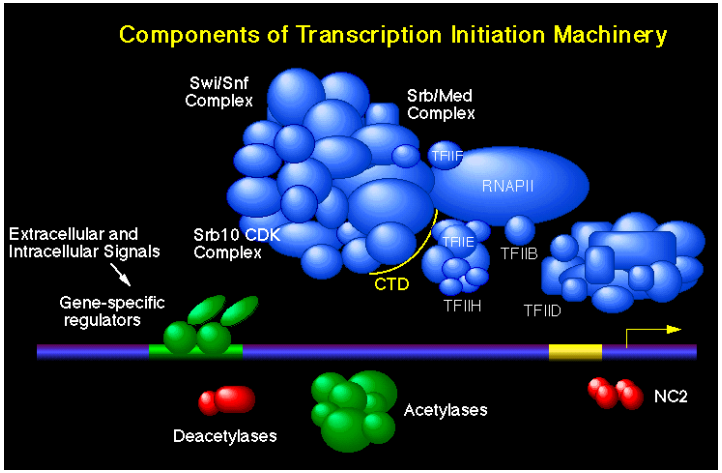
Podiplomski študij Biomedicine

Genetika: Genetski koncepti (2010-011)

Uravnavanje izražanja genov, transkriptomika in drugi pristopi po-genomskega obdobja

Prof. dr. Damjana Rozman
Damjana.rozman@mf.uni-lj.si
http://cfgbc.mf.uni-lj.si/

1. Osnove uravnavanje izražanja genov.
2. Študije genoma in transkriptoma z DNA mikromrežami (čipi).
3. Študije genoma in transkriptoma z novo generacijo sekveniranja.
4. Pomen informatike pri študijah “omov”.



1. Osnove uravnavanje izražanja genov pri evkariontih

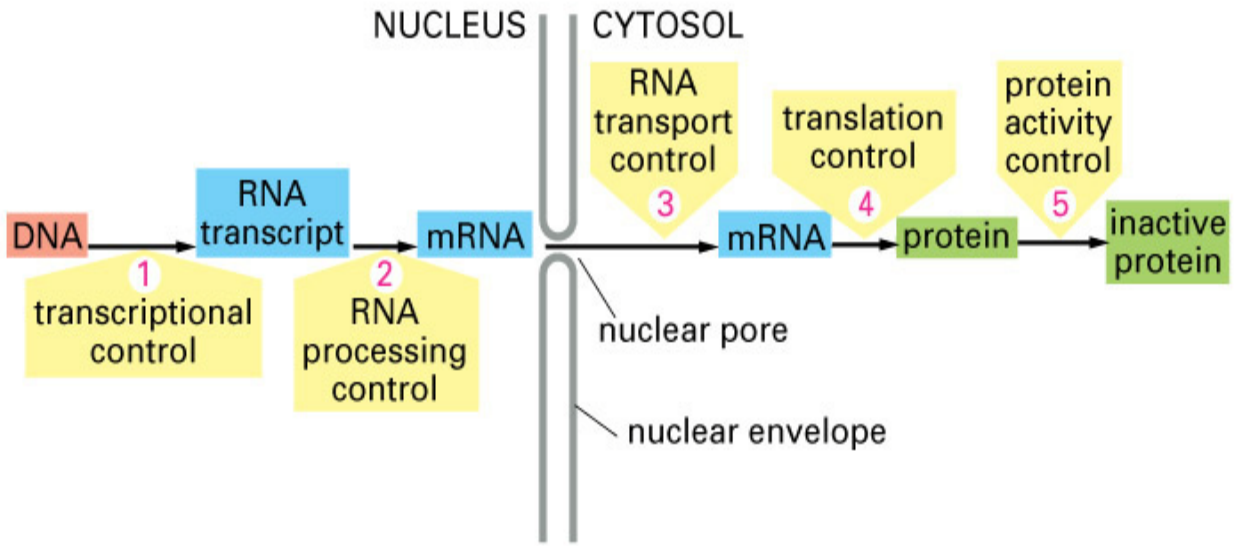


Figure 8-3 Essential Cell Biology, 2/e. (© 2004 Garland Science)

Bakterije

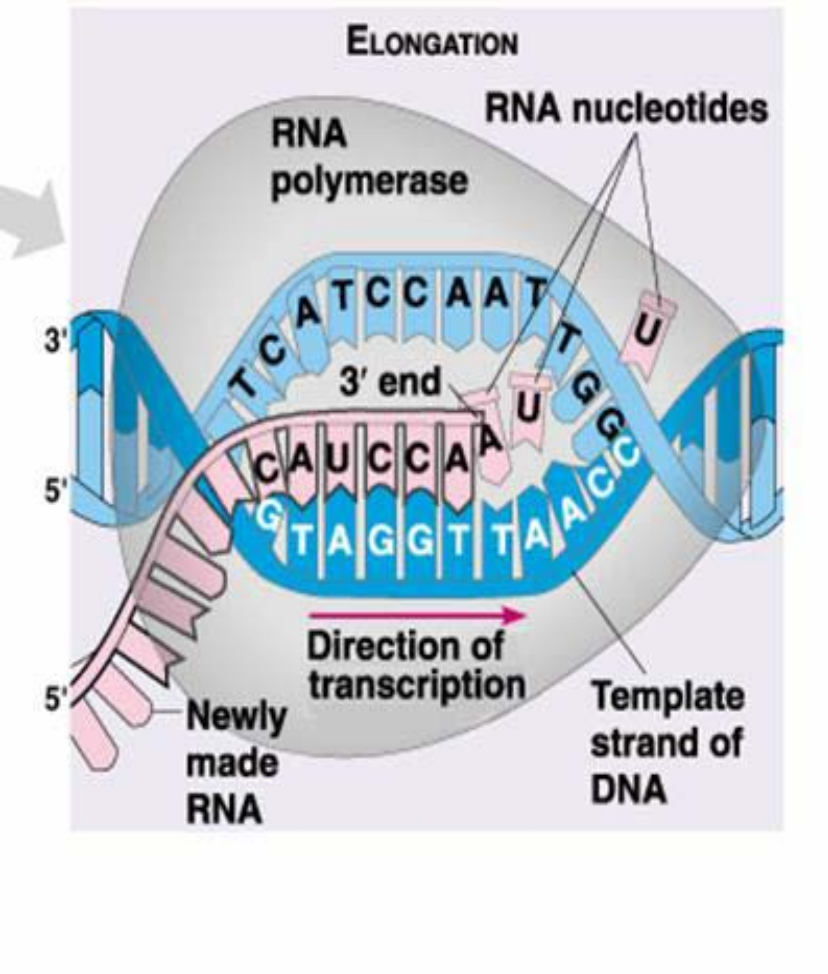
– 1 RNA polimeraza

Evkarionti

– RNA pol I (45s pre-rRNA)

-RNA pol II (mRNA kodirani geni, miRNA)

-RNA pol III (tRNA, 5S rRNA)



RNA polimeraza se lahko premika v obe smeri

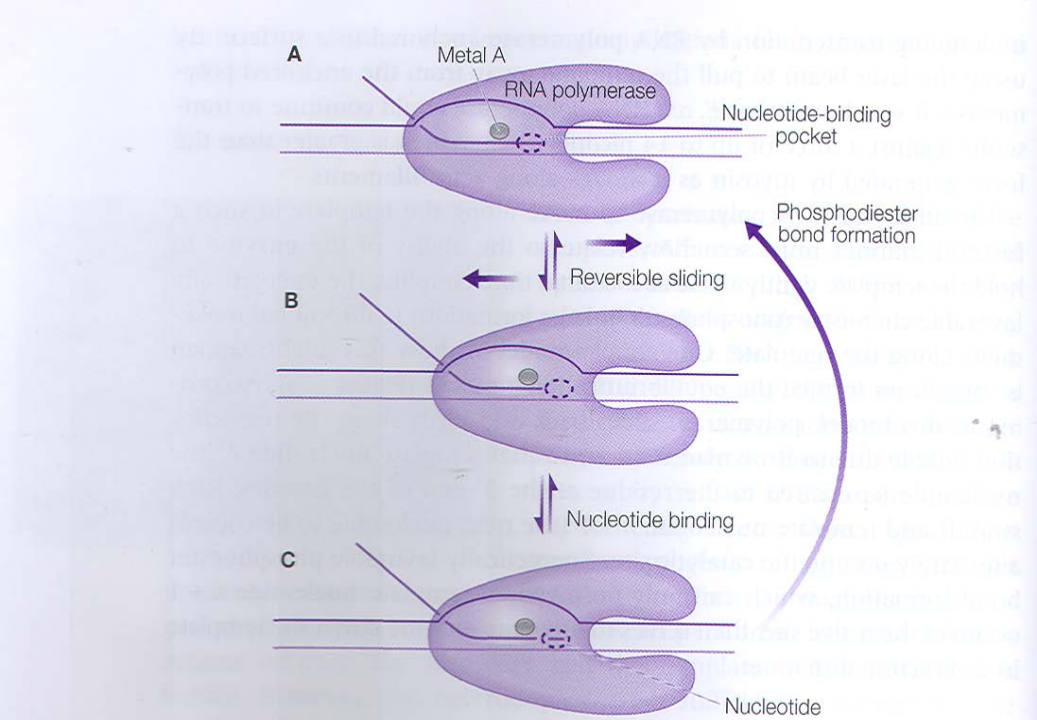
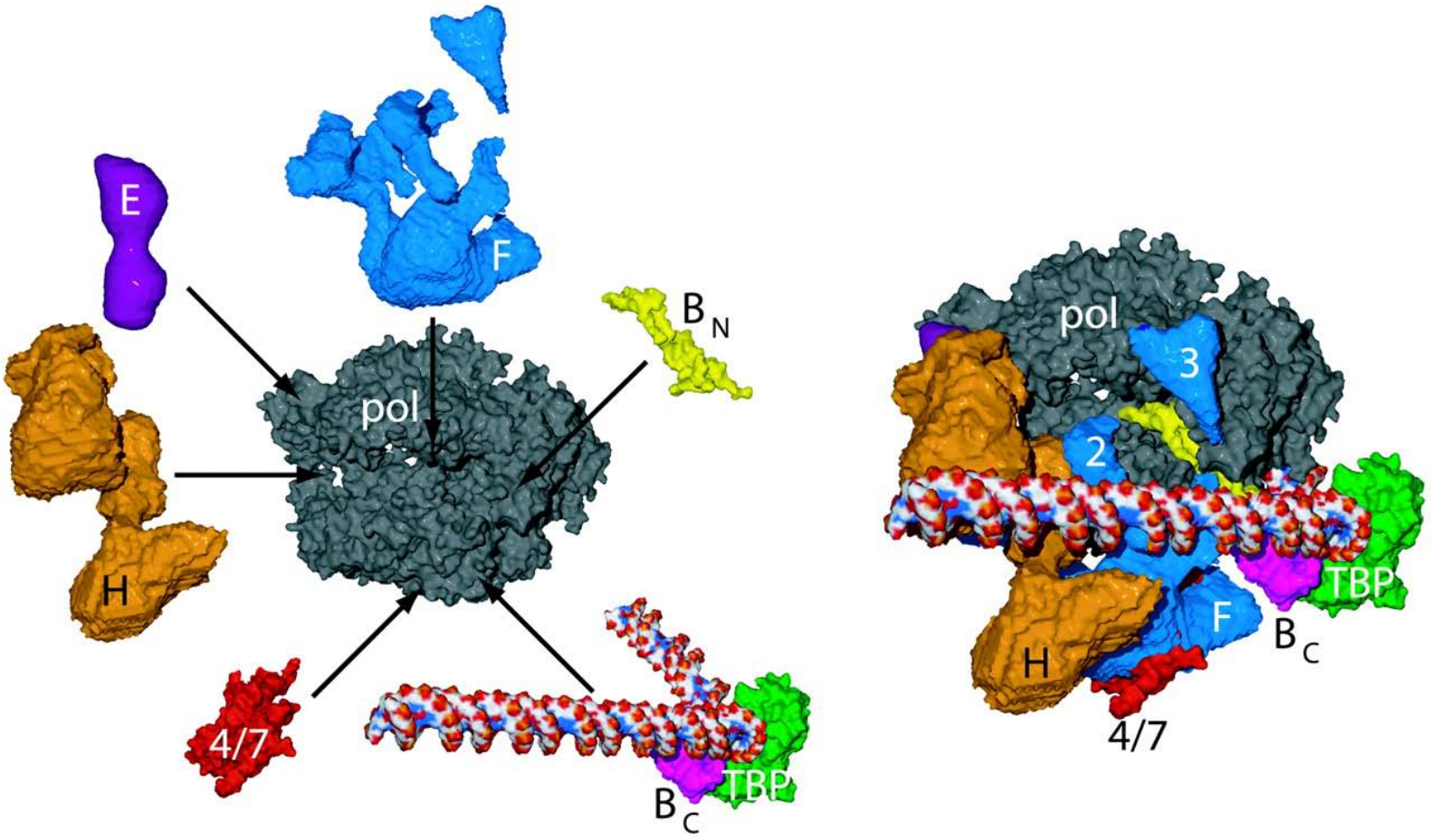


Figure 2.6 *Equilibrium sliding model for translocation.* (A) A schematic view of the elongating polymerase after a round of phosphodiester bond formation, but before translocation. (B) A view of the elongating polymerase after translocation, which opens up the nucleotide-binding



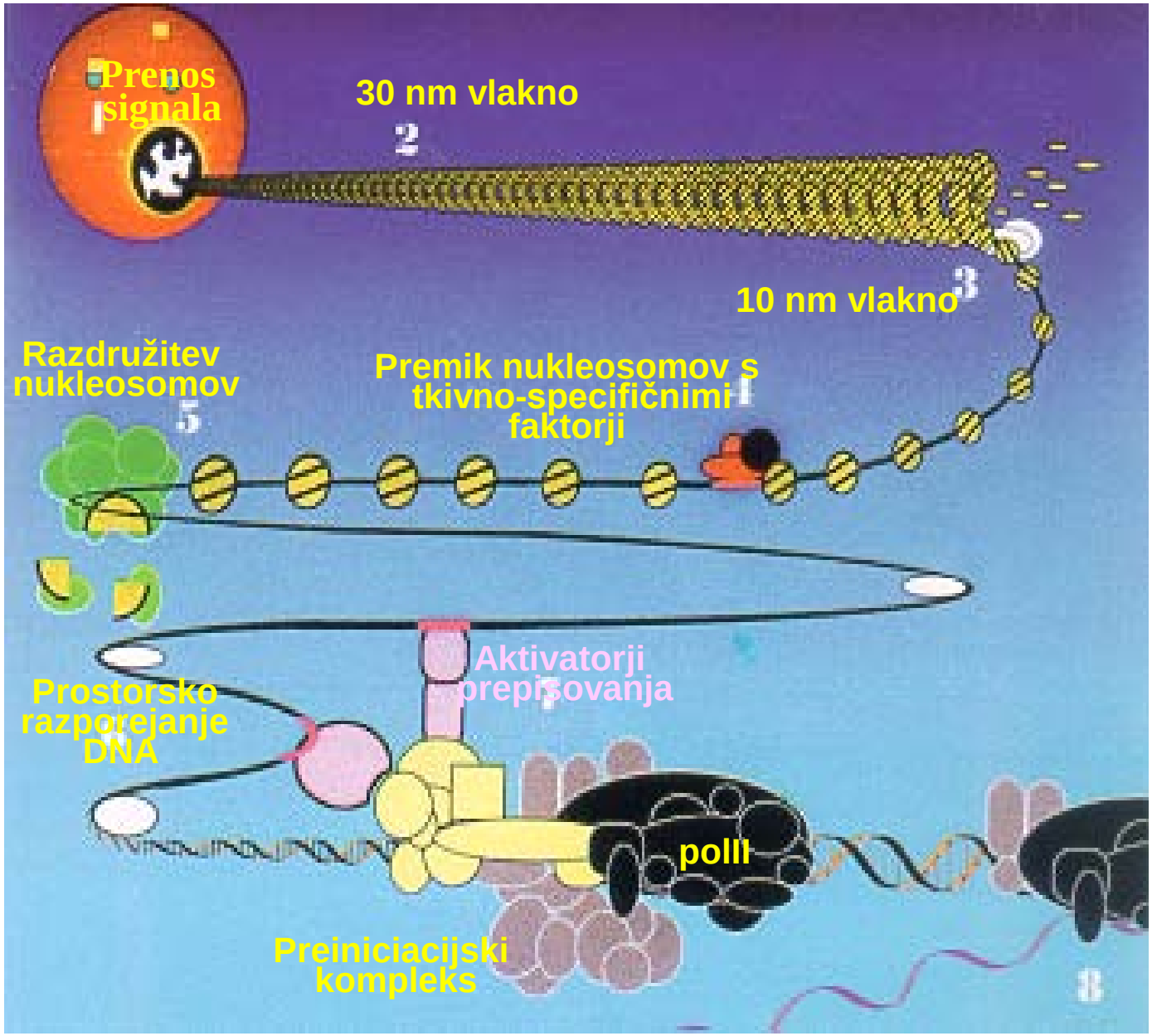
Model of the pol II transcription initiation complex with the addition of electron microscope structures of TFIIE, TFIIF, and TFIIH.

Bushnell et al., Science 303, 983-988 (2004)

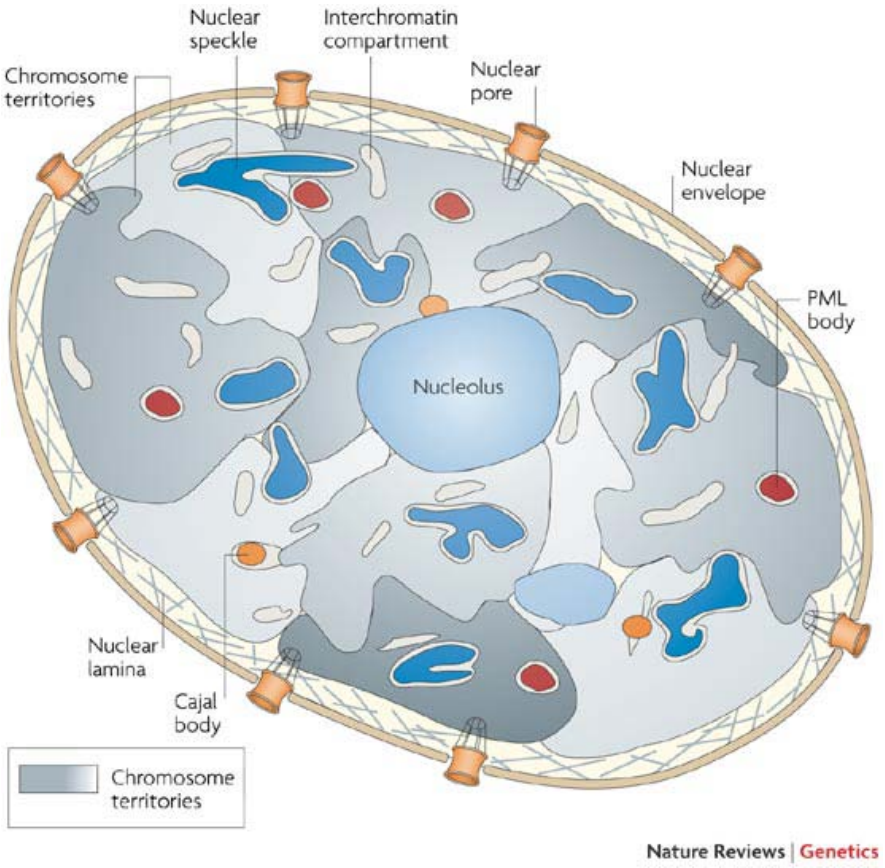
Splošni transkripcijski faktorji RNA polimeraze II

- TFIID: velik multiproteinski kompleks, ki prične sestavljanje transkripcijskega aparata. Vsebuje TBP in TAFe
- TFIIB: interagira z TBP in z DNA (analog bakterijskih σ faktorjev)
- TFIIF: udeležen pri iniciaciji kot pri podaljševanju verige.
- TFIIE: veže TFIIH
- TFIIH ima kinazno aktivnost, je helikaza, sodeluje tudi pri popravljanja DNA. Predstavlja povezavo med prepisovanjem in popravljanjem DNA in uravnavanjem celičnega cikla.
- TFIIA - po nekaterih izsledkih ne spada med splošne transkripcijske faktorje temveč med specializirane TFIID koaktivatorje. V pogojih *in vitro* veže TBP-DNA kompleks in povzroči širjenje DNA-protein kontaktov, obstajajo pa tudi dokazi, da ima pomen le pri kontaktiranju z gensko-specifičnimi transkripcijskimi faktorji.

Molekularni dogodki pri izražanju genov pri evkariontih



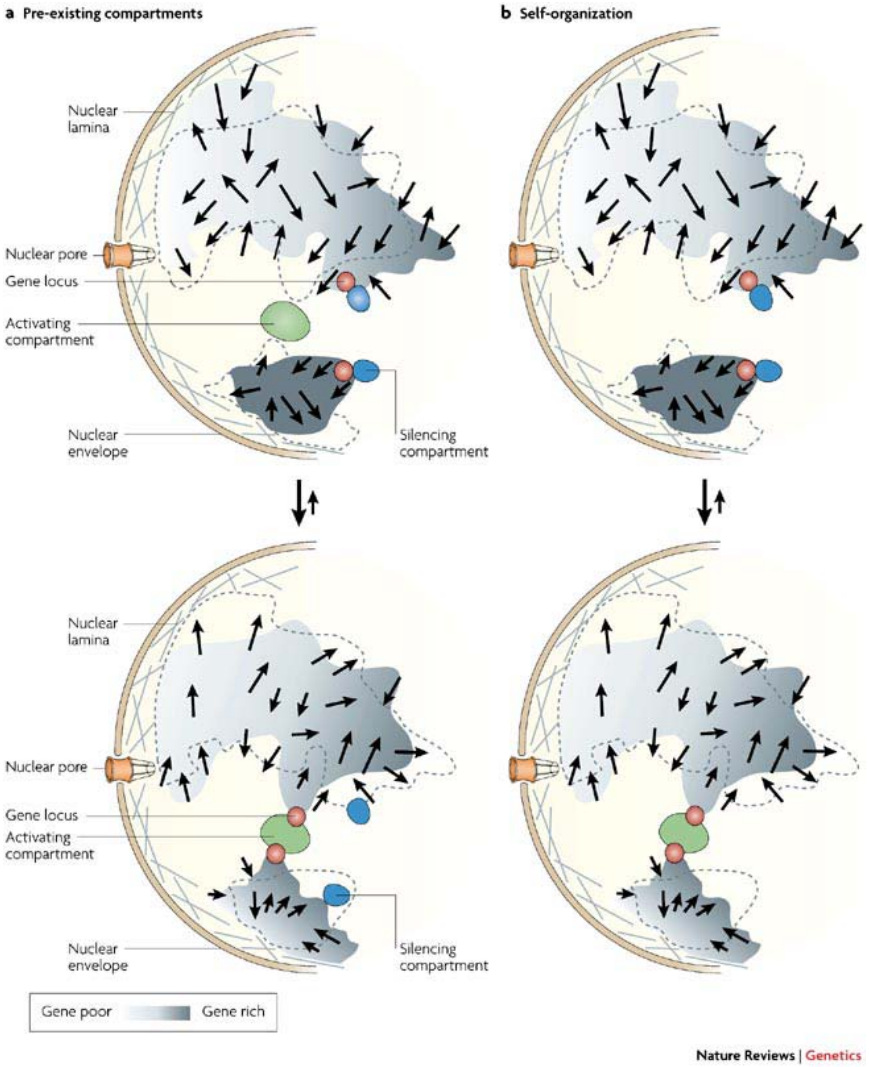
The chromosome territory–interchromatin compartment (CT–IC) model



Lanctôt *et al*. *Nature Reviews Genetics* **8**, 104–115 (February 2007) | doi:10.1038/nrg2041

Chromatin is organized in distinct CTs. Nuclear topography remains a subject of debate, expecially with regard to the extent of chromatin loops expanding into the IC and intermingling between neighbouring CTs and chromosomal subdomains.

Chromatin mobility allows dynamic interactions between genomic loci and between loci and other nuclear structures.



a) Movement of chromatin from one compartment to another leads to changes in expression of the corresponding genomic regions.

Activation is triggered by repositioning of the gene loci to an activating compartment, away from silencing compartments.

b) Compartments are transient self-organizing entities.

Gene activation leads to dissolution of the silencing compartments, changes in gene positioning and *de novo* assembly of an activating compartment.

Lanctôt *et al*. *Nature Reviews Genetics* **8**, 104–115 (February 2007) | doi:10.1038/nrg2041

nature REVIEWS GENETICS

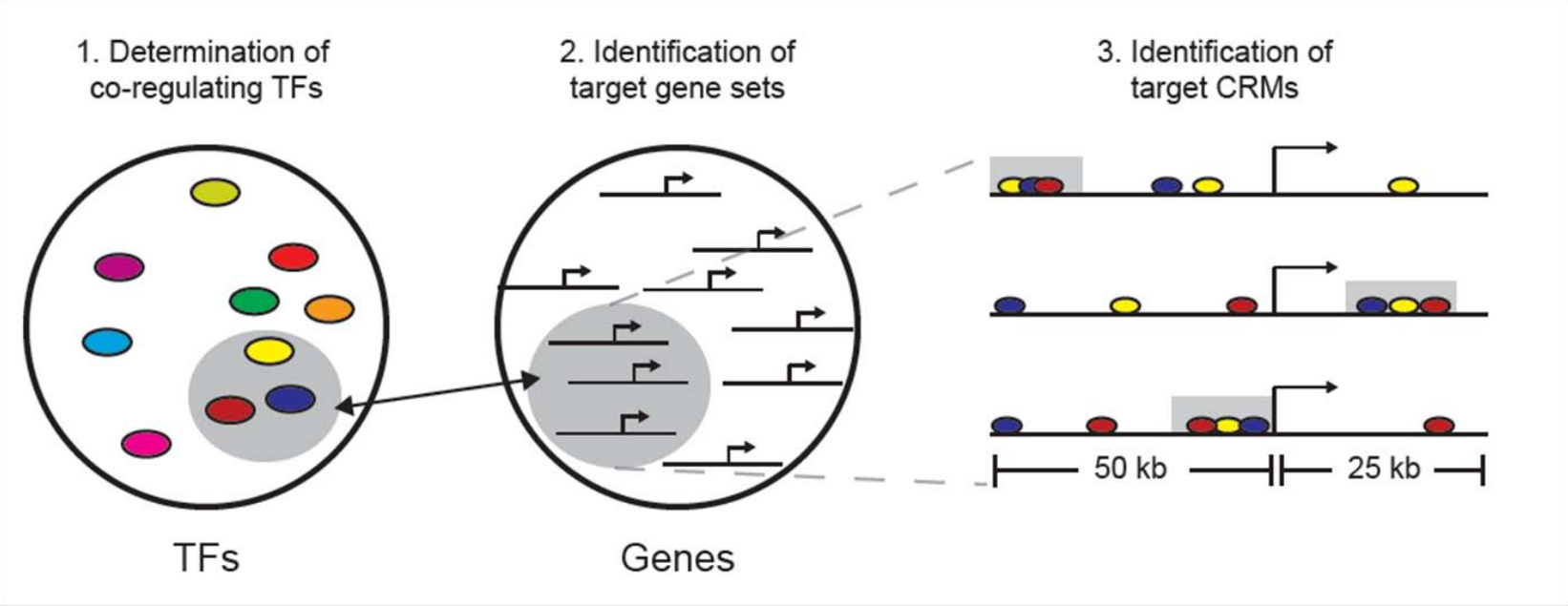
Transkripcijski aparat

•Osnovni aparat – RNA polimeraza

•Regulatorji (aktivatorji, represorji; trans-delujoči faktorji)

Cis-elementi

•Cis-regulatorni moduli (CRM); vzpodbujevalci (enhacers), utiševalci (silencers)



http://the_brain.bwh.harvard.edu/Lever/Lever.png

Struktura gena razreda pol II

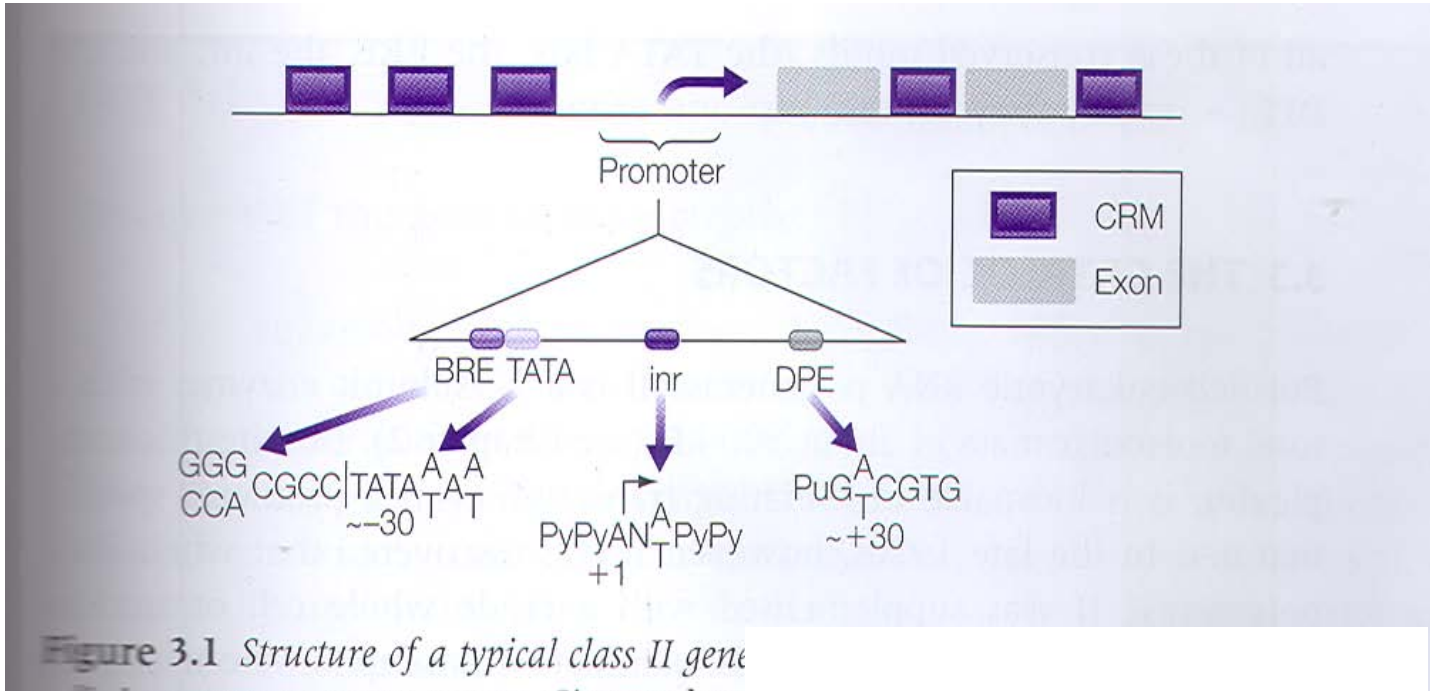
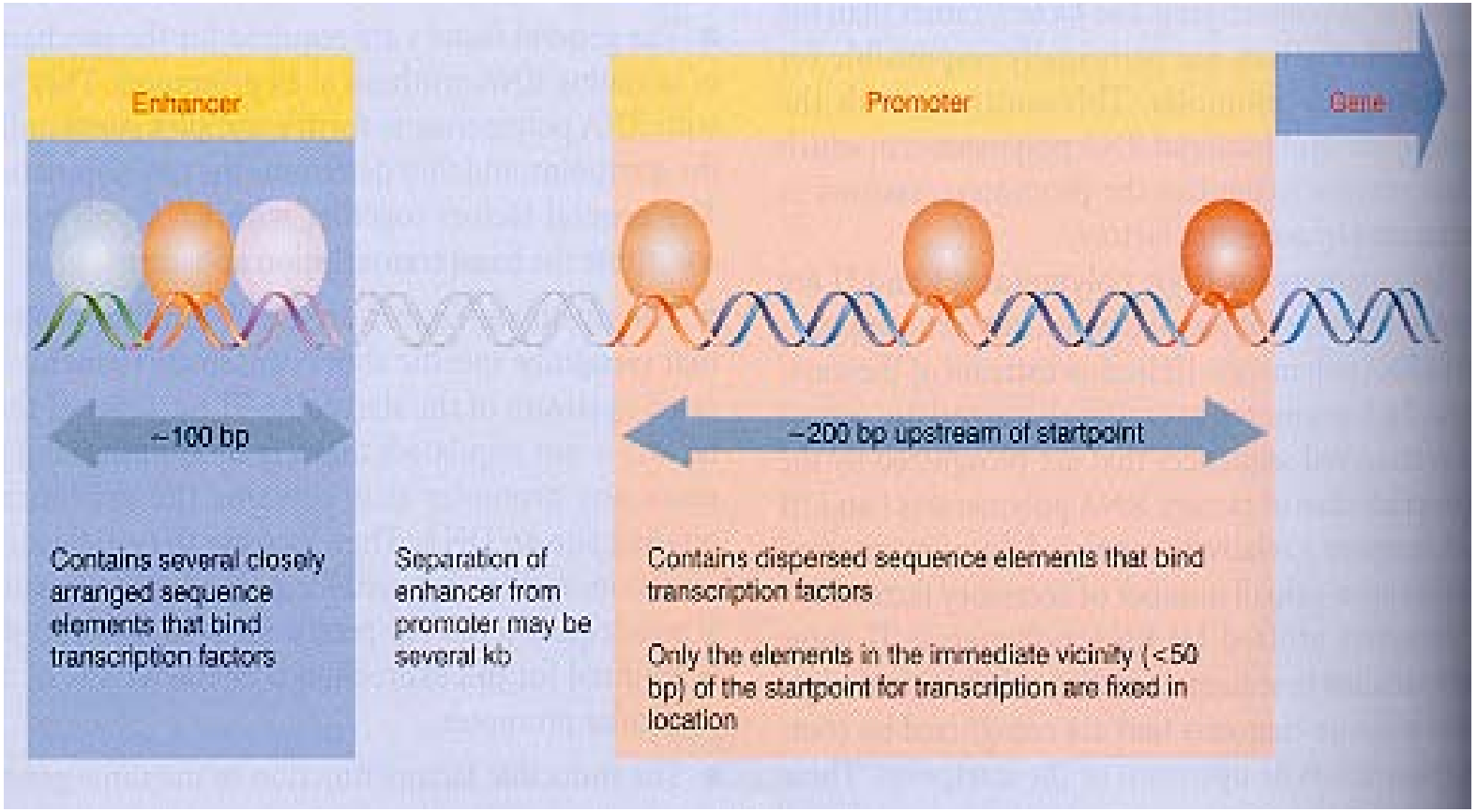


Figure 3.1 *Structure of a typical class II gene*

Tipični pol II gen ima promotor, ki zavzema področje cca 200 bp navzgor od mesta pričetka prepisovanja. Lahko vsebuje tudi pospeševalec (enhancer), ki je oddaljen poljubno.



Lewin, GenesVII

Core promoter-selective RNA polymerase II transcription

The **core promoter**:

-minimal DNA region that is sufficient to direct low levels of activator-independent (basal) transcription by RNAP II in vitro;

-typically extends approx. 40 bp up- and downstream of the start site of transcription;

-can contain several distinct core promoter sequence elements;

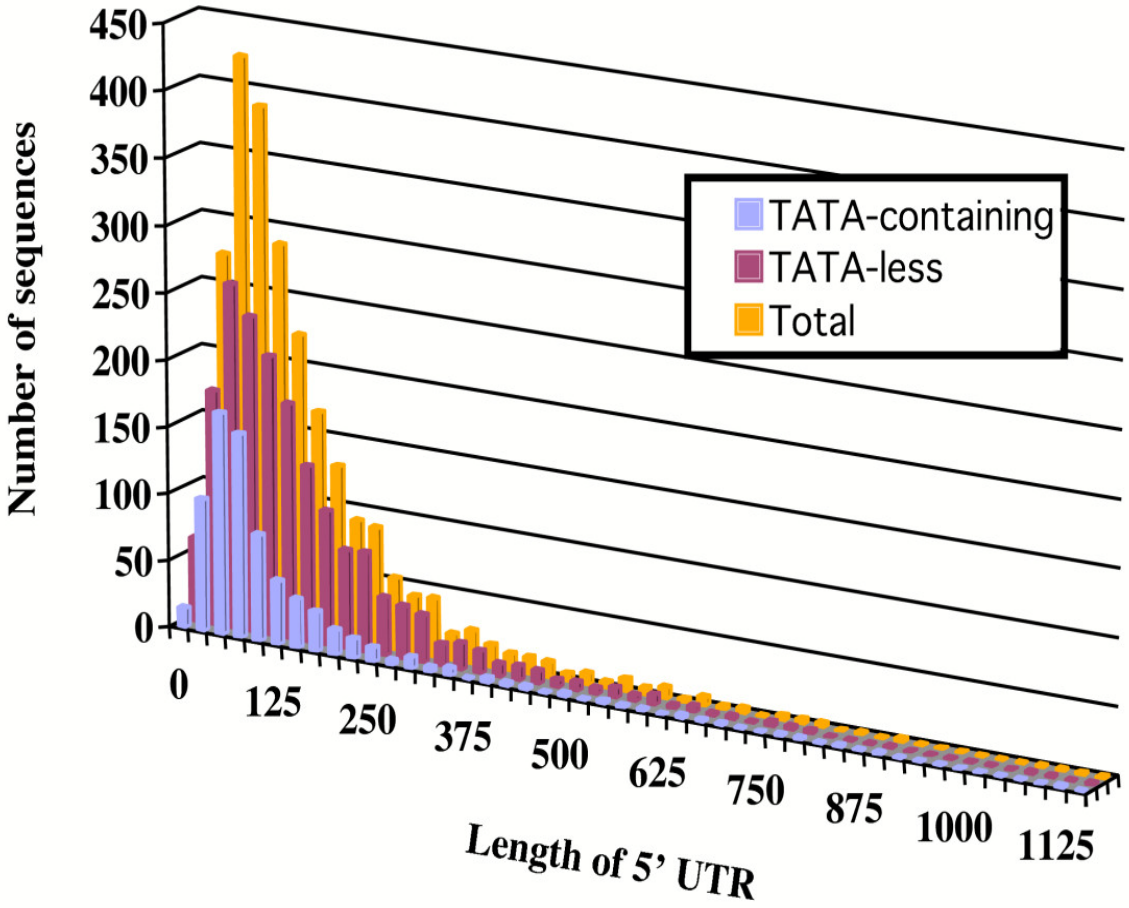
-contains TATA box or INR (initiator) element.

Computational analysis of metazoan genomes suggests that the prevalence of the TATA box has been overestimated in the past and that the majority of human genes are TATA-less.

While TATA-mediated transcription initiation has been studied in great detail and is very well understood, very little is known about the factors and mechanisms involved in the function of the INR and other core promoter elements.

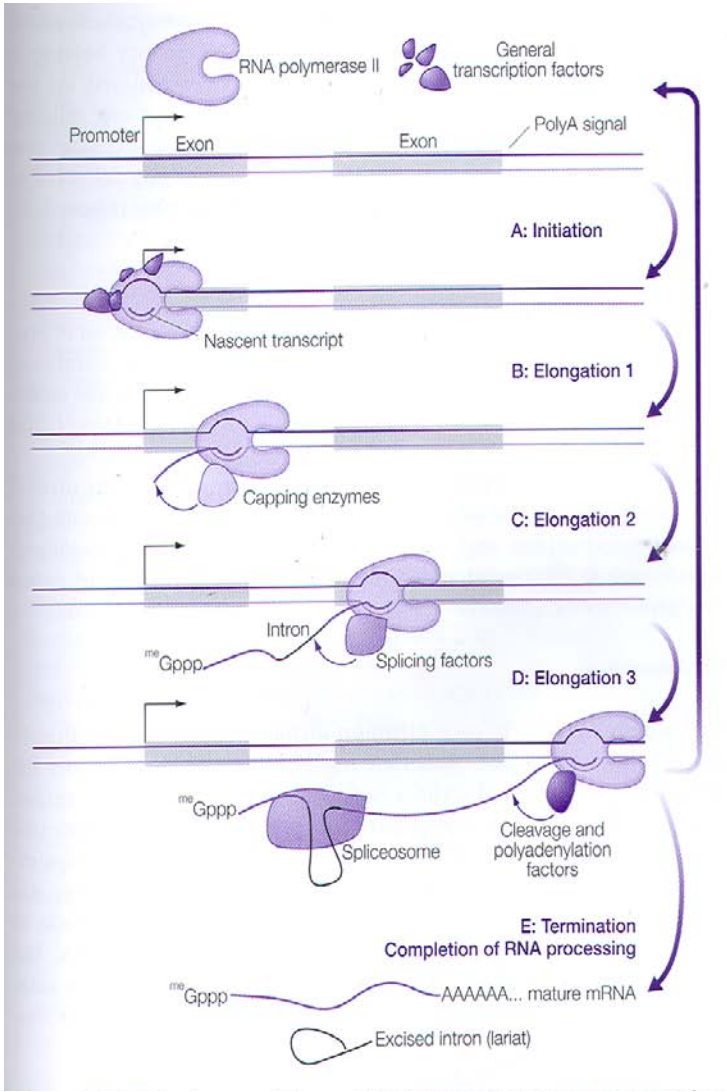
Biochem Soc Symp. 2006;(73):225-36.

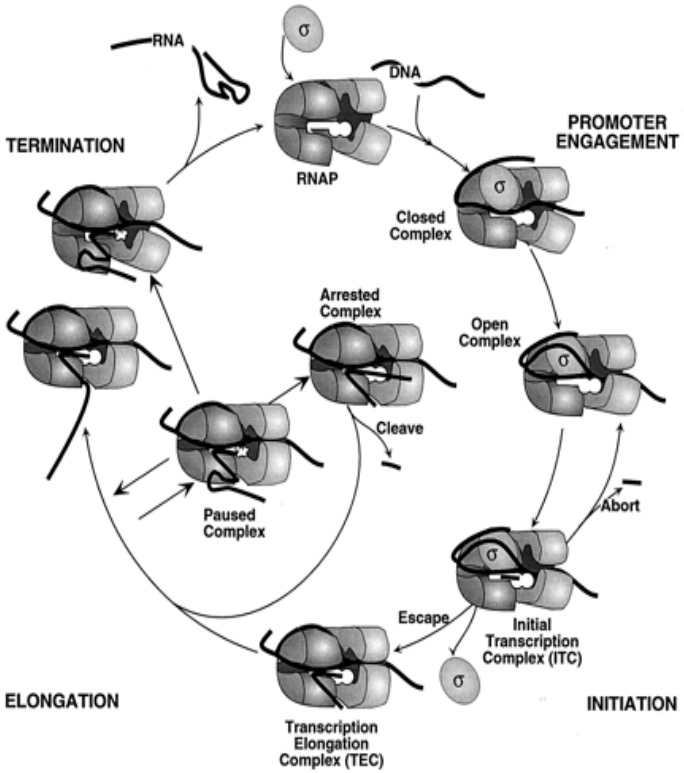
TATA-less promoters



Molina and Grotewold *BMC Genomics* 2005 **6:25**

Transkripcijski cikel





Komunikacija med CRM in promotorjem

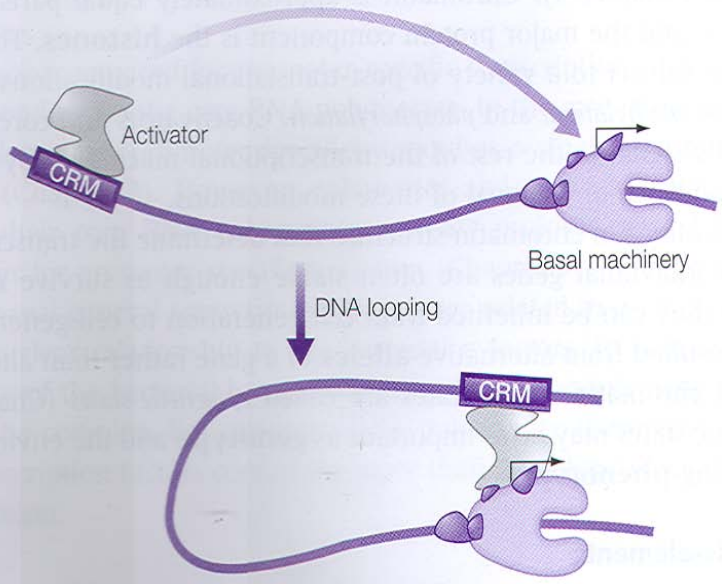
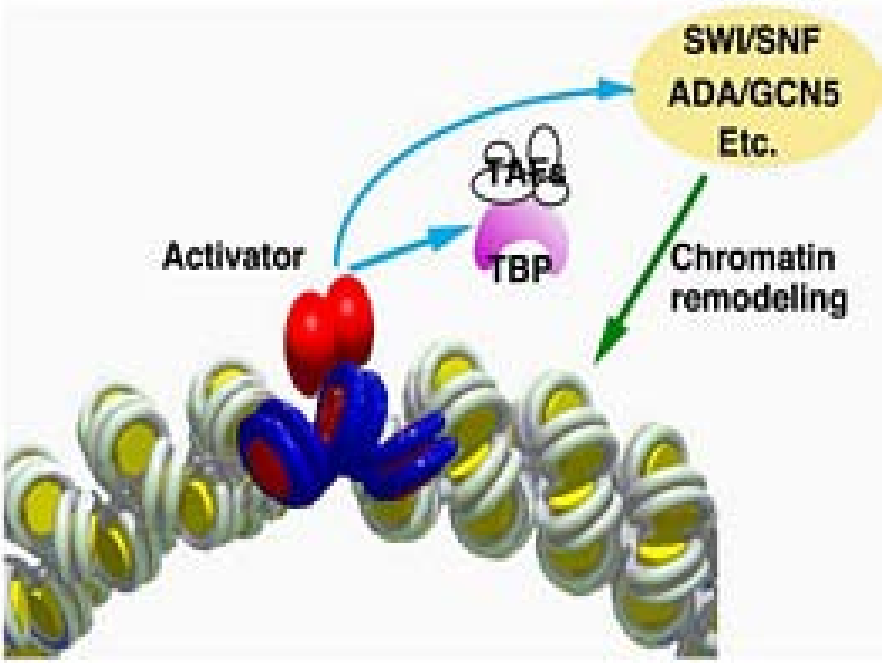


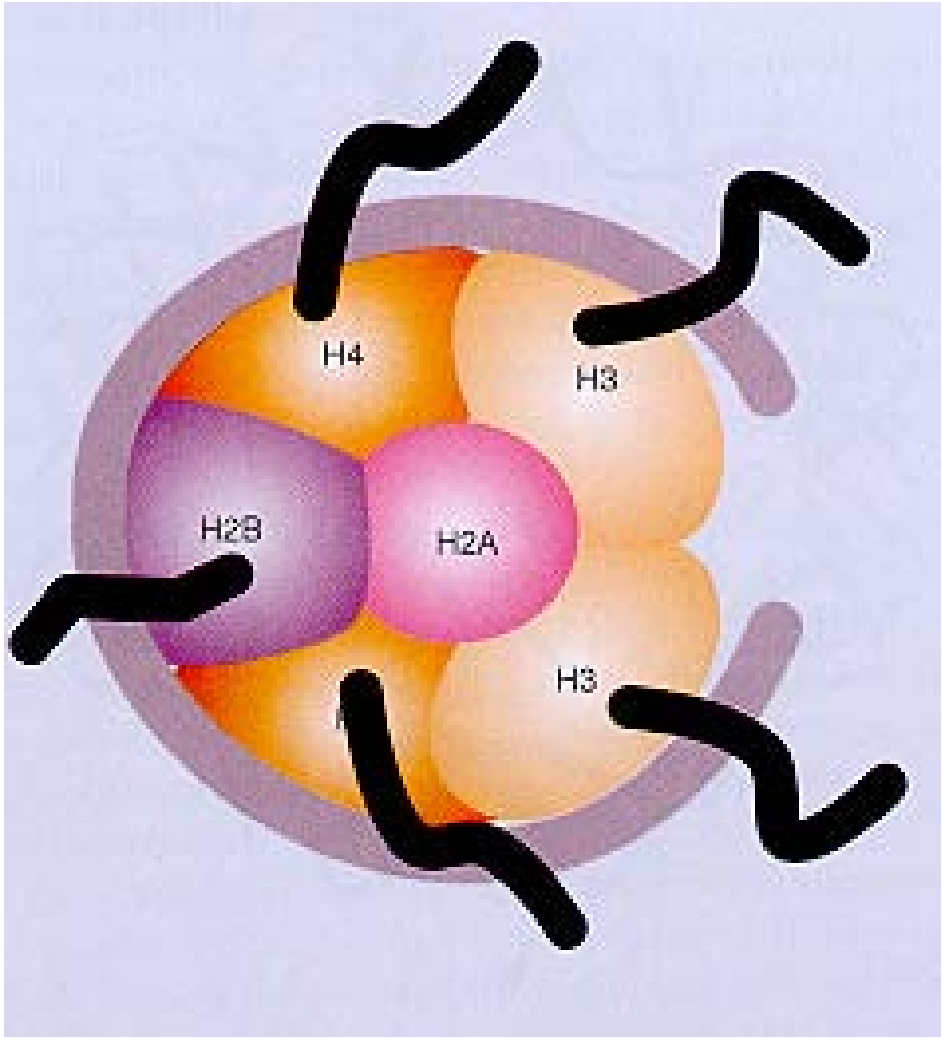
Figure 1.2 *Communication between cis-regulatory modules (CRMs) and promoters by DNA looping.* In eukaryotes, CRMs are often a long distance from the promoter. Communication between the CRM and the promoter may involve DNA looping, allowing direct contact between the regulatory factors (such as activators) bound to the CRM and the basal machinery bound to the promoter.

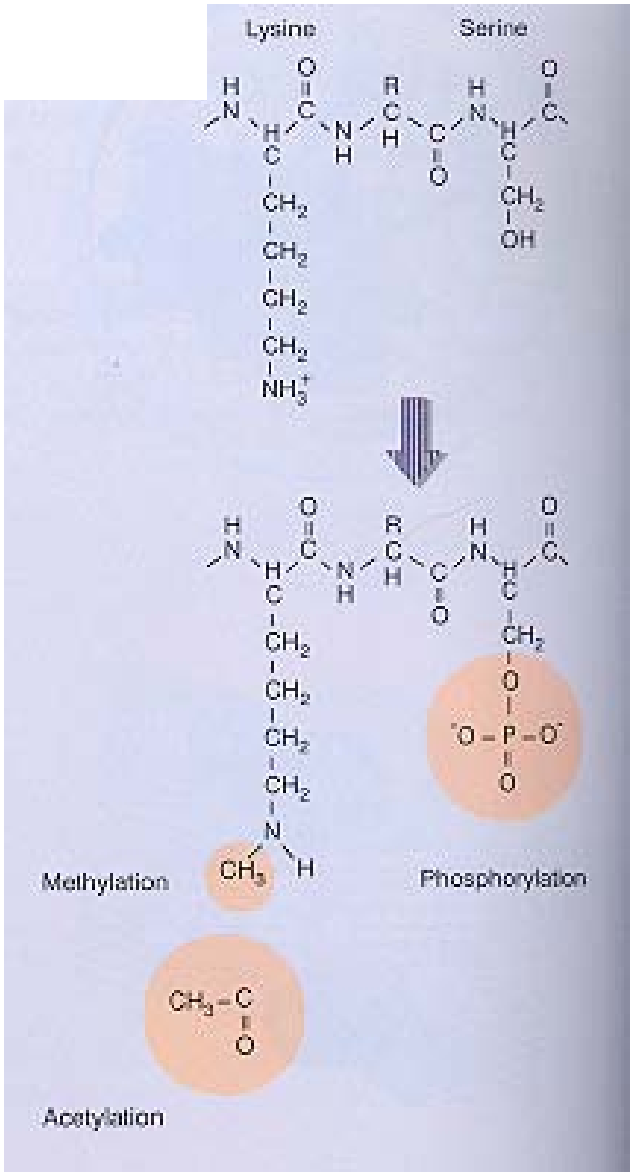
Kromatin in prepisovanje

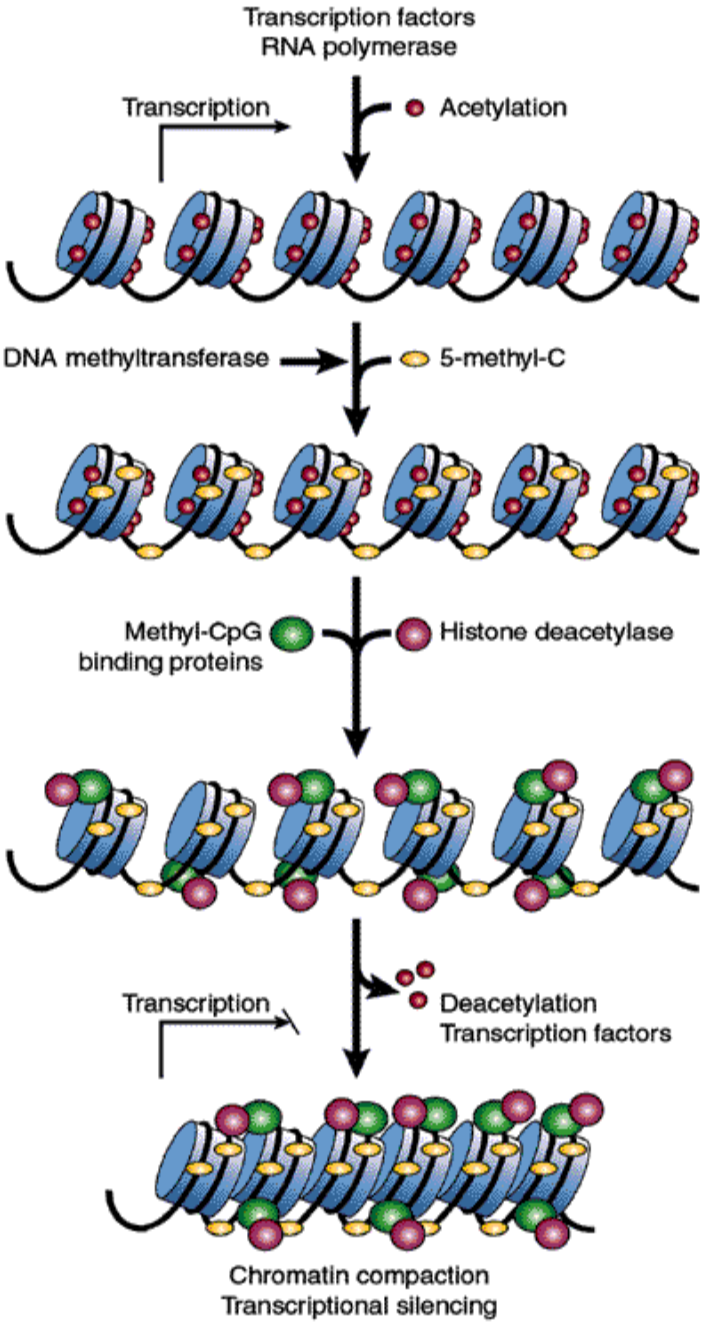


http://www.wadsworth.org/images/bio_icons/tn_transchromatin.jpg

N-terminalni repki histonov so gibljivi in obrnjeni navzven. Modifikacija histonskih repkov vodi do strukturnih sprememb, potrebnih pri pričetku prepisovanja







Potranslacijske spremembe histonov določajo strukturo kromatina

Acetilacija (HAT) pomeni aktivacijo.
Deacetilacija (HDAC) je inaktivacija.
Pomen metilacije je odvisen od konteksta.

www.med.ufl.edu/biochem/ keithr/fig1pt2.html

Povzetek I: RNA polimeraze, osnovni aparat in cikel transkripcije, kromatin

•Transkripcijo katalizirajo evolucijsko ohranjene RNA polimeraze. Evkarionti imajo RNA poI I, II in III. Potrebni pa so še drugi proteini (RNA polimeazni kompleks).

•Če polimeraza naleti na oviro, se lahko premakne nazaj (back-tracking).

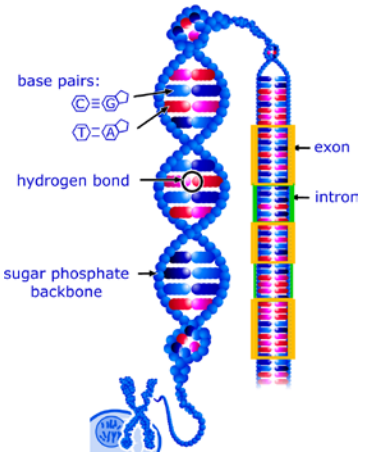
•Cis regulatorni moduli CRM preko vezave transkripcijskih faktorjev sodelujejo z RNA polimeraznim kompleksom.

•Cikel transkripcije vsebuje začetek (iniciacijo), podaljševanje (elongacijo) in zaključek (terminacijo).

•TBP je univerzalni transkripcijski faktor, ki prepozna TATA, potreben pa je tudi za prepisovanje s promotorjev brez TATA.

•Potranslacijske spremembe histonov določajo strukturo kromatina in s tem prepisovanje oziroma utišanje genov.

2. Študije transkriptoma z DNA mikromrežami (čipi), povezave z drugimi “omi”



The suffix **“*-ome-*”** originated as a *back-formation* from “gen**ome**”, a word formed in analogy with “chromosome”.

Because “genome” refers to the *complete* genetic makeup of an organism, some people have made the inference that there exists some root, *“*-ome-*”, of Greek origin referring to *wholeness* or to *completion*.

Because of the success of large-scale quantitative biology projects such as genome sequencing, the suffix "*-ome-*" has migrated to a host of other contexts. Bioinformaticians and molecular biologists figured amongst the first scientists to start to apply the "-ome" suffix widely.

http://en.wikipedia.org/

The "*ome*" world

http://en.wikipedia.org/

The *transcriptome*, the mRNA complement of an entire organism, tissue type, or cell; with its associated field *transcriptomics*

The *metabolome*, the totality of metabolites in an organism; with its associated field *metabolomics*

The *metallome*, the totality of metal and metalloid species; with its associated field *metallomics*

The *lipidome*, the totality of lipids; with its associated field *Lipidomics*

The *glycome*, the totality of glycans, carbohydrate structures of an organism, a cell or tissue type. *Glycomics*: The associated field of study.

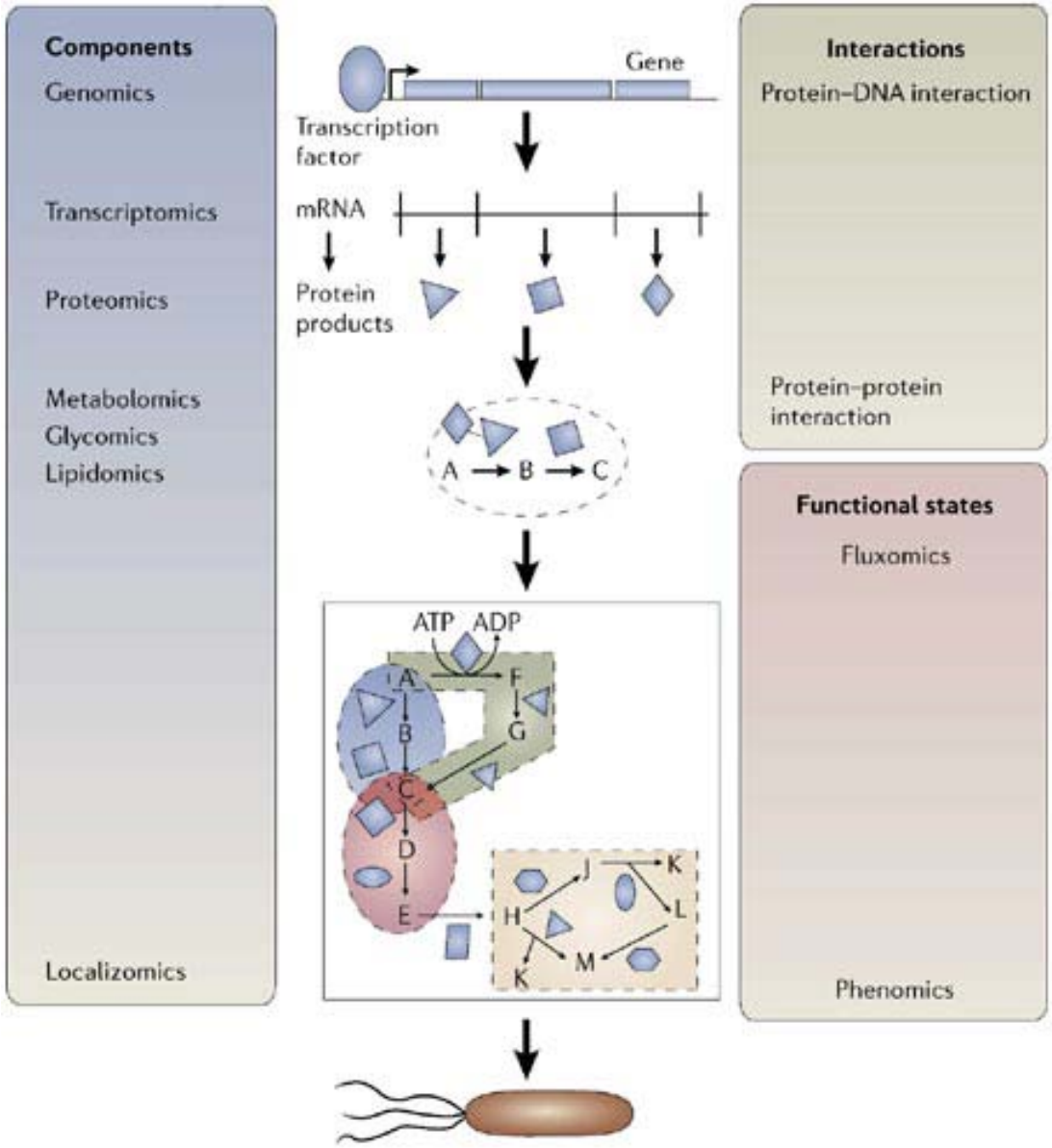
The *interactome*, the totality of the molecular interactions in an organism *a once proposed field of interactomics* has generally become known as systems biology

The *spliceome* (see spliceosome), the totality of the alternative splicing protein isoforms; with its associated field *spliceomics*.

The *ORFeome* refers to the totality of DNA sequences that begin with the initiation codon ATG, end with a nonsense codon, and contain no stop codon. Such sequences may therefore encode part or all of a protein.

The *Phenome* - the organism itself. The Phenome is to the genome what the phenotype is to the genotype. Also, the complete list of phenotypic mutants available for a species.

The *Exposome* - the collection of an individual's environmental exposures.



Nature Reviews | Molecular Cell Biology

Genomics

http://en.wikipedia.org/

Genomics is the study of an organism's entire genome.

The field includes intensive efforts to determine the entire DNA sequence of organisms and fine-scale genetic mapping efforts.

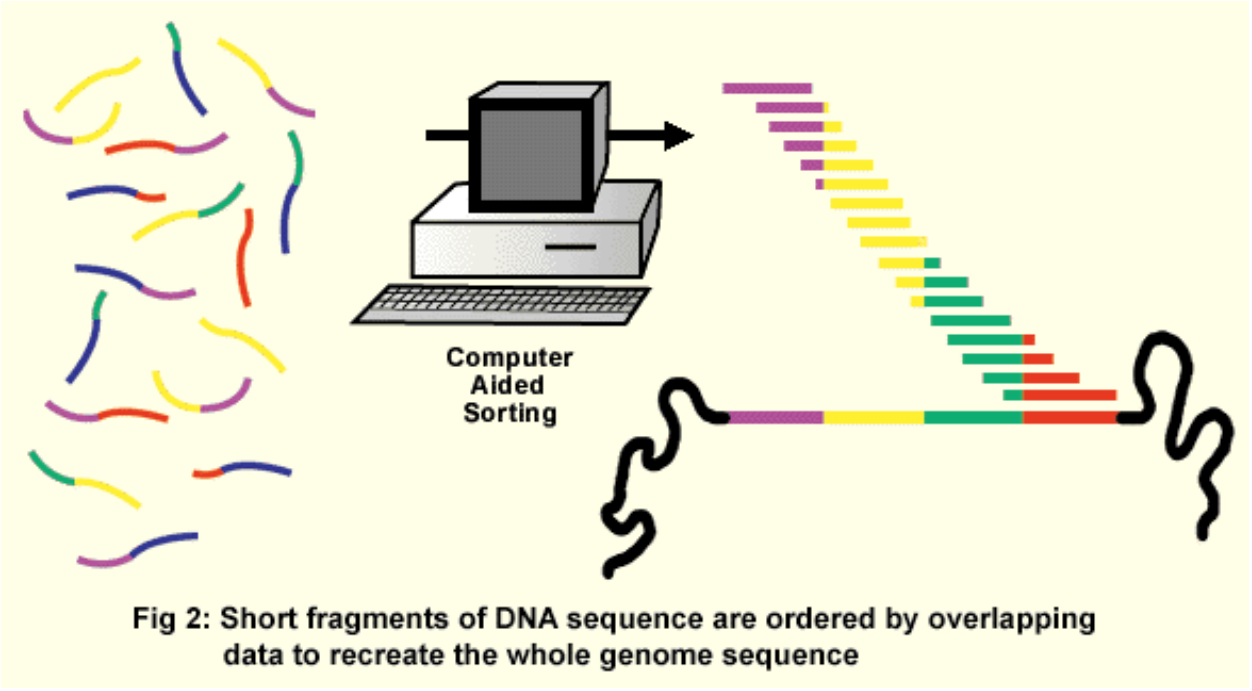


Fig 2: Short fragments of DNA sequence are ordered by overlapping data to recreate the whole genome sequence

http://www.wiley.com/legacy/college/boyer/0470003790/cutting_edge/shotgun_seq/computer.gif

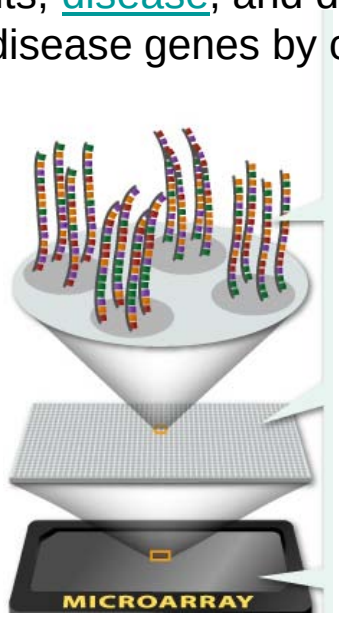
A DNA microarray

A **DNA microarray** (also commonly known as *gene chip*, *DNA chip*, or *biochip*) is a collection of microscopic DNA spots attached to a solid surface, such as glass, plastic or silicon chip forming an array for the purpose of monitoring of expression levels for thousands of genes simultaneously.

The affixed DNA segments are known as *probes* (although some sources will use different nomenclature), thousands of which can be used in a single DNA microarray.

Microarray technology evolved from Southern blotting, where fragmented DNA is attached to a substrate and then probed with a known gene or fragment.

Measuring gene expression using microarrays is relevant to many areas of biology and medicine, such as studying treatments, disease, and developmental stages. For example, microarrays can be used to identify disease genes by comparing gene expression in disease and normal cells.



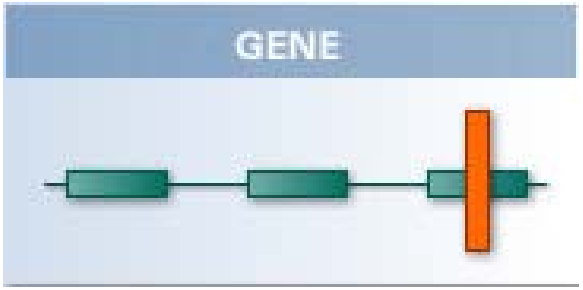
DNA čipi (DNA mikromreže)

- **Analiza genoma** (genotpizacija enojnih nukleotidnih polimorfizmov SNP, analiza variance števila kopij CNV oz. CGH).

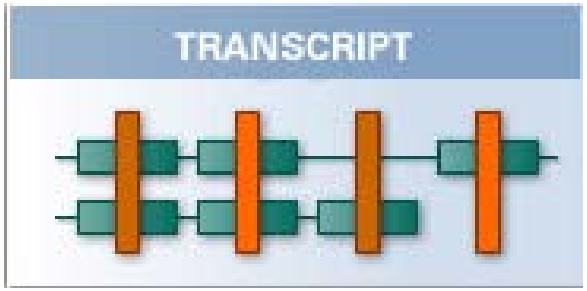
- **Analiza transkriptoma** (ekspresijsko profiliranje: 3' ekspresijski, eksonski ali genski čipi, čipi za sledenje izražanja miRNA).

- **Študije uravnavanja izražanja genov** (kromatinska imunoprecipitacija na čipu ChIP-on-Chip, mapiranje prepisov).

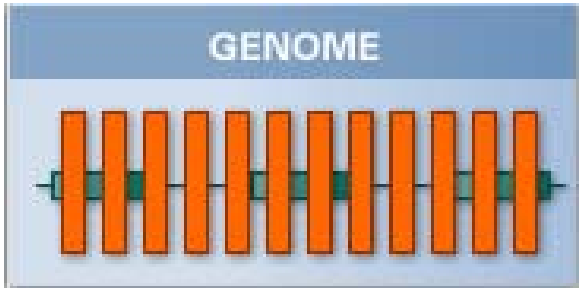
Different parts of the gene as **probes** on the array



Robust, simple representation focusing on the 3' ends.



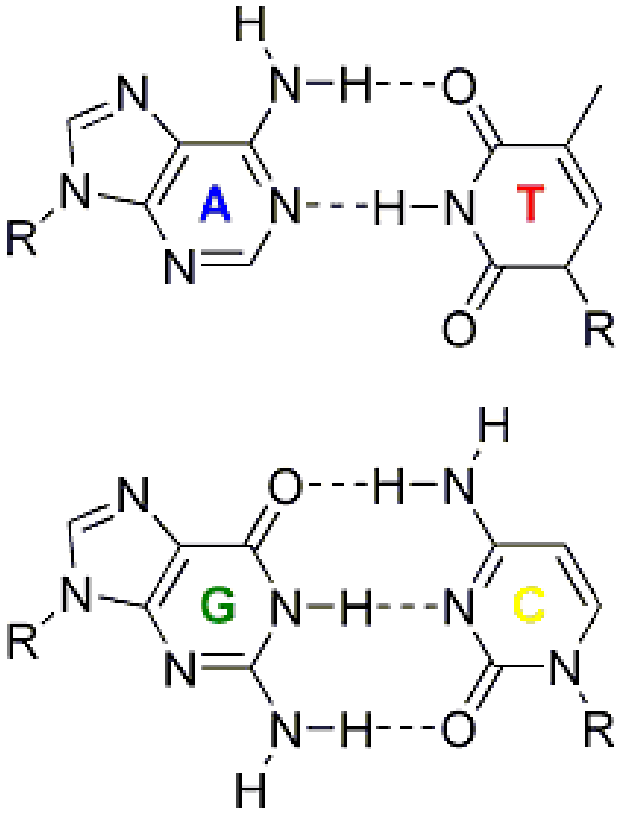
Genome-wide, exon-level analysis on a single array —
a survey of alternative splicing and gene expression.



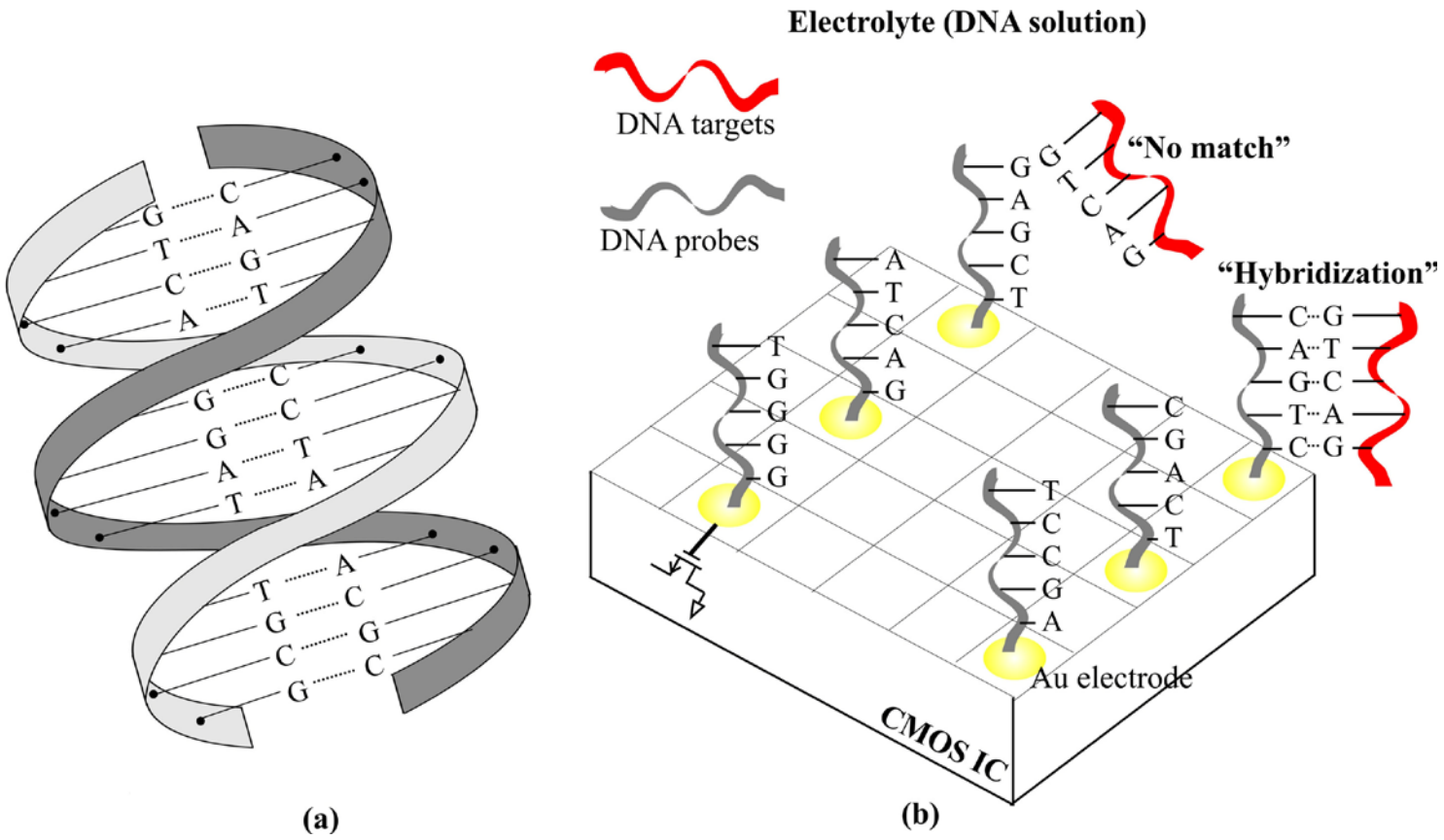
High-density tiled microarrays for transcript mapping.

http://www.affymetrix.com/products/arrays/index.affx

Hybridization: Watson-Crick base pairing; AT(U), GC

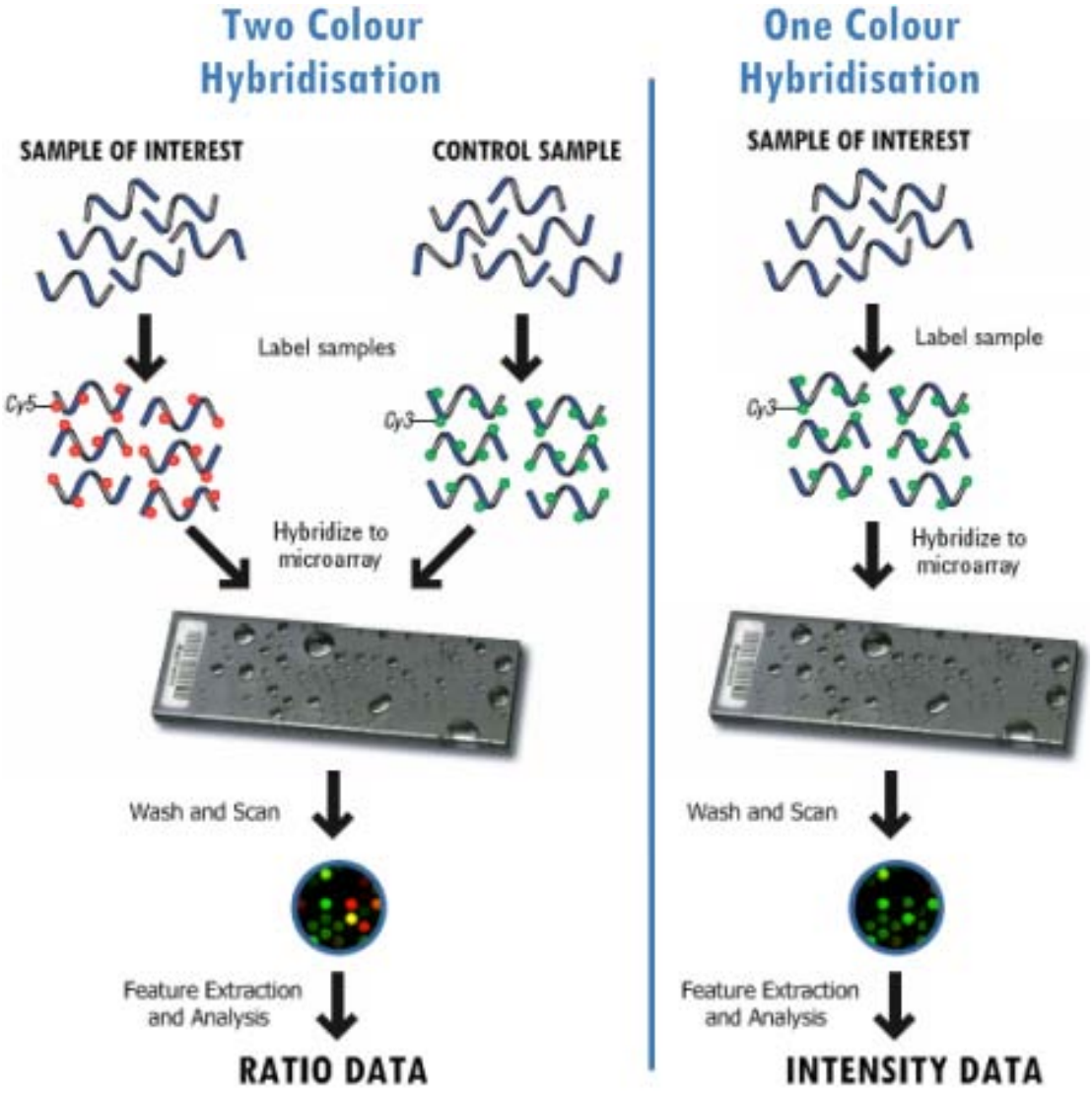


Hybridization on the microarray



http://www.ieee.org/portal/cms_docs_sscs/sscs/07Fall/HamFig1s.jpg

Detekcija signala po hibridizaciji DNA čipov



Methods of detection of differences at the genome level

- Mendelian Genetics,
- Direct DNA Sequencing,
- RFLP analysis,
  - Restriction Fragment Length Polymorphism,
- Allele Specific Oligonucleotides,
- DNA Microarrays
- Novel generation of high-throughput sequencing

Genotyping microarrays

DNA microarrays can be used to *read* the sequence of a genome in particular positions.

SNP microarrays are a particular type of DNA microarrays that are used to identify genetic variation in individuals and across populations.

Short oligonucleotide arrays can be used to identify the single nucleotide polymorphisms (SNPs) that are thought to be responsible for genetic variation and the source of susceptibility to genetically caused diseases.

Amplifications and deletions can also be detected using **comparative genomic hybridization** in conjuction with microarrays.

Resequencing arrays have also been developed to sequence portions of the genome in individuals. These arrays may be used to evaluate germline mutations in individuals, or somatic mutations in cancer.
Genome tiling arrays include overlapping oligonucleotides designed to blanket an entire genomic region of interest. Many companies have successfully designed tiling arrays that cover whole human chromosomes.

If every human genome is different, what does it mean to sequence "the" human genome?

The complete human genome sequence announced in June 2000 is a "representative" genome sequence based on the DNA of just a few individuals.

Over the longer term, scientists will study DNA from many different people to identify where and what variations between individual genomes exist. Sequencing a genome is such a Herculean task that capturing its person-to-person variability on the first pass would be next to impossible.

Since every person's genome is unique, no one person is any more or less "representative" than any other and it hardly matters whose genome is sequenced first.

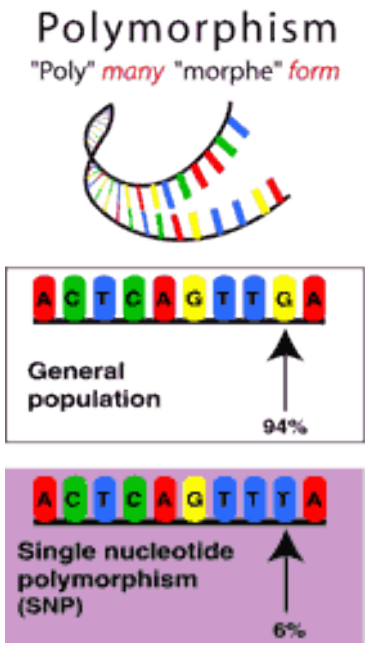
The vast majority of the genome's sequence is the same from one person to the next, with the same genes in the same places.

www.genomenewsnetwork.org/.../ Chp4_1.shtml

Why is every human genome different?

Where are genome variations found?

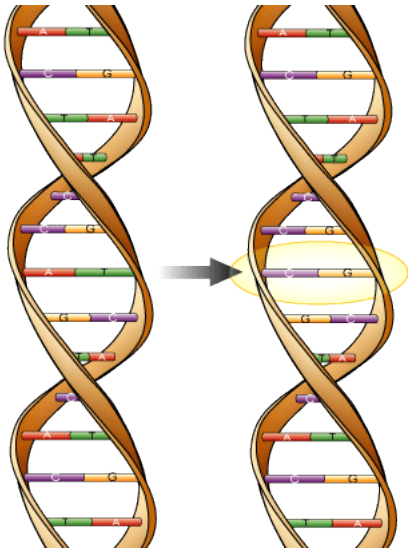
What kinds of genome variations are there?



DNA Polymorphism

…a DNA locus that has two or more sequence variations, each present at a frequency of 1% or more in a population,

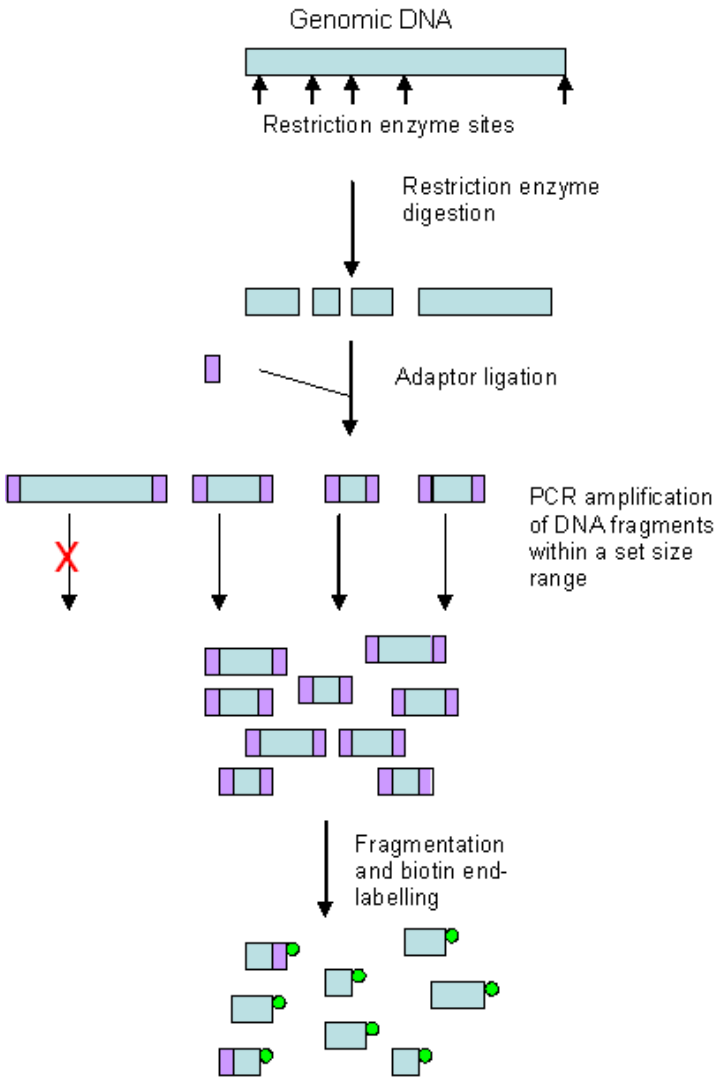
- 1 in 700 frequency common in most species,
- less than 1 million loci in humans (1 in 3,000).

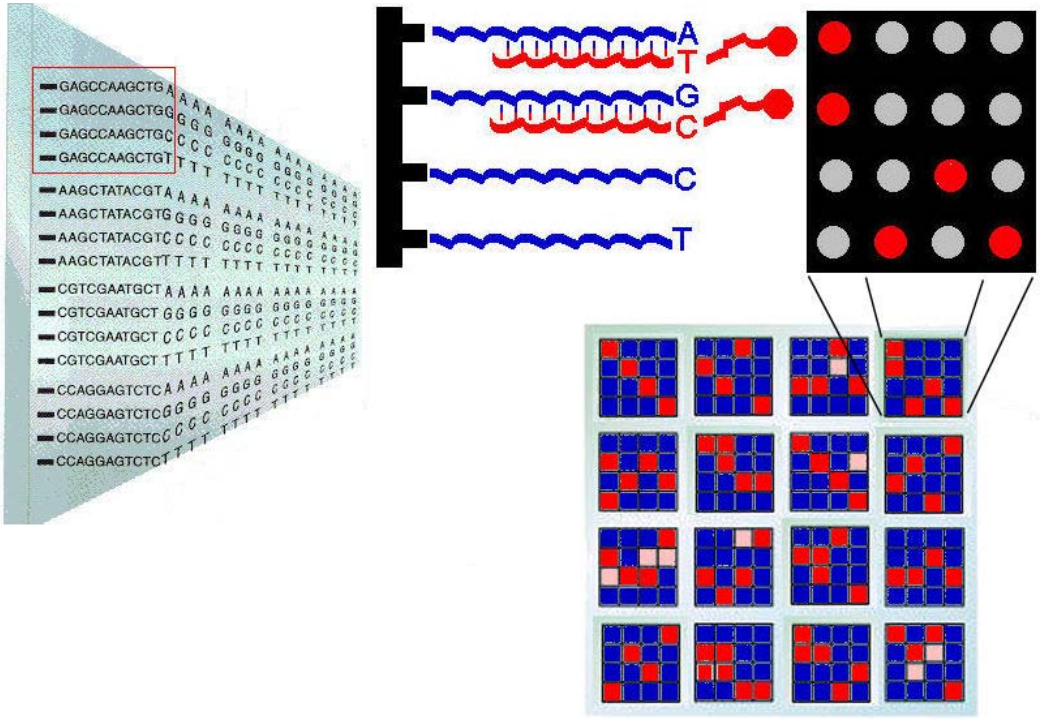


Most SNPs have only two alleles

The human genome contains more than 2 million SNPs.

DNA preparation for SNP array analysis





SNP chips

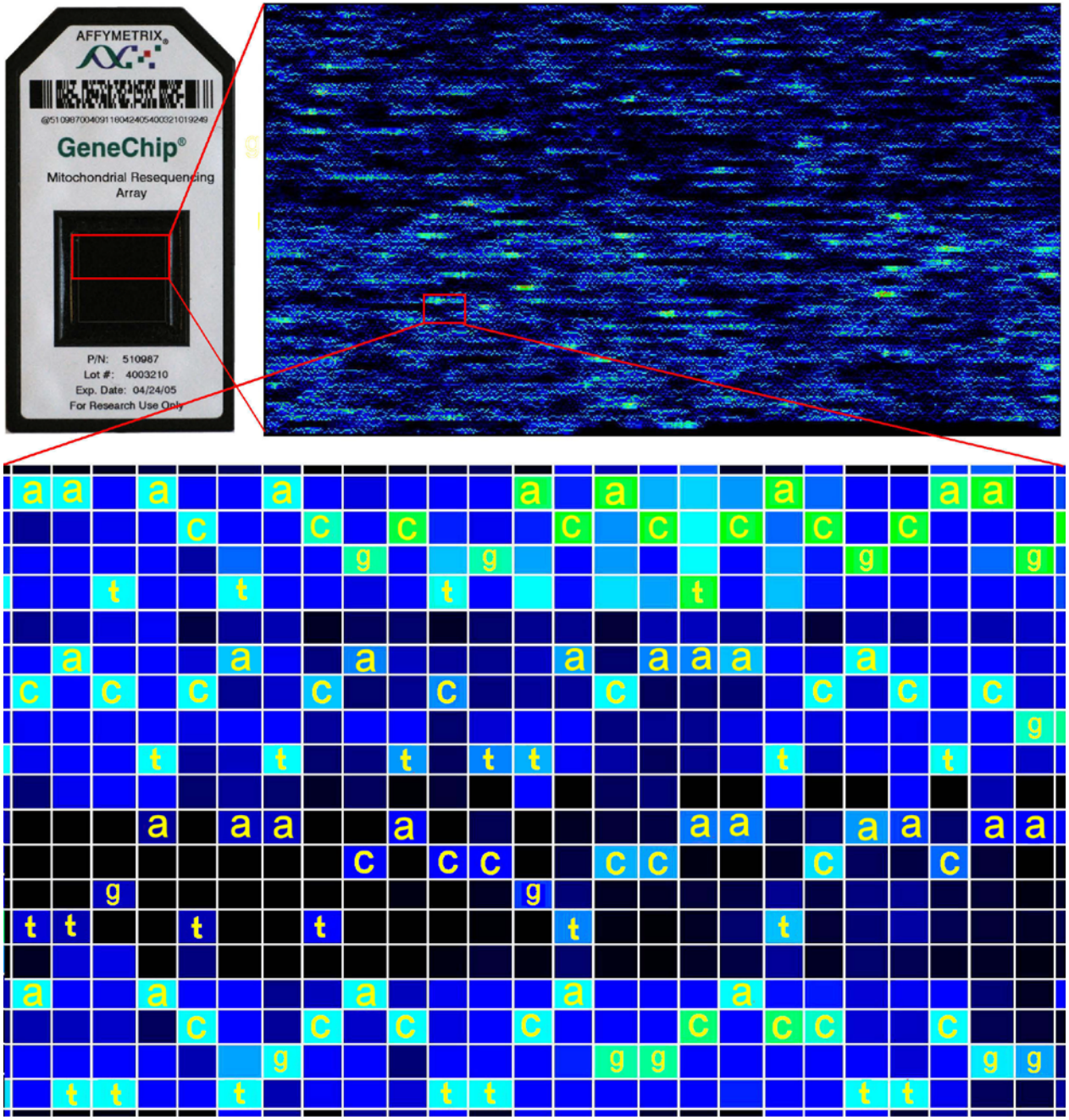
To determine which alleles are present, **genomic DNA** from an individual is isolated, fragmented, tagged with a fluorescent dye, and applied to the chip.

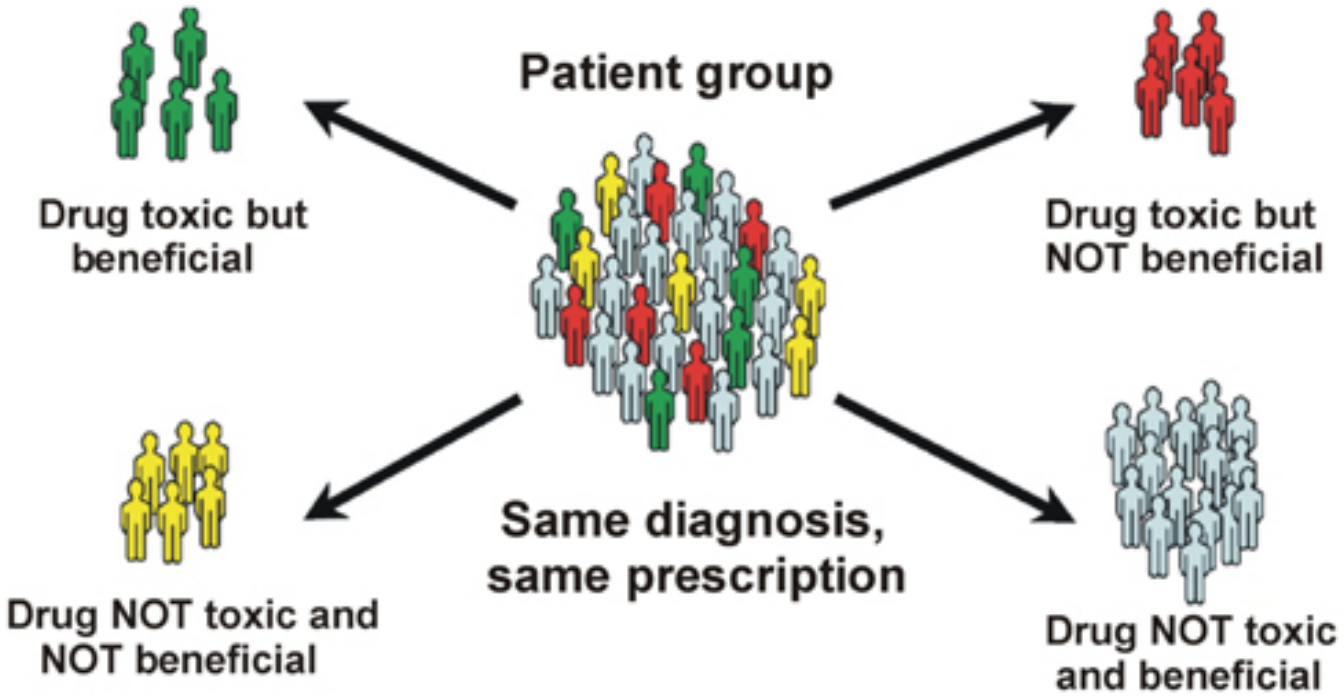
The **genomic DNA** fragments anneal only to these **oligos** to which they are perfectly complementary.

A computer reads the position of the two fluorescent tags and identifies the individual as a **C / T heterozygote**.

The *single* spots in the other three columns indicate that the individual is *homozygous* at the three corresponding **SNP** positions.

http://www.mun.ca/biology/scarr/DNA_VDA_chip_schematic.jpg


AFFYMETRIX®
GeneChip®
Mitochondrial Resequencing Array
P/N: 510987
Lot #: 4003210
Exp. Date: 04/24/05
For Research Use Only

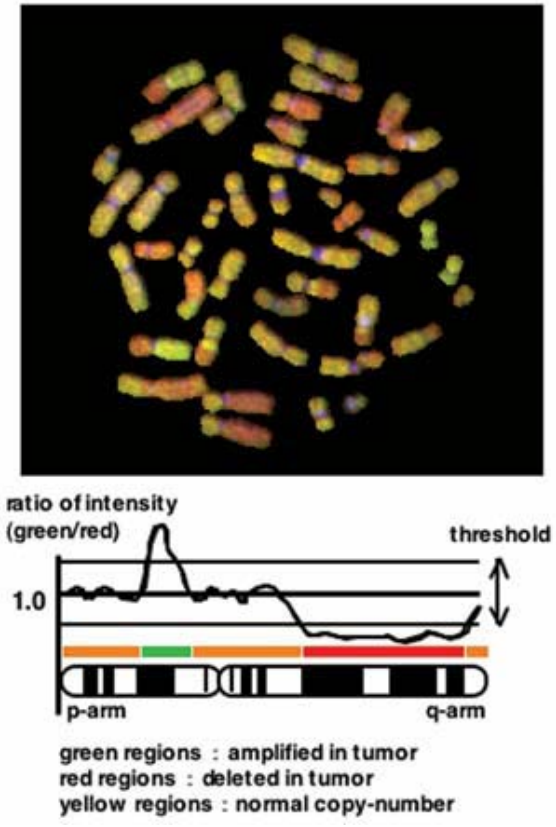


Pharmacogenomics principle: To identify patients at risk for toxicity or reduced response to therapy for optimal medication and/or dose selection

Comparative genomic hybridization

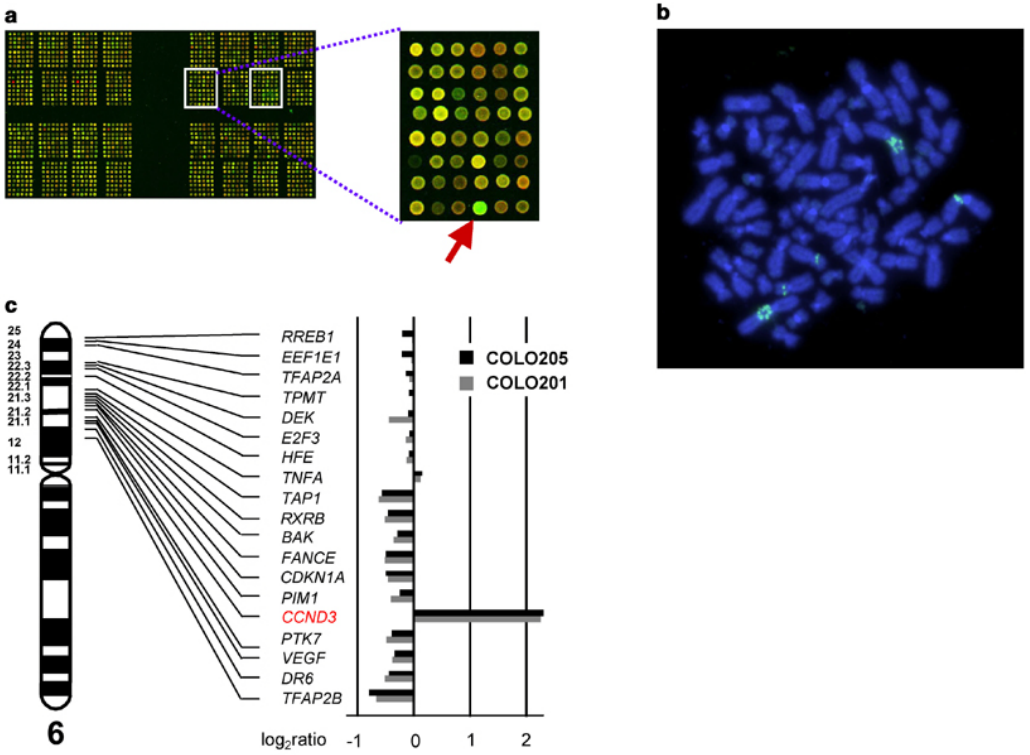
Array comparative genomic hybridization (CGH) is a tool for the assessment of DNA copy number variation (CNV) within any given DNA sample.

The technique is derived from the concept of conventional CGH where patient and reference DNA (the latter derived from a normal male/female) is differentially labelled and co-hybridized to a normal metaphase spread.



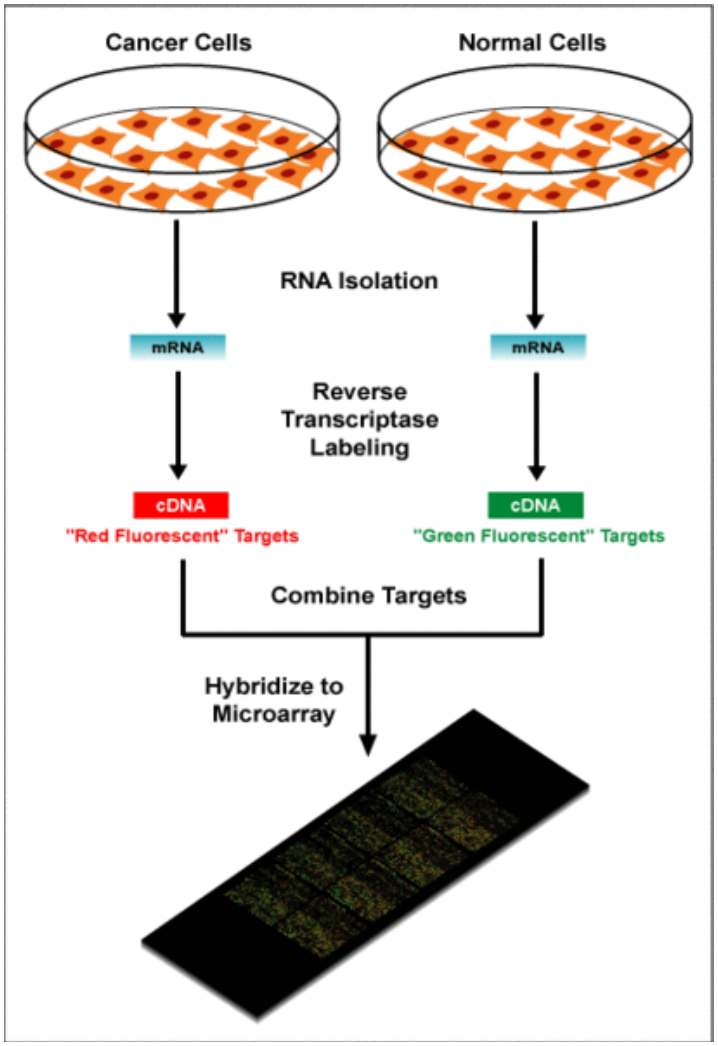
Some diseases are associated with sections of chromosomes that are erroneously replicated or deleted.

The technique of array-based genomic hybridisation (array-comparative genomic hybridization CGH or copy number variation CNV) allows high-throughput, rapid identification and mapping of these genomic DNA copy number changes, with higher resolution than is possible using traditional, non-array methods.



http://www.nature.com/labinvest/journal/v85/n9/images/3700312f2.jpg

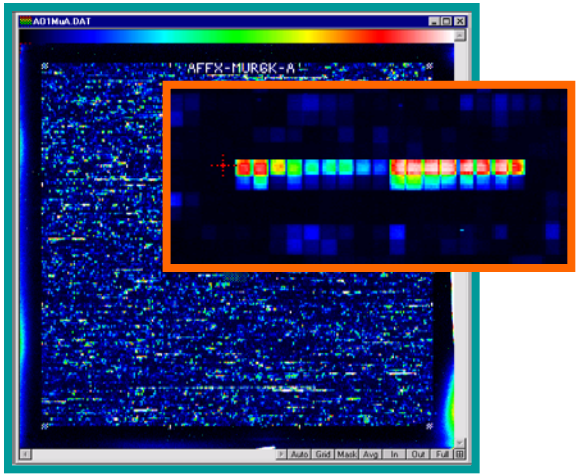
Ekspresijsko profiliranje z DNA mikromrežami

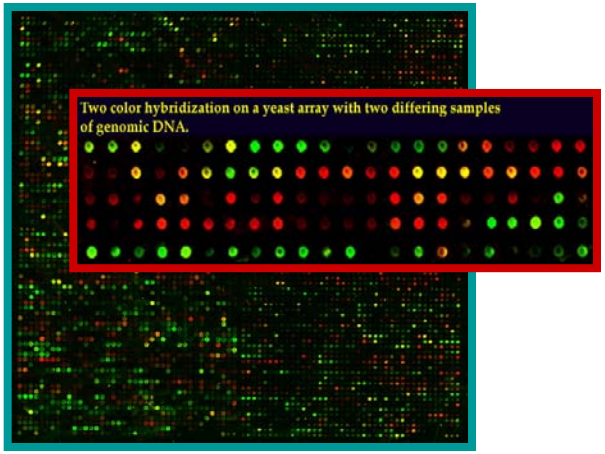


Gene Expression (expression profiling)

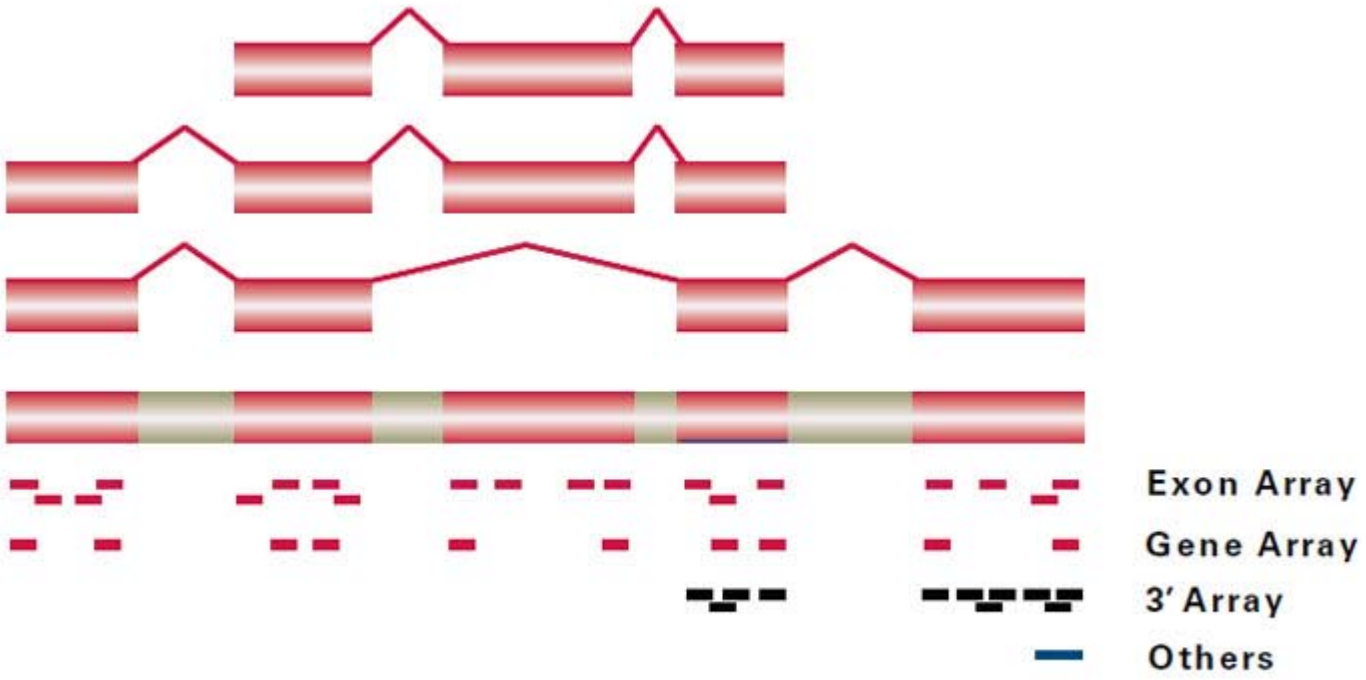
General Scheme:

1) Extract mRNA,
2) synthesize labeled cDNA,
3) Hybridize with DNA on the array,
4) Scan (image analysis)
5) look for genes that are expressed similarly (normalization, clustering, informatics)

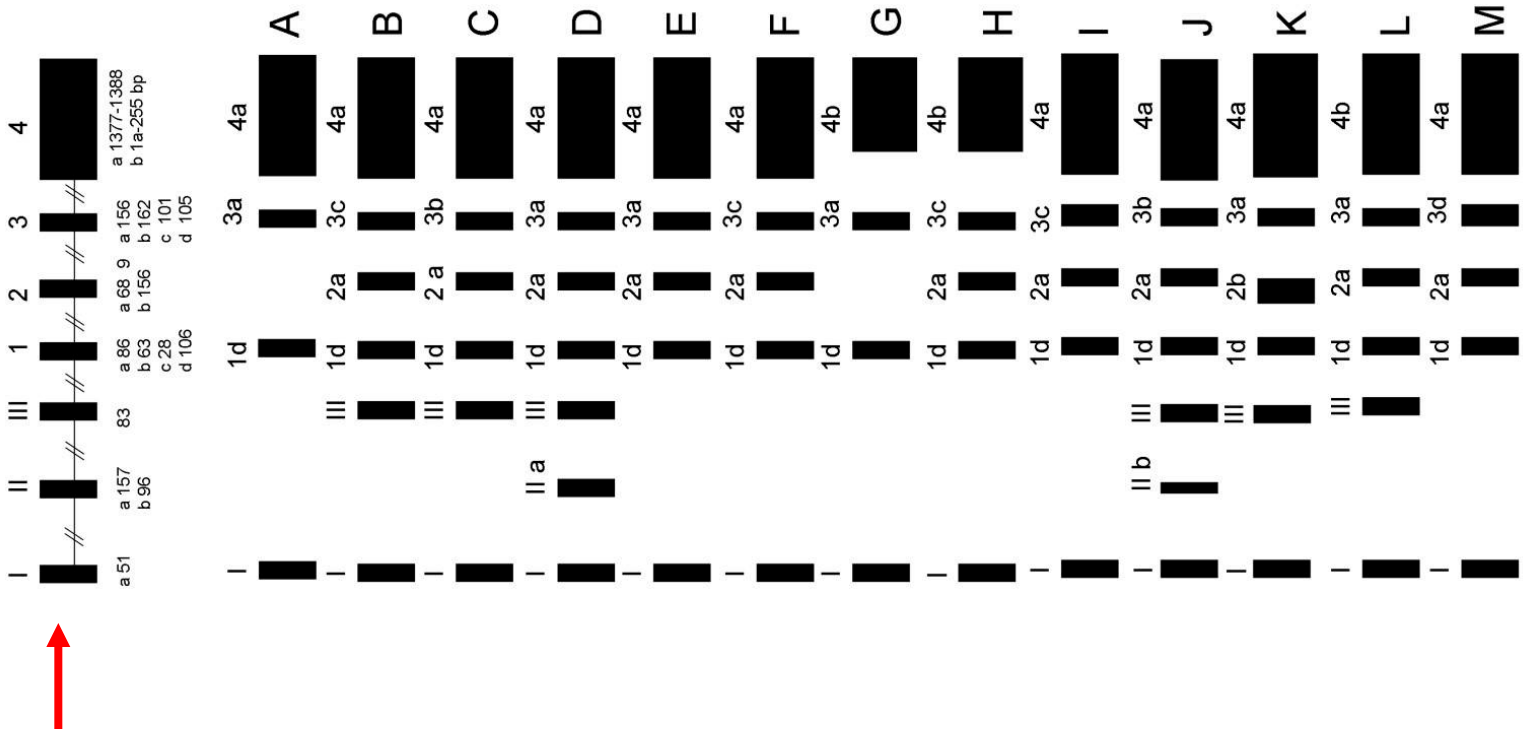




Parts of a gene represented in different types of microarrays



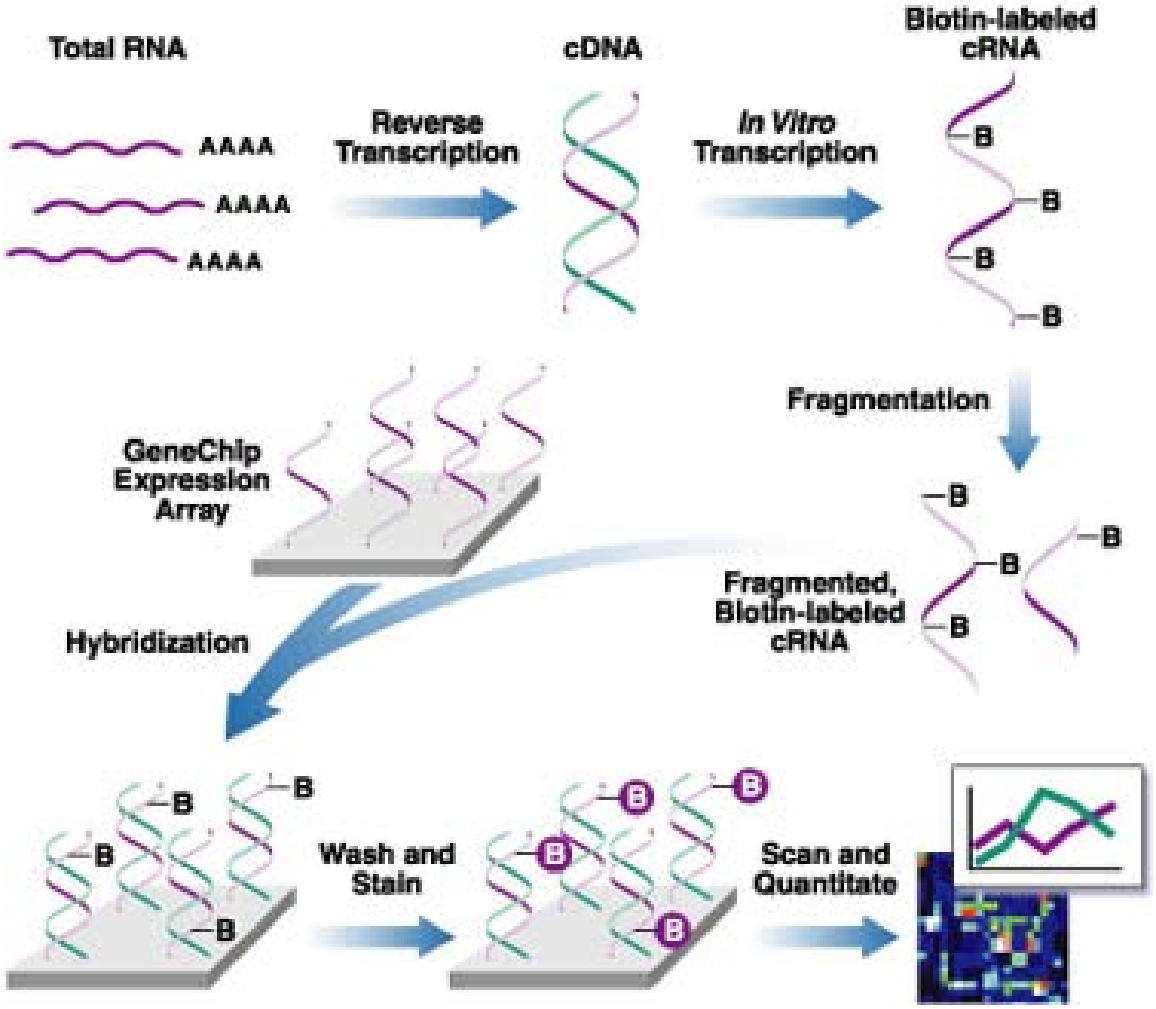
Exon arrays – example I



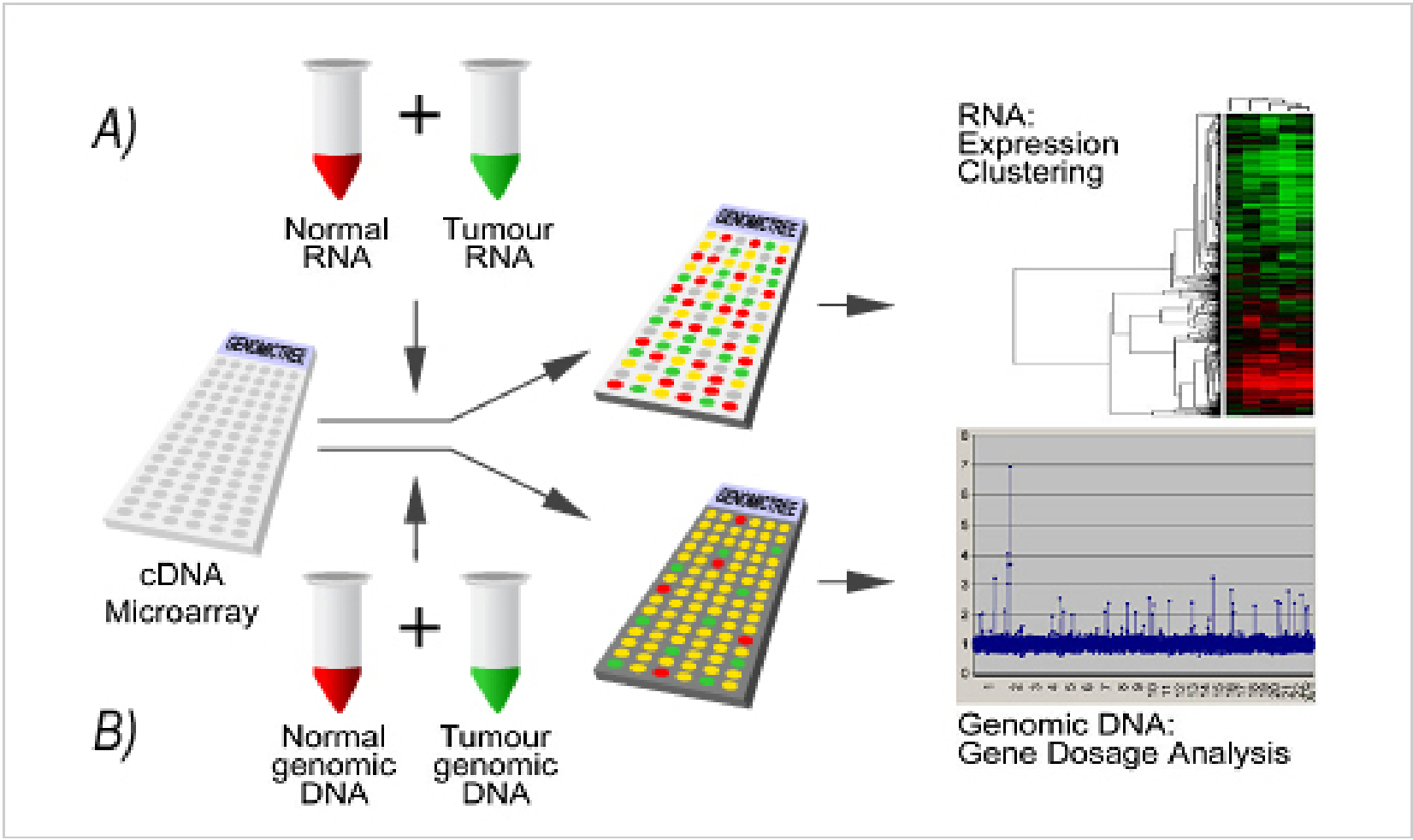
Complex organisation and structure of the ghrelin antisense strand gene GHRLOS, a candidate non-coding RNA gene
Seim I et al., *BMC Molecular Biology* 2008, **9:**95

AFFYMETRIX®

Short oligonucleotides from 3'-end



Example for insights into disease: combination of expression profiling and CGH



http://www.genomictree.com/images/bioimg09.jpg

miRNA Profiling

MicroRNAs (miRNAs) are 21-23 nucleotide long non-coding RNA molecules that have regulate the stability or translational efficiency of target messenger RNAs.

The expected number of miRNA in humans is over 800 (Sanger Institute mirBase database)

miRNA genes regulate protein production for at least 10% of human genes.

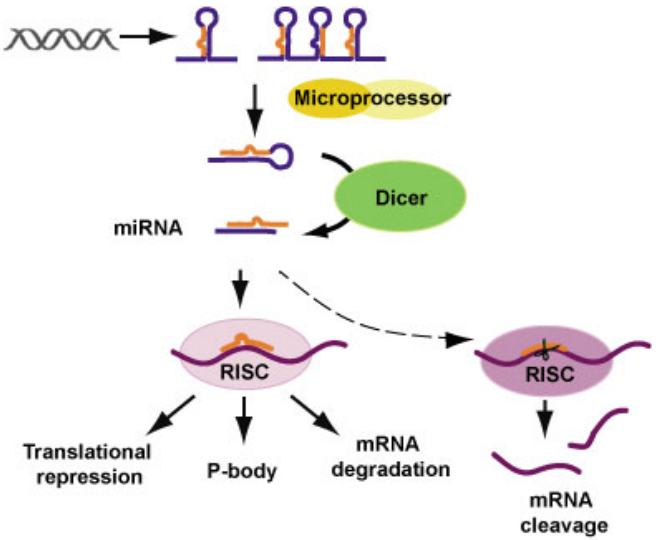
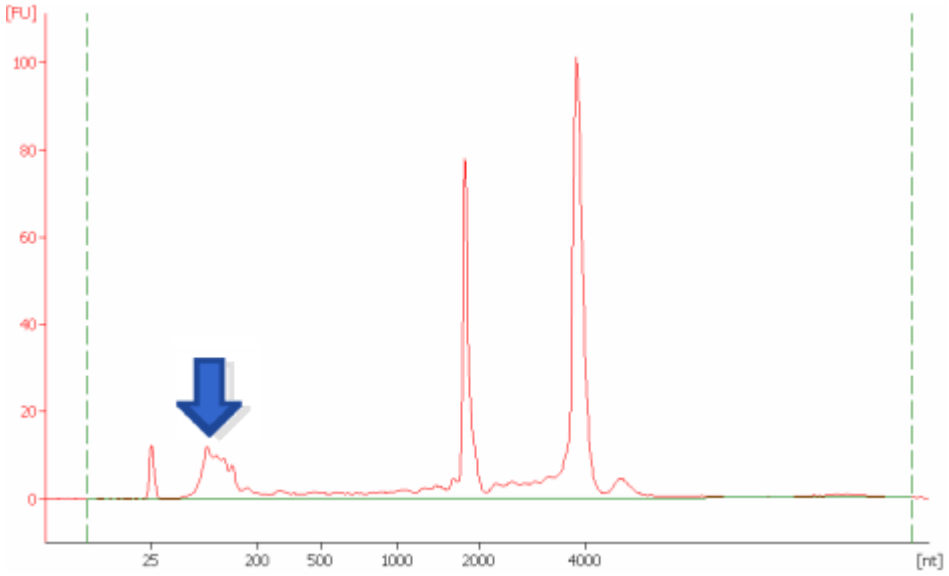


Fig1. The current model on miRNA biogenesis and posttranscriptional gene silencing.

miRNA isolation

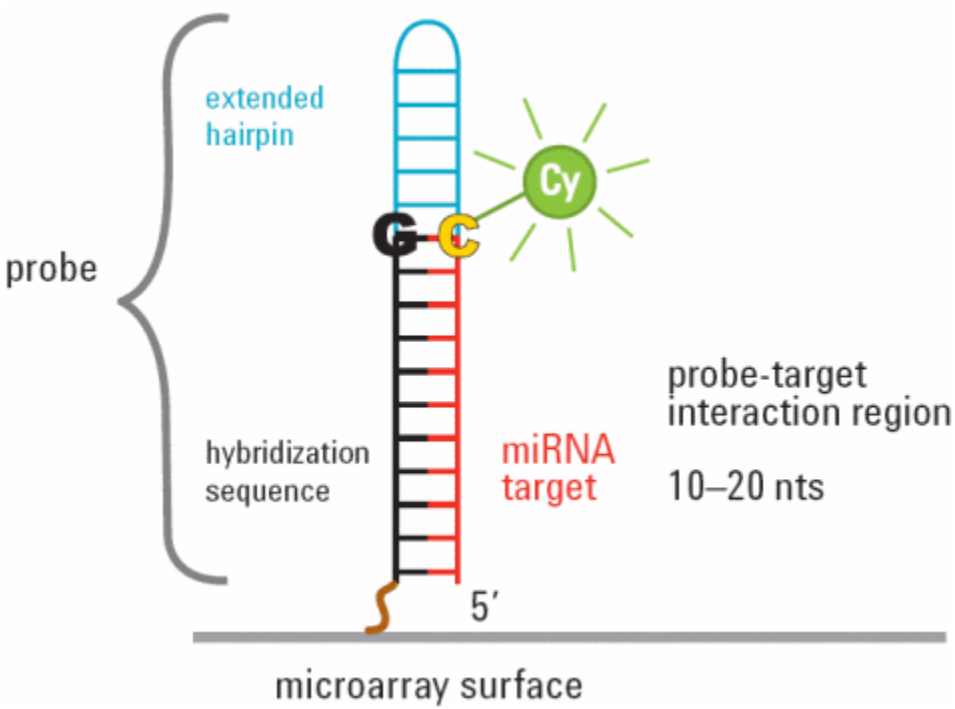
The starting material is **total RNA** (not small RNA) by isolated with i.e. TRIZOL still containing small RNA molecules. As such, **they should not have been cleaned up using columns** as this will invariably remove the small RNAs which the arrays are trying to measure. Just 100ng of total RNA is required for labelling. Once isolated, samples can be quantified using the Nanodrop Spectrophotometer.

Because miRNAs are so short, they are less prone to degradation then mRNAs and it has been tentatively shown that good data can be obtained from Formalin Fixed Parafin Embedded (FFPE) tissue. Total RNA quality can be checked on the Agilent 2100 Bioanalyser to ensure no degradation of larger RNA molecules has occurred.



Screen capture of Agilent 2100 Bioanalyzer electropherograms of a TRIZOL Reagent isolated total RNA. Note the pronounced peak of RNA at <200nt, typically not seen in column cleaned-up samples. Source: Microarray Centre

Hybridiation of miRNA microarrays

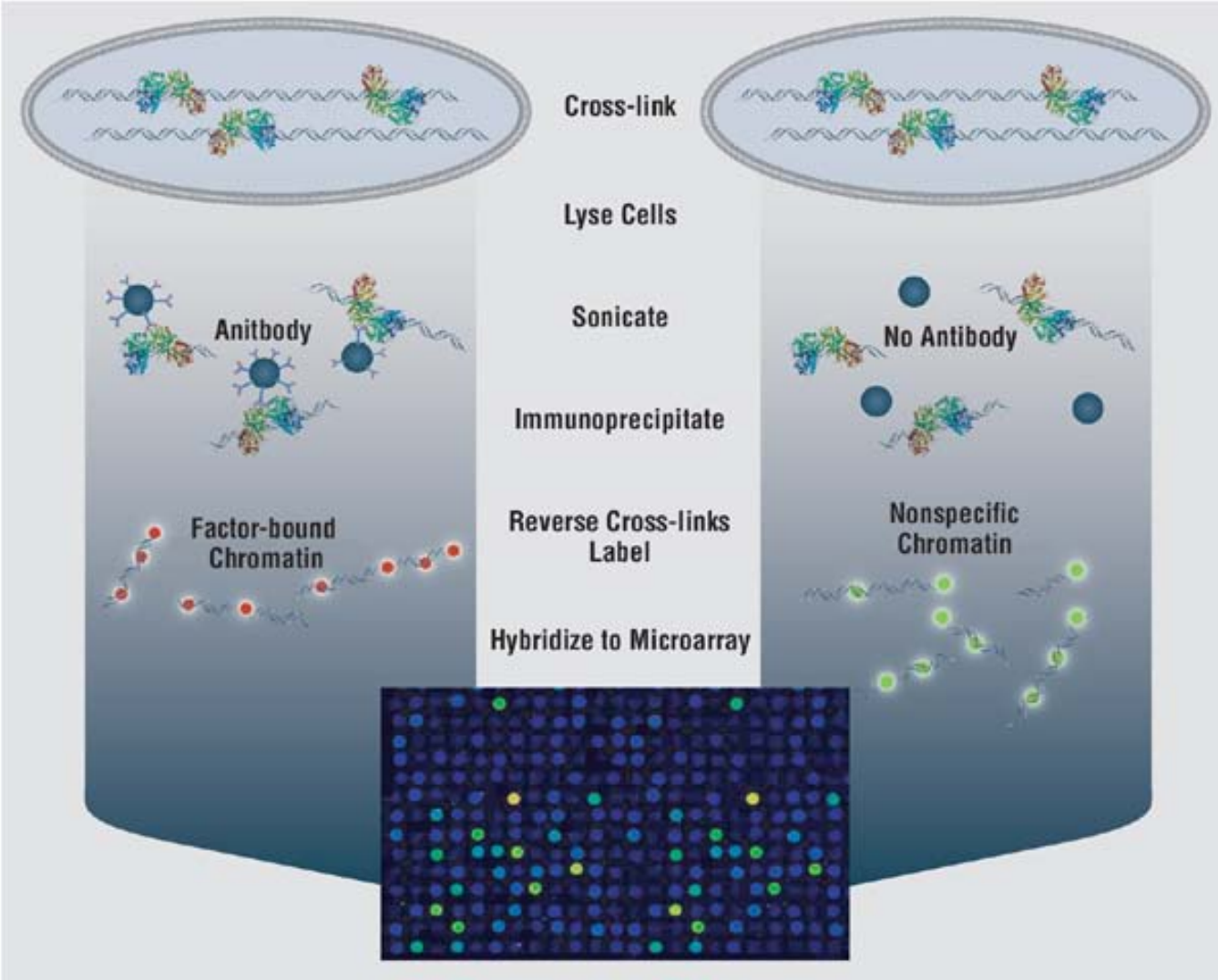


The labeling protocol adds a Cy3-Cytosine residue to the 3' end of miRNAs. This, coupled with the inclusion of Guanine residue at the 5' end of the probe, increases the stability of binding to labeled target miRNAs.

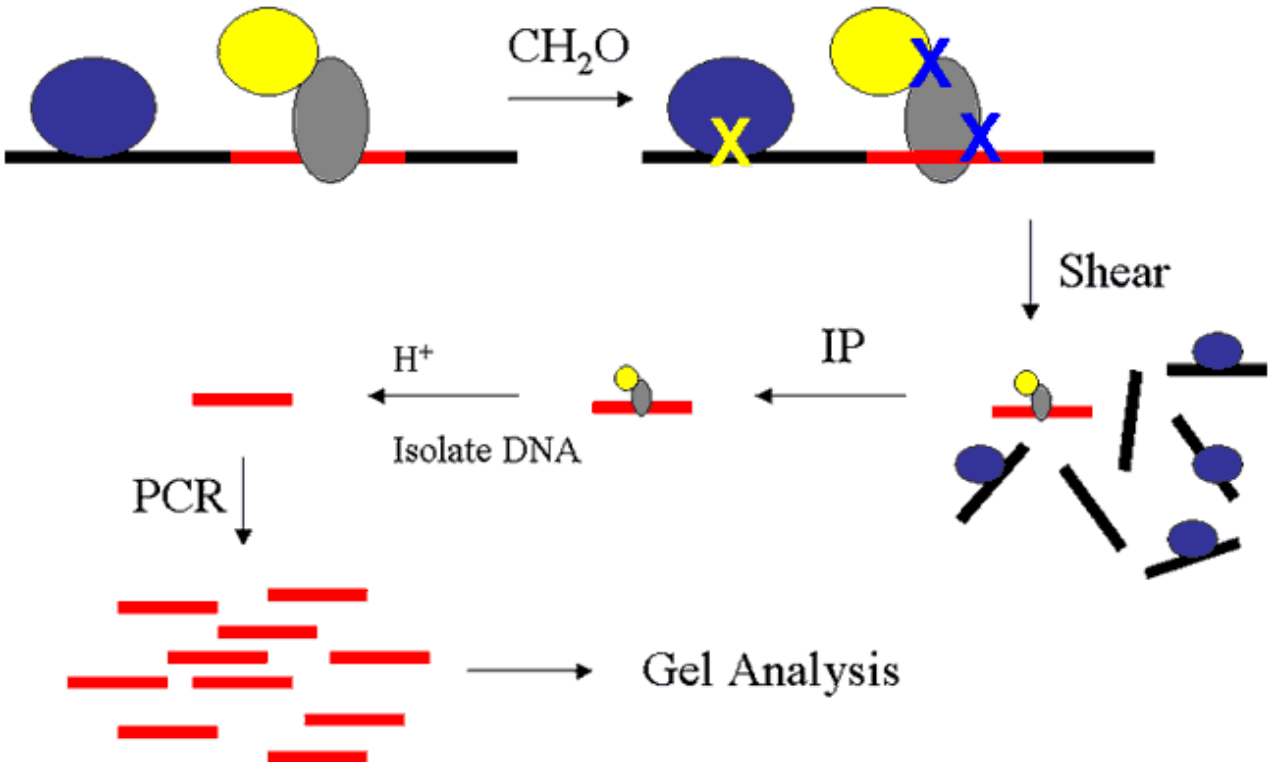
The 5' hairpin on each probe provides excellent size discrimination.

Source: Agilent

Chromatine immunoprecipitation on a chip (CHIP-Chip)



The Chip protocol



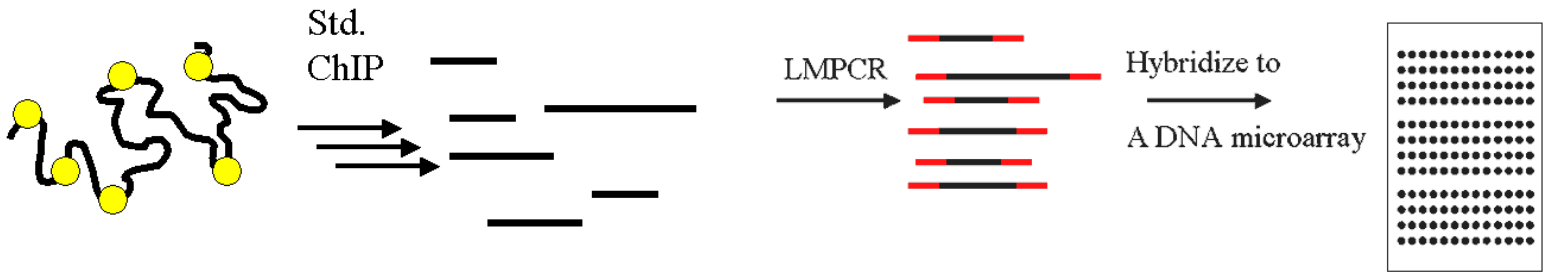
The "ChIP to Chip" Assay

The DNA fragments co-immunoprecipitated with the protein of interest are amplified by ligating linkers to the ends of the DNA fragments.

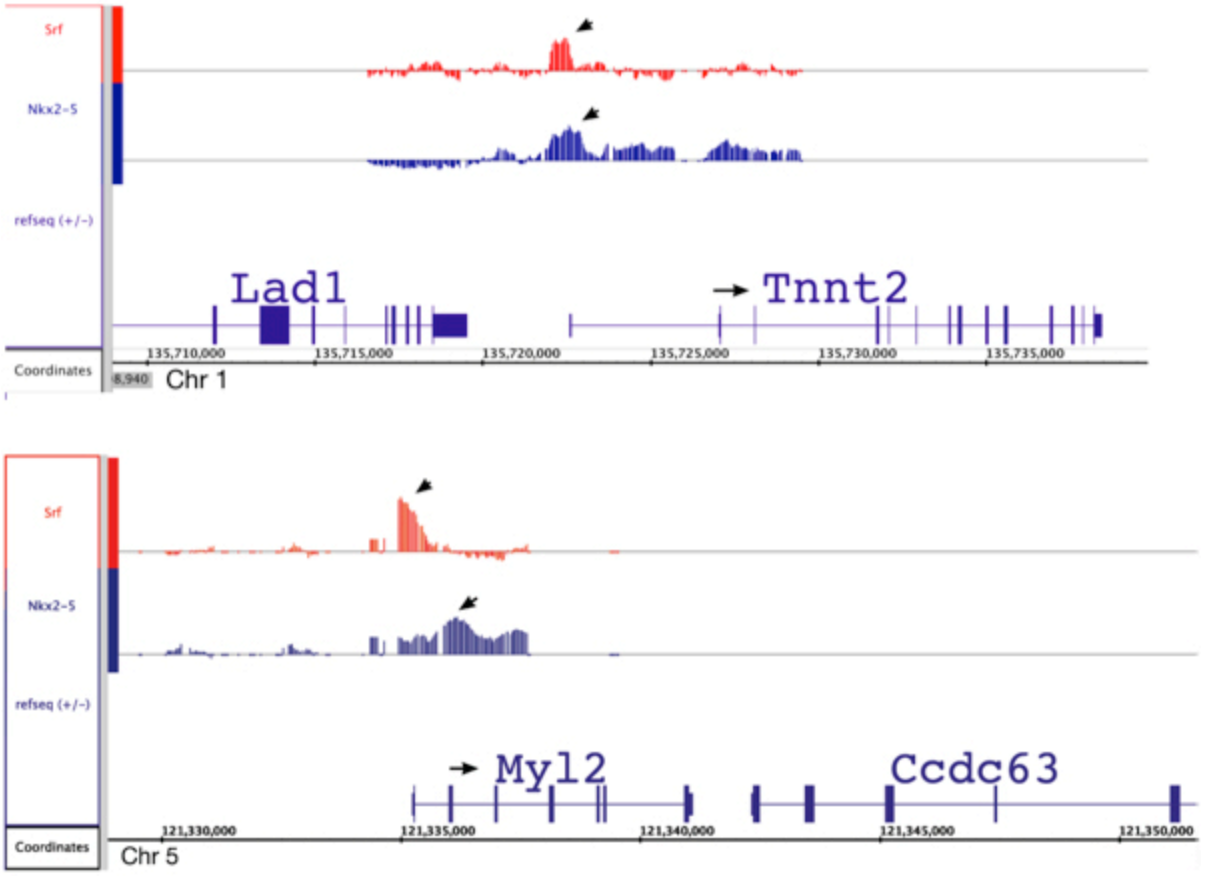
The amplified material is then annealed to a DNA microarray containing the appropriate probes.

This is often done with two colors, the other being the total DNA as a control.

Any DNA enriched in the immunoprecipitation above the level of the total is scored as a site of protein binding.



Chip-chip – example I



Detection of Nkx2-5 binding to the Actc1 promoter and intron 1 (A) and the Nppa promoter (B) using the Integrated Genome Browser

www.nimr.mrc.ac.uk/devbiol/mohun/cardiacdiff/

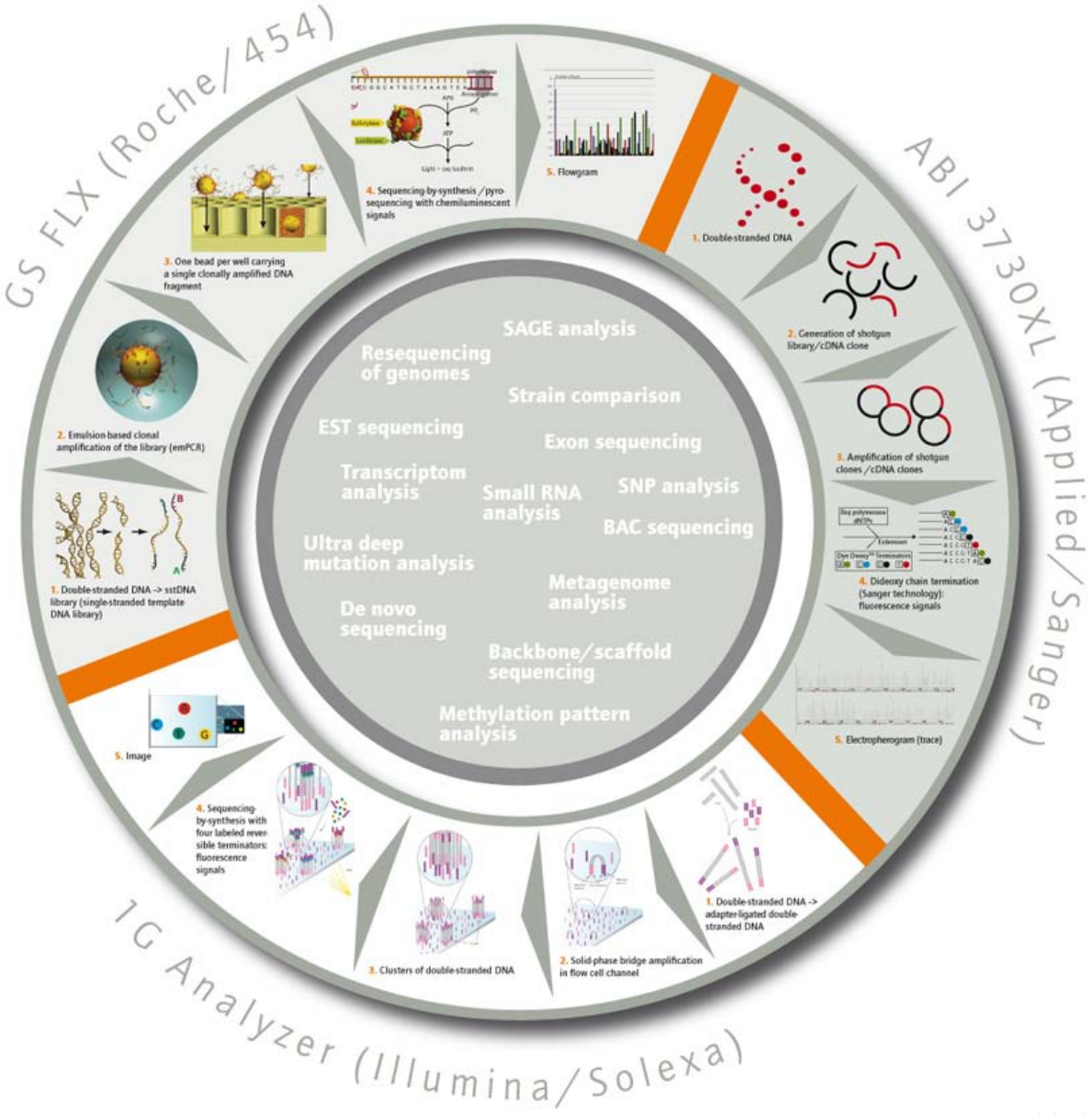
Povzetek

DNA čipi so zbirka mikroskopskih “DNA točk”, cDNA ali oligonukleotidov, pritrjenih na trdo podlago.

Uporabljajo se za:

- **Analizo genomske DNA** (genotipizacija, SNP analiza, CHG, sekvenciranje).
- **Analizo izražanja genov** (ekspresijsko profiliranje) , kjer probe lahko predstavljajo 3'–konce genov, eksone ali različne dele genov. Posebni čipi pa so za sledenje izražanja miRNA.
- **Študije uravnavanja izražanja genov** (določanje vezavnih mest transkripcijskih faktorjev, metilacija kromatina), kjer probe predstavljajo 5’-neprevedene dele in nerepetitivne dele genov.

Nova generacija sekvenciranja



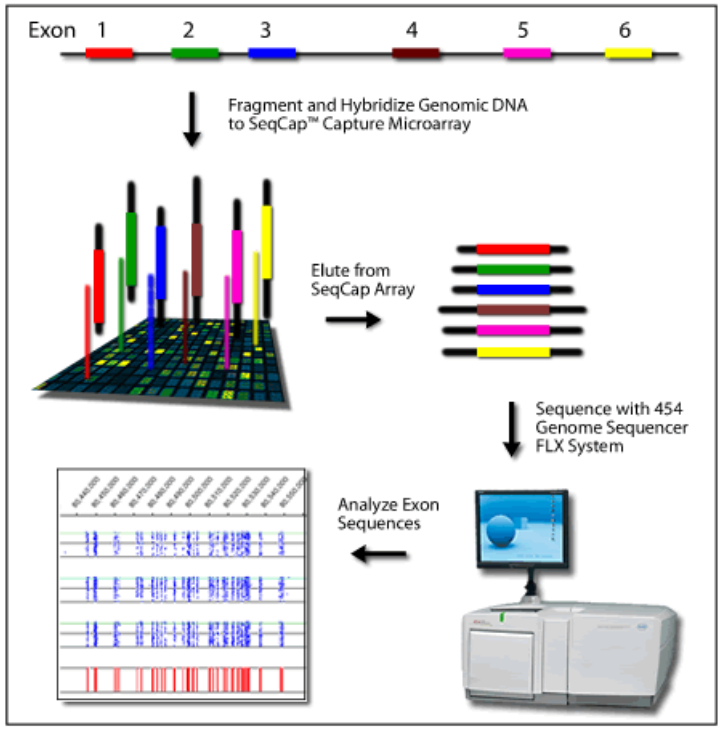
www.gatc.co.uk/cgi-bin/wPrintpreview.cgi?sour...

Nova (naslednja) generacija sekvenciranja

- Sekvenciranje človeškega genoma je trajalo več let, z uporabo cca 20 kb BAC klonov, ki so vsebovali cca 100 kb dolge tarčne fragmente, in 8-kratnega pokrivanja vsakega dela tarče. Analiza s kapilarno elektroforezo.

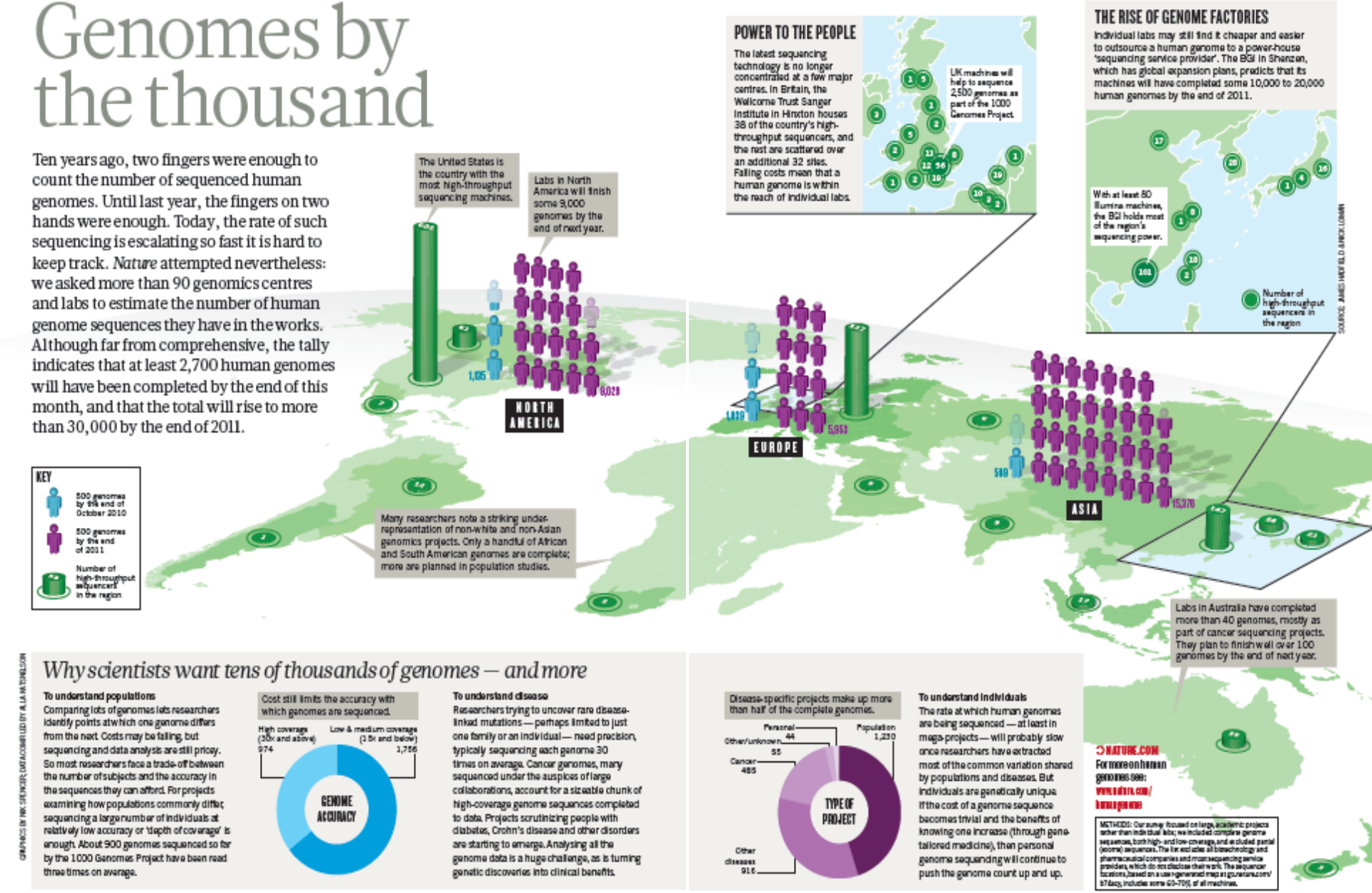
-Nadaljnji razvoj sekvenciranja je temeljil na sočasnem sekvenciranju celotnega genoma (WHS, angl. *whole genome sequencing*), ki je bil vstavljen v vektorje. Metoda je hitrejša, pušča pa velike praznine v zelo polimorfnih ali repetitivnih genomih. Analiza je potekala s kapilarno elektroforezo.

-Naslednja generacija sekvenciranja (2004) – visokozmogljivostno paralelno čitanje odsekov DNA na ravni celega genoma preko PCR pomnoževanja enoverižnih fragmentov genomske knjižnice.



Genomes by the thousand

Ten years ago, two fingers were enough to count the number of sequenced human genomes. Until last year, the fingers on two hands were enough. Today, the rate of such sequencing is escalating so fast it is hard to keep track. *Nature* attempted nevertheless: we asked more than 90 genomics centres and labs to estimate the number of human genome sequences they have in the works. Although far from comprehensive, the tally indicates that at least 2,700 human genomes will have been completed by the end of this month, and that the total will rise to more than 30,000 by the end of 2011.



Why scientists want tens of thousands of genomes — and more

To understand populations

Comparing lots of genomes lets researchers identify points at which one genome differs from the next. Costs may be falling, but sequencing and data analysis are still pricey. So most researchers face a trade-off between the number of subjects and the accuracy in the sequences they can afford. For projects examining how populations commonly differ, sequencing a large number of individuals at relatively low accuracy or 'depth of coverage' is enough. About 900 genomes sequenced so far by the 1000 Genomes Project have been read three times on average.

Cost still limits the accuracy with which genomes are sequenced.

To understand disease

Researchers trying to uncover rare disease-linked mutations — perhaps limited to just one family or an individual — need precision, typically sequencing each genome 30 times on average. Cancer genomes, many sequenced under the auspices of large collaborations, account for a sizeable chunk of high-coverage genome sequences completed to date. Projects scrutinizing people with diabetes, Crohn's disease and other disorders are starting to emerge. Analysing all the genome data is a huge challenge, as is turning genetic discoveries into clinical benefits.

Disease-specific projects make up more than half of the complete genomes.

To understand individuals

The rate at which human genomes are being sequenced — at least in mega-projects — will probably slow once researchers have extracted most of the common variation shared by populations and diseases. But individuals are genetically unique. If the cost of a genome sequence becomes trivial and the benefits of knowing one increase (through gene-tailored medicine), then personal genome sequencing will continue to push the genome count up and up.

NATURE.COM
For more on human genomes see:
www.nature.com/humangenome

METHODS: Our survey focused on large, academic projects rather than individual labs; we included complete genome sequences, both high- and low-coverage, and excluded partial (exome) sequences. The list excludes all biotechnology and pharmaceutical companies and most sequencing service providers, which do not disclose their work. The sequencer locations, based on a user-generated map at gluerium.com/bTdoy, includes some 60–70% of all machines.

GRAPHICS BY NIK SPENCER; DATA COMPILED BY ALLA KATSNELSON

Next-generation of DNA and RNA sequencing methods

- The bead-amplification sequencing (Roche/454FLX)

-Sequencing by synthesis (Illumina/Solexa Genome analyzer)

-Sequencing by ligation (Applied Biosystems SOLID System)

-Helicos Helioscope (2008)

-Pacific Biosciences SMRT (2010)

Common features:

-A compex interplay of enzymology, chemistry, software, hardware, optics engineering…)

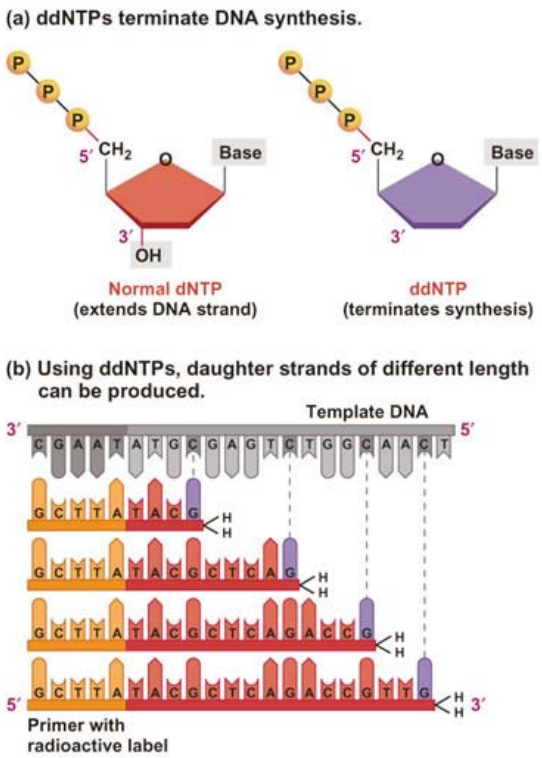
-A streamline of sample preparation prior to sequencing (time saving)

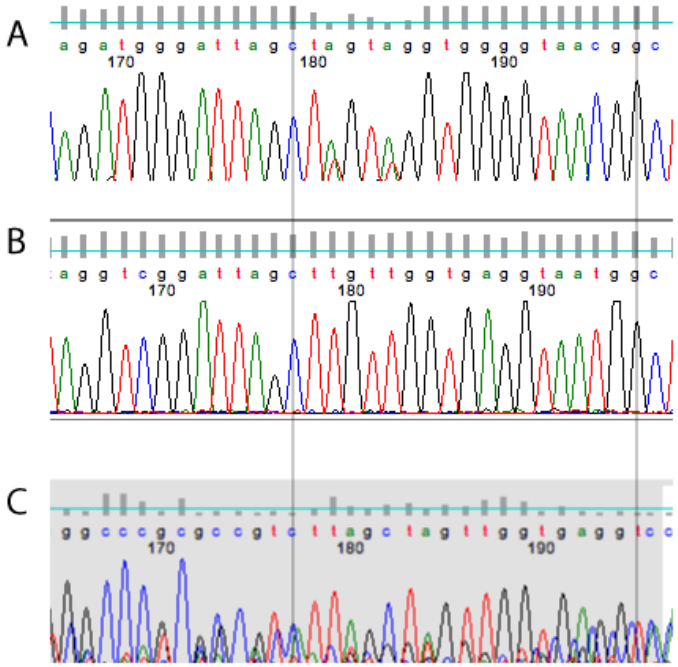
-Preparation of fragment libraries of the DNA of interest by annealing for platform-specific linkers and amplification

-Amplification of single stranded fragment library and performing sequencing on amplified fragments

-Single molecule sequencing just arrived or is under development

Classical Sanger dideoxy sequencing method





Dideoxynucleotide sequencing represents only one method of sequencing DNA. It is commonly called Sanger sequencing since Sanger devised the method. This technique utilizes 2',3'-dideoxynucleotide triphospates (ddNTPs), molecules that differ from deoxynucleotides by the having a hydrogen atom attached to the 3' carbon rather than an OH group. These molecules terminate DNA chain elongation because they cannot form a phosphodiester bond with the next deoxynucleotide.

campus.queens.edu/faculty/jannr/molecular/

General concepts for clonal-array generation and sequencing

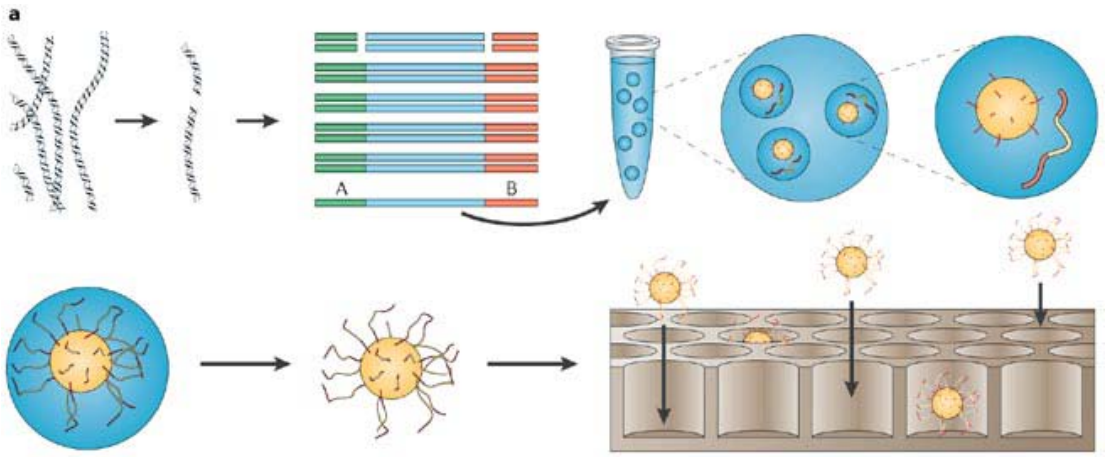
a | **Bead-chips**. Genomic DNA is fragmented and adaptors are ligated to create an insert library that is flanked by two universal priming sites. This library is cloned on beads using emulsion PCR technology. Beads with clones are affinity selected and assembled onto a planar substrate. A subsequent cycle-sequencing reaction is used to read out the sequence on the clones (illumina)

b | **Sequencing by synthesis** (SBS). A common anchor primer is annealed to a constant sequence (universal priming site) that is contained within the library clones. The sequence is read out by polymerase extension in a base-by-base fashion (pyrosequencing). After incorporation of a single base or base type, the incorporated base is identified by fluorescence (laser) or chemiluminescence (no laser required). (Roche)

c | **Sequencing by ligation**. The array set-up is similar to SBS in which a common primer is annealed to an arrayed library and used to read out the sequence through a stepwise ligation of random oligomers. After read-out of each ligation event, the primer and the ligated oligomer are stripped, a new primer reannealed and the process repeated with an oligomer that contains a query base at a different position. (ABI)

Fan *et al. Nature Reviews Genetics* **7**, 632–644 (August 2006) | doi:10.1038/nrg1901

General concepts for clonal-array generation and sequencing

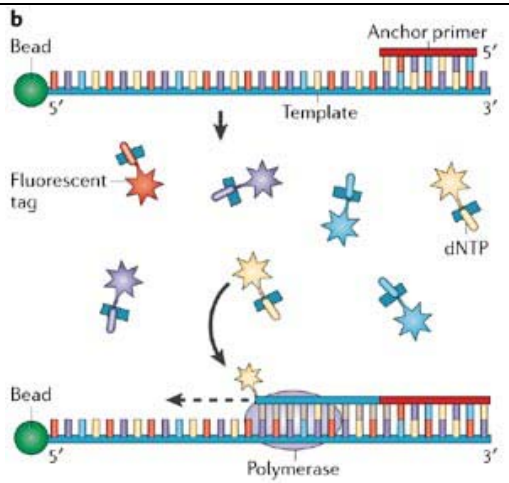


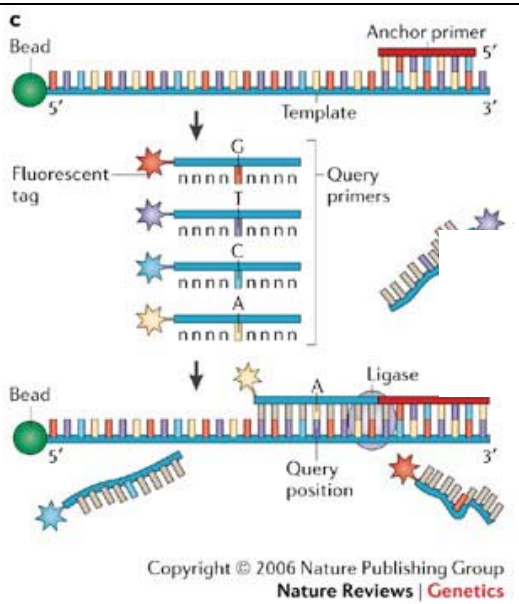
Bead chips

Roche
pyrosequencing

Sequencing by synthesis

Solexa / illumina



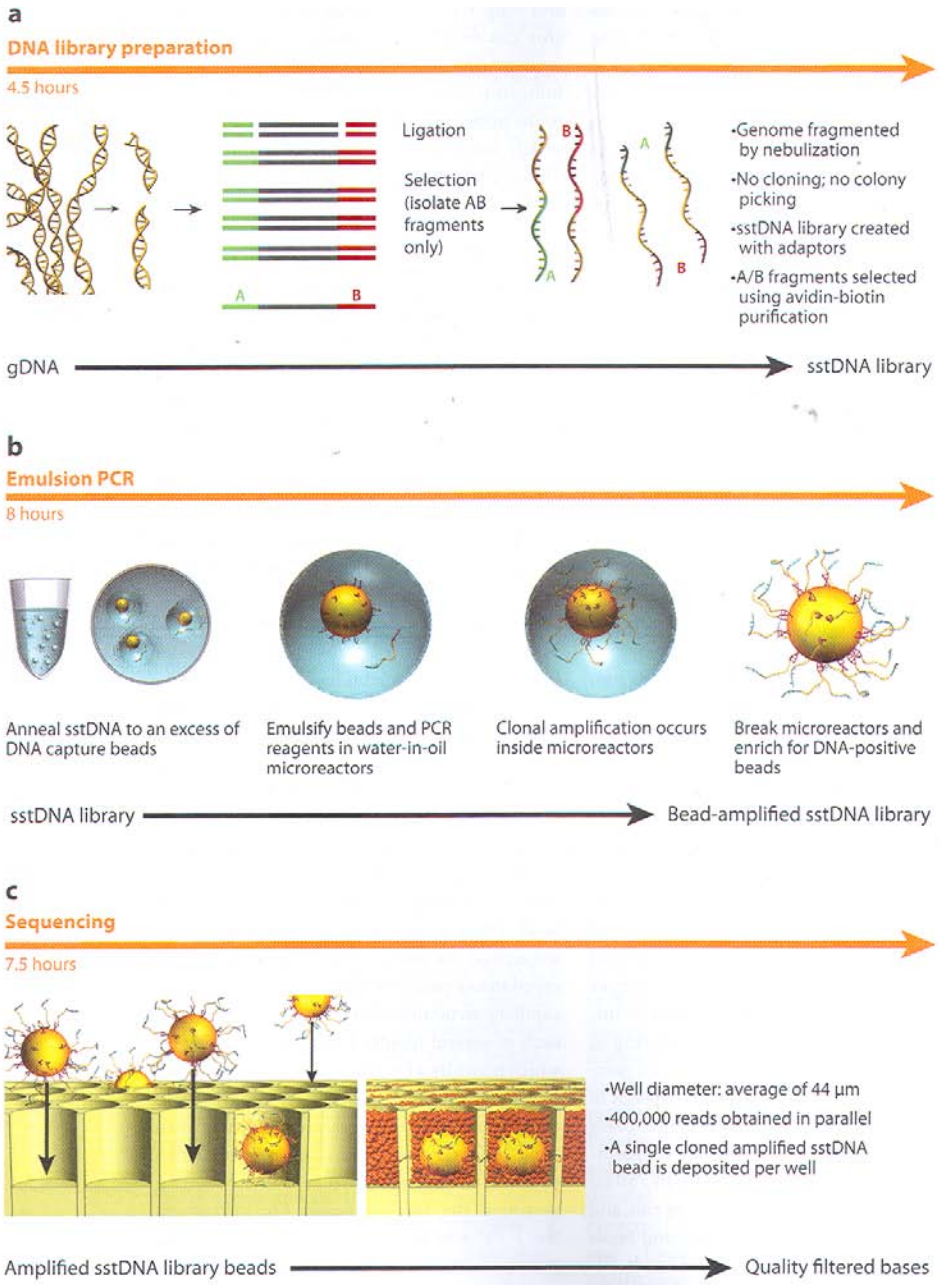


Sequencing by Ligation

Applied biosystems

Fan *et al. Nature Reviews Genetics* **7**, 632–644 (August 2006) | doi:10.1038/nrg1901

Roche/454 FLX Pyrosequencer



Library fragments are mixed with agarose beads with oligos complementary to adapter sequences on the library.

Each bead is associated with a single fragment.

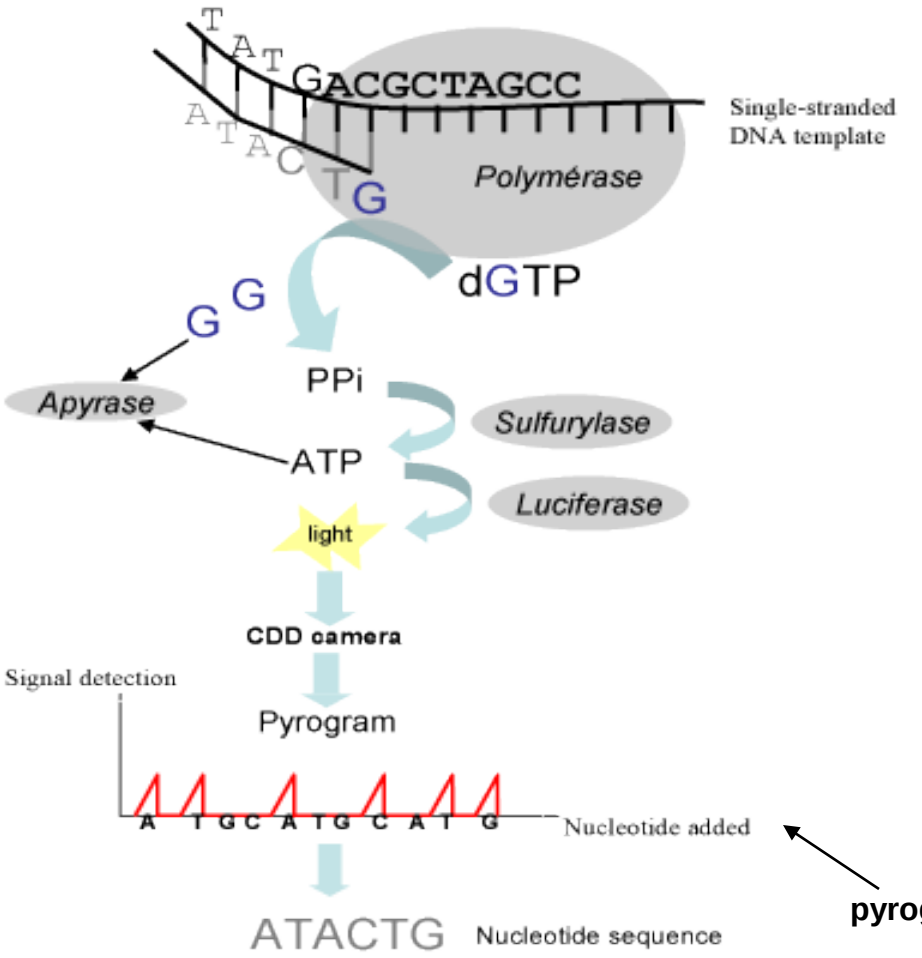
Each fragment-bead complex is isolated into individual oil:water micelles with PCR mixture.

Thermal cycling of this emulsion PCR of the micelles produces amplified unique sequences on the bead surface.

“En mass” sequencing of PCR products on picotiter plates (PTP) with single beads in each picowell.

Enzyme/substrate containing beads for the pyrosequencing reaction are added to wells that act as floww cells for addition of individual pure nucleotide solutions. The CCD camera records the light emitted at each bead.

Principles of Pyrosequencing



A sequencing primer is hybridized to a single-stranded PCR amplicon that serves as a template.

A single dNTP is added at each time.

Mixtures are incubated with the **enzymes, DNA polymerase, ATP sulfurylase**, **luciferase**, and **apyrase** as well as the **substrates**, adenosine 5' phosphosulfate (**APS**), and **luciferin**.

ATP sulfurylase converts PPi to ATP in the presence of adenosine 5' phosphosulfate (APS).

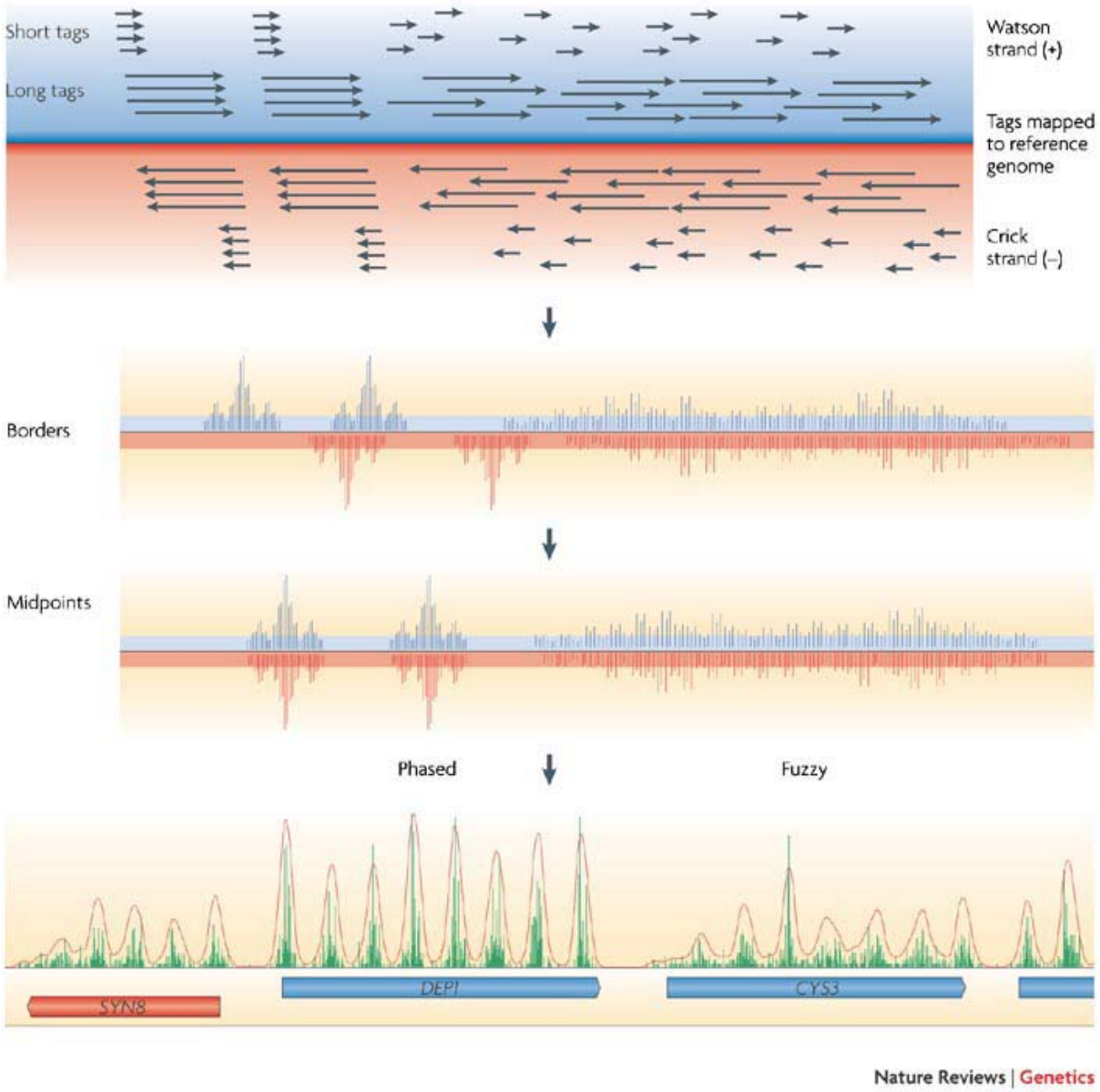
ATP drives the **luciferase-mediated conversion of luciferin to oxyluciferin** that generates visible light in amounts that are **proportional to the amount of ATP.**

Apyrase, a nucleotide-degrading enzyme, continuously degrades unincorporated nucleotides and ATP. When degradation is complete, another nucleotide is added

Light signal is detected only when the proper dNTP is incorporated,

pyrogram

Watson and Cricks pyrosequencing readout



2008

comments on this story

Stories by subject
- Biotechnology
- Business
- Cell and molecular biology
- Genetics
- Health and medicine

Stories by keywords
- Craig Venter
- Human genome sequencing
- Next-generation sequencing technology
- James Watson
- 454 LifeSciences

This article elsewhere
Blogs linking to this article

Published online 16 April 2008 | Nature | doi:10.1038/452788b

News

James Watson's genome sequenced at high speed

New-generation technology takes just four months and costs a fraction of old method.

Meredith Wadman

The first full genome to be sequenced using next-generation rapid-sequencing technology is published today (see page 872)[1], marking another milestone in the extraordinarily fastmoving field of human genome sequencing.

It took just four months, a handful of scientists and less than US$1.5 million to sequence the 6 billion base pairs of DNA pioneer James Watson. The achievement is first proof of principle that these rapid-sequencing machines can decipher large, complex genomes (see page 819)[2]. Made in this case by Connecticut-based 454 Life Sciences — a division of Roche Diagnostics — they allow many more sequencing reactions to proceed at the same time, on the same surface, than the previous generation of machines that produced the inaugural human genomes[3,4]. That change has had big pay-offs in speed, efficiency and, ultimately, cost (see Table 1).

most recent | commented

- Fat cell numbers stay constant through adult life
  05 May 2008
- "Raft" offers route for drug treatments
  02 May 2008
- Climate troubles brewing for beer makers
  02 May 2008
- Of myths and men
  02 May 2008
- 'Ocean deserts' are growing
  01 May 2008

Related stories
- Ready or not
  10 April 2008
- Watson's folly
  25 October 2007
- Common sense for our genomes
  18 October 2007
- Genomics: The personal side of genomics
  04 October 2007
- All about Craig: the first 'full' genome sequence
  06 September 2007

illumina sequencing by synthesis

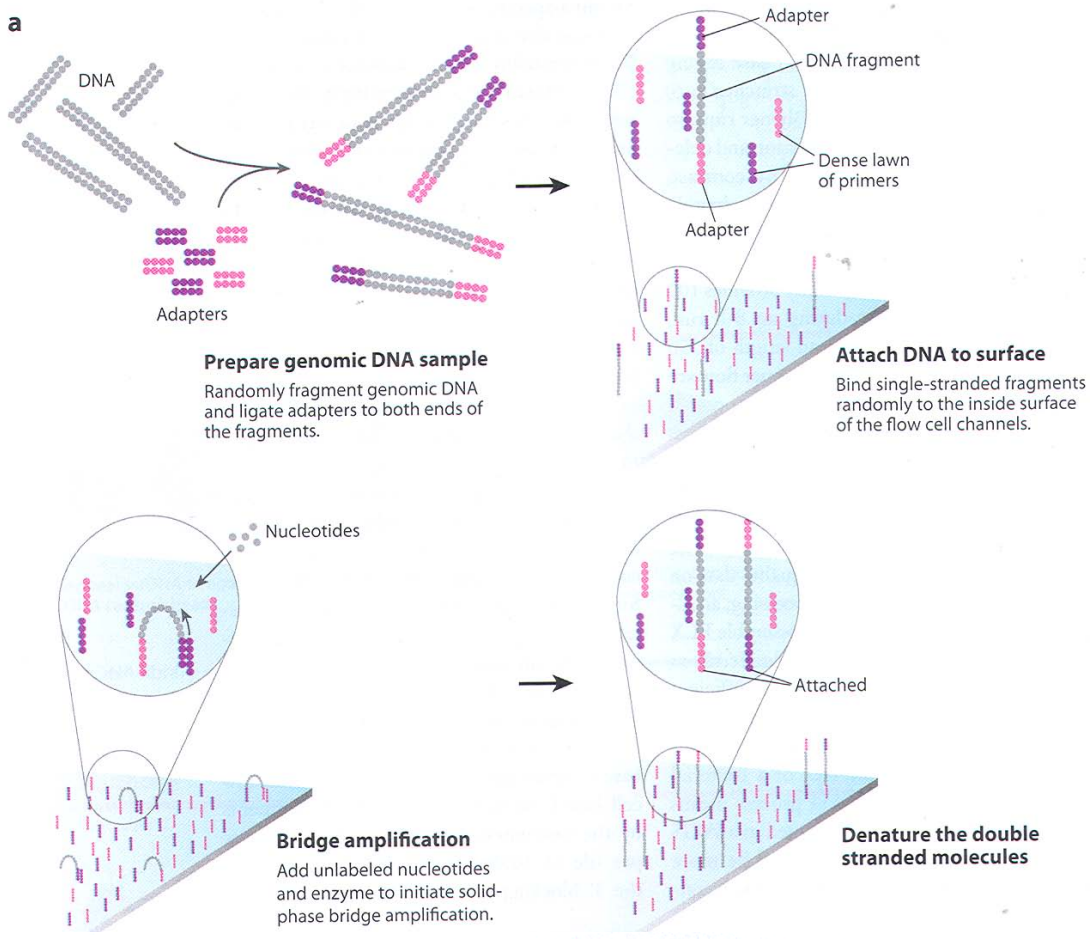
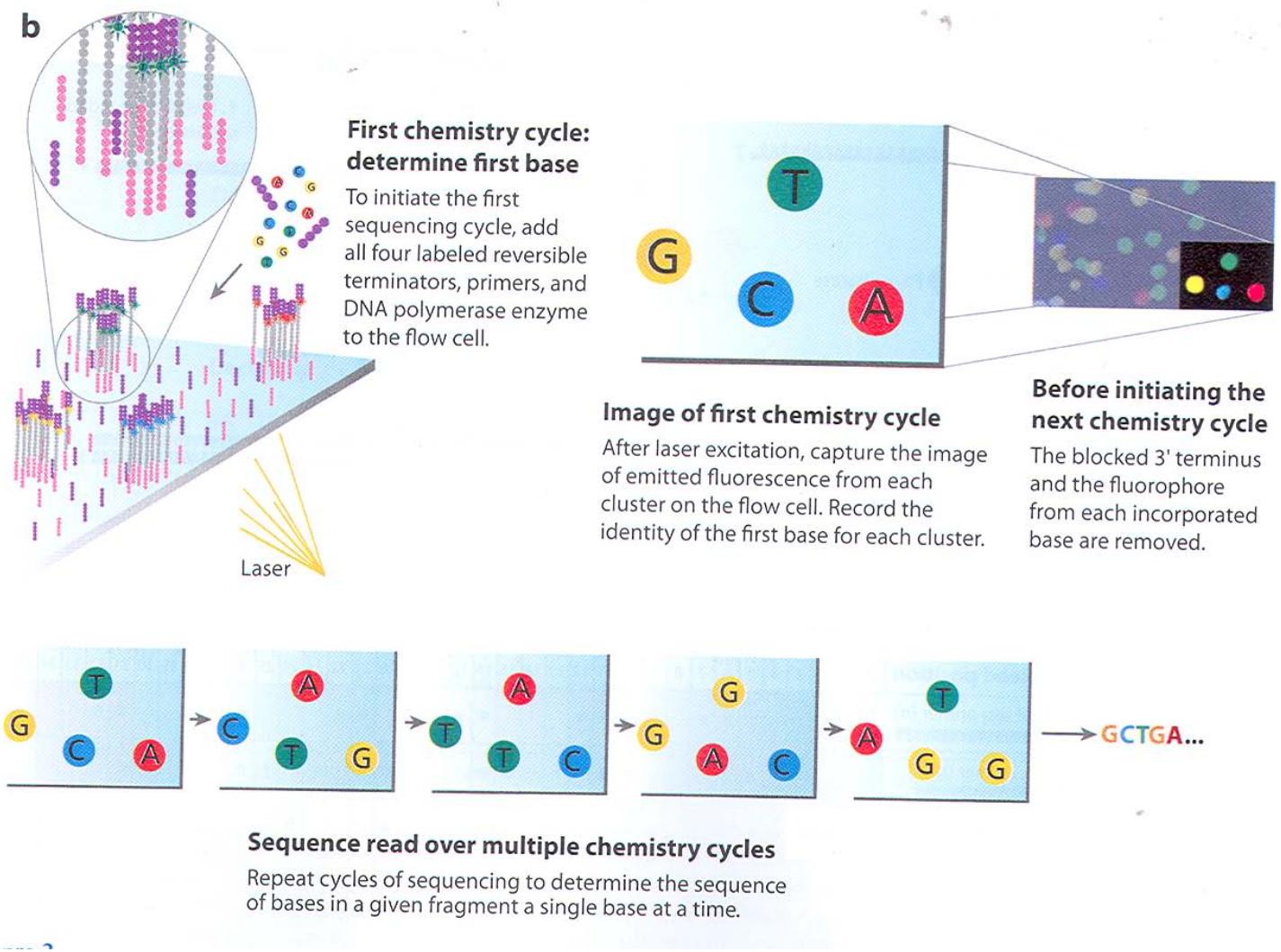


Figure 2

The Illumina sequencing-by-synthesis approach. Cluster strands created by bridge amplification are primed and all four fluorescently labeled, 3′-OH blocked nucleotides are added to the flow cell with DNA polymerase. The cluster strands are extended by one nucleotide. Following the incorporation step, the unused nucleotides and DNA polymerase molecules are washed away, a scan buffer is added to the flow cell, and the optics system scans each lane of the flow cell by imaging units called tiles. Once imaging is completed, chemicals that effect cleavage of the fluorescent labels and the 3′-OH blocking groups are added to the flow cell, which prepares the cluster strands for another round of fluorescent nucleotide incorporation.

Illumina sequence - decoding



Nova generacija sekvenciranja - povzetek

• Nova generacija visokozmogljivostnega sekvenciranje omogoča razpoznavanje zaporedij DNA na ravni celega genoma, z resolucijo posameznega baznega para.

• Iz vsakega vzorca se pripravi z adaptorji ligirana knjižnica, ki vsebuje vse v vzorcu prisotne fragmente DNA ali RNA (cDNA).

• Vse platforme bazirajo na ligaciji adaptorjev in pomnoževanju, imajo pa različne pristope sekvenciranja:

-Pirosekvenciranje (Roche-Nimblegen)

-Sekvenciranje s sintezo (Illumina-Solexa)

-Sekvenciranje z ligacijo (ABI)

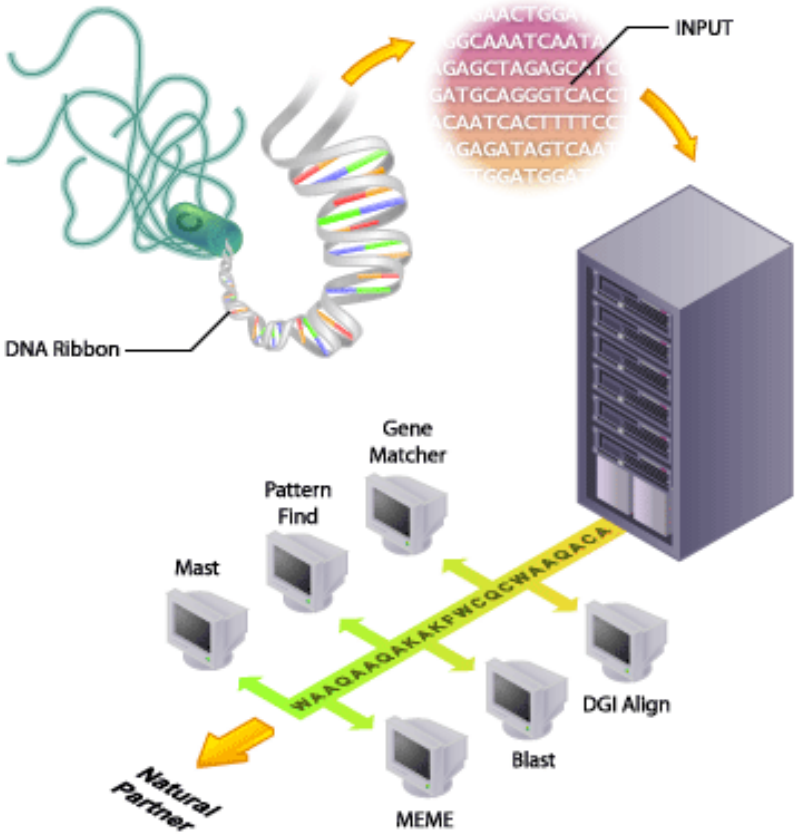
•Razvijajo se tudi metode, ki pred sekvenciranjem ne potrebujejo pomnoževanja.

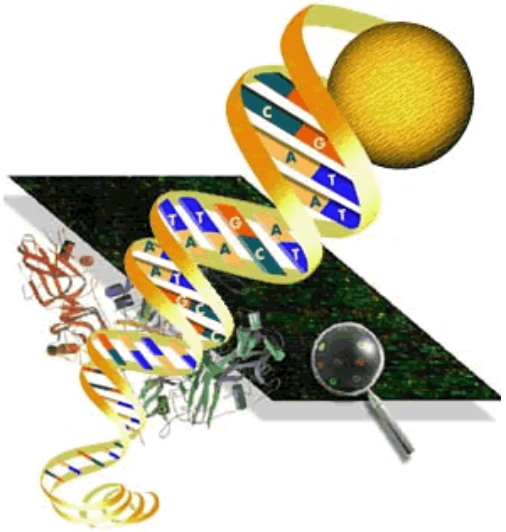
• Aplikacije so enake kot pri klasičnih mikromrežah (ekspresijsko profiliranje oz. SAGE, genotipizacija SNP in CNV, kroamtinska imunoprecipitacija, metilacija kromatina, itd.).

•Prednost pred klasičnimi mikromrežami je v preprosti pripravi vzorca in zmožnosti procesiranja velikega števila vzorcev v kratkem času.

•Procesiranje velikega števila vzorcev na eni ali več aparaturah lahko upravlja en človek.

Bioinformatika v raziskavah “omov”



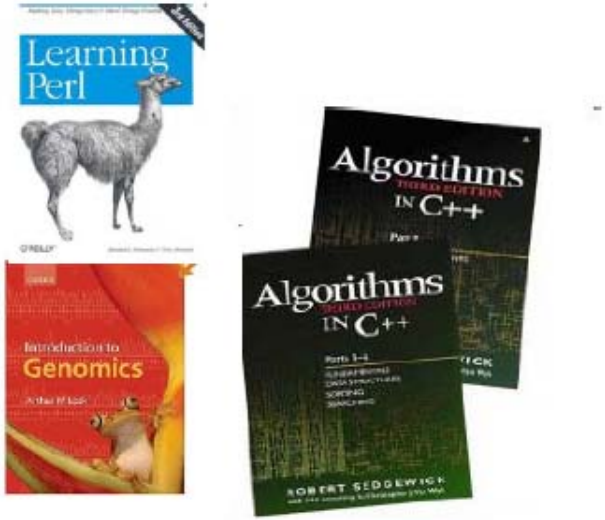


www.gwumc.edu

http://bioinformatics.ubc.ca/about/what_is_bioinformatics/images/computer.gif

Be warned...

Skills required for DNA sequencing projects

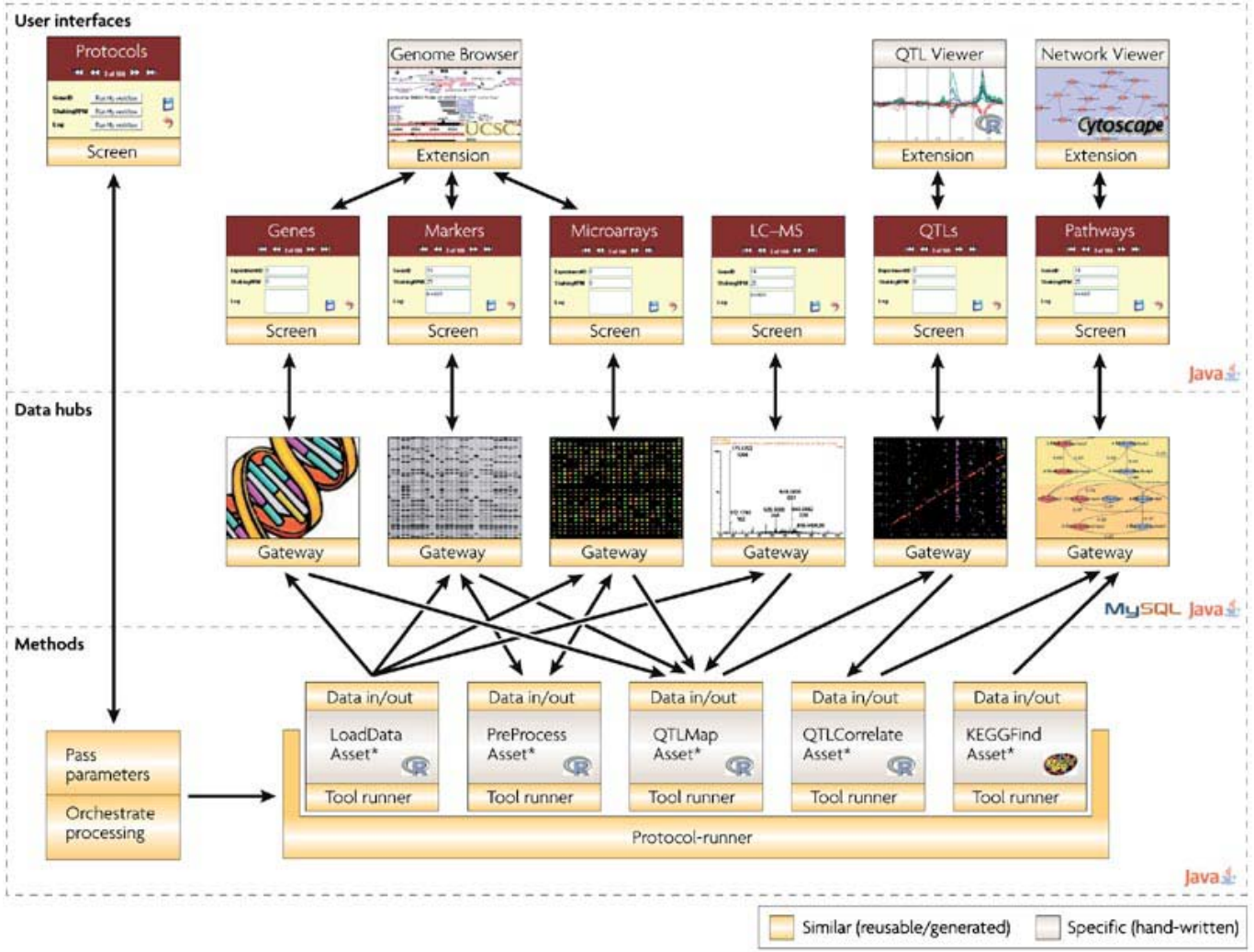


1 %

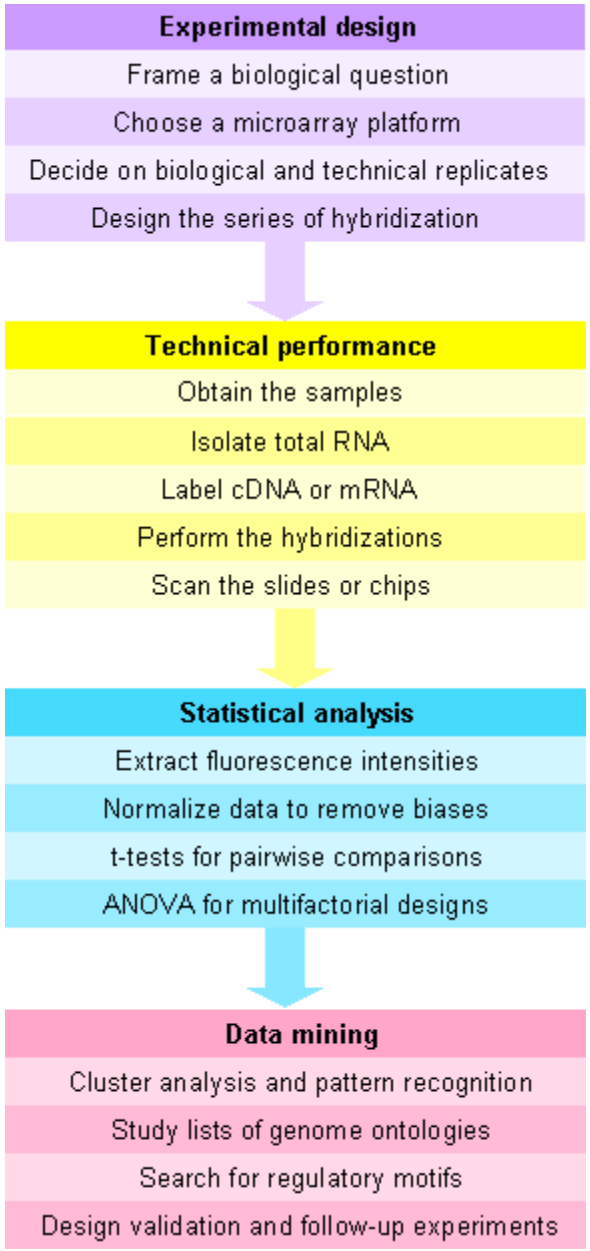
99 %

http://www.uni-leipzig.de/

Obtaining Reliable Data from “ome” Experiments – Standardization and beyond



Steps in post-genome experiments



Consult bioinformatician

Requested bioinformatician

Consult bioinformatician

irfgc.irri.org/cropbioportal/index.php?option...

Bioinformatics

Wikipedia

Making sense of the huge amounts of DNA data produced by gene sequencing projects.

Bioinformatics and **computational biology** involve the use of techniques from applied mathematics, informatics, statistics, and computer science to solve biological problems.

Research in computational biology often overlaps with systems biology.

Major research efforts in the field include sequence alignment, gene finding, genome assembly, protein structure alignment, protein structure prediction, prediction of gene expression and protein-protein interactions, and the modeling.

The terms *bioinformatics* and *computational biology* are often used interchangeably, although the former typically focuses on algorithm development and specific computational methods, while the latter focuses more on hypothesis testing and discovery in the biological domain.

Microarrays and bioinformatics

Standardization

The lack of standardization in arrays presents an interoperability problem in bioinformatics, which hinders the exchange of array data.

Various projects are attempting to facilitate the exchange and analysis of data produced with non-proprietary chips.

The "Minimum Information About a Microarray Experiment" (MIAME) XML based standard for describing a microarray experiment is being adopted by many journals as a requirement for the submission of papers incorporating microarray results.

http://www.mged.org/Workgroups/MIAME/miame.html

MGED SOCIETY

Search

HOME | MEETINGS | WORKGROUPS | MISSION | MGED BOARD | SITE MAP

>> Quick Links

Home › Workgroups › MIAME

Minimum Information About a Microarray Experiment - MIAME

MIAME describes the **Minimum Information About a Microarray Experiment** that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment. [Brazma et al, Nature Genetics]

The six most critical elements contributing towards MIAME are:

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

For more details, see MIAME 2.0.

MIAME does not specify a particular format, however, obviously the data are more usable, if it is encoded in a way that the essential information specified by MIAME can be accessed easily. MGED recommends the use of MAGE-TAB format, which is based on spreadsheets, or MAGE-ML.

MIAME also does not specify any particular terminology, however for automated data exchange the use of standard controlled vocabularies and ontologies are desirable. MGED recommends the use of MGED Ontology for the description of the key experimental concepts, and where possible ontologies developed by

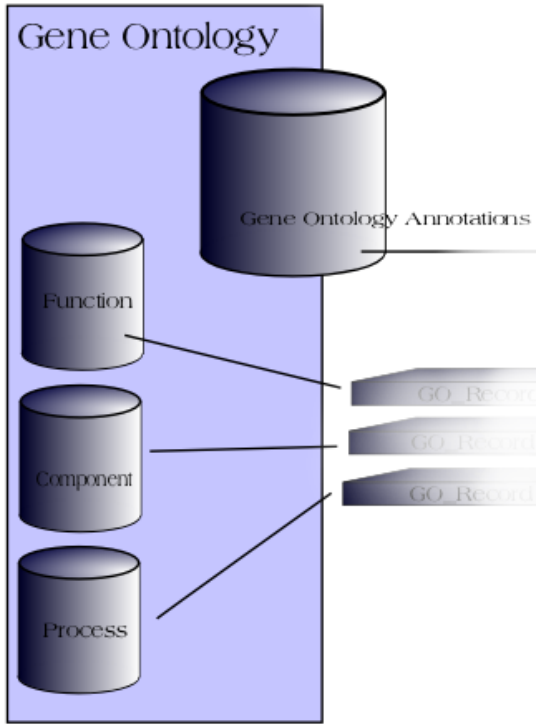
Statistical analysis

The analysis of DNA microarrays poses a large number of statistical problems, including the normalisation of the data.

From a hypothesis-testing perspective, the large number of genes present on a single array means that the experimenter must take into account a multiple testing problem: even if each gene is extremely unlikely to randomly yield a result of interest, the combination of all the genes is likely to show at least one or a few occurrences of this result which are false positives.

Data mining includes gene ontology

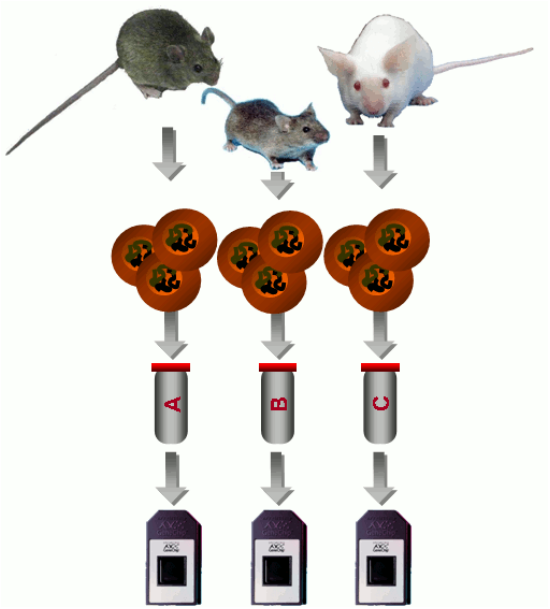
Gene ontology is a controlled vocabulary used to describe the biology of a gene product in any organism. There are 3 independent sets of vocabularies, or ontologies, that describe the:
-molecular function of a gene product,
- the biological process in which the gene product participates,
-and the cellular component where the gene product can be found.

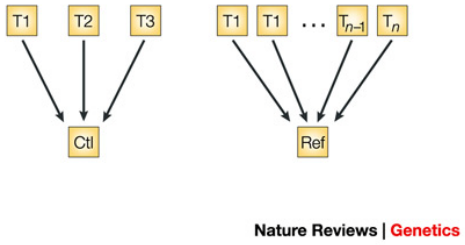


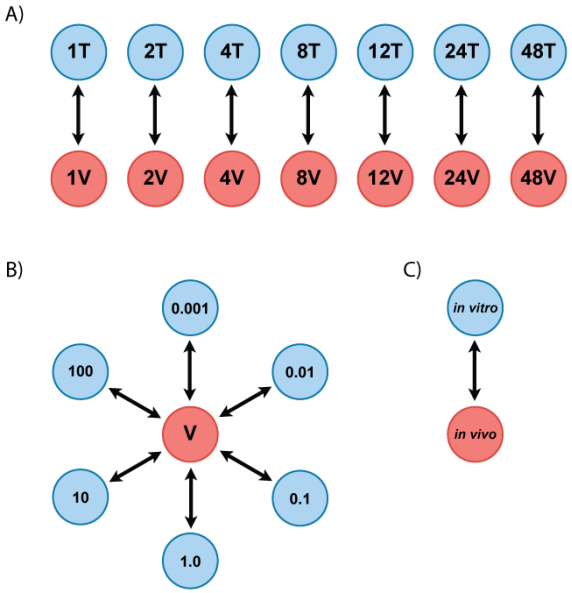
Experimental design is crucial for a succesfull “omic” experiment

A proper experimental design is crucial for obtaining useful conclusions from a project. The choice of design ideally includes an assessment of the biological variation, the technical variation, the cost and duration of the experiment, and the availability of biological material. The experimental design can also depend on the methods that will be used to analyse the data afterwards. In certain cases, the parameters needed to find the optimal design must be obtained by a pilot experiment.

A related problem is the comparison of different competing experimental methods or devices. Here, a proper test design is crucial as well to be able to make a firm conclusion in favor of one or the other method.



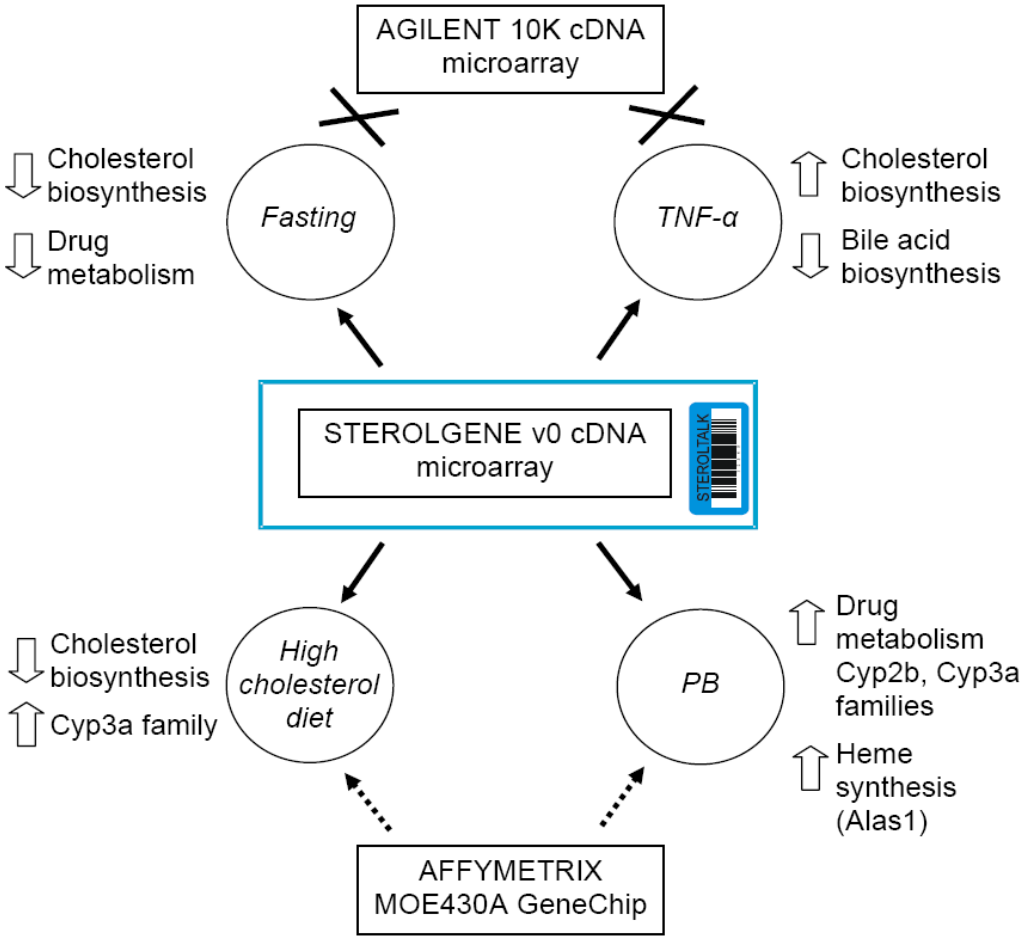




Dere E. et al., *BMC Genomics* 2006, **7:**80doi

lunabiosciences.com/experimentaldesign.html

Experimenal design – an example



Changes in cholesterol homeostasis and drug metabolism caused by different factors in mouse liver

T. Rezen et al., *BMC Genomics* 2008, 9:76

Pomen bioinformatike v raziskavh “omov” - povzetek

- Matematično-informatični pristopi v bioloških poskusih pogenomske dobe (bioinformatika) omogočajo razbiranje biološkega smisla v enormni količini podatkov.

-Bioinformatika (računska biologija, matematična biologija) za reševanje bioloških problemov uporablja metode uporabne matematike, informatike, statistike in računalniških znanosti.

-Dobro načrtovanje bioloških poskusov zahteva sodelovanje eksperimentatorjev in informatikov že od vsega začetka.

-Zasnova poskusa zahteva definicijo biolškega vprašanja, izbiro platforme za analizo, določitev števila bioloških (in tehničnih) ponovitev in načrt serije poskusov (hibridizacij ali sekvenciranj).

-Po tehnični izvedbi poskusa sledita statistična in informatična obdelava ter rudarjenje podatkov.

- Statistično-informatična obdelava obsega ekstrakcijo intenzitet signala, normalizacijo podatkov ter različne statistične teste, da pridobimo listo diferencialno izraženih genov ali listo zaporedij DNA.

- Sekundarna informatična analiza in rudarjenje podatkov obsegata gručanje in razpoznavanje vzorcev, študije seznama genov z genskimi ontologijami, iskanje regulatornih vzorcev, kot tudi načrtovanje validacijskih eksperimentov.

-Matematično-informatične metode za povezovanje podatkov različnih “omov” (npr. transkriptom, proteom, metabolom) so še v razvoju.