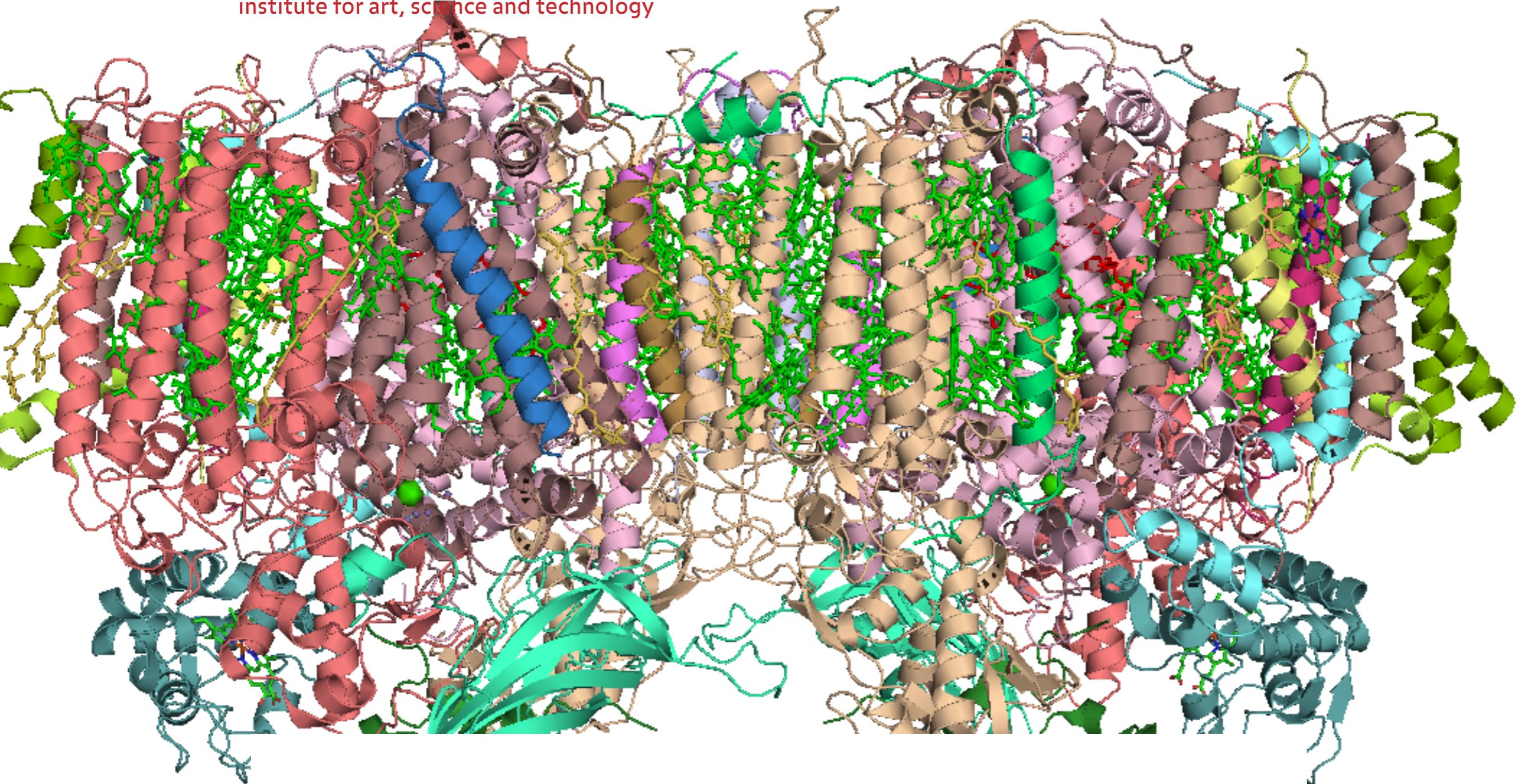




waag society

institute for art, science and technology



Bio Informatics

- US US

Neveu,Curtis CC-BY-SA 3.0



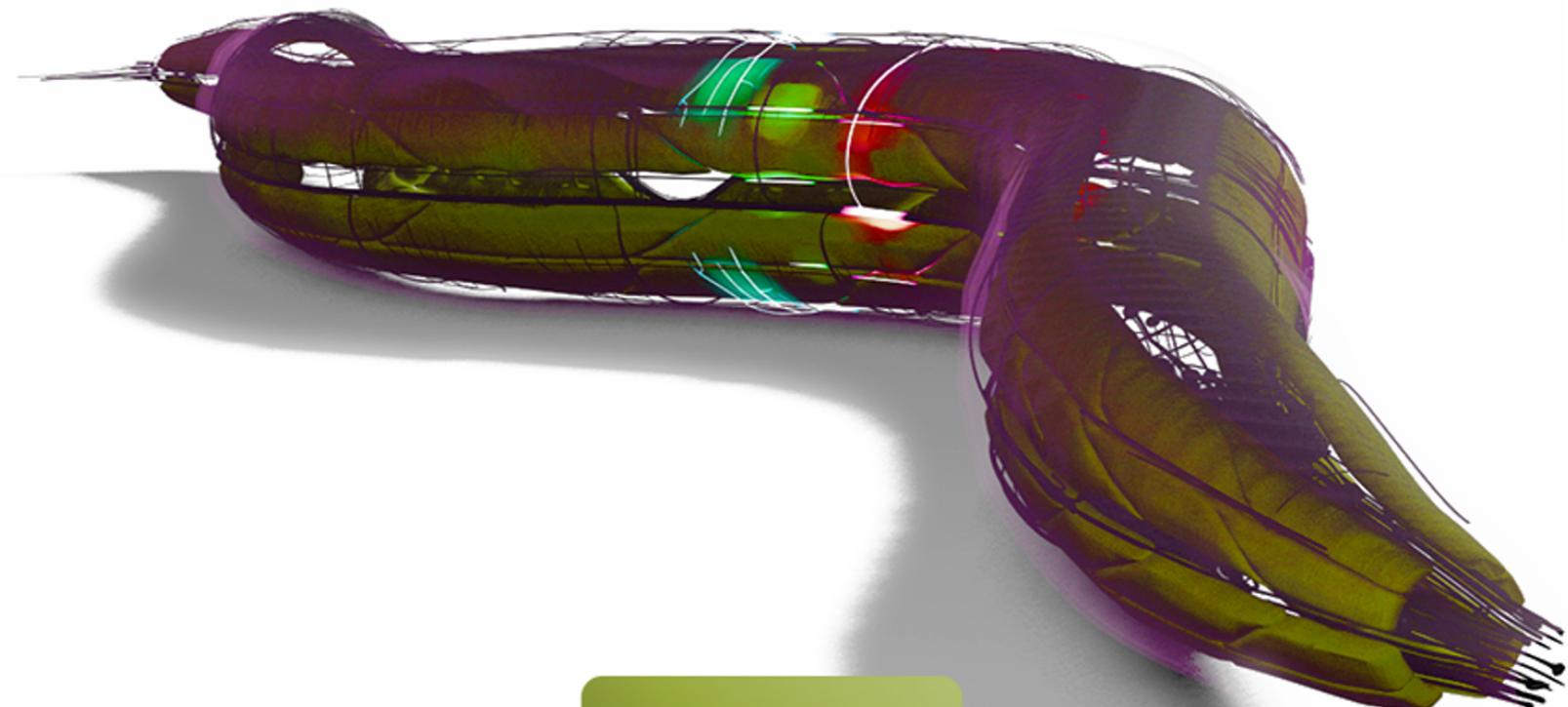
The information / “Omics” age

- “Genomics” DNA sequence analysis
- “Transcriptomics” DNA expression analysis
- “Proteomics” Protein (structure) prediction / analysis
- “Interactomics” Protein – Protein, DNA – Protein interaction
- “Metabolomics” Metabolism modeling



What is it used for

- Optimizing yield
- Predicting organisms behaviour
- Medical diagnostics
 - Personal medicine
- Drug discovery





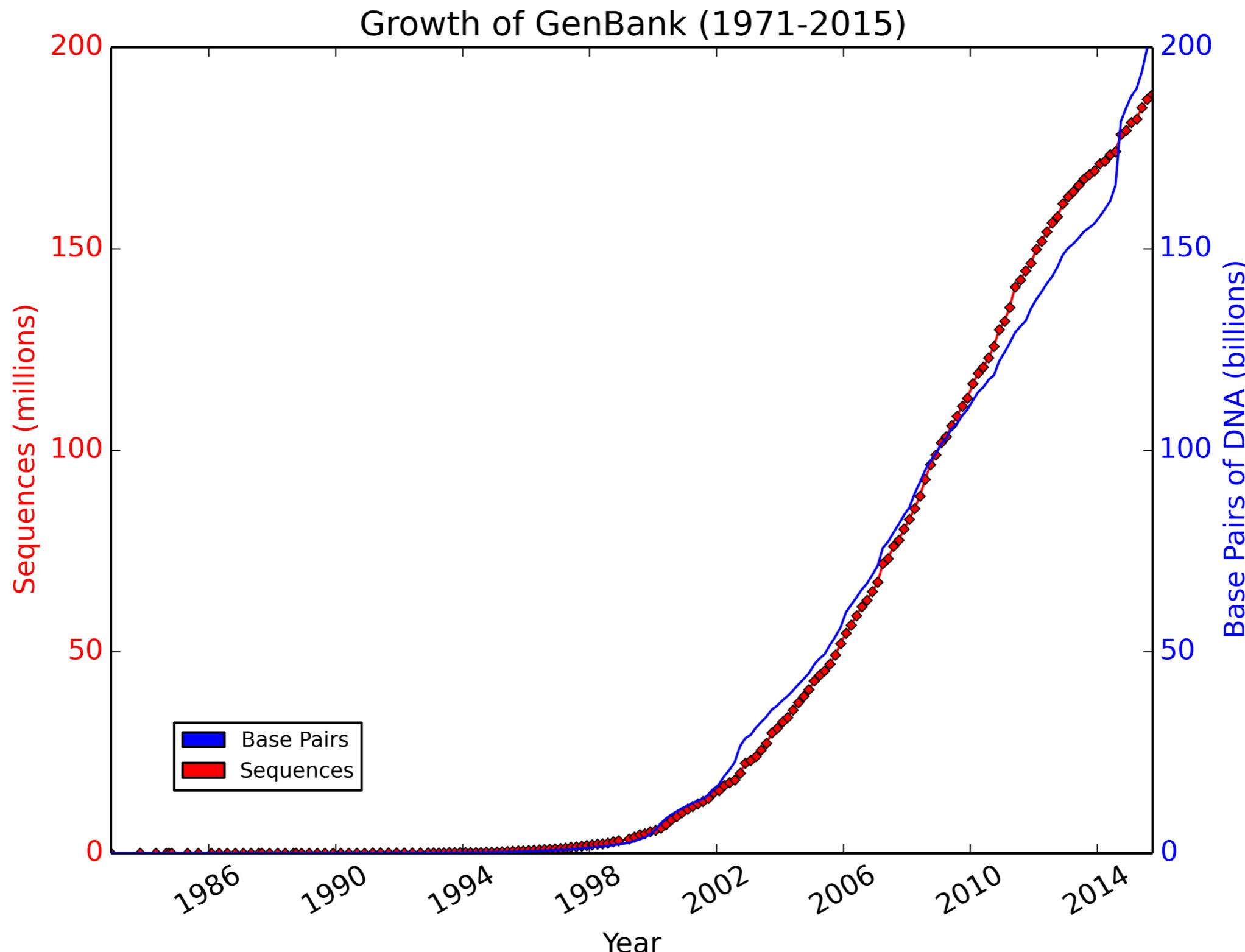
Genomics

- Functional genomics
- Metagenomics
- Personal Genomics
- Epigenomics





DNA database GenBank





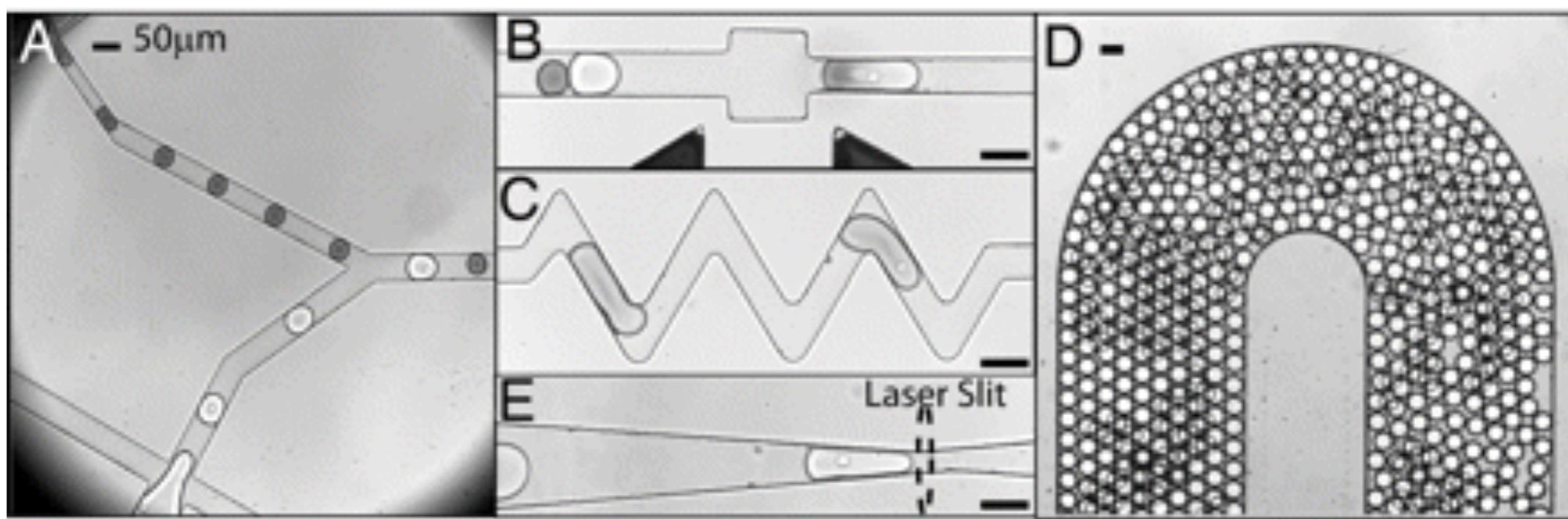
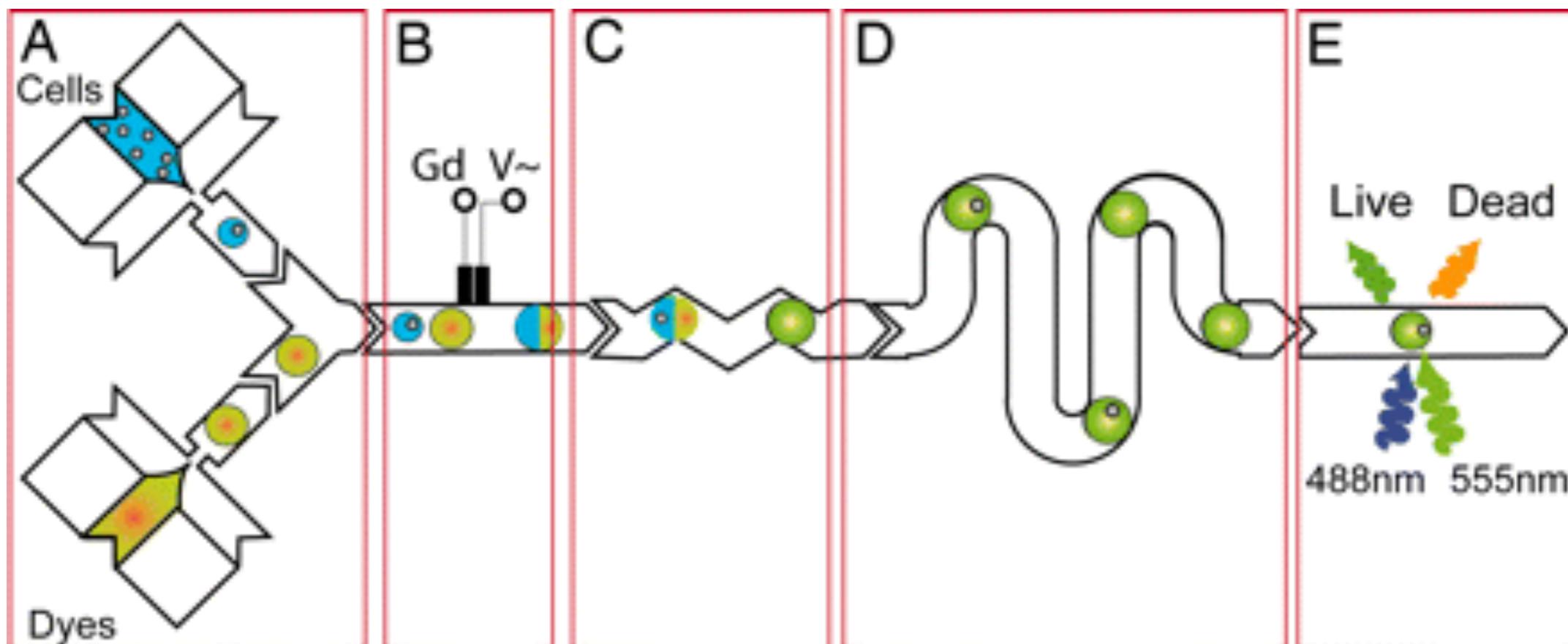
Drivers

- “High Throughput Research”
 - Robotics
 - Databases
 - Visualisation
- Public tools
- Open data



High Throughput Screening

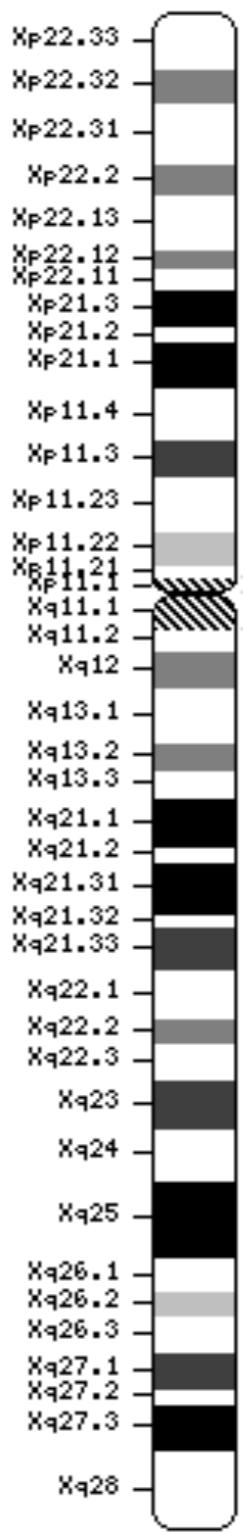
Eric Brouzes et al. PNAS 2009;106:14195-14200





Gene annotations

Ideogram → X





Sequence Alignment

14	SIKLWPPSQTTTRLLLVERMANNLST..PSIFTRK..YGSLSKEEAREN AKQIEEVACSTANQ.....HYEKEPDGDGGSAVQLYAKECSKLILEVLK	101
13	SIKLWPPSESTRIMLVDRMTNNLST..ESIFSRK..YRLLGKQEAHENAKT IEELCFALADE.....HFREEPDGDSAVQLYAKETSKMMLEVLK	100
23	VFKLWPPSQGTREAVRQKMALKLSS..ACFESQS..FARIELADAQE HARAIEEVAFGAAQE.....ADSGGDKTGSAGVVMVYAKHASKLMLET LR	109
13	SVKLWPPGQSTRMLVERMTKNFIT..PSFISRK..YGLLSKEEA EEDAKKIEEVAFAAANQ.....HYEKQPDGDGSSAVQIYAKESSRLM LEVLK	100
30	SFSIWPPPTQRTRDAVVRRLVDTLGG..DTILCKR..YGAVPA ADAEPAAARGIEAEAFDAAAA..SGEAAAATASVEEGIKALQLYS KEVSRRLLDFVK	120
44	SLSIWPPSQRTRDAVVRRLVQTLVA..PSILSQR..YGAVPEAE AGRAAAAVEAEAYAAYTES..SSAAAAPASVEDGI EVLQAYSKEVSRRLLELAK	135
56	SFSIWPPPTQRTRDAIIISRLIETLST..TSVLSKR..YGTIPKEE EASEASRRRIEEEAFSGAST.....VASSEKDGL EVLQLQLYSKEISKRMLETVK	141
29	SFAVWPPTRRTRDAVVRRLVA VLSGDTTTALRKRYRYGAVPA ADAERAARAVEAQAFDAASA.....SSSSSS VEDGIETLQLQLYSREVSNRLLAFVR	121
13	SIKLWPPSESTRMLVERMTDNLSS..VSFFSRK..YGLLSKEE AAENAKRIEETAFLAAND.....HEAKEPNLDDSSVV QFYAREASKLMLEALK	100
57	SLRIWPPTQKTRDAVLNRLIETLST..ESILSKR..YGT LKSDDATTVA KLIEEEEAYGVASN.....AVSSDDD GIKILELYSKEISKRMLESVK	142
25	NYSIWPPKQRTRDAVKNRLIETLST..PSVLT K..YGTMSADE ASAAAIQIEDEAF SVANA.....SSSTSNDNV TILEVYSKEISKRM IETVK	110
28	SFKIWPPPTQRTR EAVVRRLVETLTS..QSVLSKR..YGV IPEEDAT SAARIIEEEAF SVASV..ASAA ASTGGRPE DEWI EVLHIY SQEIX QRV VESAK	119
25	SFSIWPPPTQRTRDAVINRLIESLST..PSILSKR..YGTLPQ DEASET ARLIEEEEAF AAAGS.....TASDADD GIEILQV YSKEISKRM IDTVK	110
14	SVKMWPPSKSTRMLVERMTKNITT..PSIFSRK..YGLLSVE EAEQDAKRIEDLA FATANK.....HFQNEPDGDGT SAVHVY AKESSKL MLDVIK	101
13	SIKLWPPSLPTRKALIERITNNFSS..KTIFTEK..YGS LTKDQATEN AKRIEDIA FSTANQ.....QFEREPDG DGGSAVQLY AKECSKL ILEVLK	100
48	SLSIWPPPTQRTRDAVITRLIETLSS..PSVLSKR..YGT ISHDE AESARR IEDEAF GVANT.....ATS AEDDG LEILQL YSKEISRR MLDTV K	133



BLAST: Basic Local Alignment Search Tool

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

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/ BLAST/ blastn suite Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#) Query subrange [?](#)

From
To

Or, upload file [Choose File](#) no file selected [?](#)

Job Title
Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

Choose Search Set

Database Human genomic + transcript Mouse genomic + transcript Others (nr etc.):
Nucleotide collection (nr/nt) [?](#)

Organism [Optional](#) Enter organism name or id--completions will be suggested Exclude [+](#)
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude [Optional](#) Models (XM/XP) Uncultured/environmental sample sequences

Limit to [Optional](#) Sequences from type material

Entrez Query [Optional](#) YouTube [Create custom database](#)
Enter an Entrez query to limit search [?](#)

Program Selection

Optimize for Highly similar sequences (megablast)
 More dissimilar sequences (discontiguous megablast)
 Somewhat similar sequences (blastn)
Choose a BLAST algorithm [?](#)

BLAST Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)
 Show results in a new window



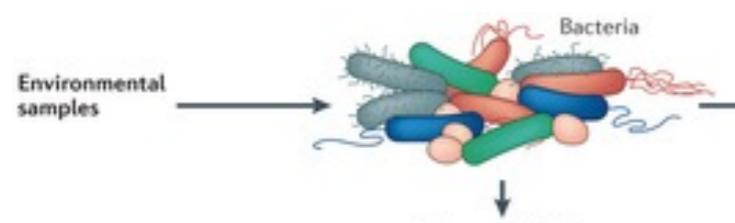
Scoring Matrix BLOSUM

(BLOcks SUbstitution Matrix)

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5				
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

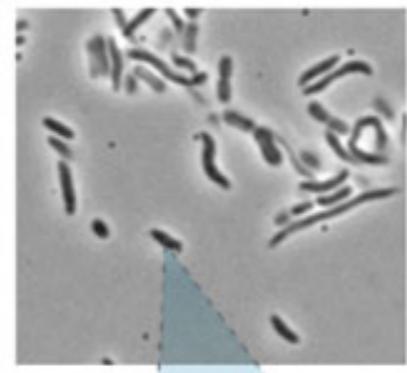


Environmental DNA analysis





16S RNA

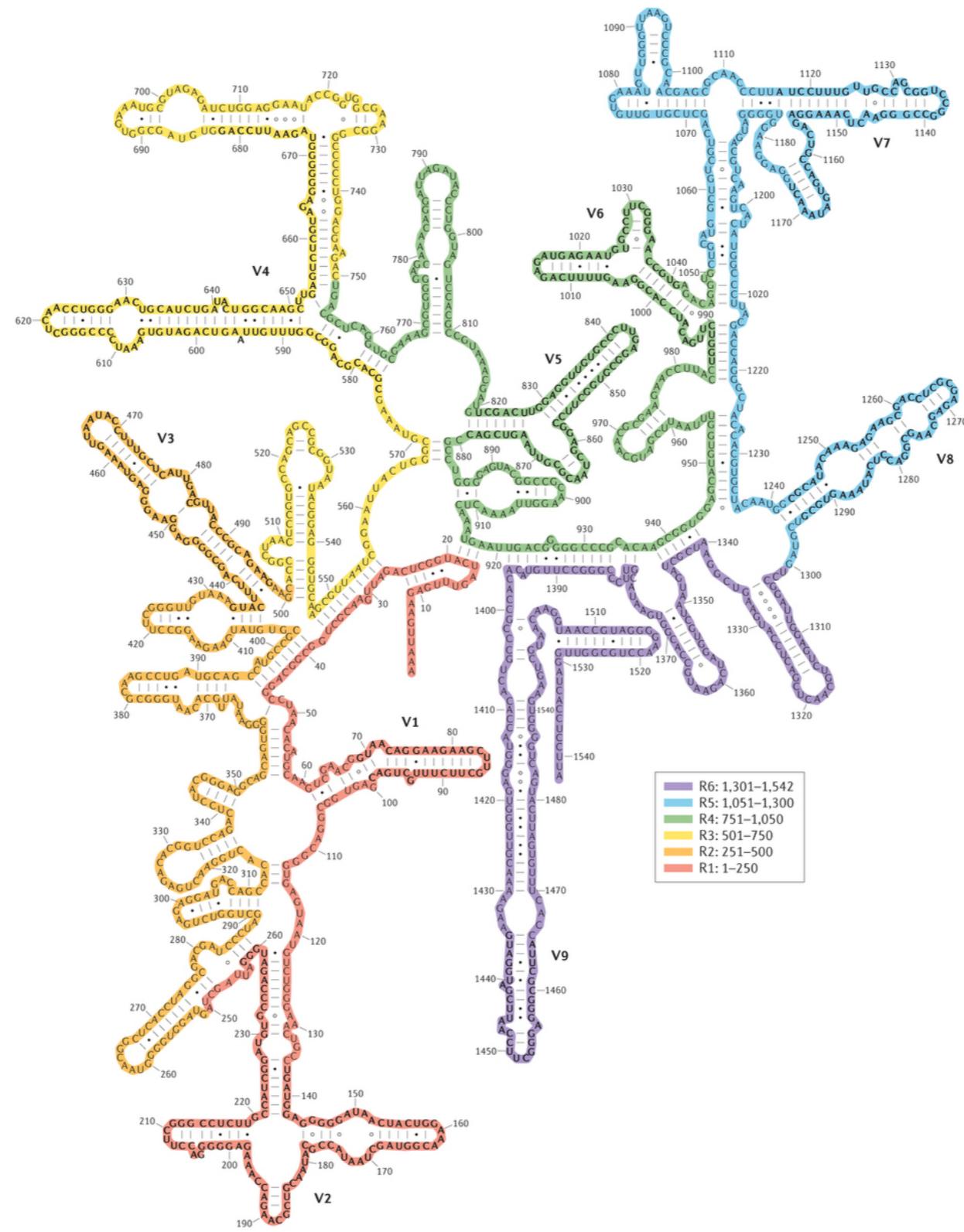


Environmental samples

DNA extraction



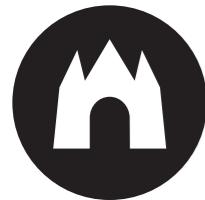
16S RNA molecule





Challenges

- Cross linking data / Data mining
 - Relate Genomics to Transcriptomics, Proteomics
 - Relate to structure
 - Relate to disease



Bio Informatics for the public

The screenshot shows the 23andMe website. At the top, there is a navigation bar with links for "welcome" (which is highlighted in green), "health", "ancestry", "how it works", "research", "buy", "help", and a search icon. To the right of the navigation are icons for "sign in", "register kit", a shopping cart with "0" items, and a magnifying glass. On the left side of the main content area, there is a large image of a DNA collection kit box. The box is white with a colorful, abstract geometric pattern on the left and the text "welcome to you" in the center. Below the box, the text "DNA Collection Kit" and the 23andMe logo are visible. To the right of the box, the text "Get to know you. Health and ancestry start here." is displayed in large, bold, dark gray font. Below this text is a bulleted list of benefits:

- View reports on over 100 health conditions and traits
- Find out about your inherited risk factors and how you might respond to certain medications
- Discover your lineage and find DNA relatives

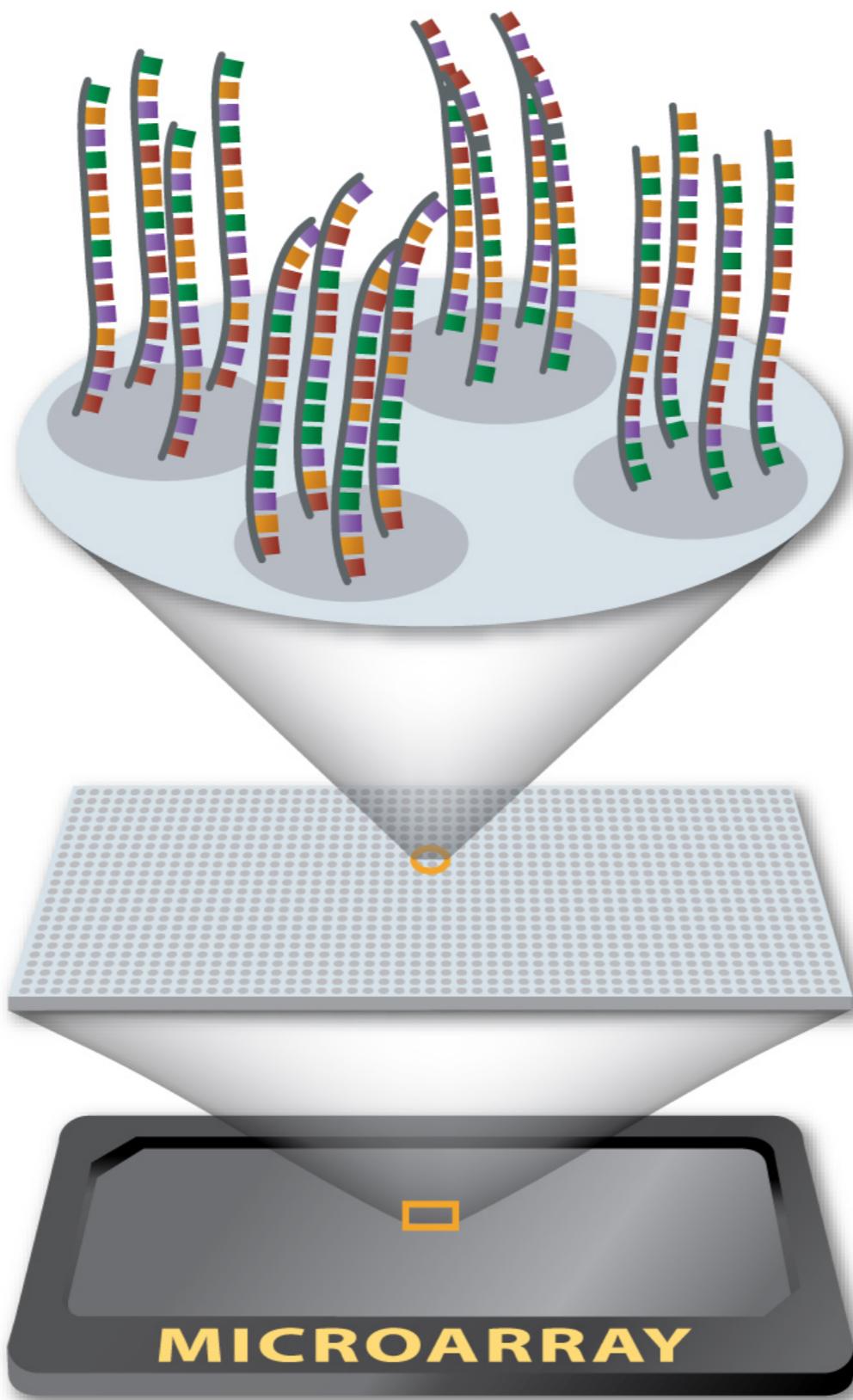
At the bottom right of the main content area, there is a pink button with the text "order now" and a price of "€169" followed by the smaller text "shipping included".

What your DNA says about you.

Find out how your genetics relate to things like abnormal blood clotting, cystic fibrosis or response to certain medications. You can also see if your body metabolises caffeine quickly or if you're likely lactose intolerant. We believe the more you know about



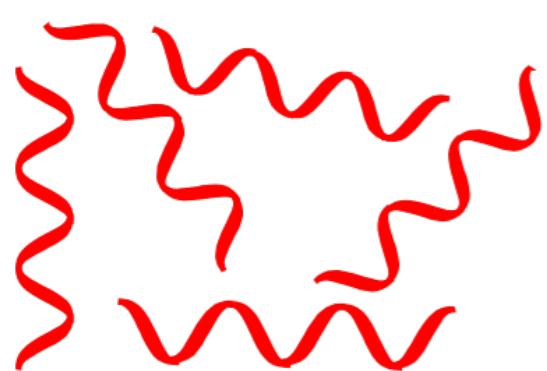
DNA Microarray



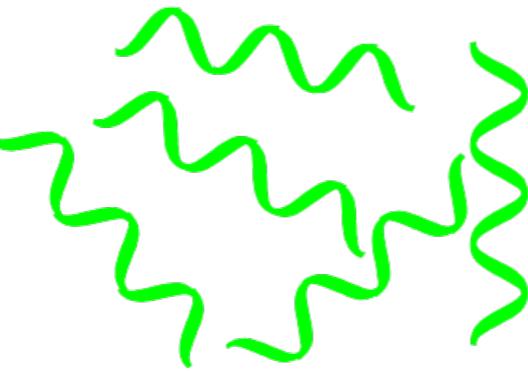


DNA Microarray Expression Analysis

DNA extraction and digestion



Test = Tumoral DNA
labeled with **Cy5**



Reference = Normal DNA
labeled with **Cy3**



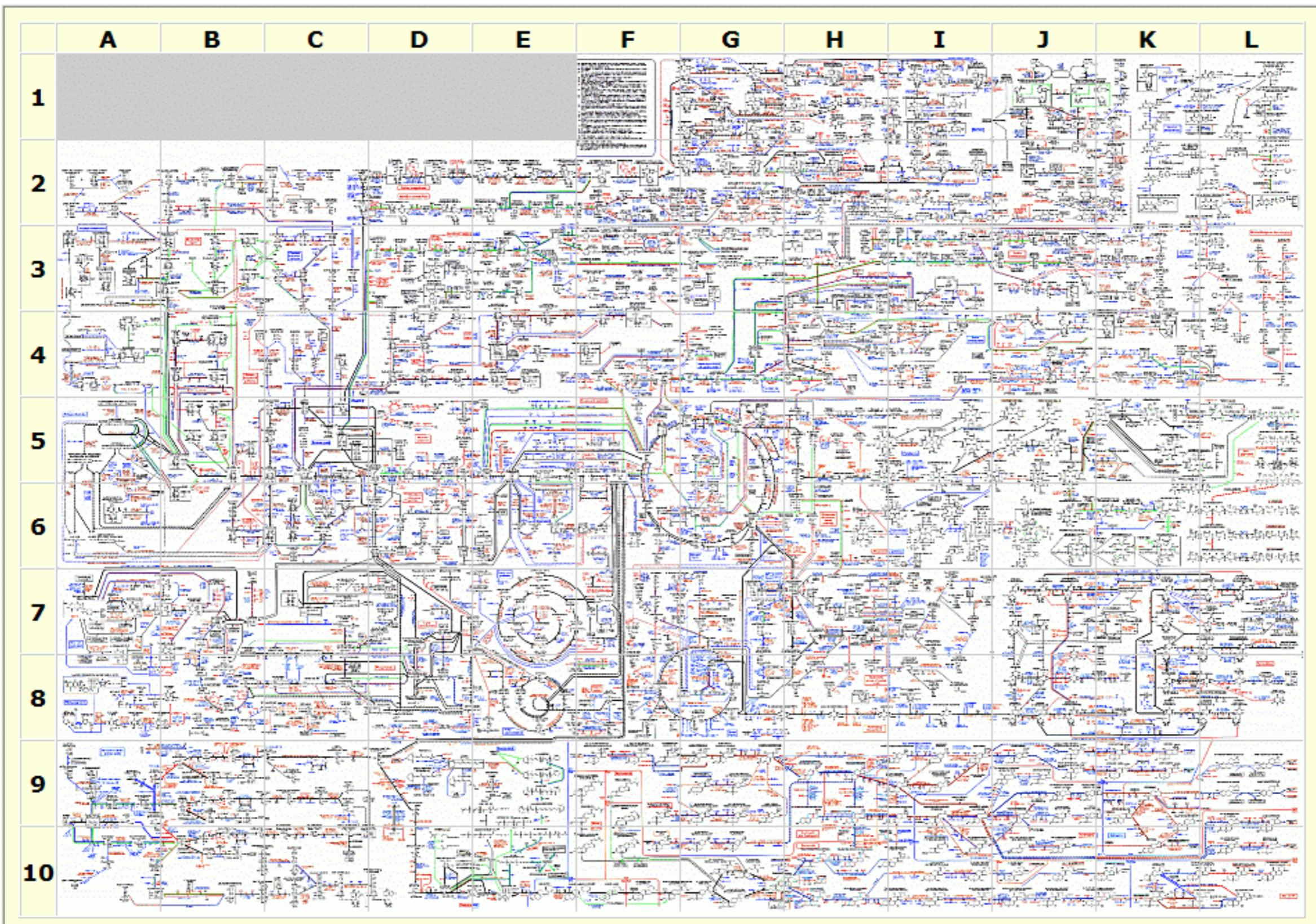
waag society

institute for art, science and technology

Proteins

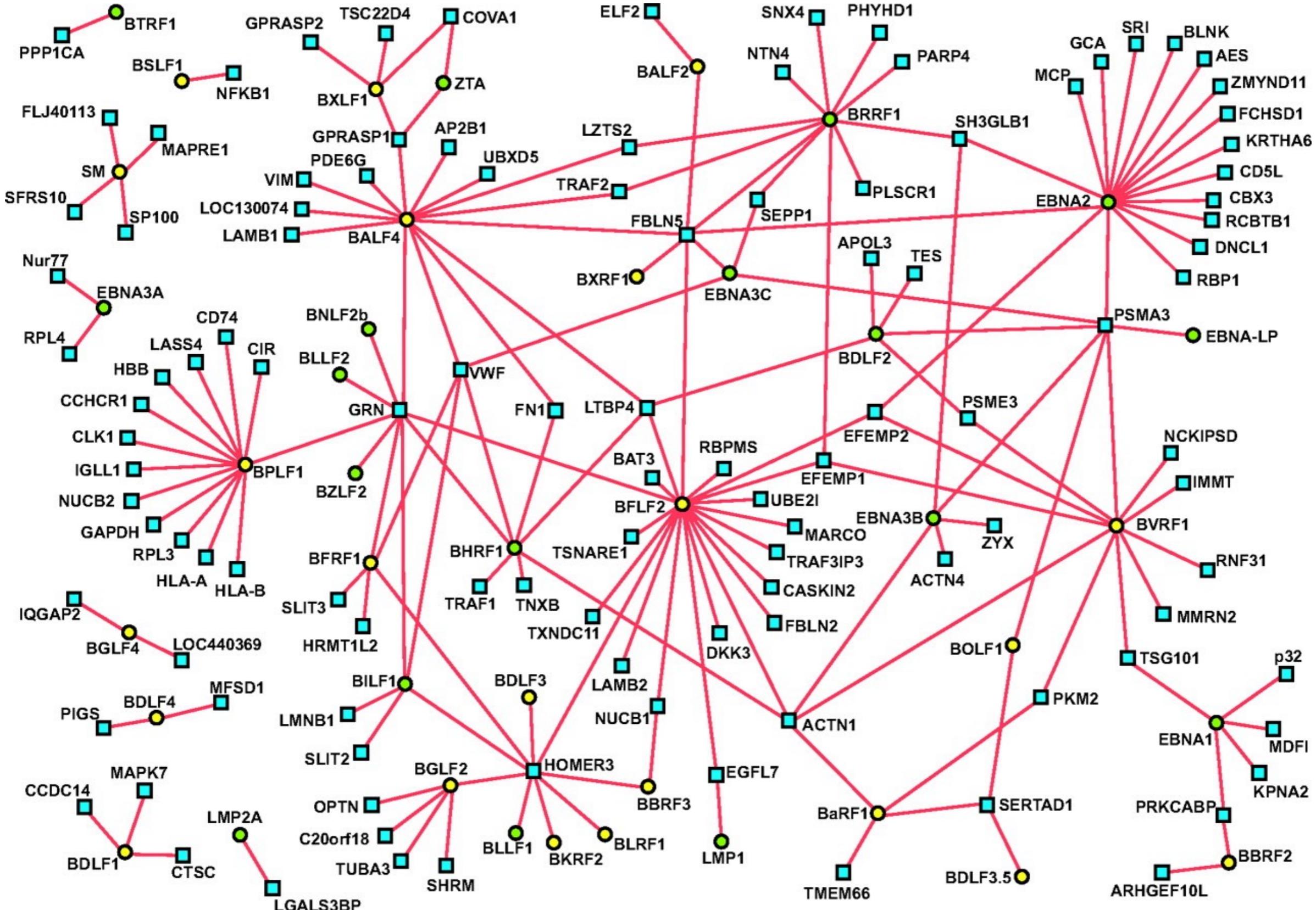


Biochemical Pathways of the Cell





Protein interaction mapping: MS





What is this?

Simon Eugstar - CC-BY-SA 3.0



Debstart - CC-BY-SA 3.0





NMR Machines

MartinSaunders

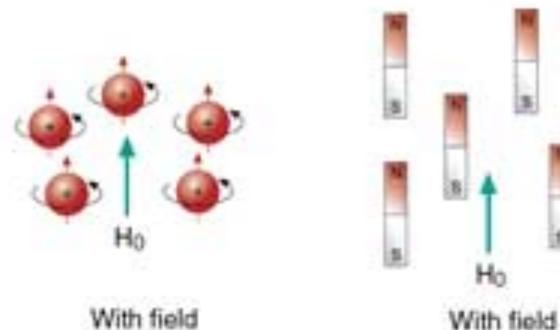
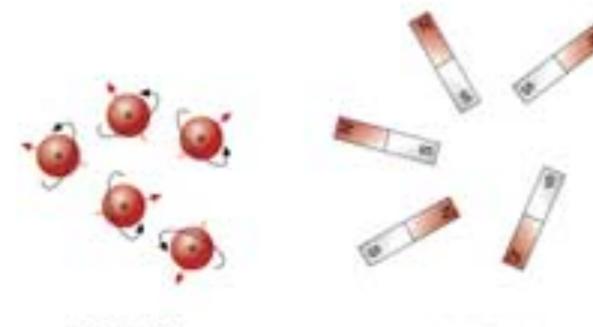
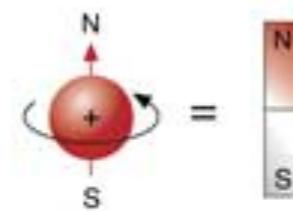
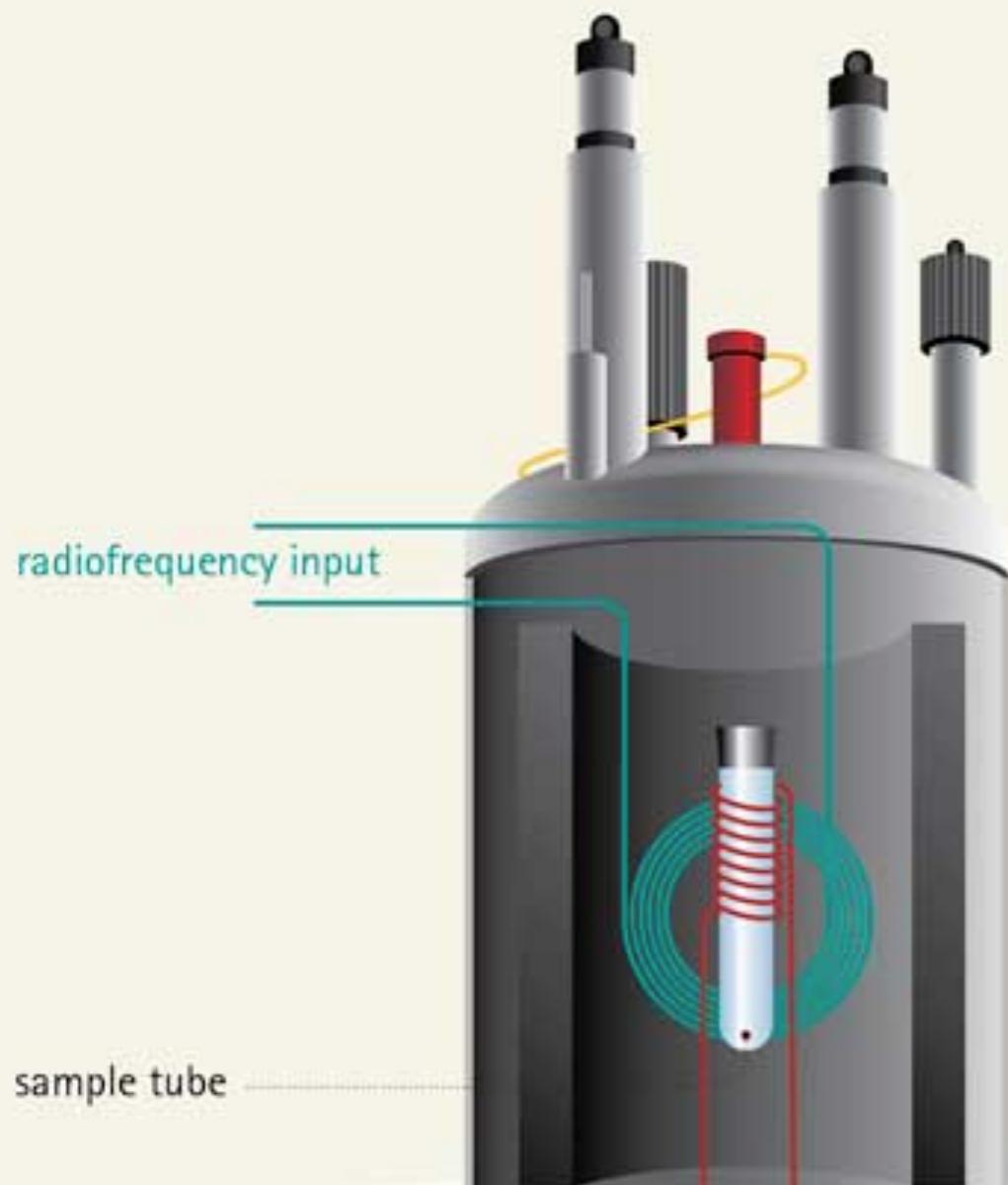


Public Domain





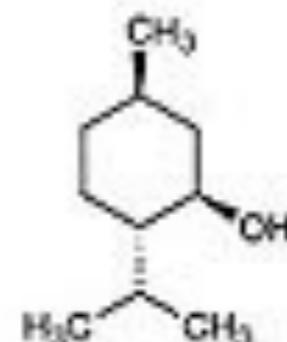
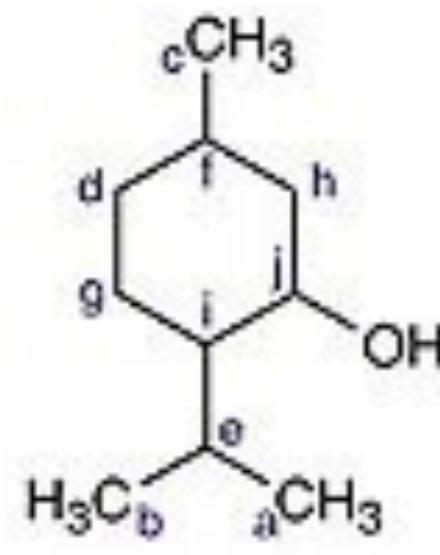
NMR principles



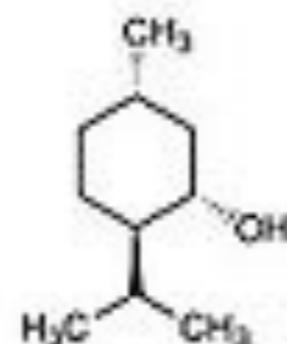


NMR Spectrum

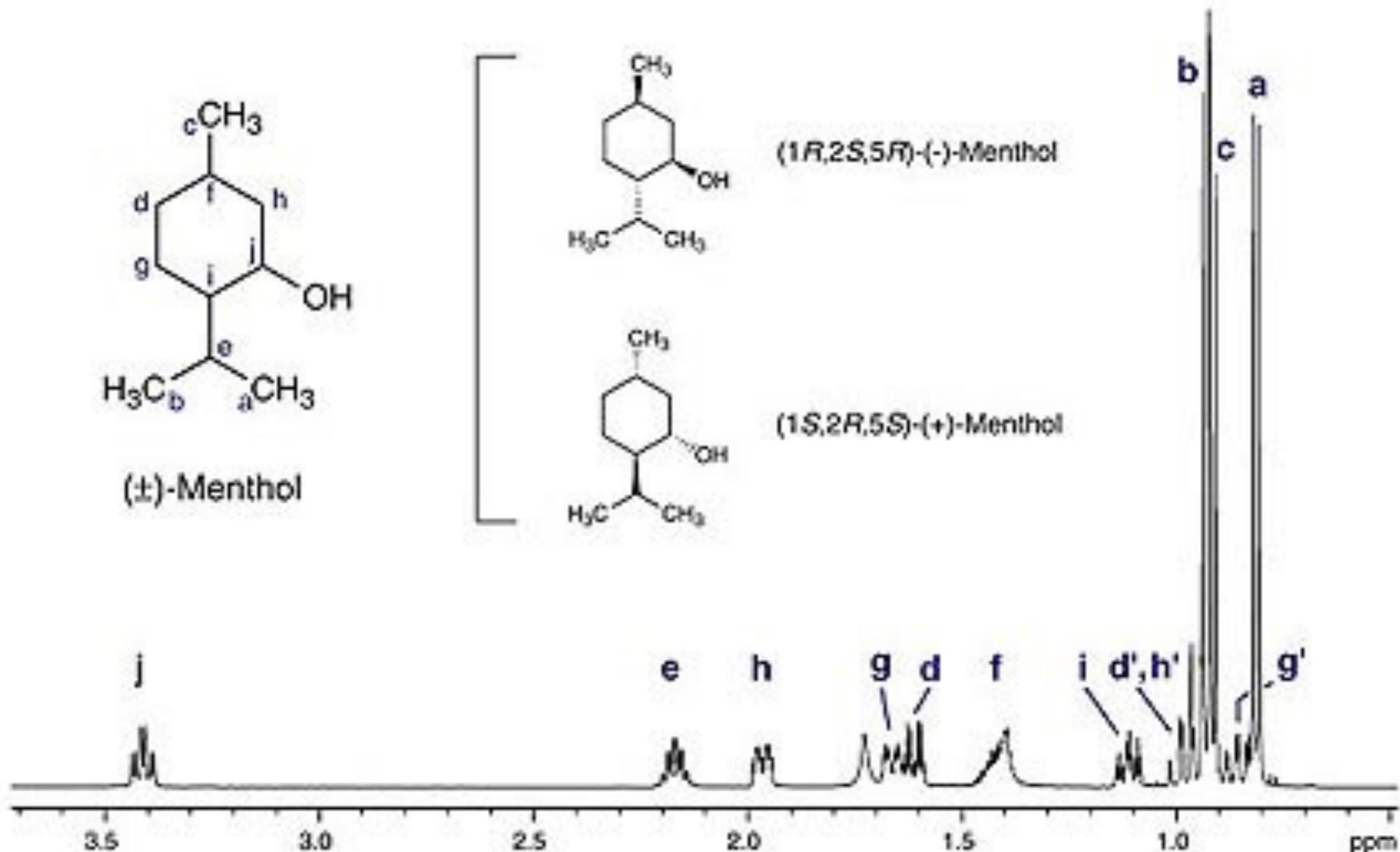
1D PROTON SPECTRUM



(1*R*,2*S*,5*R*)-(-)-Menthol



(1*S*,2*R*,5*S*)-(+)-Menthol

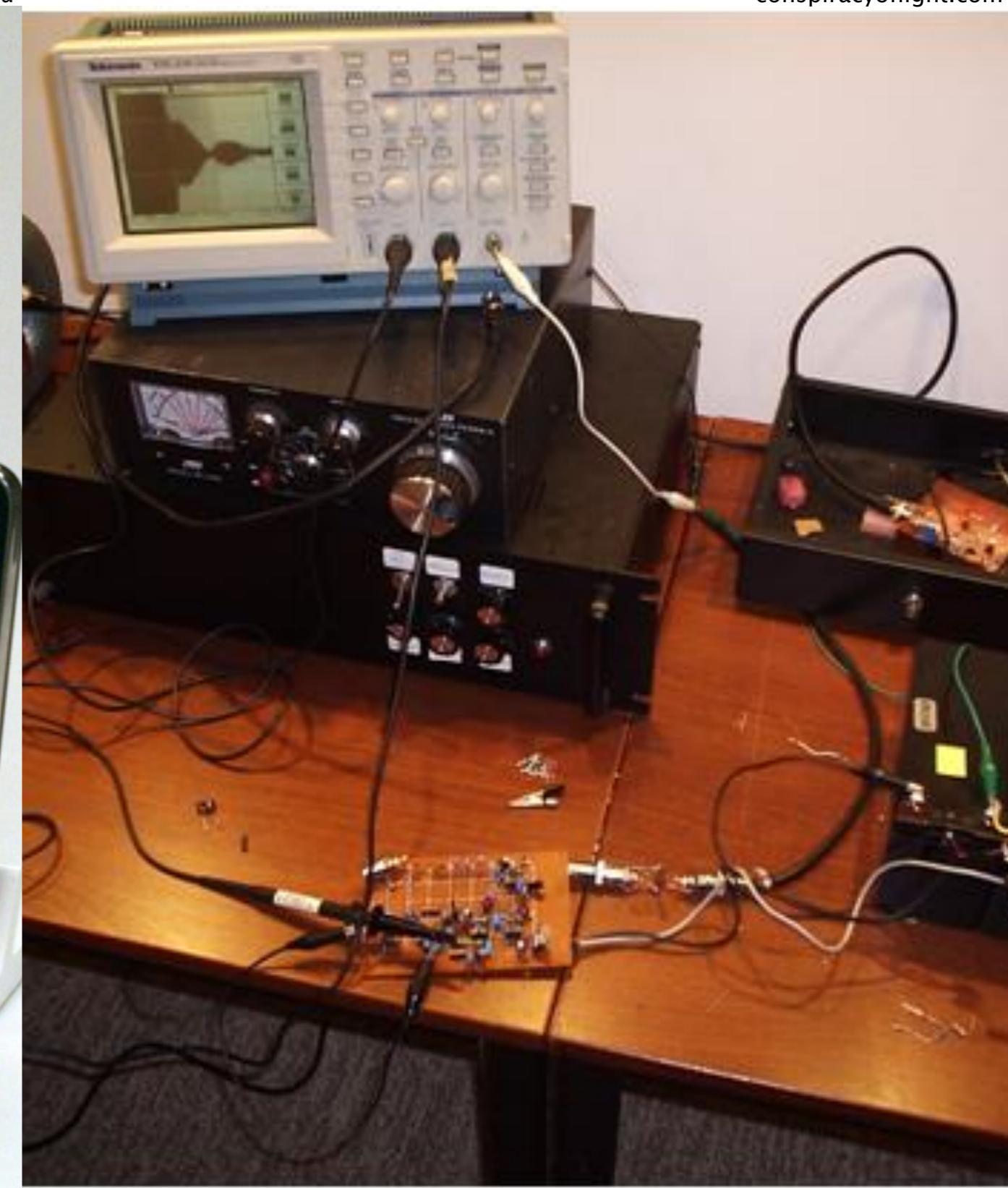
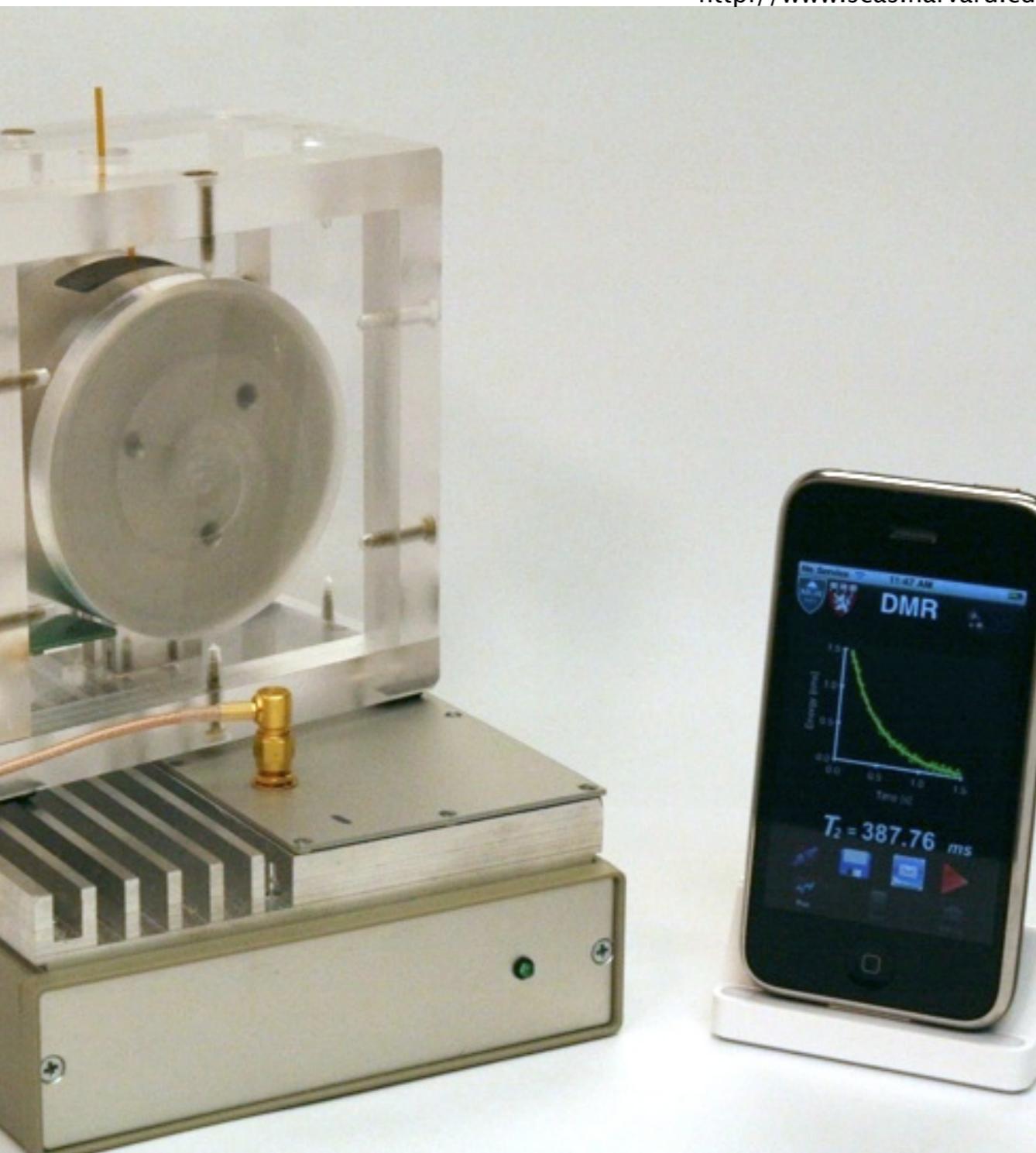




DIY NMR?

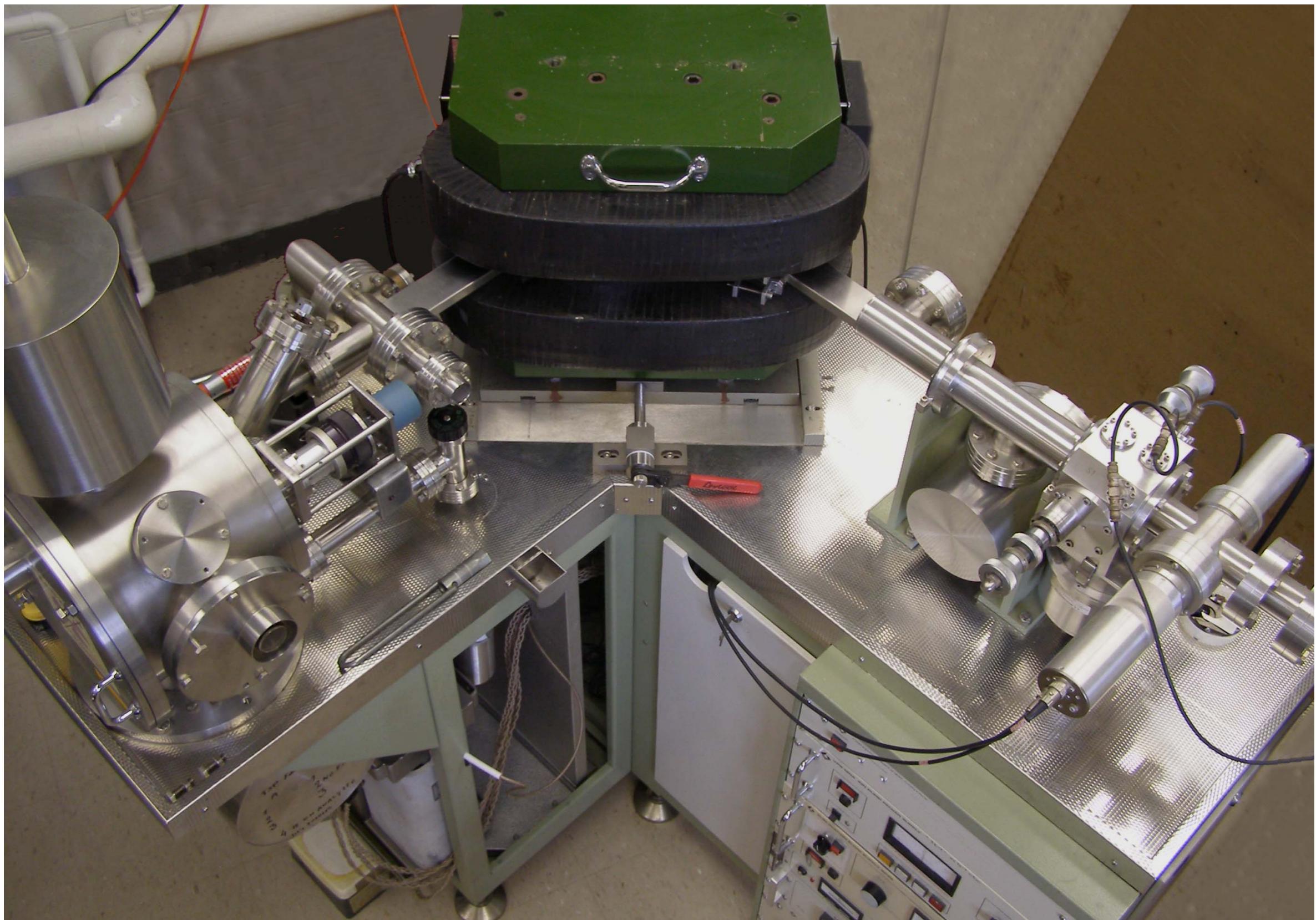
<http://www.seas.harvard.edu>

conspiracyoflight.com



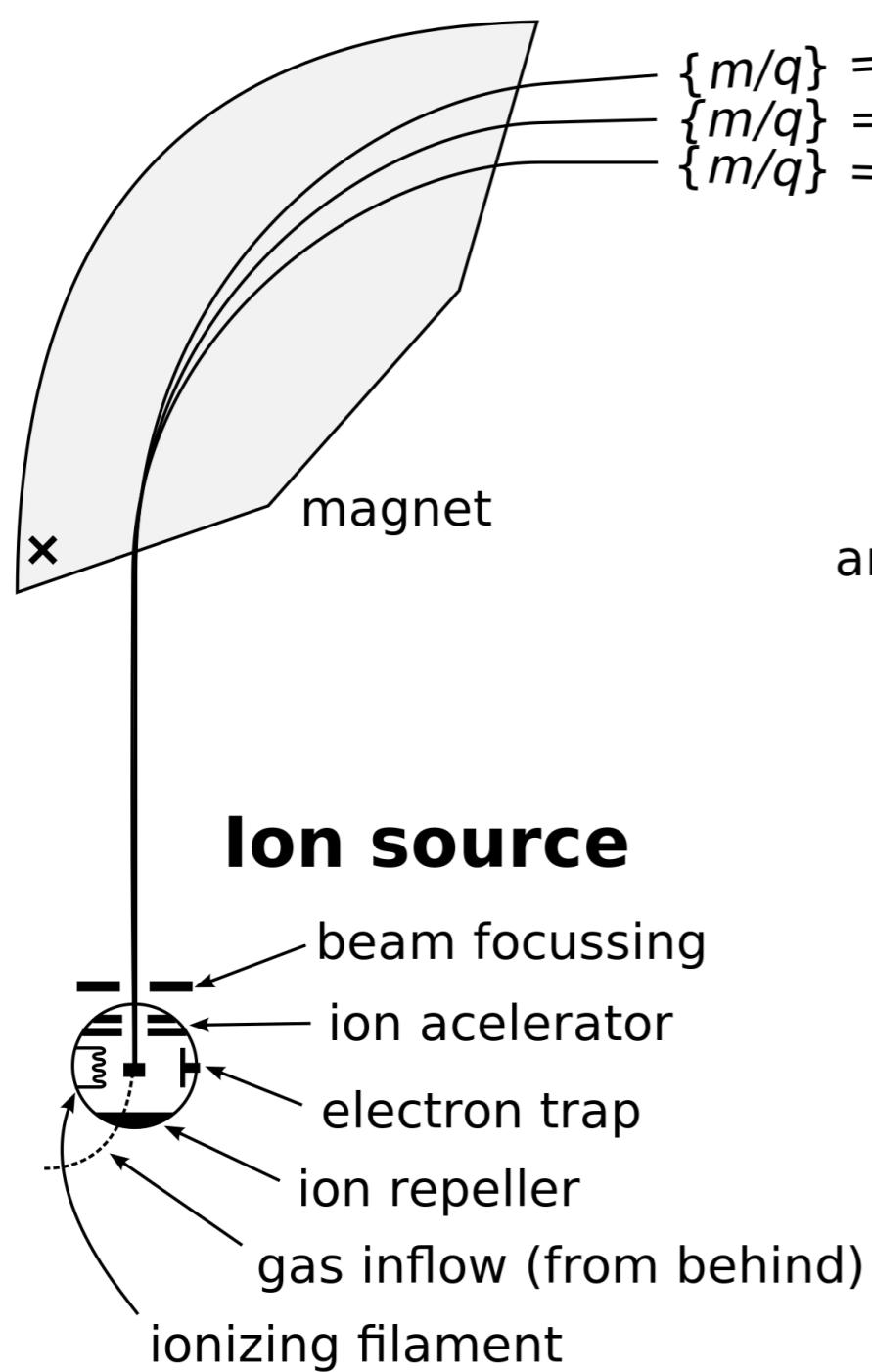


Mass Spectrometer

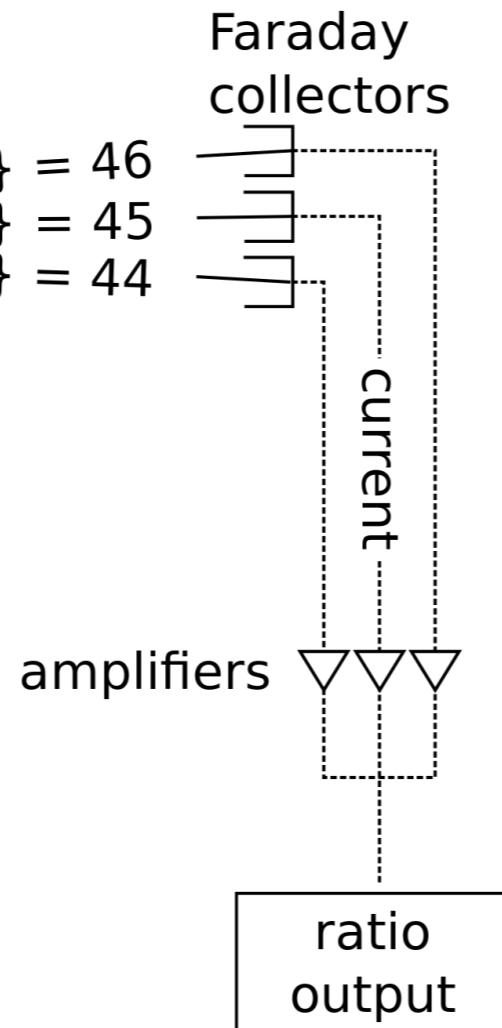




Simple Mass Spectrometry



Detection



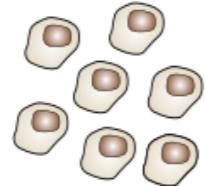
High mass = high m/q

Big charge = low m/q

legend:
 m ... ion mass
 q ... ion charge



Mass Spectrometry



cells or tissue

TOF

MALDI

Tandem



Procedure

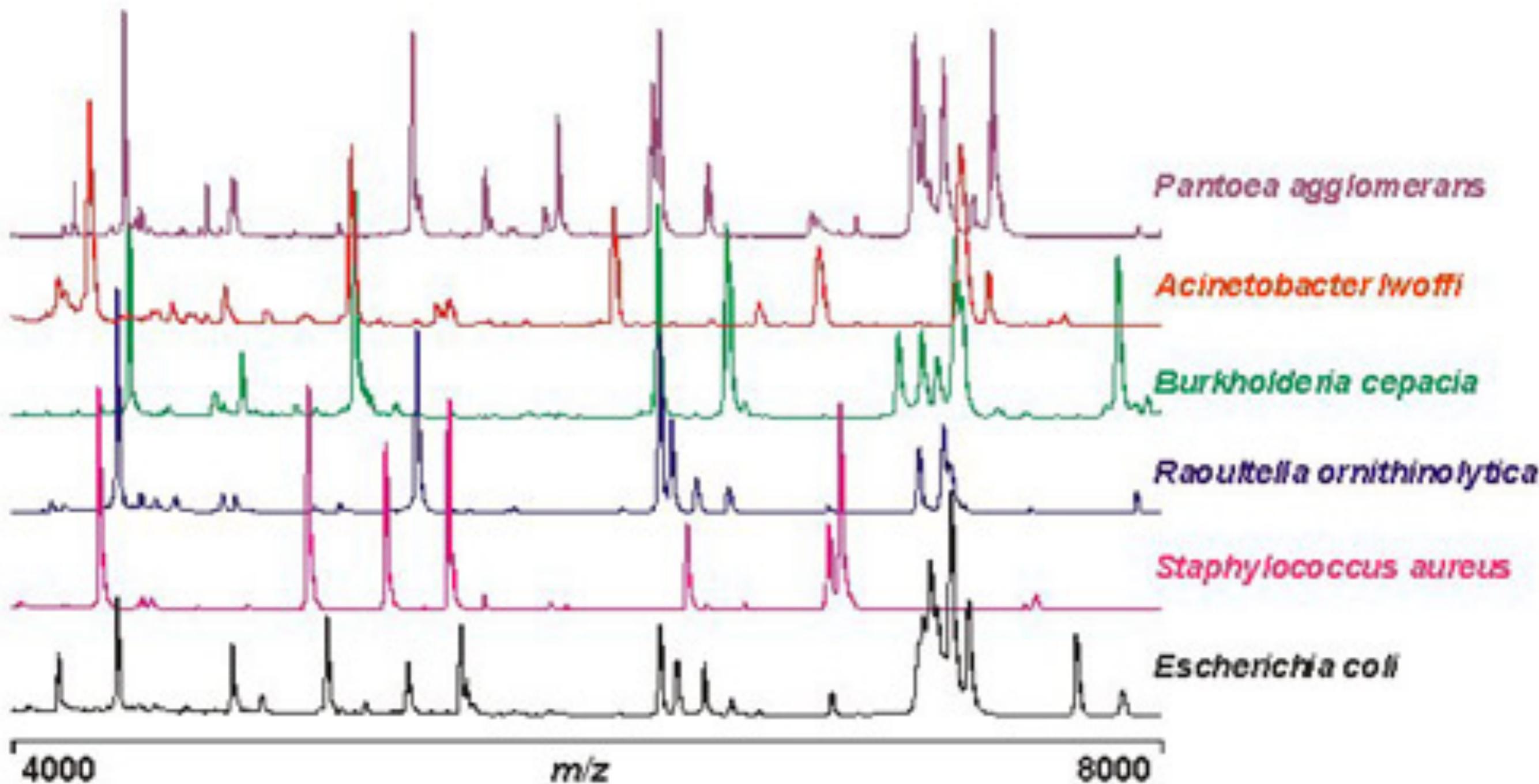


Sampling the colony

± Extraction of intracellular protein in 70% formic acid and absolute ethanol

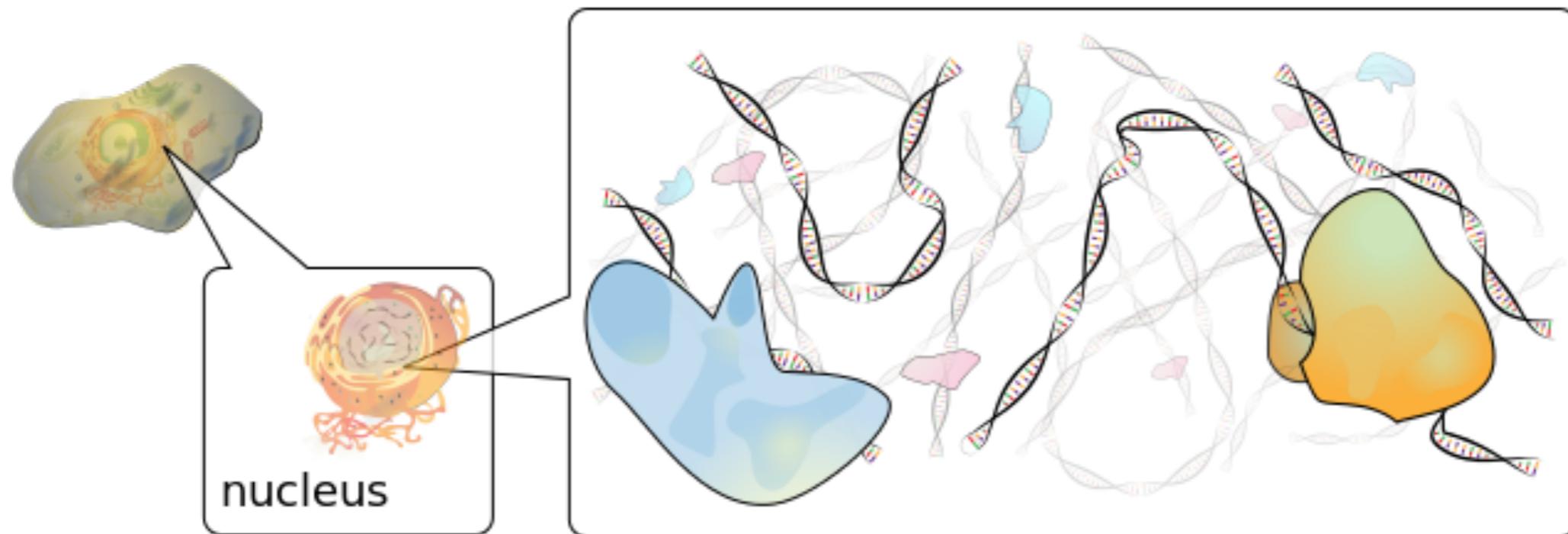


Bacterial profiles

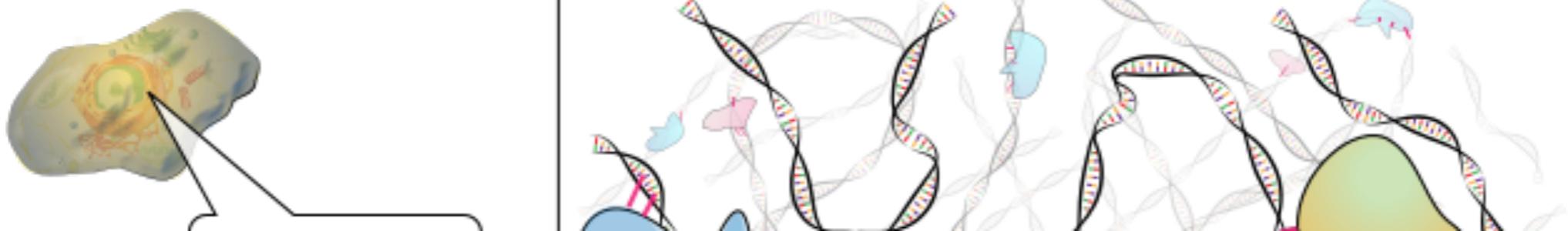




Chromatin ImmunoPrecipitation ChIP

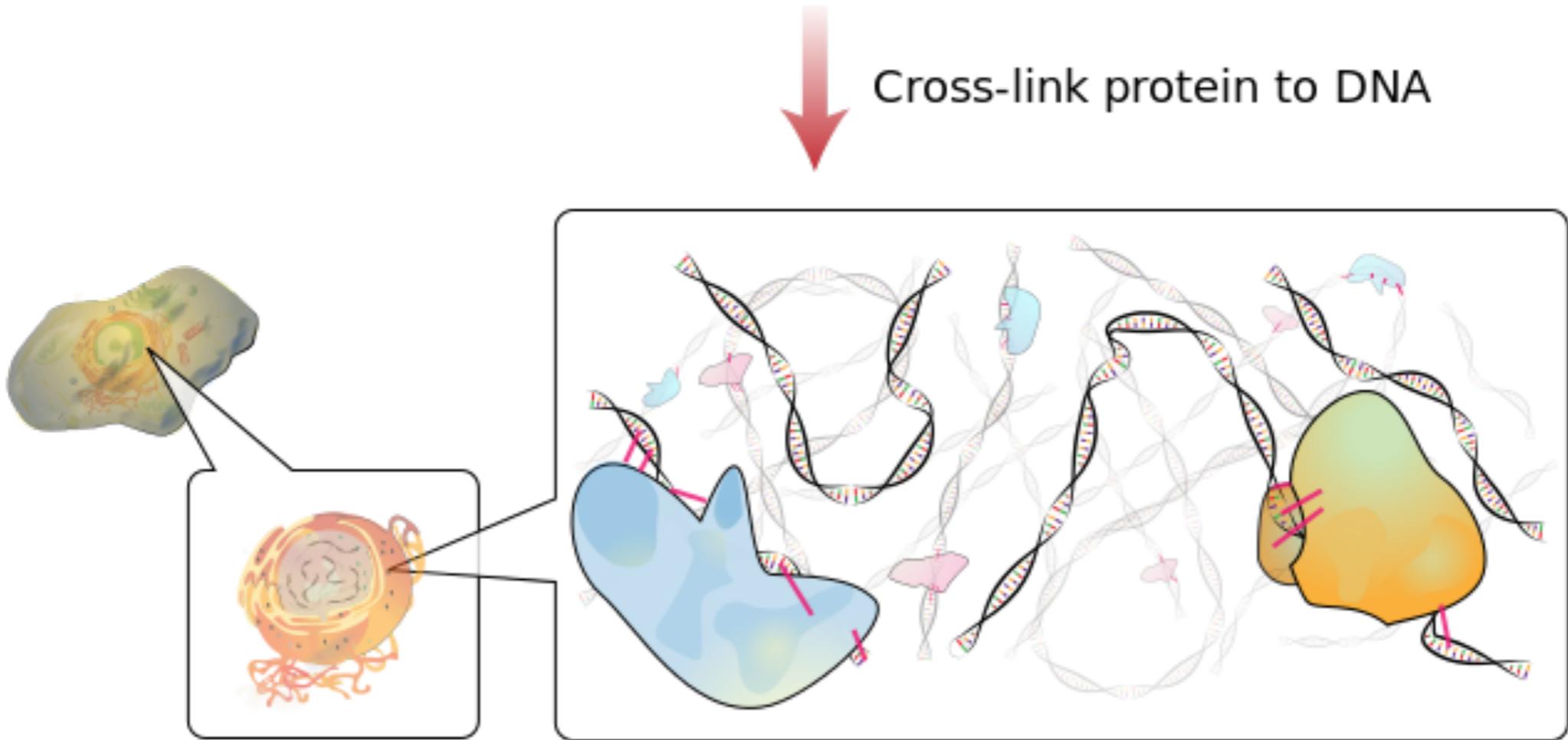


Cross-link protein to DNA





Chromatin ImmunoPrecipitation ChIP

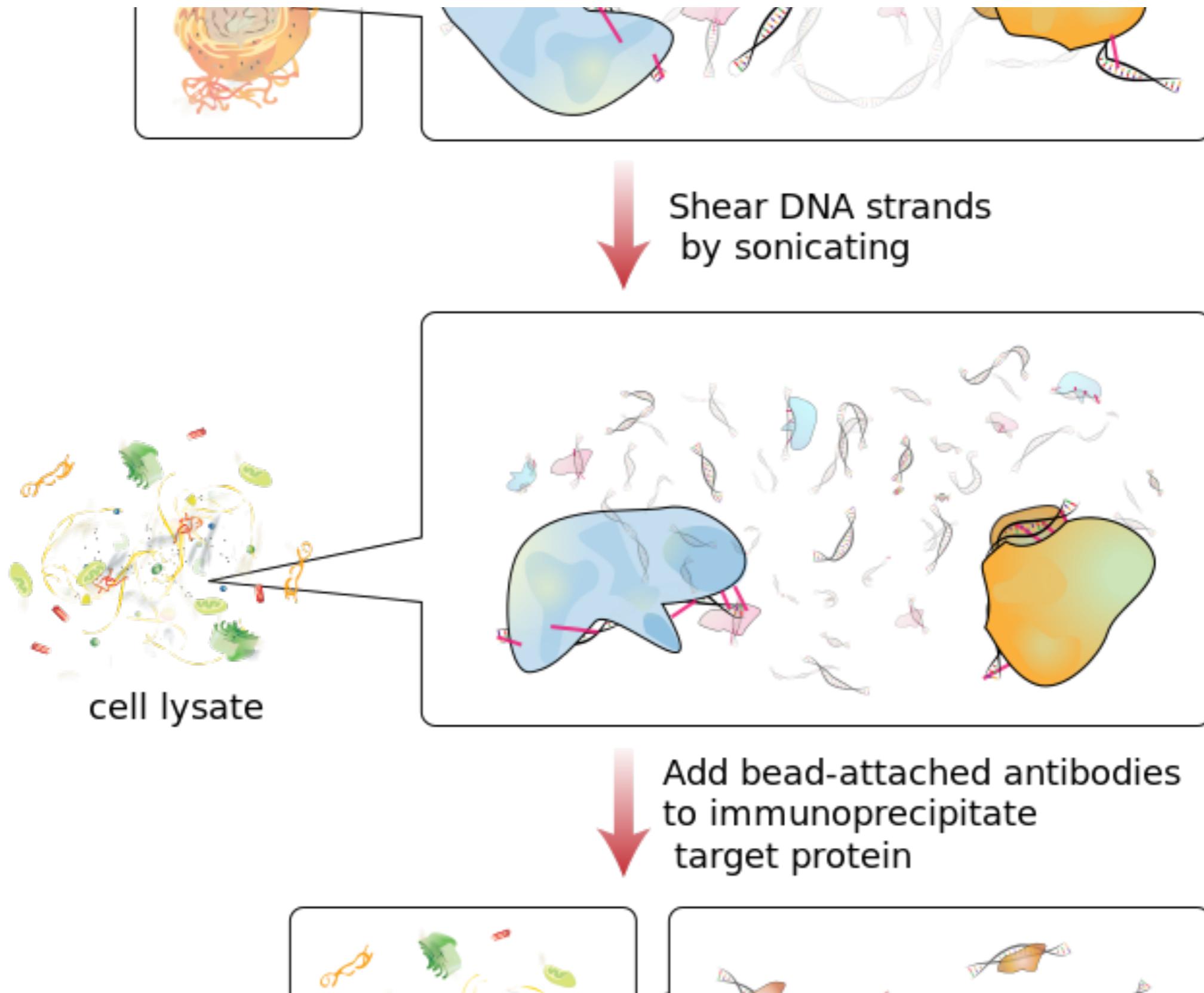


Cross-link protein to DNA

Shear DNA strands
by sonicating

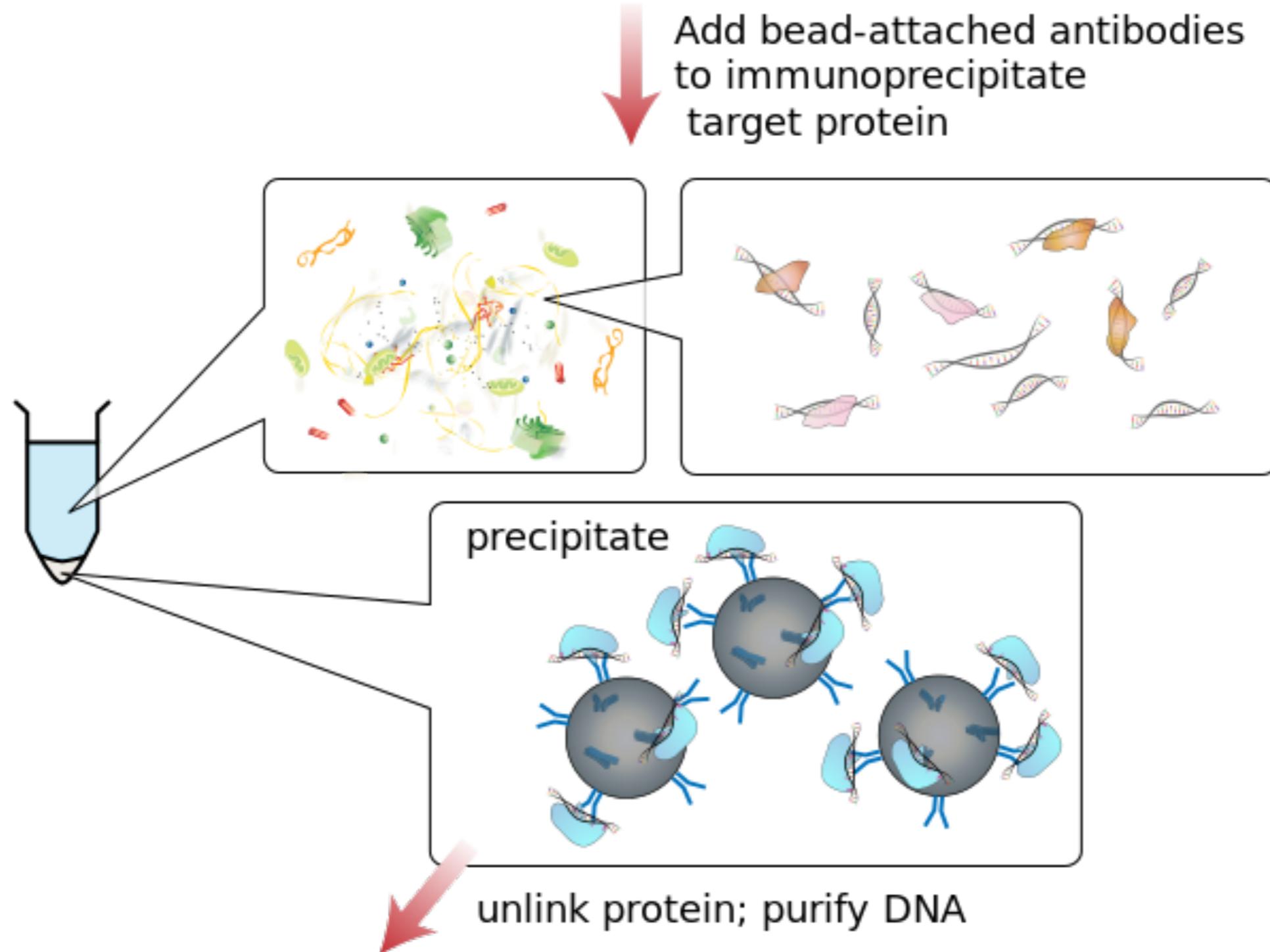


Chromatin ImmunoPrecipitation ChIP



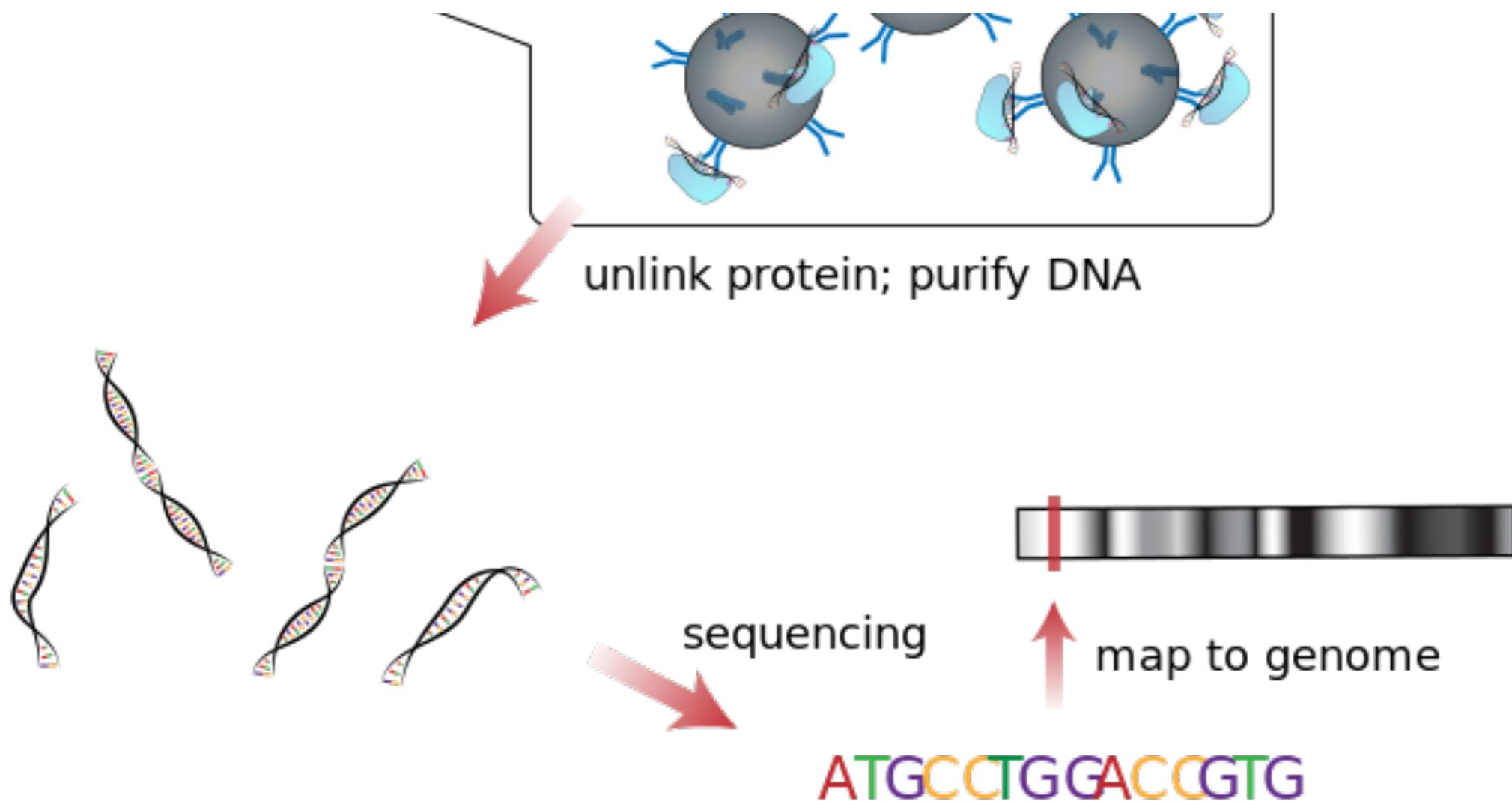


Chromatin ImmunoPrecipitation ChIP





Chromatin ImmunoPrecipitation ChIP





From analysis to synthesis

- Genomics: DNA sequence data
- Transcriptomics: Gene expression data
- Proteomics: Protein composition data
- Can we reverse this process and design our own bioproducts?



Bridging protein and dna data

20n

Open-source Platform

Bioreachables Service

Blog

Contact

20n is open sourcing its platform for synthetic biology

Over the last 4 years we have developed a better way to bioengineer organisms. We are now open sourcing our entire software stack, 20n/act. Find it at <https://github.com/20n/act>.

The stack will enumerate all bio-accessible chemicals, called *reachables* ([20n/act/reachables](#)). For each of those chemicals, it will design DNA blueprints. These DNA blueprints can bioengineering organisms with un-natural function. E.g., build organisms to make chemicals that were previously only sourced through petrochemistry.

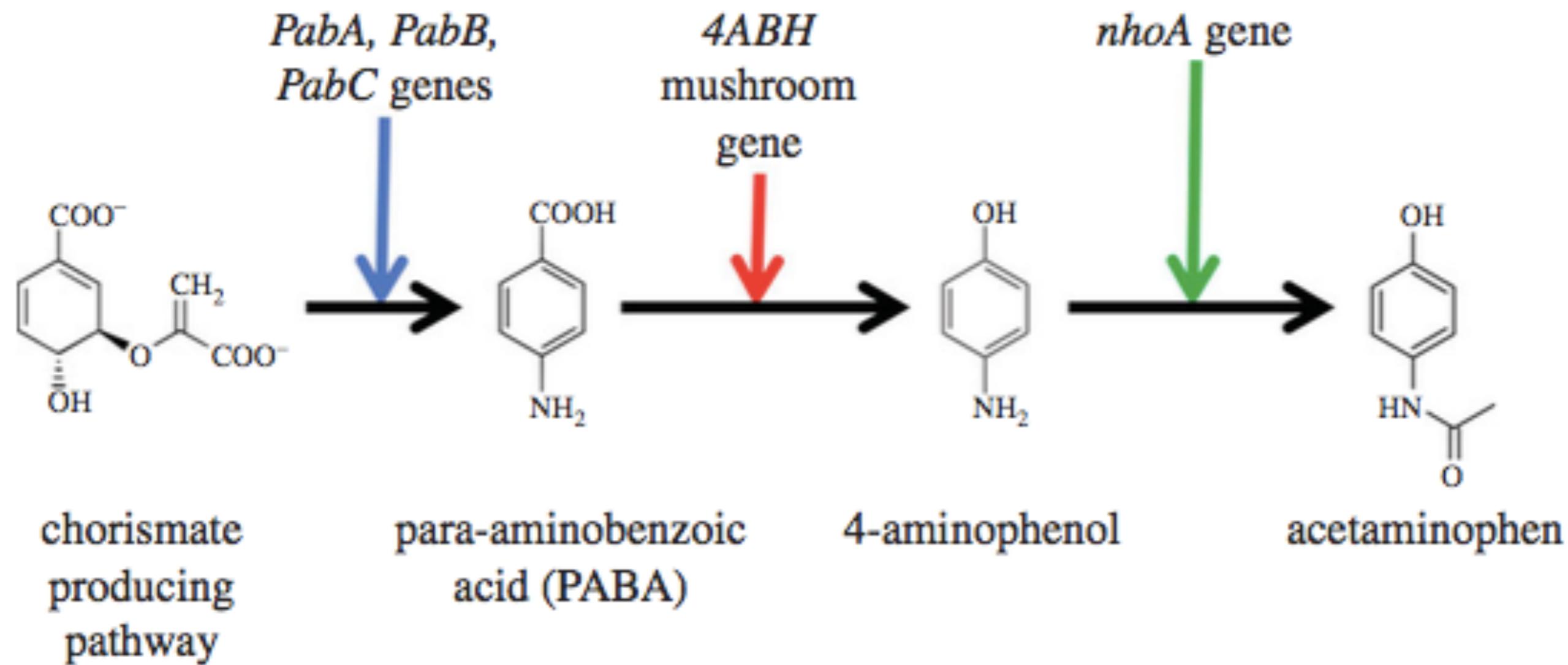
To do that, the stack contains many modules built from scratch in-house. Some of them: mine raw biochemical data, integrate heterogenous sources, learn rules of biochemistry, automatically clean bad data, mine patents, mine plain text, bioinformatic identification of enzymes with desired function.

Once the suggested DNA is used to create new engineered cells those cells can be analyzed with LCMS for function. Our [deep learning-based untargeted metabolomics](#) stack processes the raw data and enumerates all side-effects of the changed genomic structure of the cell. Some would be expected, as the organism making the desired chemical, and some unexpected metabolic changes are highlighted.

We are also releasing a [economic cost model for bioproduction](#). This economic cost model maps the desired the market price of the biological product to the "science needed" to get there. The "science needed" is measured in fermentation metrics, yield,



Predicted pathway





Issues

- Ethical
 - Who owns bio data?
 - Who decides what to use data for?
 - Is de-personalized bio information possible?
- Imagine:
 - You are one of the only persons immune to Zika virus. Are you entitled to royalties on the vaccine derived from your blood?



**some
rights
reserved**



Visualisation PyMol

