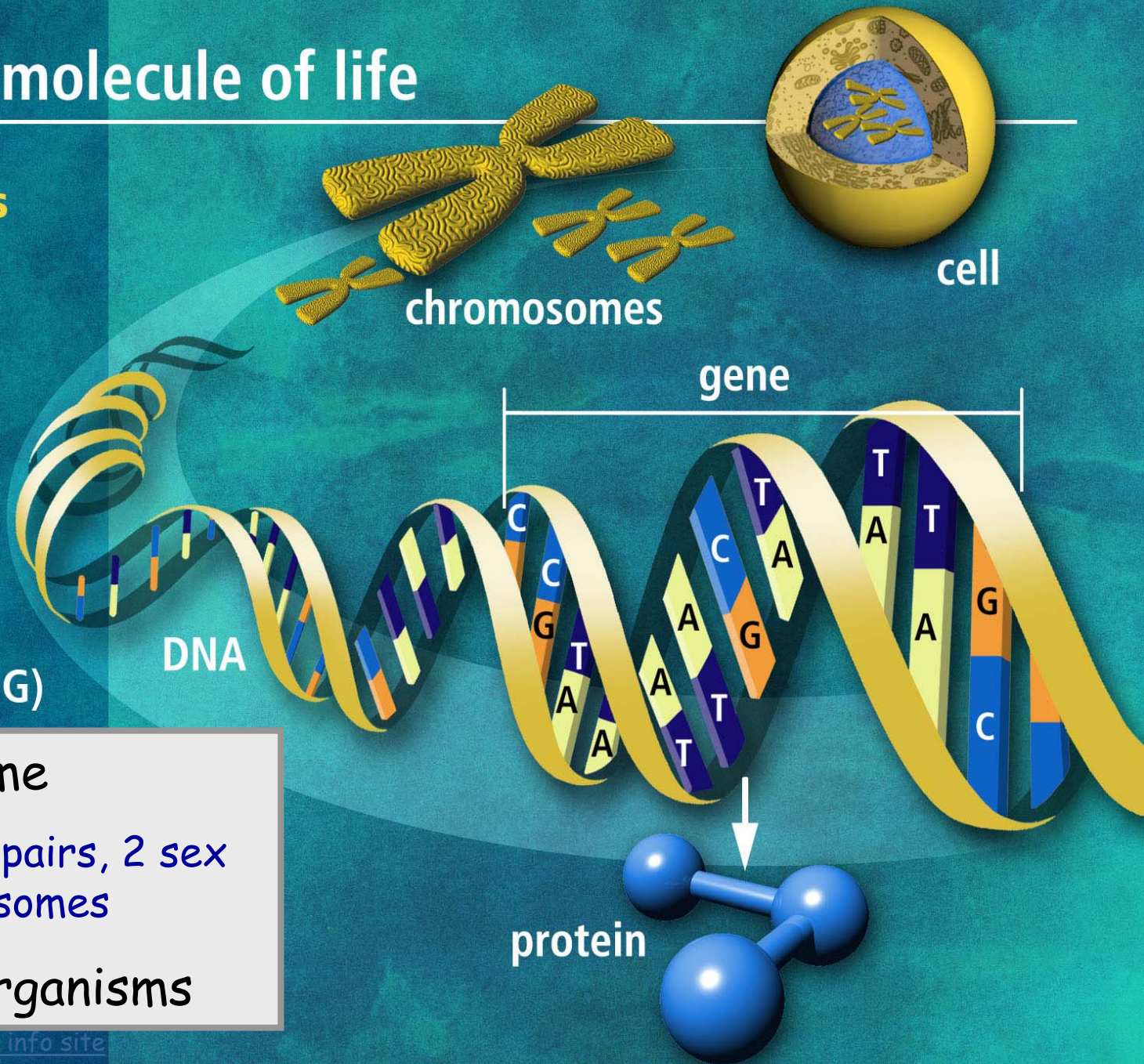
Each cell:

- 46 human chromosomes
- 2 meters of DNA
- 3 billion DNA subunits (the bases: A, T, C, G)

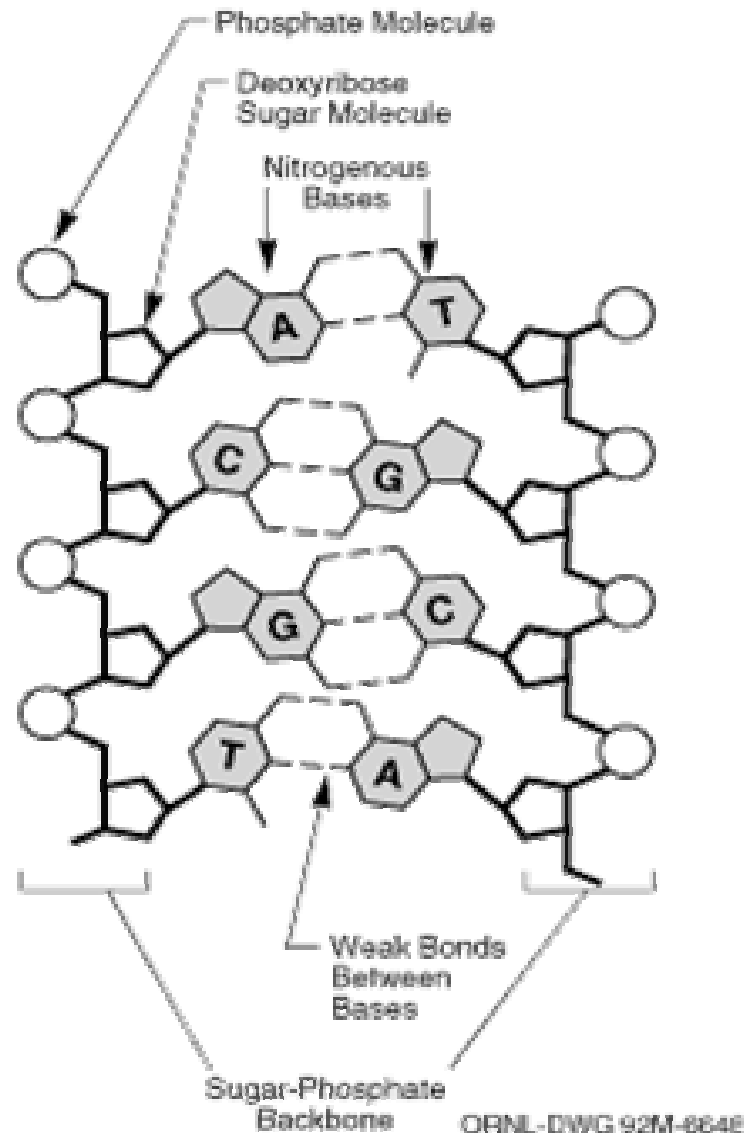
Human Genome

23 autosomal pairs, 2 sex chromosomes

Eukaryotic organisms



Essential Structure of DNA

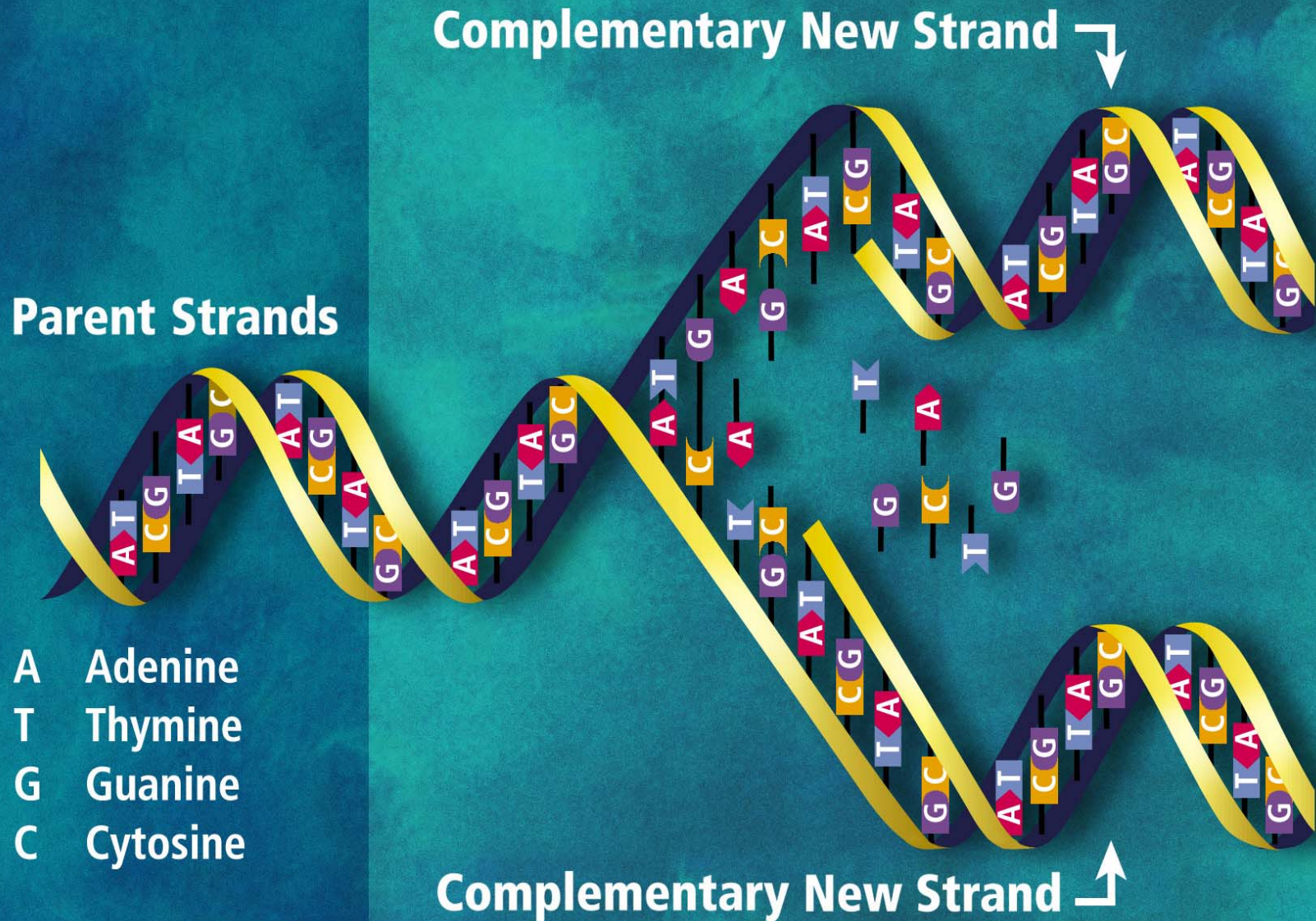


Nucleotide Bases:

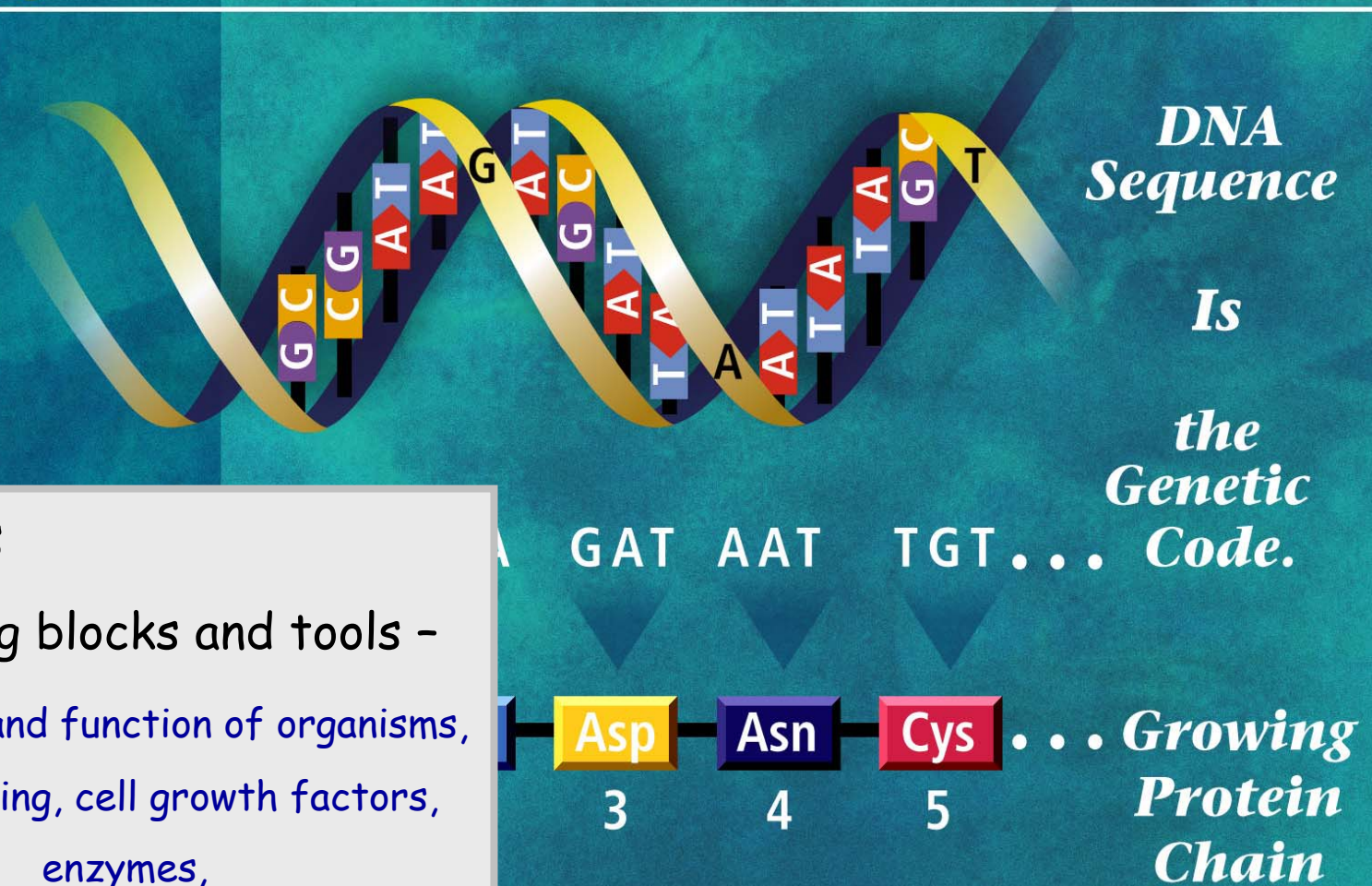
Adenine
Thymine
Guanine
Cytosine

+

DNA Replication Prior to Cell Division

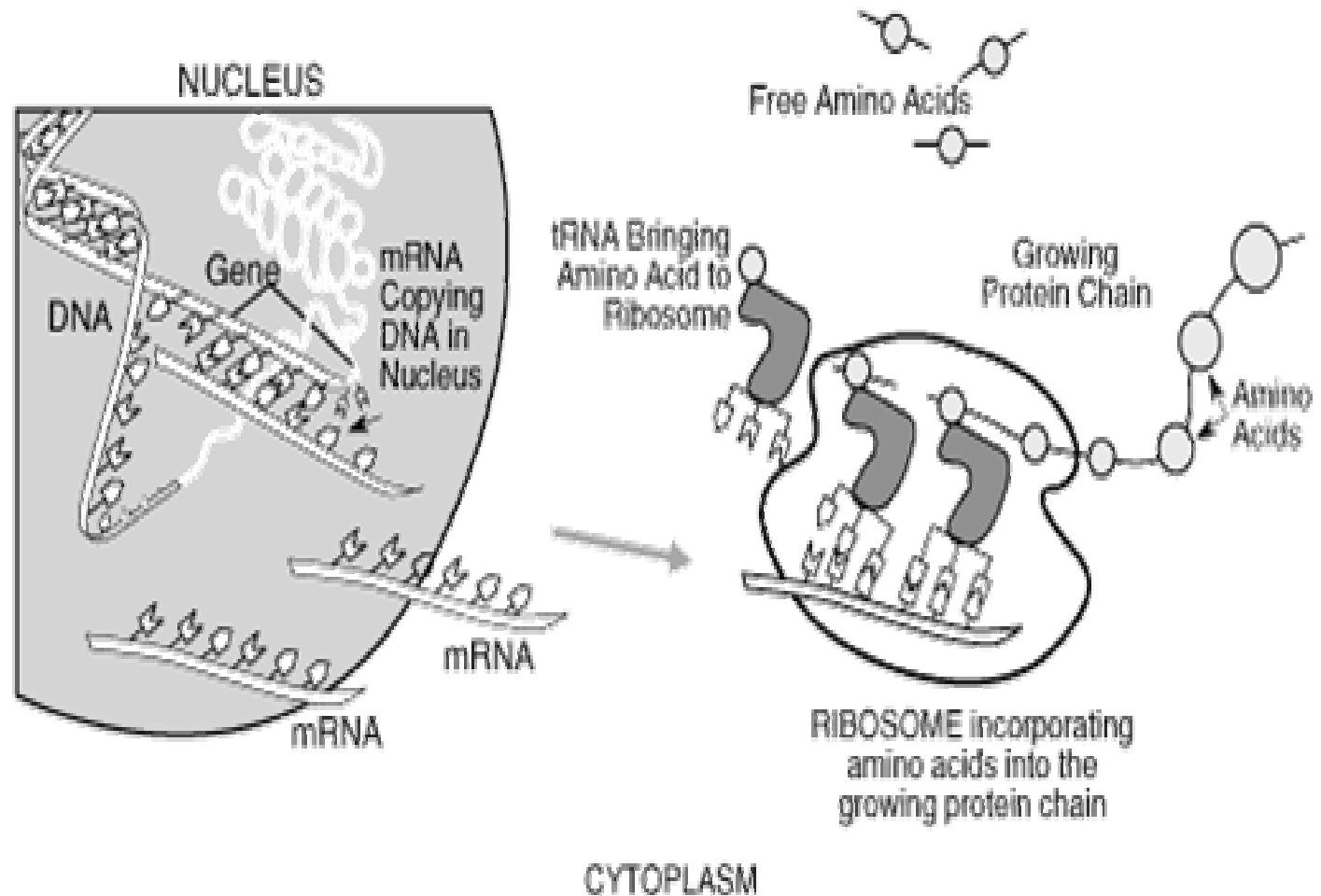


DNA Genetic Code Dictates Amino Acid Identity and Order



Transcription & Translation

ORNL-DWG 91M-17360

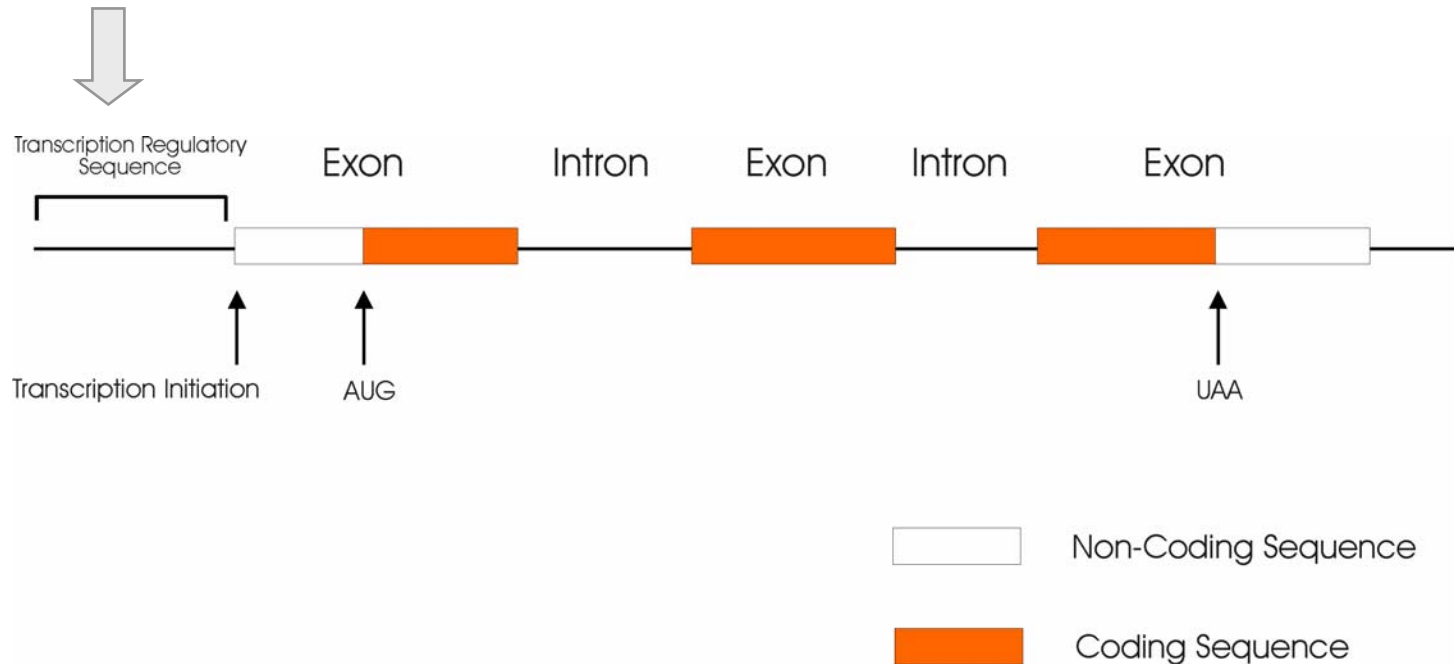


but, what IS a gene?

Estrogen Receptor Gene (ESR1 or ERα)

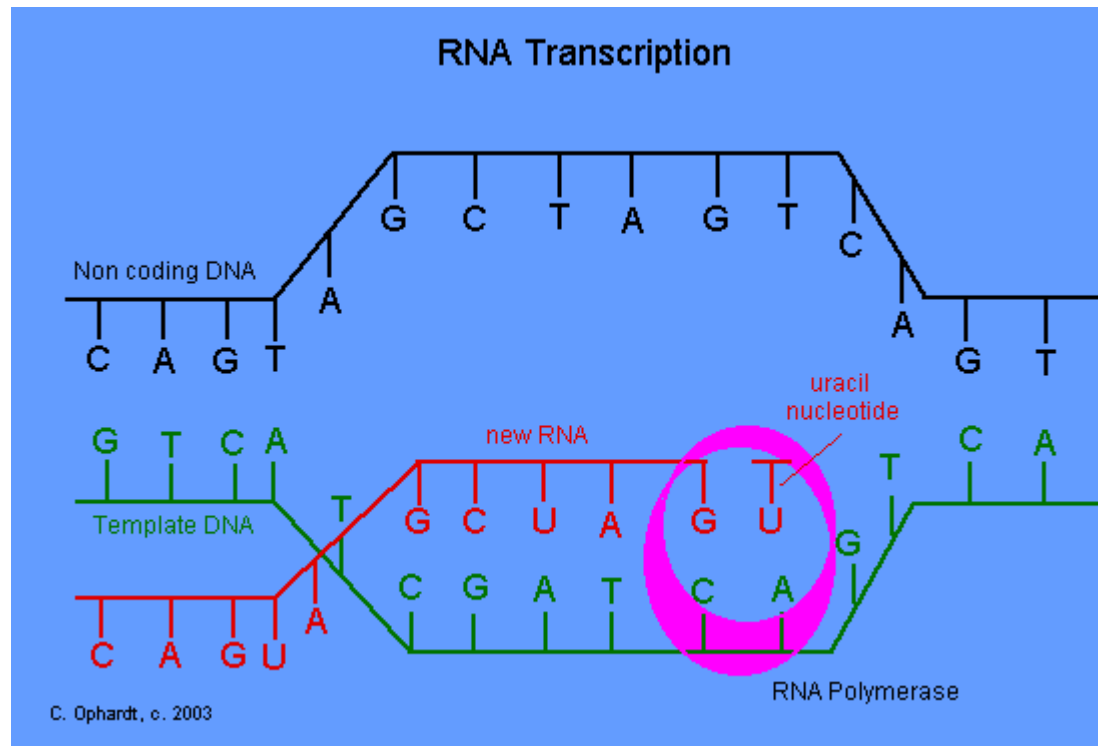
gaattccaaa atttgtatgt ttcttgtatt tttgatgaag gagaataact gtaatgatca ctgtttacac tatgtacact ttaggccagc
cctttgtagc gttatacaaa ctgaaagcac accggaccgc caggctcccg gggcagggcc ggggccagag ctgcgctgtc ggcgggacat
gcgctgcgtc gcctctaacc tcgggctgtg ctctttttcc aggtggcccg ccggtttctg agccttctgc cctgcgggga cacggtctgc
accctgcccg cggccacgga ccatgaccat gaccctccac accaaagcat ctgggatggc cctactgcat cagatccaag ggaacgagct
ggagcccctg aaccgtccgc agctcaagat cccctggag cggcccctgg gcgaggtgta cctggacagc agcaagcccg ccgtgtacaa
ctaccccgag ggcgccgcct acgagttcaa cgcgcgggc gccgccaacg cgcaggtcta cggtcagacc ggcctcccct acggccccgg
gtctgaggct gcggcgttcg gtccaacgg cctgggggggt ttccccccac tcaacagcgt gtctccgagc ccgtgatgc tactgcaccc
gccgccgcag ctgtcgccct tctgcagcc ccacggccag caggtgccct actacctgga gaacgagccc agcggctaca cggtgcgca
ggccggcccg ccggcattct acaggccaaa ttacagataat cgacgccagg gtggcagaga aagattggcc agtaccaatg acaaggaag
tatggctatg gaatctgcca aggagactcg ctactgtgca gtgtgcaatg actatgcttc aggtaccat tatggagtct ggtcctgtga
gggctgcaag gccttcttca agagaagtat tcaaggacat aacgactata tgtgtccagc caccaaccag tgcaccattg ataaaaacag
gaggaagagc tgccaggcct gccggctccg caaatgctac gaagtgggaa tgatgaaagg tgggatacga aaagaccgaa gaggagggag
aatgttgaaa cacaagcgcc agagagatga tggggagggc aggggtgaag tggggtctgc tggagacatg agagctgcca accttggcc
aagcccgtc atgatcaaac gtctaaagaa gaacagcctg gccttgtccc tgacggccga ccagatggtc agtgccttgt tggatgctga
gcccccata ctctattccg agtatgatcc taccagaccc ttacagtgaag cttcgatgat gggcttactg accaacctgg cagacagga
gctggttcac atgatcaact gggcgaagag ggtgccaggc tttgtggatt tgacctcca tgatcaggte caccttctag aatgtgctg
gctagagatc ctgatgattg gtctcgtctg gcgtccatg gagcaccag tgaagctact gtttgcctct aacttgctct tggacaggaa
ccagggaaaa tgtgtagagg gcatggtgga gatcttcgac atgctgctgg ctacatcacc tcggttccgc atgatgaatc tgcagggaga
ggagtthgtg tgcctcaaat ctattattht gcttaattct ggagtgtaca catttctgtc cagcacctg aagtctctgg aagagaagga
ccatatccac cgagtcttg acaagatcac agacactthg atccacctga tggccaaggc aggcctgacc ctgcagcagc agcaccagcg
gctggcccag ctctctctca tctctctcca catcaggcac atgagtaaca aaggcatgga gcatctgtac agcatgaagt gcaagaacgt
ggtgccctc tatgacctgc tgcaggagat gctggacgcc caccgcctac atgcgccac tagccgtgga ggggcatccg tggaggagac
ggaccaaagc cacttggcca ctgcgggctc tacttcatcg cattccttgc aaaagtatta catcacgggg gaggcagagg gtttccctgc
cacagtctga gagtccctg gc ...

Promotor sequence
Start region for transcription
to create RNA

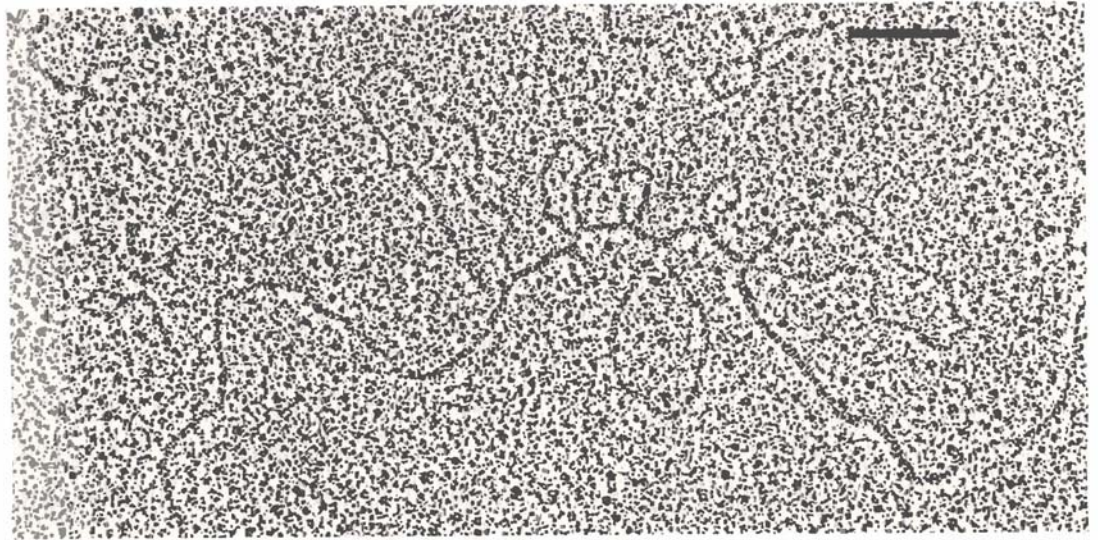


Coding sequence
Codes for proteins

RNA - Transcription

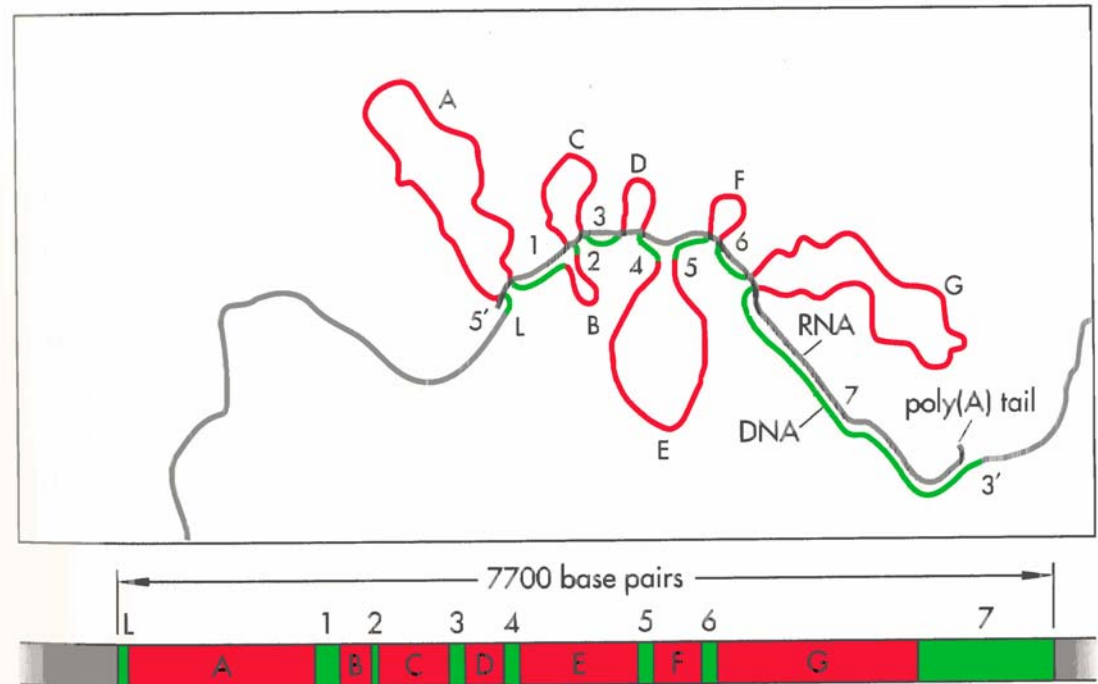


The message: mRNA
Splice exons
- coding only -



Alternative splicing
Variants of same gene
and/or gene function
-Protein -

Gene variants
DNA SNPs, variants



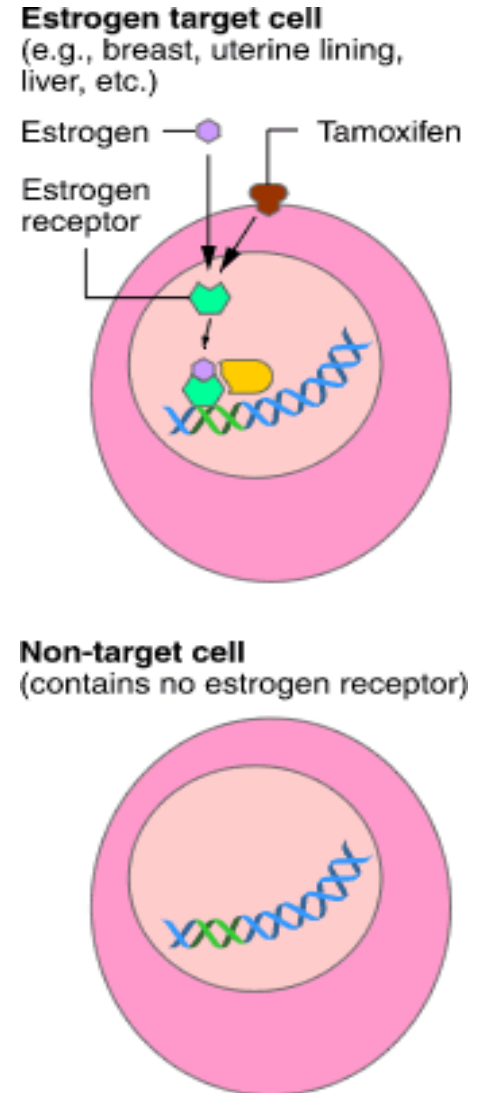
A Gene, Protein and some of its Function in Cell Biology and Cancer

(O)Estrogen Receptor

Estrogen hormone -
blood-borne chemical messenger
produced in ovaries

Estrogen Receptor (ER) -
protein in cells - docking station for
estrogen

"switch on" growth signal to nucleus for
cell growth, division proliferation, ...



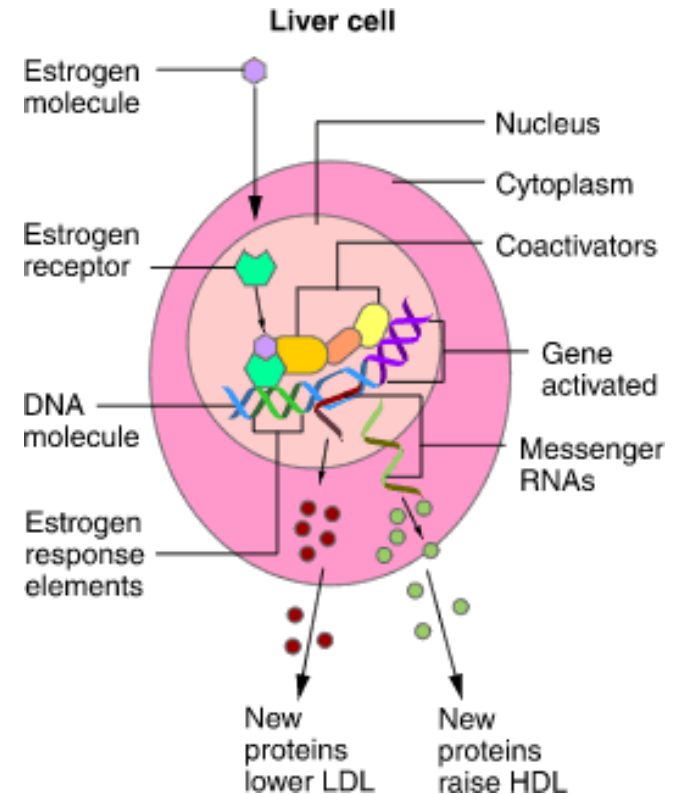
Estrogen Receptor: A Transcription Factor

ER triggers gene activation

e.g.,

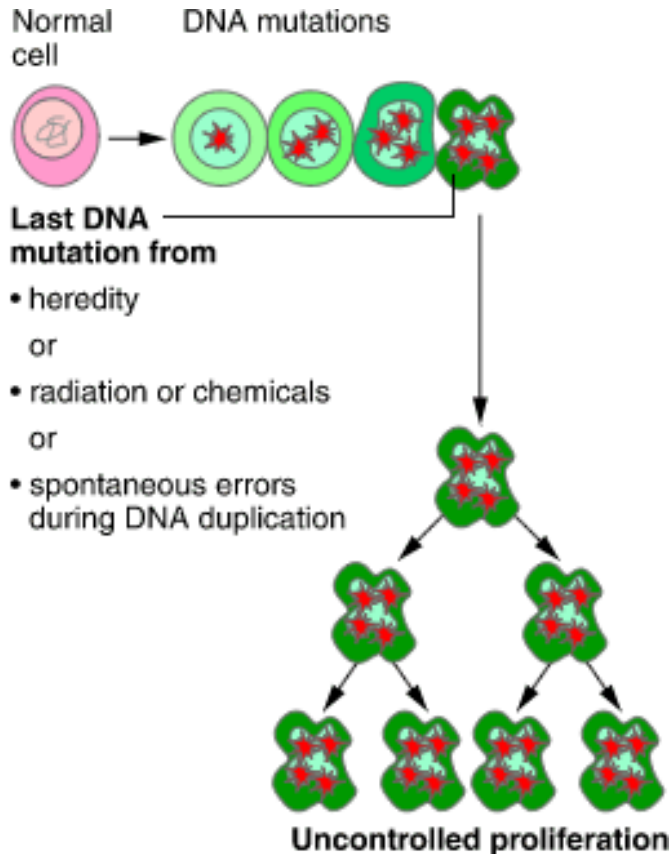
- Genes that influence cholesterol
- Growth signals to nucleus:
cell division (mitosis) and proliferation

- mammary milk gland cells -
menstrual cycle: ER responsive cell
growth, death
- endometrial cells in uterus
- ...



ER and Cancer

Cancer: growth of cells with DNA mutations



ER - normal cell growth regulator

Over-expression of ER -
potential oncogenic

Some cancers: Estrogen dependent
breast cancer: 60%-75%

Breast cancer neoadjuvant drugs:

- block ER receptor (Tamoxifen)
- inhibit estrogen production -
aromatase inhibitors (Letrozol)

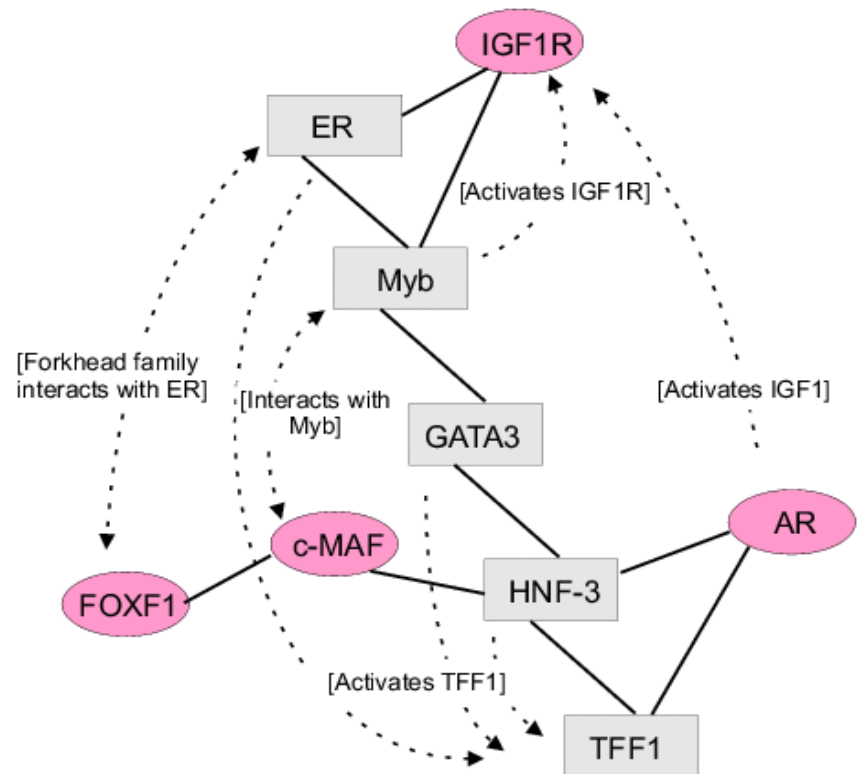
Transcription Networks & Expression

ER regulates multiple genes
and hence their mRNA and
protein products

ER is regulated by, and
co-regulated with, multiple
genes

mRNA expression levels

- Abundance of gene
- Related across genes in
transcriptional interaction



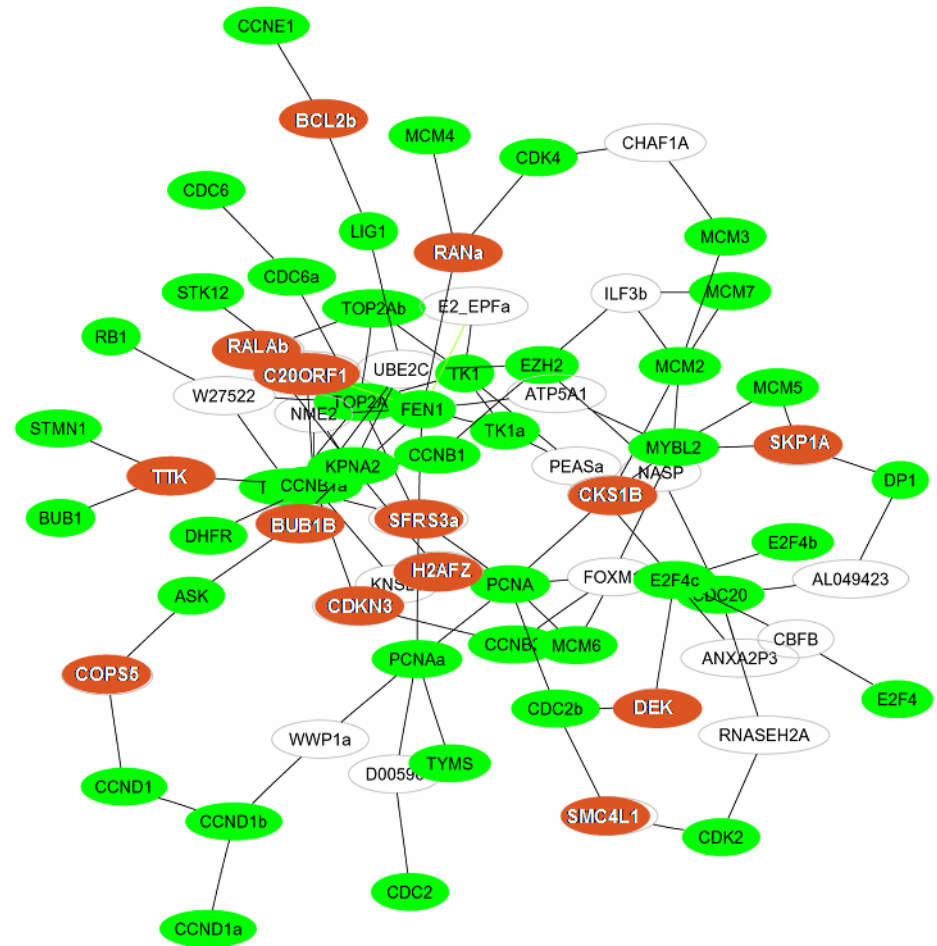
Transcription Networks & Expression

ER regulates multiple genes
and hence their mRNA and
protein products

ER is regulated by, and
co-regulated with, multiple
genes

mRNA expression levels

- Abundance of gene
- Related across genes in transcriptional interaction



Data and Information Resources on Genes

Gene Ontologies - web servers in US,EU

(GenBank, Unigene, Ensembl, AmiGO,...)

Key start point: *LocusLink > Entrez Gene*

US NLM/NCBI Entrez Gene Web Site

Our system: integrates genomics
databases, information servers,
statistical & graphical tools

Duke Integrated Genomics (DIG)

Gene - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene> Search Print

Home Bookmarks Google CAGP Duke Calendar Duke Directory DUMC SAMSI mozilla.org Duke ASC WebMail DIG

NCBI

Entrez Gene

Entrez PubMed BioRxiv Protein Genome Structure PMC Taxonomy SRA

Search Gene for ESR1 Go Clear ☒ current records only

Limits Preview/Index History Clipboard Details

Entrez

SITE MAP
Entrez Help

Gene
Search
Gene Help

FAQ
FTP site

Related sites
Entrez Genome
Genomic Biology
HomoloGene
LocusLink
Map Viewer
OMIM
RefSeq
UniGene

Feedback
Help Desk
Corrections
Submit GeneRIFs

Subscriptions
RefSeq
Gene
Map Viewer

- Enter one or more search terms.
- More information about available fields is available [here](#).
- Consider use of the limits and preview/index functions.
- Remember, boolean operators (AND, OR, NOT) must be in uppercase.

Gene

Background

Gene provides a unified query environment for genes defined by sequence and/or in NCBI's Map Viewer. You can query on names, symbols, accessions, publications, GO terms, chromosome numbers, E.C. numbers, and many other attributes associated with genes and the products they encode.

Because Gene is now an Entrez database, all the familiar and useful functions are now available, including Preview/Index, History, and LinkOut.

Please note: Entrez Gene is under active development. We welcome your suggestions.

Getting started

Sample queries

Look for genes by name part and multiple species
transporter AND ("Drosophila melanogaster"[orgn] OR "Mus musculus"[orgn]) more...

Look for genes by chromosome and symbol
(II[chr] OR 2[chr]) AND adh*[sym] more...

What's new?

March 24, 2004 A small set of tab-delimited files became available for transfer by ftp. These include Gene/RefSeq and Gene/PubMed reports.

December 15, 2003 Gene became accessible via the [Entrez cross-database search](#) mechanism. An ftp site is still under development.

November 20, 2003 Gene became available in a limited mode. When more functions are implemented, Gene will be fully integrated with other Entrez databases, including global query.

Restrictions on Use | Write to the Help Desk
NCBI | NLM | NIH

May 12 2004 6:43:09

Done

start Inbox - Microsoft Out... Trento Slides Gene - Mozilla Microsoft PowerPoint ... 8:47 PM

Gene - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=Retrieve&dopt=Graphics&list_uids=2099 Search Print

Home Bookmarks Google CAGP Duke Calendar Duke Directory DUMC SAMSI mozilla.org Duke ASC WebMail DIG

NCBI

Entrez Gene

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books OMIM

Search Gene for Go Clear ☒ current records only

Limits Preview/Index History Clipboard Details

Display Graphics Show: 5 Send to Text

☐ 1: **ESR1** **estrogen receptor 1** [*Homo sapiens*]
 GeneID: 2099 Locus tag: H6NC:3467; MIM: 133430 updated 19-May-2004

Transcripts and products: RefSeq below

NC_000006

[152159677 ▶ 5' 3' 152485397 ▶]
 NM_000125 - coding region - untranscribed region NP_000116

Genomic context: chromosome: 6; Maps: 6q25.1

Gene type: protein coding
Gene name: ESR1
Gene description: estrogen receptor 1
RefSeq status: Provisional
Organism: *Homo sapiens*
Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo
Gene aliases: ER; ESR; Era; ESRA; NR3A1
Summary: The estrogen receptor (ESR) is a ligand-activated transcription factor composed of several domains important for hormone binding, DNA binding, and activation of transcription. Alternative splicing results in several ESR1 mRNA transcripts, which differ primarily in their 5-prime untranslated regions. The translated receptors show less variability.[supplied by OMIM]
General protein information:
 Names: estrogen receptor 1
 estrogen receptor 1 (alpha)
Bibliography:
 PubMed links
GeneRifs:

1. The ligand-binding domain of estrogen receptor alpha has been expressed, purified, and crystallized to yield high amounts of soluble protein, in order to solve the structure in its native form without renaturation or modification steps.
2. ERα but not ERβ is present in human preadipocytes
3. Association of a T262C transition in exon 1 of estrogen-receptor-alpha gene with skeletal responsiveness to estrogen in post-menopausal women.
4. To determine whether receptor-induced changes in DNA structure are related to transactivation, we compared the abilities of ER alpha and ER beta to activate transcription and induce distortion and bending in DNA.
5. Estrogen receptors are found in brain areas involved in regulation of food intake. The anorexic effects of estrogen are accentuated by stress, thus that variation in the estrogen receptors may contribute to the genetic susceptibility to AN in females.
6. Reduction of coactivator expression by antisense oligodeoxynucleotides inhibits ERalpha transcriptional activity and MCF-7 proliferation
7. mutations targeted to predicted helix in the extreme carboxyl-terminal region alter its response to estradiol and 4-hydroxytamoxifen
8. expression of estrogen receptor alpha and estrogen receptor beta were studied in leiomyomas and homologous myometrium from women in the proliferative phase of the menstrual cycle and from women treated with a gonadotropin-releasing hormone analogue
9. Estrogen receptors play a role in the activation of amino acid transport system A by estrogen.

start Inbox - Microsoft Out... Trento Slides Gene - Mozilla Microsoft PowerPoint ... 8:47 PM

Query The DIG Data - Mozilla


File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop

https://dig.cgt.duke.edu/index.php

Search Print

Home Bookmarks Google CAGP Duke Calendar Duke Directory DUMC SAMSI mozilla.org Duke ASC WebMail DIG



Duke Integrated Genomics Database

ABOUT TRY A QUERY DATASETS FEATURES PDQ_MED DOWNLOAD

Data updated
25 Jul 2003

Login

mw

Password

Save password?

☐ No ☐ Yes

Login

(Log in will open new browser window)
v. 0.9.0

Please use a current version of Internet Explorer, Netscape, Safari, or Mozilla to access your workspace.

Human gene network based on Medline "baseline" 2003 — 18 May 2004

A network of human genes derived from the 2003 "baseline" Medline dataset is available for [download](#). The files are formatted for use in [GraphExplore](#), which is freely available for download. The human gene network is based on the co-occurrence of gene names and their variants in title and sentences within abstracts from the Medline database ([NCBI](#))

GO Quirkiness Fixed / Duplicate Results for "NM_" Genbank IDs — 04 May 2004

Queries that requested GO annotation failed if the number of gene symbols or Genbank accession numbers exceeded seven items. This failure has been fixed. If other quirkinesses in the server responses appear, please contact Mark DeLong ([delon008\[at\]mc.duke.edu](mailto:delon008[at]mc.duke.edu)).

Also, Genbank accession numbers beginning with "NM_" appear to have been duplicated, so they may appear twice in some results. This is a database build error.


About DIG

The Duke Integrated Genomics database (DIG) has been developed to provide annotation of genes identified in various types of biological experiments, including the analysis of DNA microarray data. The database integrates various gene annotation sources including UniGene, LocusLink, OMIM, and Homologene. A key feature is the ability to perform batch searches of these sources of information to generate a series of associated links to these sources of annotation for a collection of genes.

The DIG system is building comprehensive indices of the published literature. Currently we have compiled an index of gene symbols, protein symbols, and their variants in the Medline database of about 12,000,000 article abstracts. Analysis tools are also under development that digest and summarize results of these literature queries.

Searches conducted in DIG can be saved in user's private "workspaces" for review and execution at a later date. See the [screenshots](#). The Duke University installation of DIG includes features that are not available in the "public" version, including document and data sharing, work group support, and messaging among group members. The Duke research community also has access to PDQ_MED which does pairwise searches through all Medline records. This application makes it possible to retrieve published literature that links genes with one another or with user-defined terms. PDQ_MED was licensed from Inpharmix, Inc. and has been integrated within the Duke University DIG system to facilitate transfer of gene lists into the PDQ-MED search interface.

You are welcome to try a query just to see how things work. But query results are only part of the story. To use all the features of DIG, a personal account is required. If you are a Duke University researcher and want an account, contact Mark DeLong (668.1651, email: delon008@mc.duke.edu).



start

Inbox - Microsoft Out... Trento Slides Query The DIG Data ... Microsoft PowerPoint ...

8:40 PM

Molecular Information – Genome Technologies

Human condition – disease studies

Analysis of tissue, blood, etc

- Gene expression (microarrays)
- DNA copy number (CGH)
- DNA methylation
- Protein expression
- Metabolic expression

Analysis of the 'host'

- Genotypes – DNA sequence
- Serum protein expression
- Serum metabolic expression
- Serum gene expression?

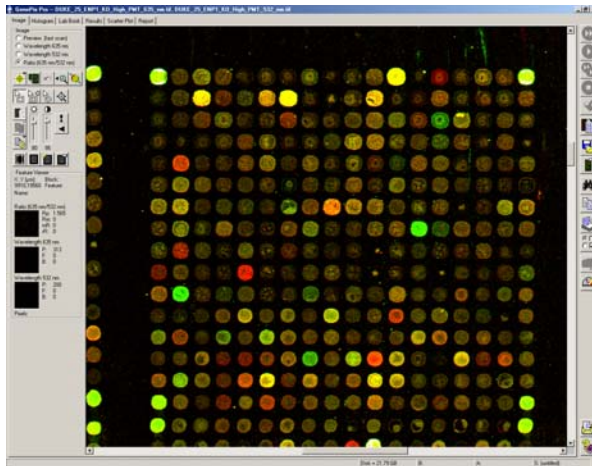
Gene Expression - DNA microarrays

High-throughput gene expression level 'snapshot'

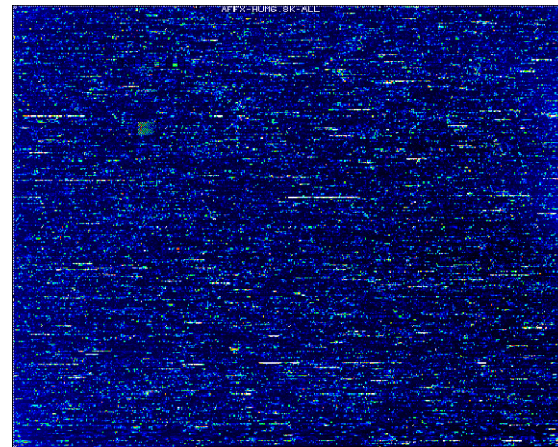
Transcript - mRNA - message

Spots : probes

- millions of copies of DNA for a gene
- nucleotide sequence
- oligonucleotides



Custom spotted array



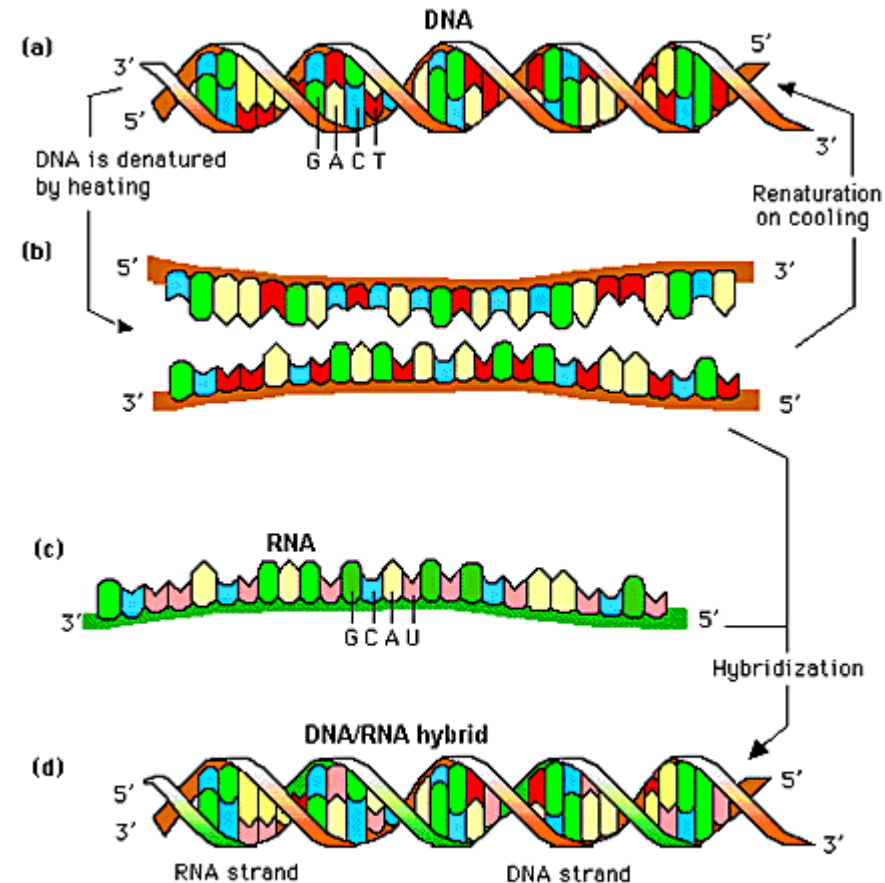
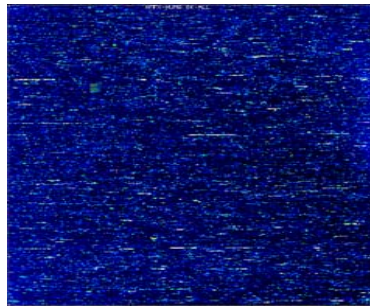
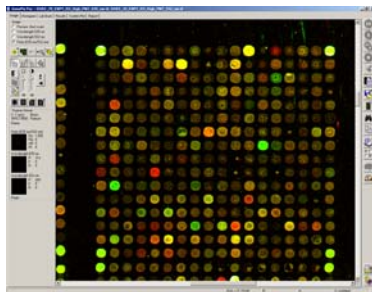
Affymetrix GeneChip

Hybridization technology

free RNA 'targets' from
biological sample hybridize to
DNA probes on array

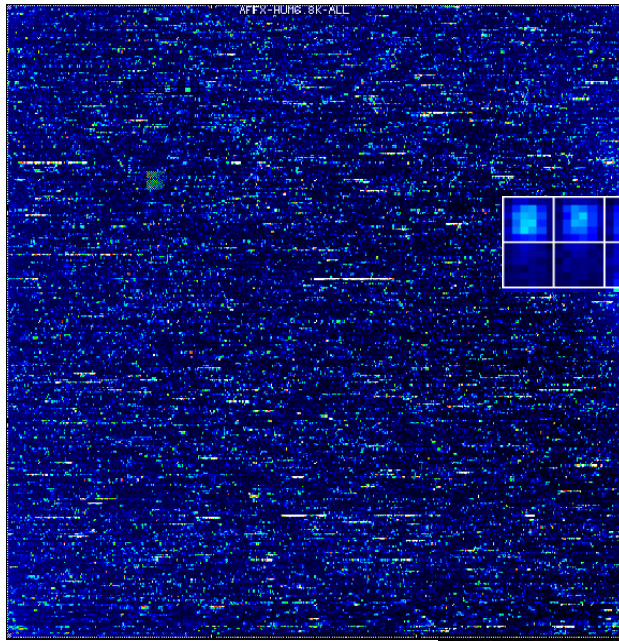
Measuring Expression

Target : fluorescent labels
Scan fluorescent intensity image

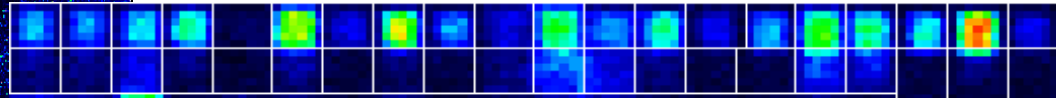


Nucleic Acid Hybridization

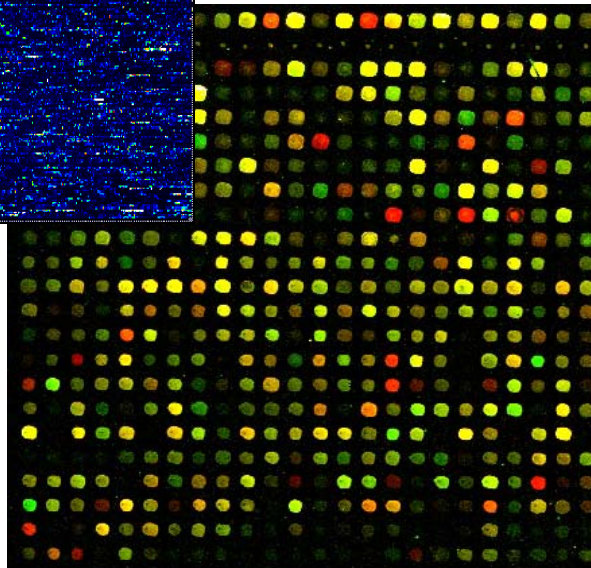
Array Fabrication



Affymetrix GeneChip arrays



mRNA target



Custom spotted arrays

cDNA target

- reverse transcription from RNA
(U → T)
- relative expression
(**Target** vs **Reference**)

Sources of Variation & Noise

Biology

Species
Strains - Genotypes
Animals - Individuals
Tissues
Tissue heterogeneity
Time points

Process

Quality of experimental sample
RNA quality
RNA extraction, amplification
Labelling effects
Hybridization process
Probe design
Cross-hybridization
Background effects
Scanning, filtering



Next up:

Biological Phenotypes