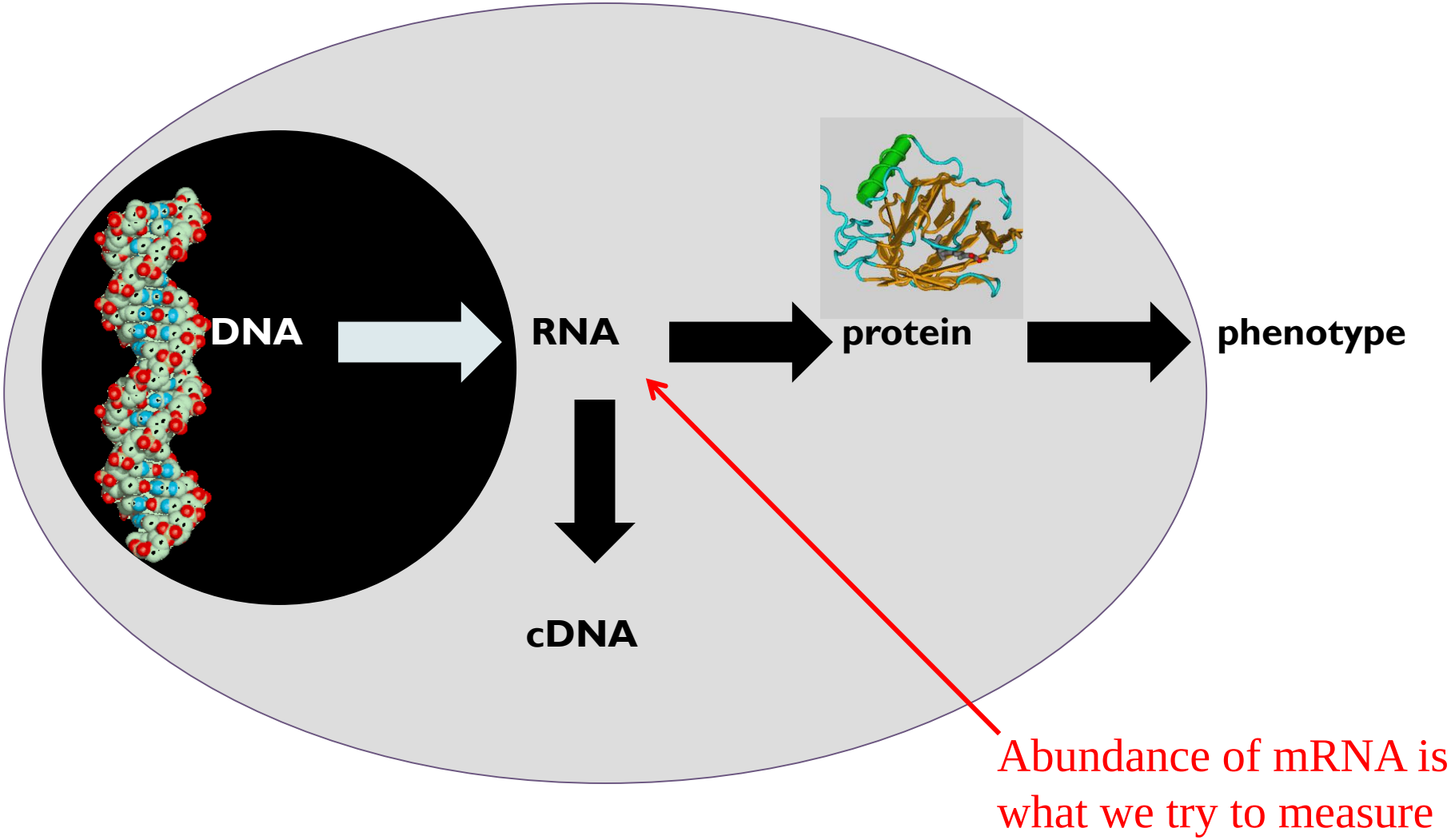
Expression Databases

Manpreet S. Katari

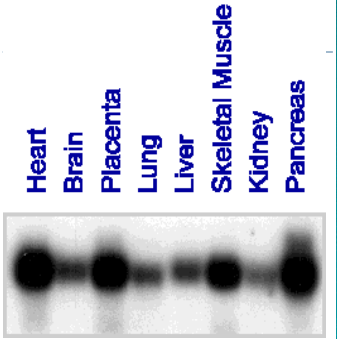

DNA
RNA
protein
phenotype
cDNA
Abundance of mRNA is
what we try to measure

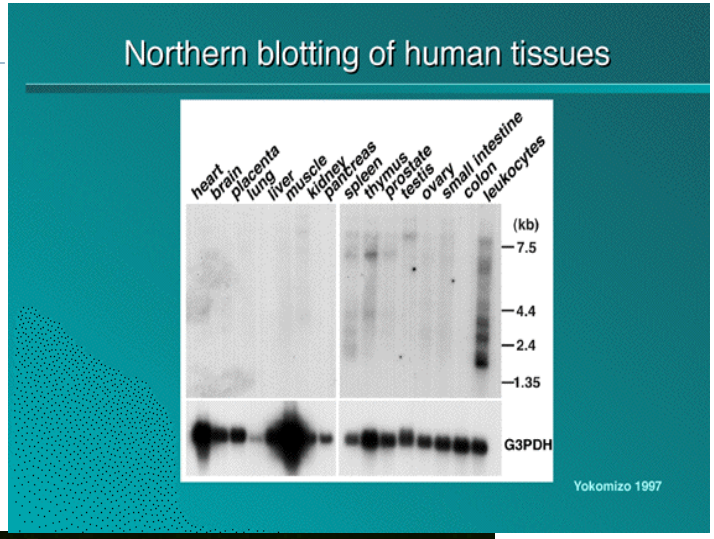
Questions that can be addressed with genome-wide expression analysis:

- What genes have similar function?
- What regulatory pathways exist?
- Can we subdivide experiments or genes into meaningful classes?
- Can we correctly classify an unknown experiment or gene into a known class?
- Can we make better treatment decisions for a cancer patient based on his or her gene expression profile?

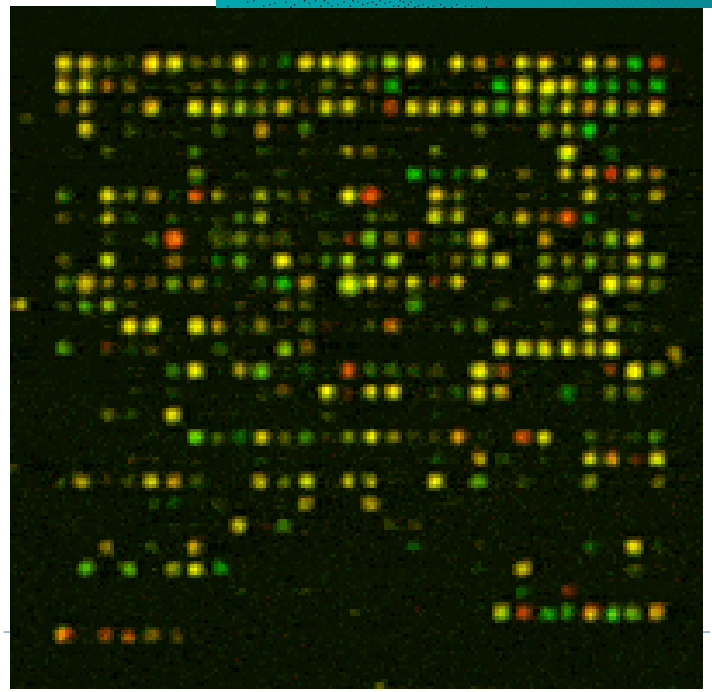
Microarrays vs Northern blots: from Gene to Genome Science

- Northern blot: limited by number of lanes in gel

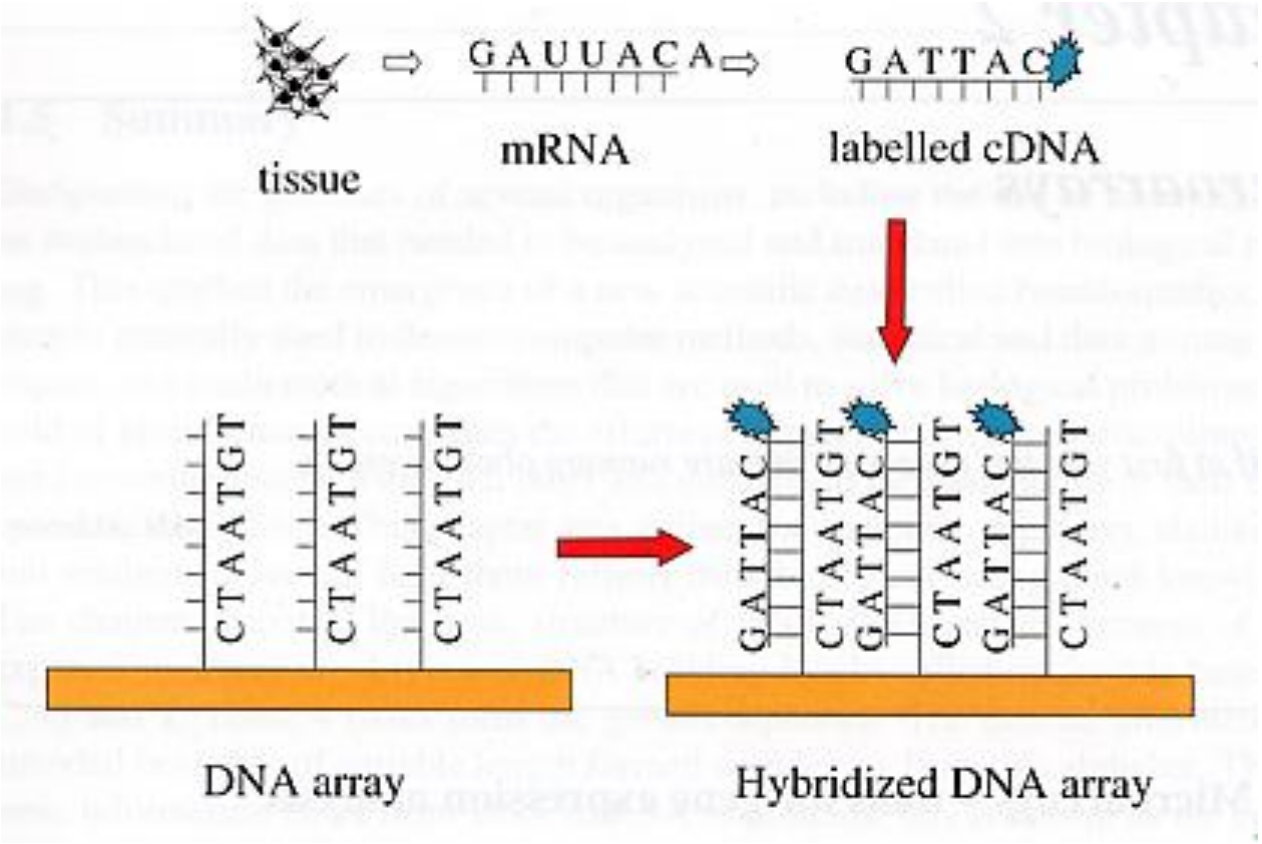




- Microarray: A large number of DNA fragments are attached in a systematic way to a solid substrate, can measure mRNA levels for thousands of genes (~ every gene in a genome) in parallel

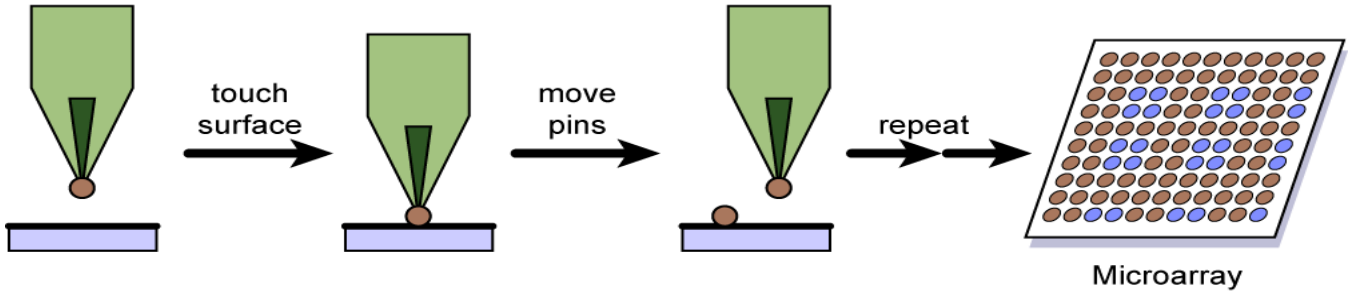


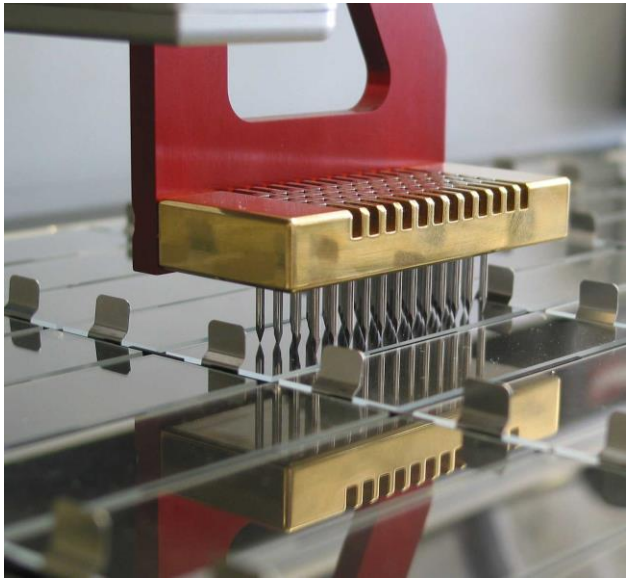
What is a microarray?



Spotted microarrays: how they are made

- DNA mechanically placed on glass slide
- Need to deliver nanoliter to picoliter volumes (too small for normal pipetting devices)
- Robot "prints," or "spots," DNA in specific places

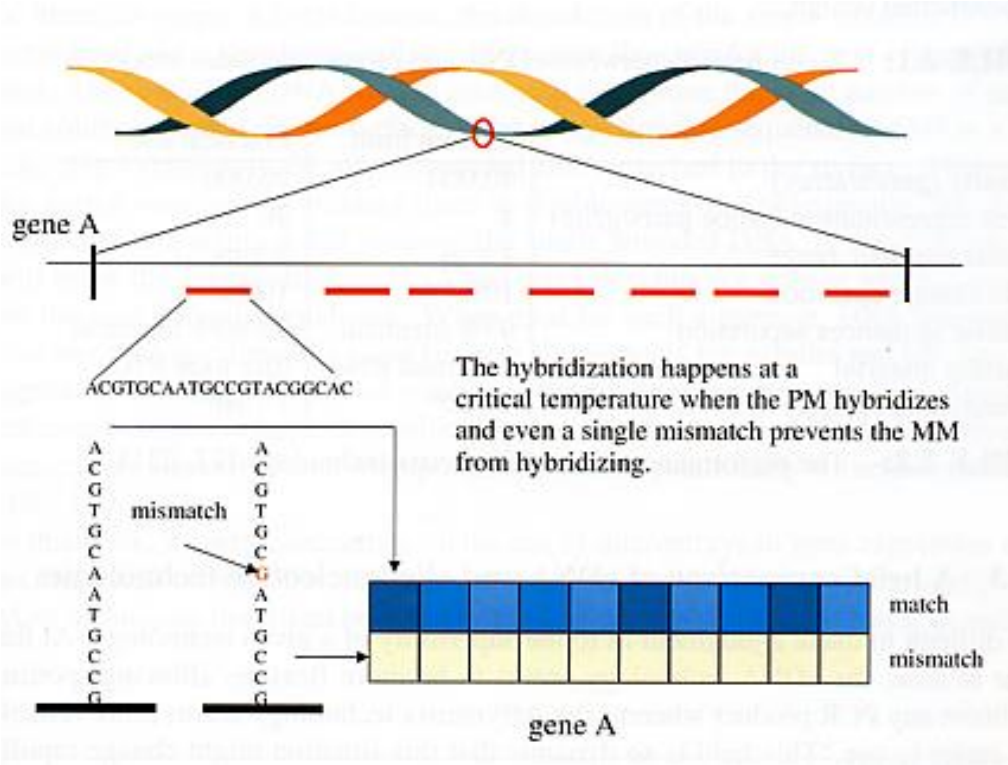




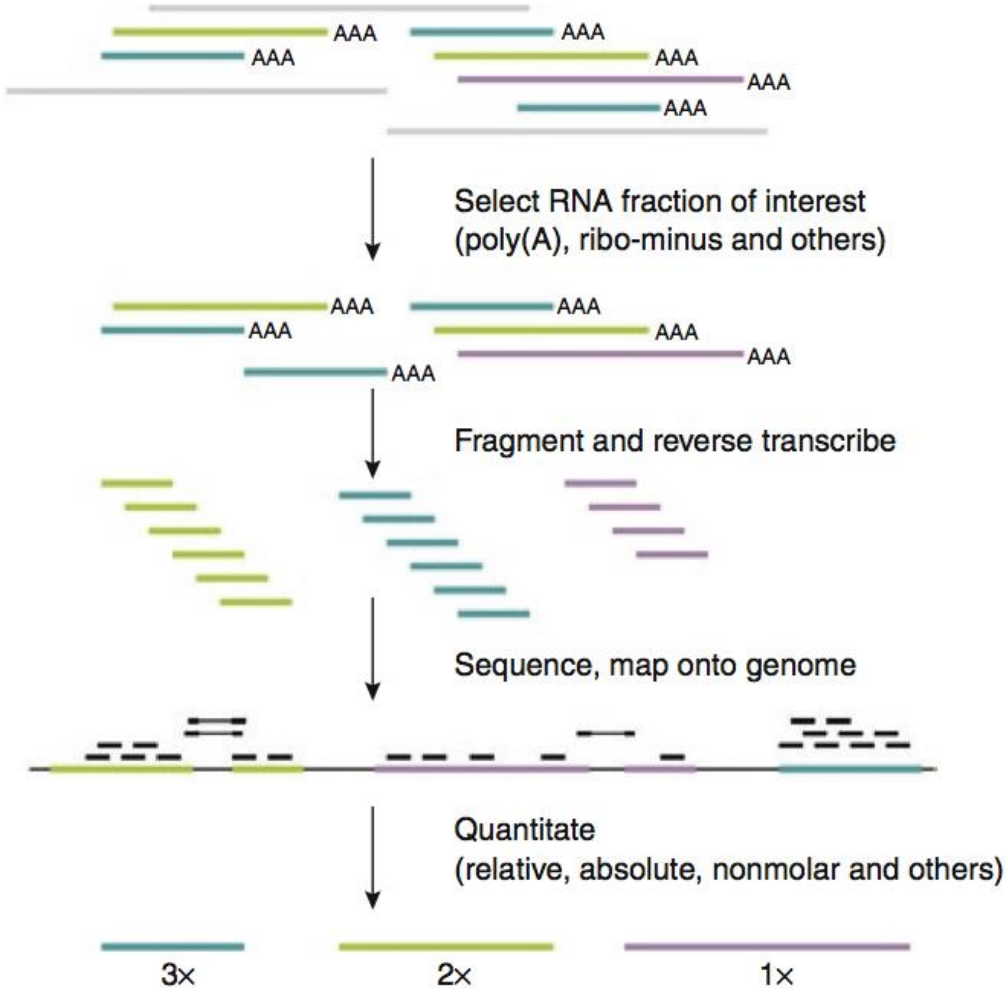
- DNA spotting usually uses multiple pins
- DNA in microtiter plate
- DNA usually PCR amplified

Affymetrix gene chip

Affymetrix Array

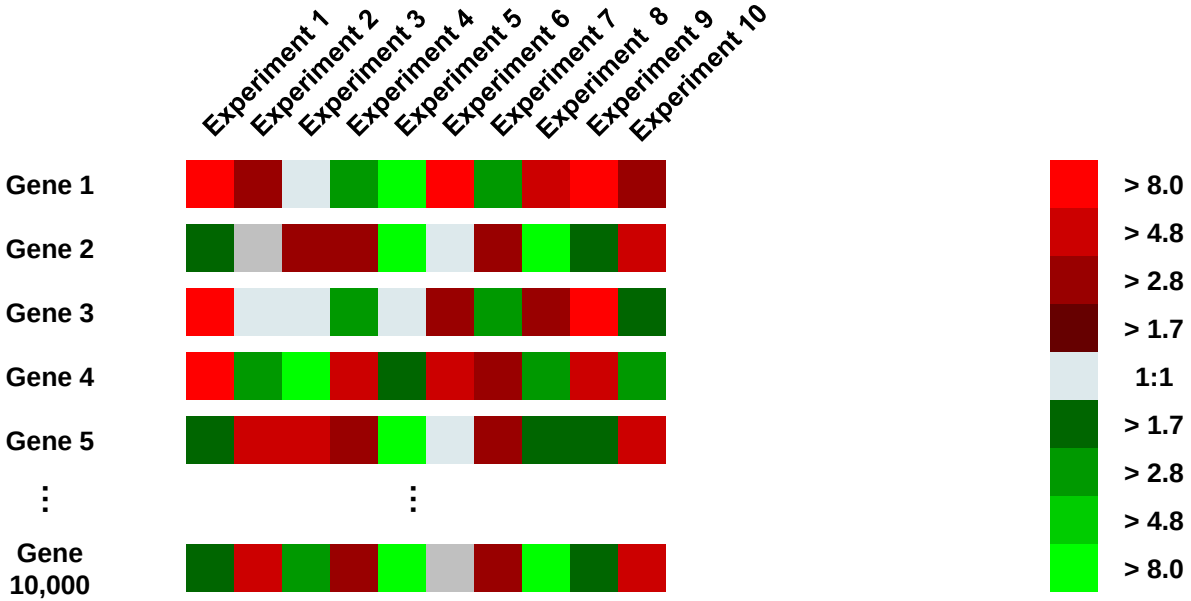


Transcriptomics using RNA-seq



Shirley Pepke[1], Barbara Wold[2] & Ali Mortazavi[2]

Gene expression can be assayed across many different conditions



A separate microarray experiment is performed using mRNA isolated from each different "condition", e.g.:

- Developmental time course
- Time course after exposure to some environmental stimulus (chemical, light/dark, etc.)
- Different tissues
- Normal vs. diseased tissue

Standard Format (MIAME)

GEO and MIAME (Minimum Information About a Microarray Experiment)

The MIAME guidelines outline the minimum information that should be included when describing a microarray experiment. Many journals and funding agencies require microarray data to comply with MIAME. GEO deposit procedures enable and encourage submitters to supply MIAME compliant data.

More information and background regarding GEO and MIAME are discussed in this **Nature Biotechnology correspondence**.

MIAME compliance is not related to the submission format or route, but rather to the content provided

The six most critical elements contributing towards MIAME are:

- The raw data for each hybridization (e.g., CEL or GPR files)
- The final processed (normalized) data for the set of hybridizations in the experiment (study) (e.g., the gene expression data matrix used to draw the conclusions from the study)
- The essential sample annotation including experimental factors and their values (e.g., compound and dose in a dose response experiment)
- The experimental design including sample data relationships (e.g., which raw data file relates to which sample, which hybridizations are technical, which are biological replicates)
- Sufficient annotation of the array (e.g., gene identifiers, genomic coordinates, probe oligonucleotide sequences or reference commercial array catalog number)
- The essential laboratory and data processing protocols (e.g., what normalization method has been used to obtain the final processed data)

http://www.ncbi.nlm.nih.gov/geo/info/MIAME.html

Common Databases

- Microarray Data
  - NCBI GEO (Gene Expression Omnibus)
  - ArrayExpress
- RNA-seq
  - NCBI SRA (Sequence Read Archive)
  - ENA (European Nucleotide Archive)

Gene Expression Omnibus Repository (GEO)

- http://www.ncbi.nlm.nih.gov/geo/

- Platform – information regarding the technology used (GPLXXXX)
- Sample – information submitted by the experimenter regarding the conditions and manipulations (GSMXXXX)
- Series – Samples are linked with series which defines the entire dataset (GSEXXX)
- GEO Dataset – GEO sample information collected by GEO staff. Can be used to compare with other datasets.

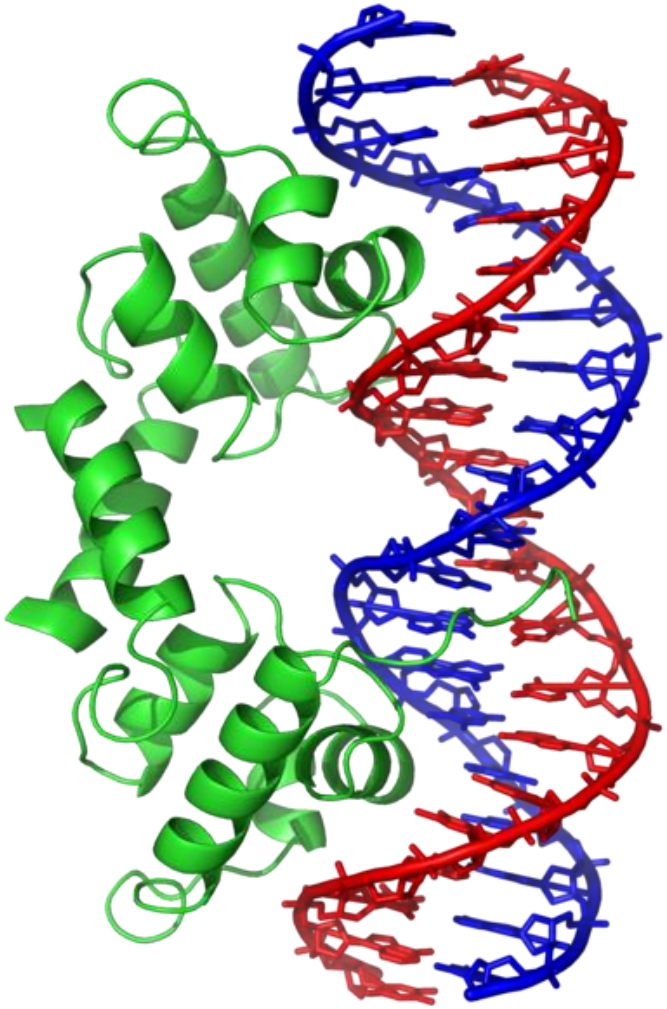
GEOquery Package

```
>source("http://bioconductor.org/biocLite.R")
>biocLite("GEOquery")
>library("GEOquery")

>gds<-getGEO("GDS2084")
>Meta(gds)
>head(Table(gds))
>Columns(gds)

>gsm<-getGEO("GSM114841")
>Meta(gsm)
>Columns(gsm)
>head(Table(gsm)
```

Different Utilities for Deep-Seq

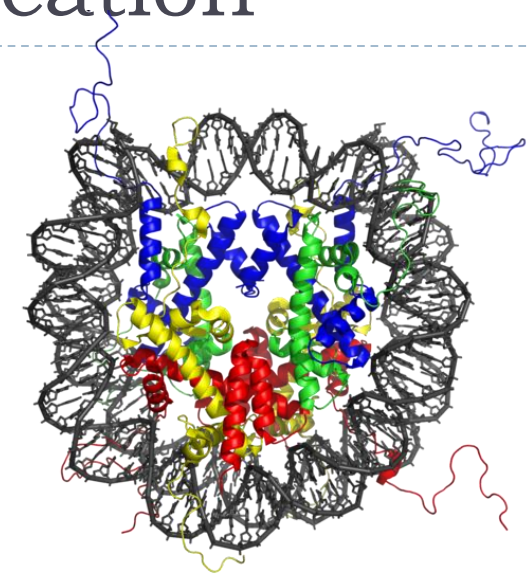
- RNA-seq
  - Sequencing equivalent to Microarray data
    - Method for discovering new RNA molecules
    - Method for quantifying RNA abundance
- CHIP-seq (Chromatin Immuno Precipitation)
  - Method for indentifying region of genome where a particular protein is binding.
    - Transcription Factors
    - Histone Modifications

Transcription Factors



http://en.wikipedia.org/wiki/DNA_binding_protein

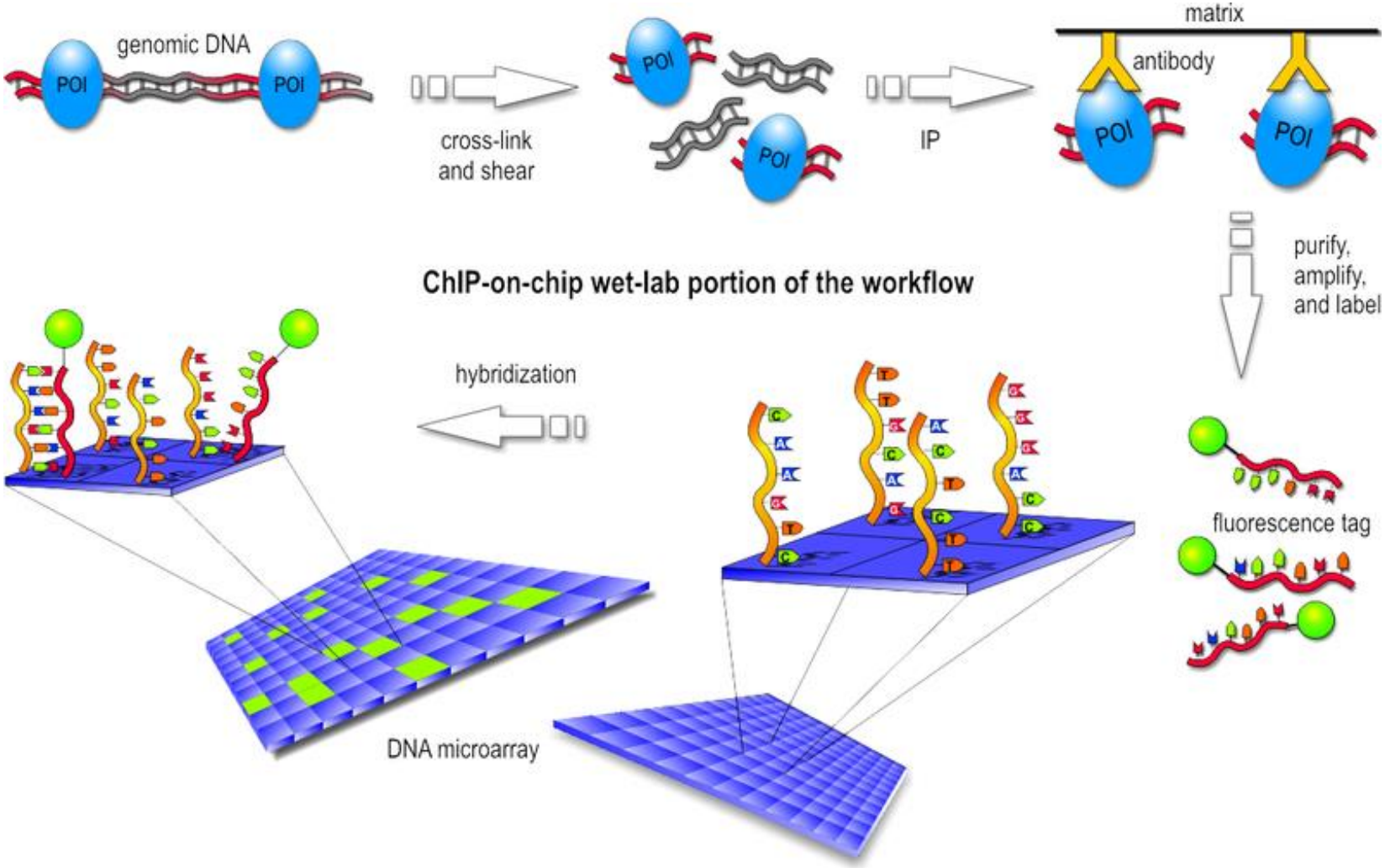
Histone Modification



Type of modification						
[11][12][13][14]	H3K4	H3K9	H3K27	H3K79	H4K20	H2BK5
monomethylation	activation[12]	activation[11]	activation[11]	activation[11][13]	activation[11]	activation[11]
dimethylation				activation[13]		
trimethylations	activation[14]	repression[11]	repression[11]	repression[11]activation[13]		
	H3K9	H3K14				
acetylation	activation[14]	activation[14]				

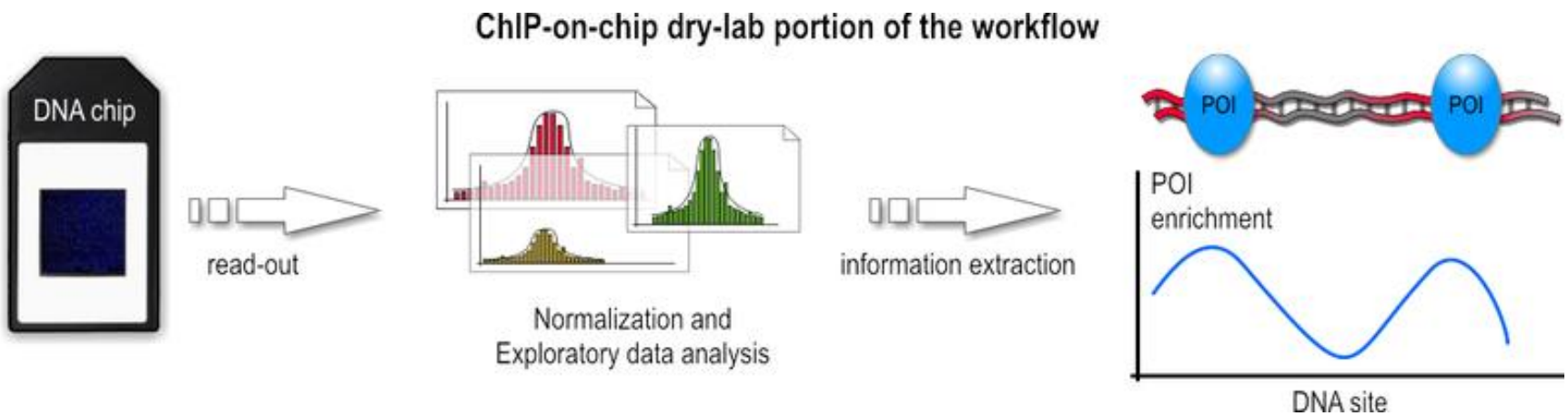
http://en.wikipedia.org/wiki/Nucleosome

CHIP-chip

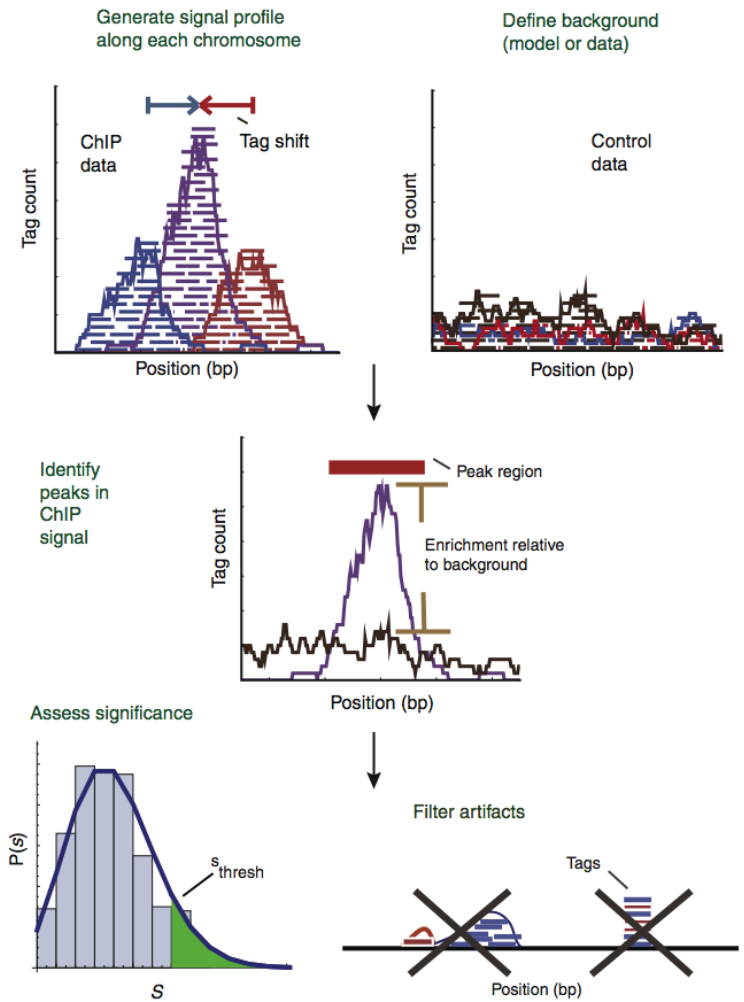


http://en.wikipedia.org/wiki/ChIP-on-chip

Regions of the genome that are enriched for sequences are areas where protein interacts

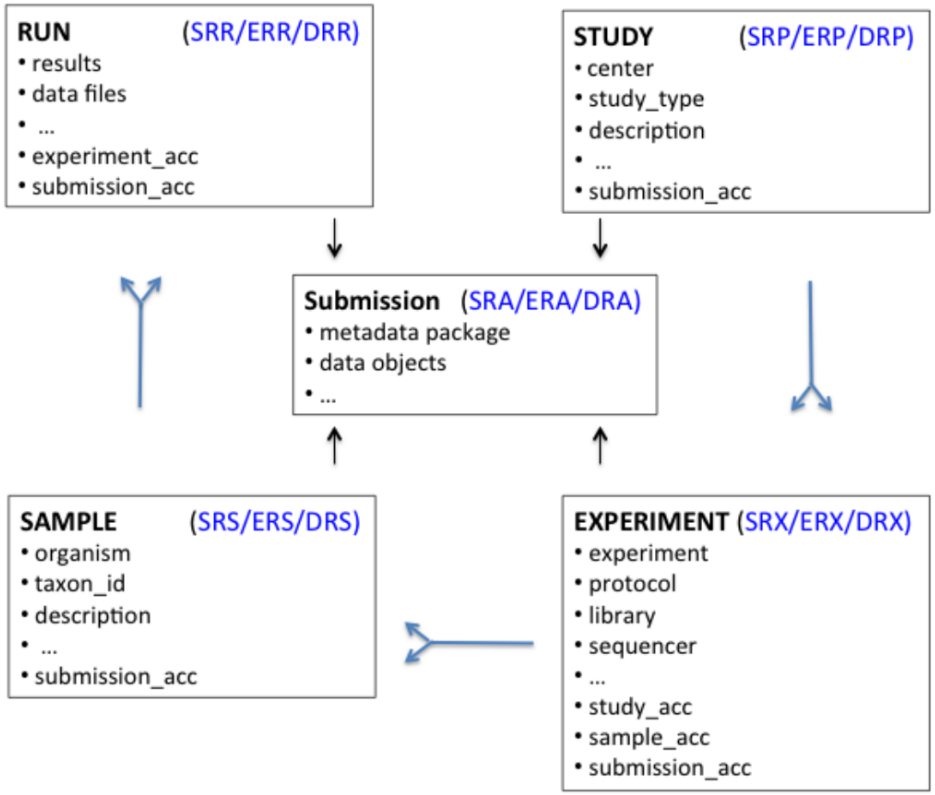


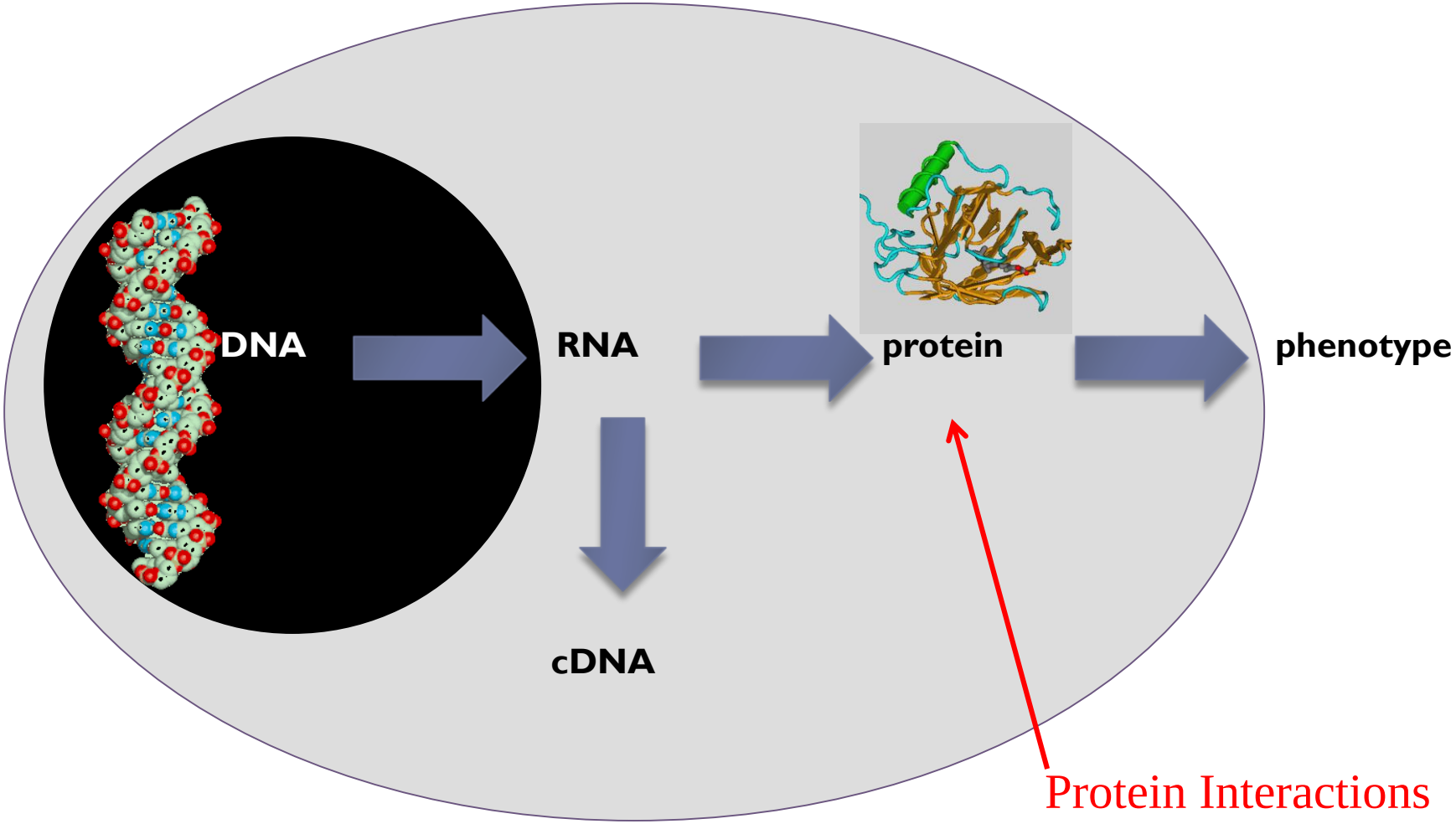
CHIP-seq: same idea but sequencing instead of hybridization.



Connecting to the SRA database

- SRA database : http://www.ncbi.nlm.nih.gov/sra




DNA
RNA
protein
phenotype
cDNA
Protein Interactions

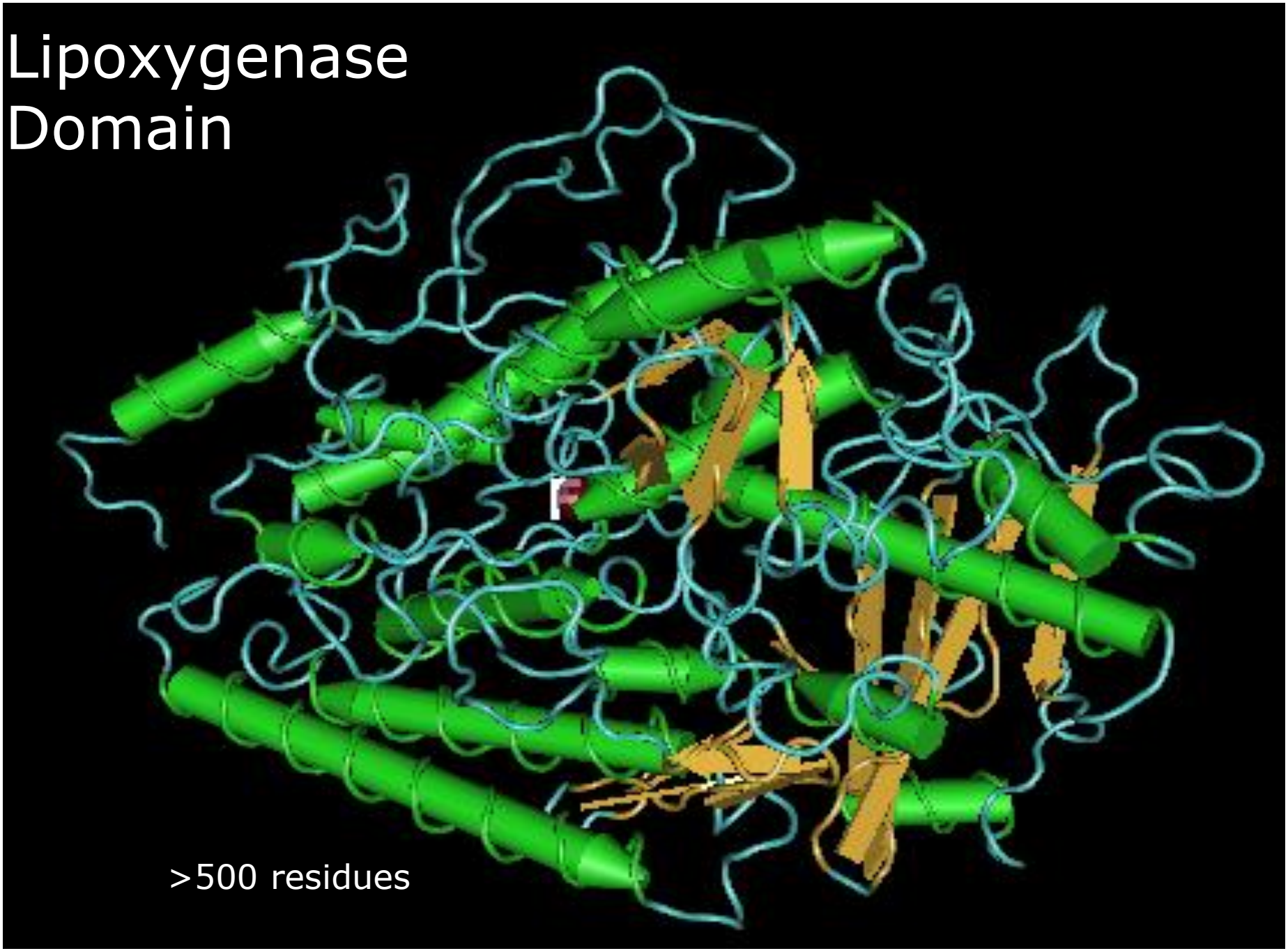
Protein Structural Elements

- 2° Structural Elements
  - α-Helix
  - β-Sheet
  - Globular regions
- Domains
  - SH2
  - Leucine Zipper

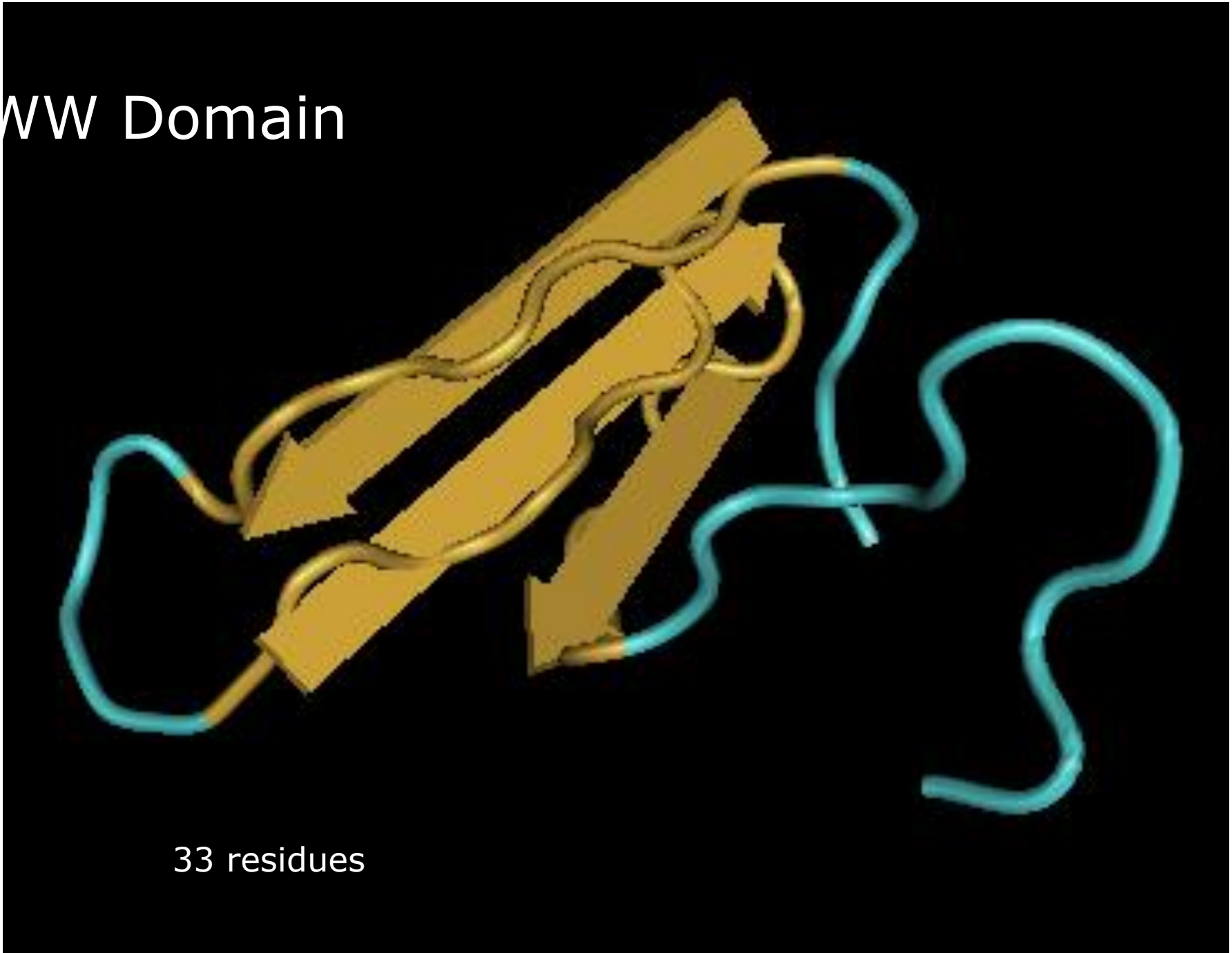
Domains

- Discrete structural units
- Can infer boundaries from sequence analysis
- 25 – 500 residues long
- Most < 200 residues
- Less than 50 residues usually stabilized by S–S bonds or metal ions

Lipoxygenase Domain



WW Domain



33 residues

Domain Determination

- Internal duplications
  - Detect with a dotplot
- Transmembrane segments
  - Hydrophobic, 15–35 residues
  - Segments easy to predict
  - Topology and multiple segments harder to predict
  - PHD, TMHMM, TMpred
- Low complexity segments
  - Composition typically "non-random"
  - Non-compact folds: coiled coils, rods, flexible domain linkers
  - Complexity function (SEG)
  - Small-pitch overlapping repeats (XNU)

Protein Sequence Databases

- GenPept
- Swiss-Prot
- TrEMBL

Protein Domain Databases

- Pfam
- PROSITE
- BLOCKS
- PRINTS
- CDD
- ProDom
- SMART
- InterPro

- HMM family profiles constructed by hand
- Structural data in alignments
- No hierarchy
- No specific compositional bias
- Good graphical output

Interproscan

EMBL-EBI | Enter Text Here | Find | Help | Feedback

Databases | Tools | Research | Training | Industry | About Us | Help | Site Index

- InterProScan
  - Help
  - Programmatic Access
- Download

- InterPro
  - Text Search
  - Databases
  - Documentation
  - FTP Site

- Database Information
  - UniProt
  - UniParc

- Similar Applications

InterProScan related literature

Search for InterProScan related literature in Medline...
more

EBI › Tools › Protein Functional Analysis › InterProScan

InterProScan Sequence Search

This form allows you to query your sequence against InterPro. For more detailed information see the documentation for the perl stand-alone InterProScan package (Readme file or FAQ's), or the InterPro user manual or help pages.

Use this tool

STEP 1 - Enter your input sequence

Enter or paste a **PROTEIN** sequence in any supported format:

Or, upload a file: Browse...

STEP 2 - Select the applications to run

Select All Clear All

☑ BlastProDom	☑ FPrintScan	☑ HMMPIR	☑ HMMPfam	☑ HMMSmart
☑ HMMTigr	☑ ProfileScan	☑ HAMAP	☑ PatternScan	☑ SuperFamily
☑ SignalPHMM	☑ TMHMM	☑ HMMPanther	☑ Gene3D	

STEP 3 - Submit your job

☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

If you plan to use these services during a course please contact us.

- http://www.ebi.ac.uk/Tools/pfa/iprscan/

BRCA1 protein sequence from NCBI in FASTA format

Display Settings: FASTA

BRCA1 [Homo sapiens]

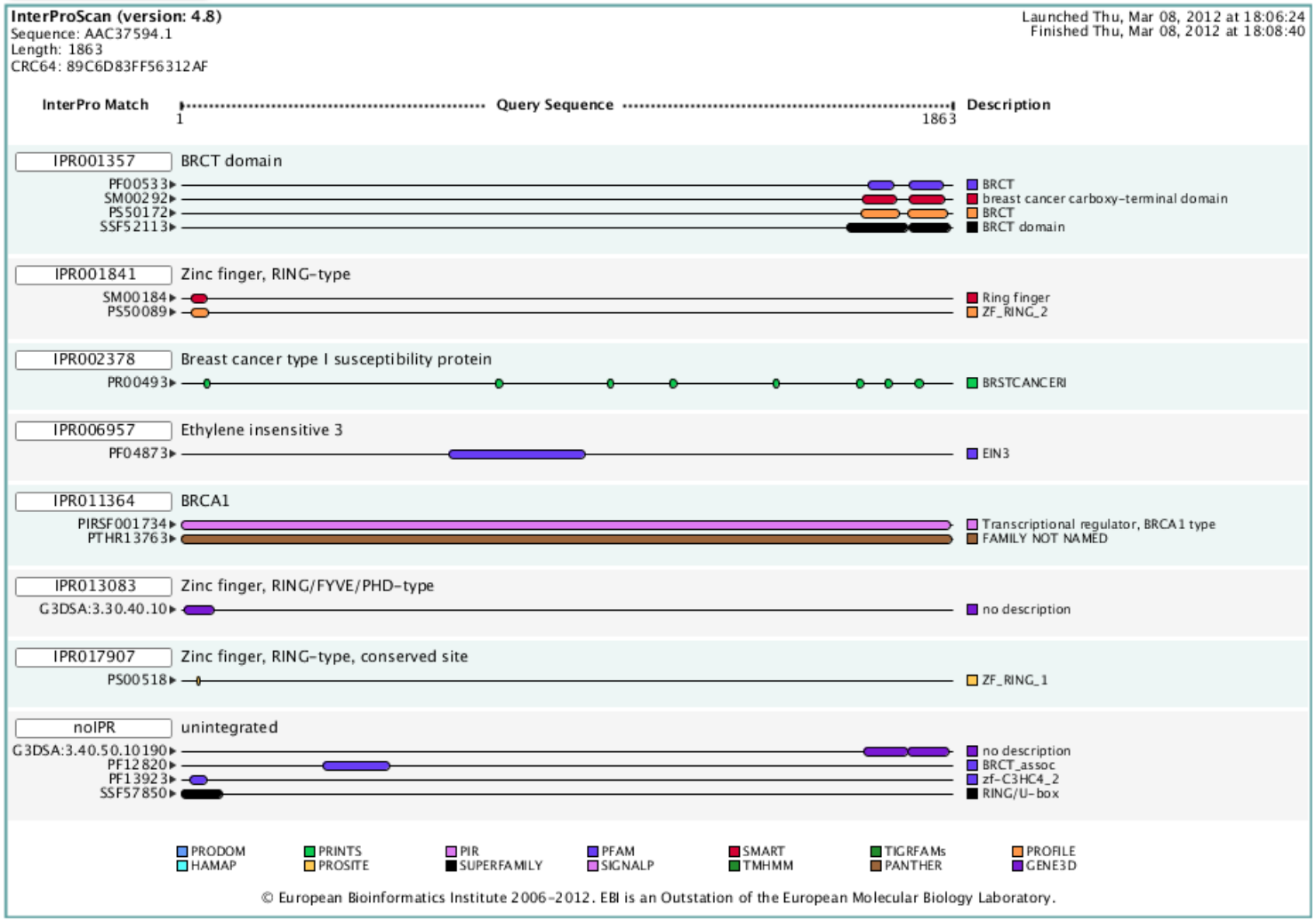
GenBank: AAC37594.1

GenPept Graphics

```
>gi|1698399|gb|AAC37594.1| BRCA1 [Homo sapiens]
MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIFCKFCMLKLLNQKKGPSQCPLCKNDITK
RSLQESTRFSQLVEELLKIICAFQLDTGLEYANSYNFAKKENNSPEHLKDEVSIIQSMGYRNRAKRLLQS
EPENPSLQETSLSVQLSNLGTVRTLRTKQRIQPQKTSVYIELGSDSSEDTVNKATYCSVGDQELLQITPQ
GTRDEISLDSAKKAACEFSETDVTNTEHHQPSNNDLNTTEKRAAERHPEKYQGSSVSNLHVEPCGTNTHA
SSLQHENSSLLLTKDRMNVEKAEFCNKSKQPGLARSQHNRWAGSKETCNDRRTPSTEKKVDLNADPLCER
KEWNKQKLPCSENPRDTEDVPWITLNSSIQKVNEWFSRSDELLGSDDSHDGESESNAKVADVLDVLNEVD
EYSGSSEKIDLLASDPHEALICKSERVHSKSVESNIEDKIFGKTYRKKASLPNLSHVTENLIIGAFVTEP
QIIQERPLTNKLKRKRRPTSGLHPEDFIKKADLAVQKTPEMINQGTNQTEQNGQVMNITNSGHENKTKGD
SIQNEKNPNPIESLEKESAFKTKAEPISSSISNMELELNIHNSKAPKKNRLRRKSSTRHIHALELVVSRN
LSPPNCTELQIDSCSSSEEIKKKKYNQMPVRHSRNLQLMEGKEPATGAKKSNKPNEQTSKRHDSDTFPEL
KLTNAPGSFTKCSNTSELKEFVNPSLPREEKEEKLETVKVSNNAEDPKDLMLSGERVLQTERSVESSSIS
LVPGTDYGTQESISLLEVSTLGKAKTEPNKCVSQCAAFENPKGLIHGCSKDNRNDTEGFKYPLGHEVNHS
RETSIEMEESELDAQYLQNTFKVSKRQSFAPFSNPGNAEEECATFSAHSGSLKKQSPKVTFECEQKEENQ
GKNESNIKPVQTVNITAGFPVVGQKDKPVDNAKCSIKGGSRFCLSSQFRGNETGLITPNKHGLLQNPYRI
PPLFPIKSFVKTKCKKNLLEENFEEHSMSPEREMGNENIPSTVSTISRNNIRENVFKEASSSNINEVGSS
TNEVGSSINEIGSSDENIQAELGRNRGPKLNAMLRLGVLQPEVYKQSLPGSNCKHPEIKKQEYEEVVQTV
NTDFSPYLISDNLEQPMGSSHASQVCSETPDDLLDDGEIKEDTSFAENDIKESSAVFSKSVQKGELSRSP
SPFTHTHLAQGYRRGAKKLESSEENLSSEDEELPCFQHLLFGKVNNIPSQSTRHSTVATECLSKNTEENL
LSLKNSLNDCSNQVILAKASQEHHLSEETKCSASLFSSQCSELEDLTANTNTQDPFLIGSSKQMRHQSES
QGVGLSDKELVSDDEERGTGLEENNQEEQSMDSNLGEAASGCESETSVSEDCSGLSSQSDILTTQQRDTM
QHNLIKLQQEMAELEAVLEQHGSQPSNSYPSIISDSSALEDLRNPEQSTSEKAVLTSQKSSEYPISQNPE
GLSADKFEVSADSSTSKNKEPGVERSSPSKCPSLDDRWYMHSCSGSLQNRNYPSQEELIKVVDVEEQQLE
ESGPHDLTETSYLPRQDLEGTPYLESGISLFSDDPESDPSEDRAPESARVGNIPSSTSALKVPQLKVAES
AQSPAAAHTTDTAGYNAMEESVSREKPELTASTERVNKRMSMVVSGLTPEEFMLVYKFARKHHITLTNLI
TEETTHVVMKTDAEFVCERTLKYFLGIAGGKWVVSYFWVTQSIKERKMLNEHDFEVRGDVVNGRNHQGPK
RARESQDRKIFRGLEICCYGPFTNMPTDQLEWMVQLCGASVVKELSSFTLGTGVHPIVVVQPDAWTEDNG
FHAIGQMCEAPVVTREWVLDSVALYQCQELDTYLIPQIPHSHY
```

- http://www.ncbi.nlm.nih.gov/protein/1698399?report=fasta

Interproscan results

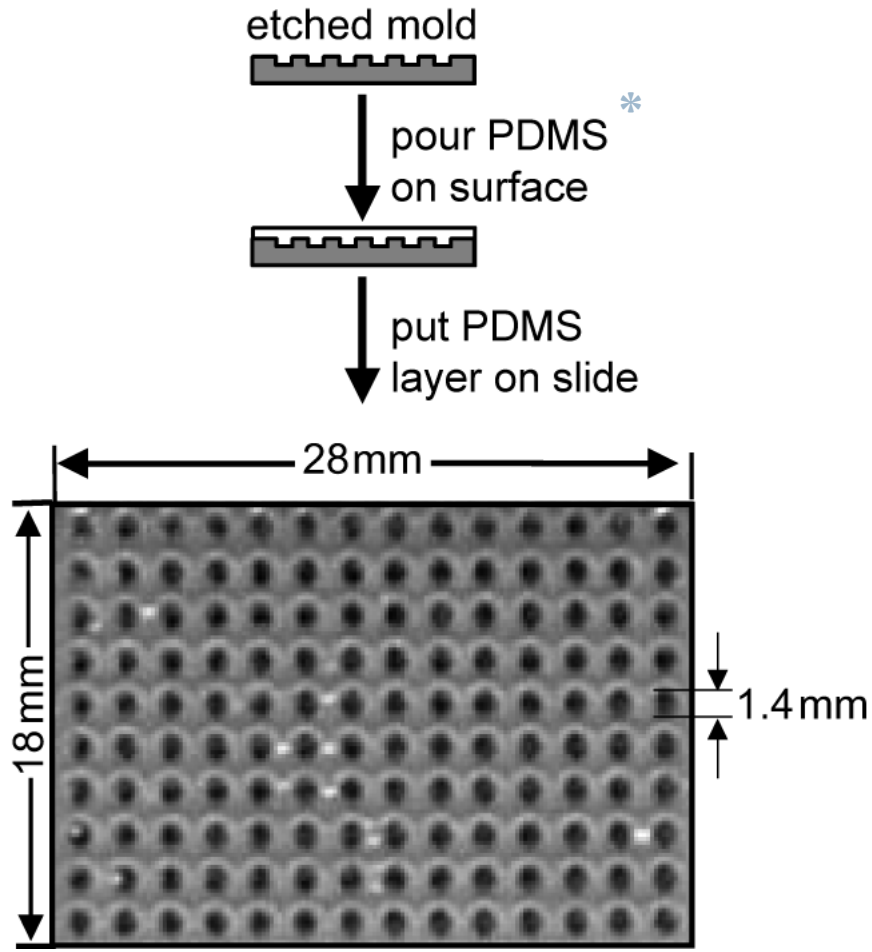


Protein Interaction Databases

- Generating Data
  - Protein chips
  - Y2H
- Databases
  - String
  - BioGRID

Fabricating protein chips

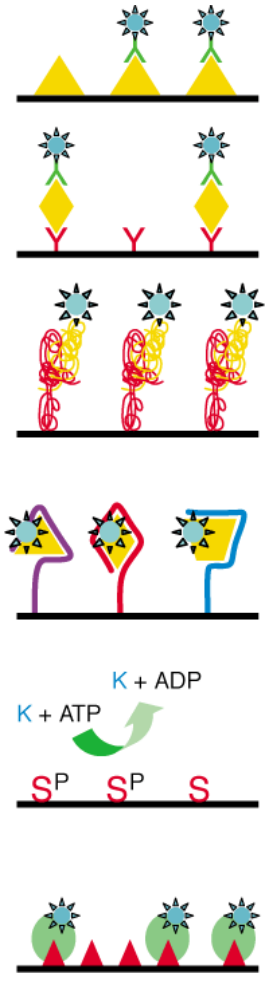
- Protein substrates
  - Polyacrylamide or agarose gels
  - Glass
  - Nanowells
- Proteins deposited on chip surface by robots



* ***polydimethoxylsiloxane***
flexible silicon-based polymer (elastomer)

Classes of capture molecules

- Different capture molecules must be used to study different interactions
- Examples
  - Antibodies (or antigens) for detection
  - Proteins for protein-protein interaction
  - Enzyme-substrate for biochemical function



Antigen–antibody

Protein–protein

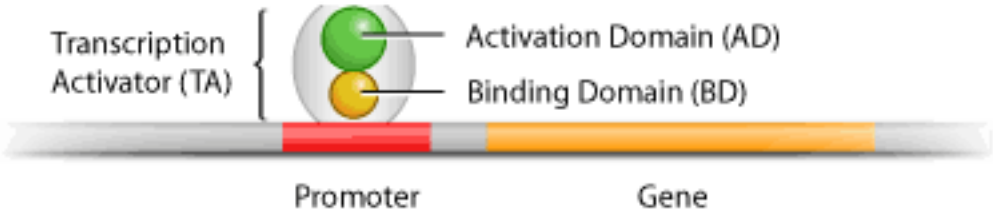
Aptamers

Enzyme–substrate

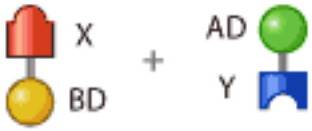
Receptor–ligand

The yeast two-hybrid system (Y2H)

A two-domain transcriptional activator



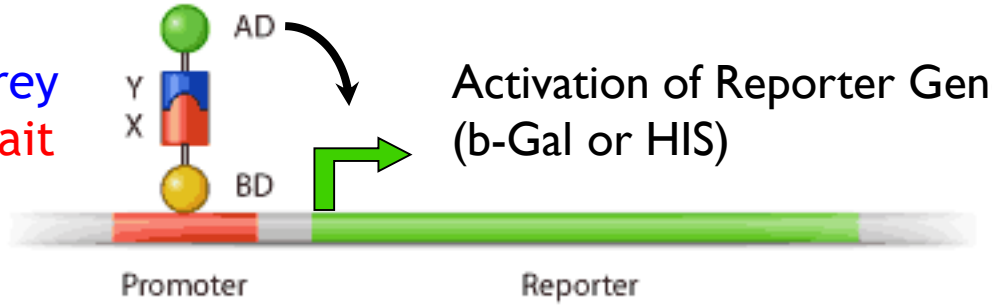
"Bait" gene X fused to BD



"Prey" gene Y fused to AD

When Bait & Prey Bind

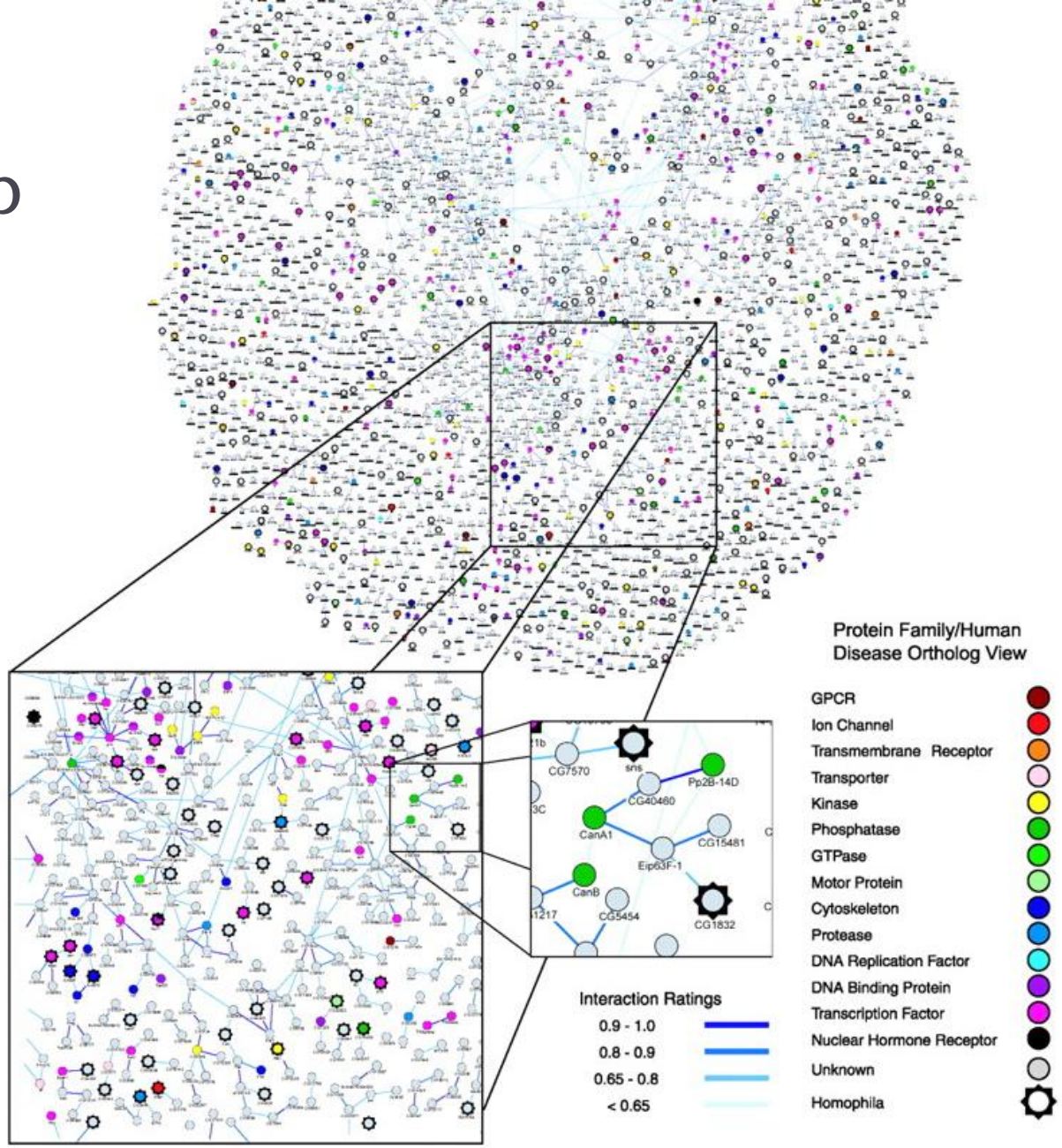
Prey
Bait



Activation of Reporter Gene (b-Gal or HIS)

If X and y physically interact, BD and AD are brought together and can activate transcription of a "reporter" gene (such as ¤-galactosidase or an auxotrophic selection marker (e.g. HIS))

Drosophila interaction map



BIOGRID: searching for BRCA1 interactions

- http://thebiogrid.org

BRCA1 *Homo sapiens*

PPP1R53, RNF53, IRIS, BRCC1, PSCP, PNCA4, BRCAI, BROVCA1

breast and ovarian cancer sususceptibility protein 1

GO Process: 37 Terms | **GO Function**: 10 Terms | **GO Component**: 12 Terms

EXTERNAL DATABASE LINKOUTS

HGNC | Ensembl | VEGA | HPRD | OMIM | Entrez Gene | RefSEQ | GenBank | UniprotKB

Download 198 Associations For This Protein

Stats & Filters

Current Statistics **Publications:** 204

High Throughput		Low Throughput
9 (2%)	570 Physical Interactions	561 (98%)
0 (0%)	19 Genetic Interactions	19 (100%)

Search Filters Customize how your results are displayed...

No Filter: Show All Associations

Switch View: Summary | Sortable Table

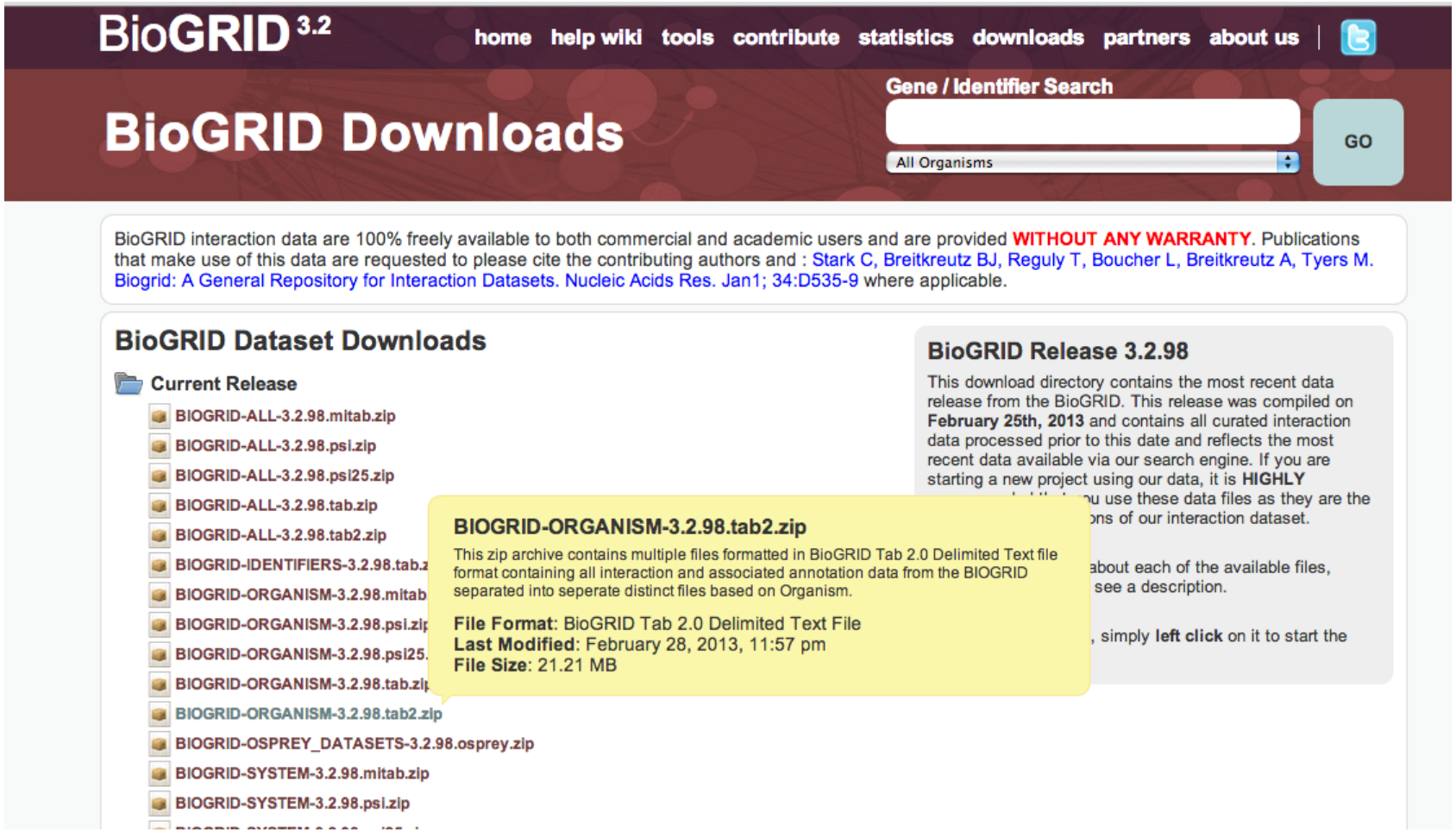
Displaying **198** total unique interactors

BARD1 51 [details]
BRCA1 associated RING domain 1 isoform epsilon

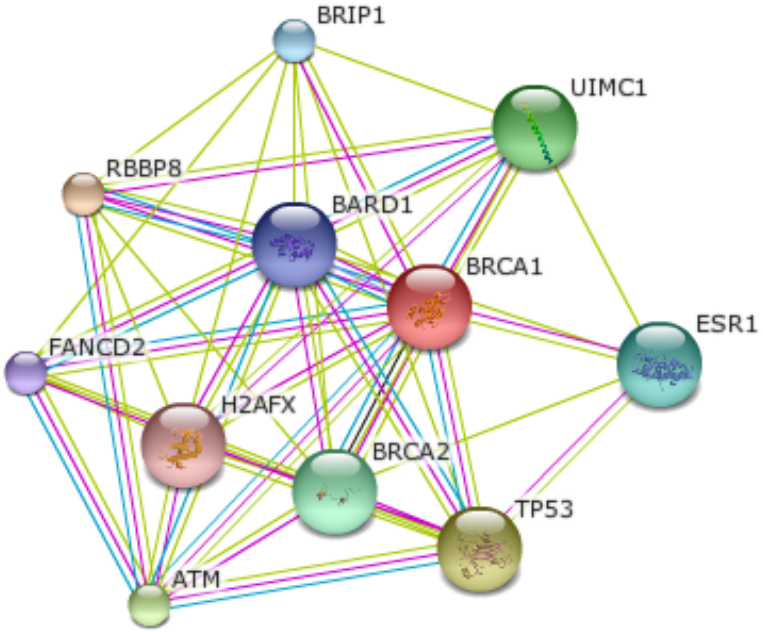
BRIP1 | BACH1, FANCJ, OF 18 [details]
BRCA1 interacting protein C-terminal helicase 1

RBBP8 | RIM, CTIP, SAE2 17 [details]
sporulation in the absence of SPO11 protein 2 homolog

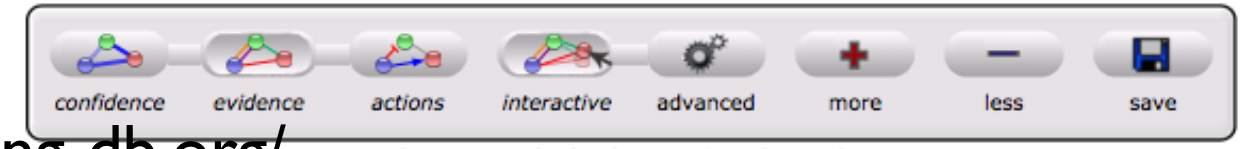
Download interactions



String: searching for BRCA1 interactions



This is the **evidence view**. Different line colors represent the types of evidence for the association.



(requires Flash player 10 or better)

- http://string-db.org/