Molecular Biology Techniques

Andreas Hein

First Edition, 2012

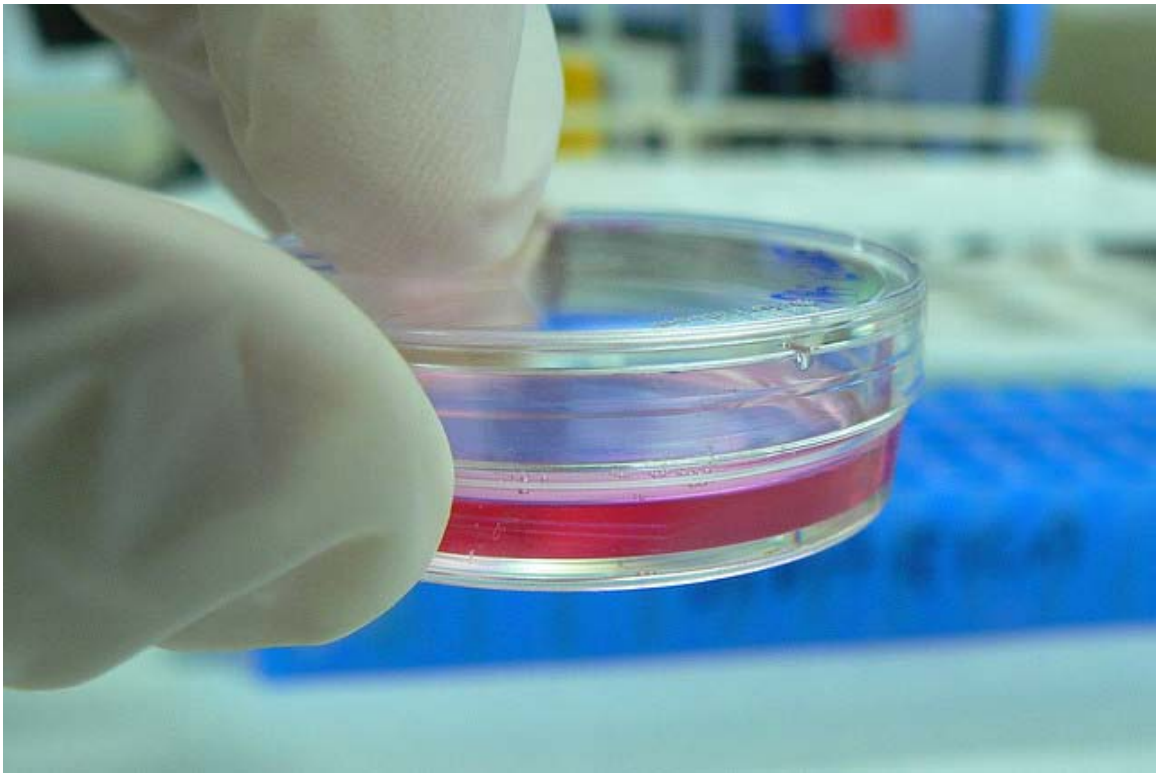
ISBN 978-81-323-2880-3

Published by:
Orange Apple
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

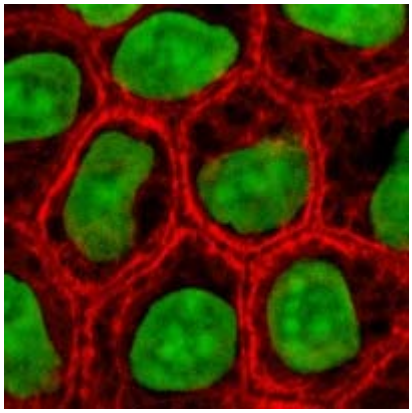
Table of Contents

Chapter- 1

Cell Culture



Cell culture in a Petri dish



Epithelial cells in culture, stained for keratin (red) and DNA (green)

Cell culture is the complex process by which cells are grown under controlled conditions. In practice, the term "cell culture" has come to refer to the culturing of cells derived from multicellular eukaryotes, especially animal cells. However, there are also cultures of plants, fungi and microbes, including viruses, bacteria and protists. The historical development and methods of cell culture are closely interrelated to those of tissue culture and organ culture.

Animal cell culture became a common laboratory technique in the mid-1900s, but the concept of maintaining live cell lines separated from their original tissue source was discovered in the 19th century.

History

The 19th-century English physiologist Sydney Ringer developed salt solutions containing the chlorides of sodium, potassium, calcium and magnesium suitable for maintaining the beating of an isolated animal heart outside of the body. In 1885 Wilhelm Roux removed a portion of the medullary plate of an embryonic chicken and maintained it in a warm saline solution for several days, establishing the principle of tissue culture. Ross Granville Harrison, working at Johns Hopkins Medical School and then at Yale University, published results of his experiments from 1907–1910, establishing the methodology of tissue culture.

Cell culture techniques were advanced significantly in the 1940s and 1950s to support research in virology. Growing viruses in cell cultures allowed preparation of purified viruses for the manufacture of vaccines. The injectable polio vaccine developed by Jonas Salk was one of the first products mass-produced using cell culture techniques. This vaccine was made possible by the cell culture research of John Franklin Enders, Thomas Huckle Weller, and Frederick Chapman Robbins, who were awarded a Nobel Prize for their discovery of a method of growing the virus in monkey kidney cell cultures.

Concepts in mammalian cell culture

Isolation of cells

Cells can be isolated from tissues for *ex vivo* culture in several ways. Cells can be easily purified from blood, however only the white cells are capable of growth in culture. Mononuclear cells can be released from soft tissues by *enzymatic digestion* with enzymes such as collagenase, trypsin, or pronase, which break down the extracellular matrix. Alternatively, pieces of tissue can be placed in growth media, and the cells that grow out are available for culture. This method is known as *explant culture*.

Cells that are cultured directly from a subject are known as ***primary cells***. With the exception of some derived from tumors, most primary cell cultures have limited lifespan. After a certain number of population doublings (called the Hayflick limit) cells undergo the process of senescence and stop dividing, while generally retaining viability.

An established or immortalised cell line has acquired the ability to proliferate indefinitely either through random mutation or deliberate modification, such as artificial expression of the telomerase gene. There are numerous well established cell lines representative of particular cell types.

Maintaining cells in culture

Cells are grown and maintained at an appropriate temperature and gas mixture (typically, 37°C, 5% CO_2 for mammalian cells) in a cell incubator. Culture conditions vary widely for each cell type, and variation of conditions for a particular cell type can result in different phenotypes being expressed.

Aside from temperature and gas mixture, the most commonly varied factor in culture systems is the growth medium. Recipes for growth media can vary in pH, glucose concentration, growth factors, and the presence of other nutrients. The growth factors used to supplement media are often derived from animal blood, such as calf serum. One complication of these blood-derived ingredients is the potential for contamination of the culture with viruses or prions, particularly in biotechnology medical applications. Current practice is to minimize or eliminate the use of these ingredients wherever possible and use chemically defined media, but this cannot always be accomplished. Alternative strategies involve sourcing the animal blood from countries with minimum BSE/TSE risk such as Australia and New Zealand, and using purified nutrient concentrates derived from serum in place of whole animal serum for cell culture.

Plating density (number of cells per volume of culture medium) plays a critical role for some cell types. For example, a lower plating density makes granulosa cells exhibit estrogen production, while a higher plating density makes them appear as progesterone producing theca lutein cells.

Cells can be grown in *suspension* or *adherent* cultures. Some cells naturally live in suspension, without being attached to a surface, such as cells that exist in the bloodstream. There are also cell lines that have been modified to be able to survive in suspension cultures so that they can be grown to a higher density than adherent conditions would allow. Adherent cells require a surface, such as tissue culture plastic or microcarrier, which may be coated with extracellular matrix components to increase adhesion properties and provide other signals needed for growth and differentiation. Most cells derived from solid tissues are adherent. Another type of adherent culture is *organotypic culture* which involves growing cells in a three-dimensional environment as opposed to two-dimensional culture dishes. This 3D culture system is biochemically and physiologically more similar to *in vivo* tissue, but is technically challenging to maintain because of many factors (e.g. diffusion).

Cell line cross-contamination

Cell line cross-contamination can be a problem for scientists working with cultured cells. Studies suggest that anywhere from 15–20% of the time, cells used in experiments have been misidentified or contaminated with another cell line. Problems with cell line cross contamination have even been detected in lines from the NCI-60 panel, which are used routinely for drug-screening studies. Major cell line repositories including the American Type Culture Collection (ATCC) and the German Collection of Microorganisms and Cell Cultures (DSMZ) have received cell line submissions from researchers that were misidentified by the researcher. Such contamination poses a problem for the quality of research produced using cell culture lines, and the major repositories are now authenticating all cell line submissions. ATCC uses short tandem repeat (STR) DNA fingerprinting to authenticate its cell lines.

To address this problem of cell line cross-contamination, researchers are encouraged to authenticate their cell lines at an early passage to establish the identity of the cell line. Authentication should be repeated before freezing cell line stocks, every two months during active culturing and before any publication of research data generated using the cell lines. There are many methods for identifying cell lines including isoenzyme analysis, human lymphocyte antigen (HLA) typing, Chromosomal analysis, Karyotyping, Morphology and STR analysis.

One significant cell-line cross contaminant is the immortal HeLa cell line.

Manipulation of cultured cells

As cells generally continue to divide in culture, they generally grow to fill the available area or volume. This can generate several issues:

- Nutrient depletion in the growth media
- Accumulation of apoptotic/necrotic (dead) cells.
- Cell-to-cell contact can stimulate cell cycle arrest, causing cells to stop dividing known as contact inhibition or senescence.

- Cell-to-cell contact can stimulate cellular differentiation.

Among the common manipulations carried out on culture cells are media changes, passaging cells, and transfecting cells. These are generally performed using tissue culture methods that rely on sterile technique. Sterile technique aims to avoid contamination with bacteria, yeast, or other cell lines. Manipulations are typically carried out in a biosafety hood or laminar flow cabinet to exclude contaminating micro-organisms. Antibiotics (e.g. penicillin and streptomycin) and antifungals (e.g. Amphotericin B) can also be added to the growth media.

As cells undergo metabolic processes, acid is produced and the pH decreases. Often, a pH indicator is added to the medium in order to measure nutrient depletion.

Media changes

In the case of adherent cultures, the media can be removed directly by aspiration and replaced.

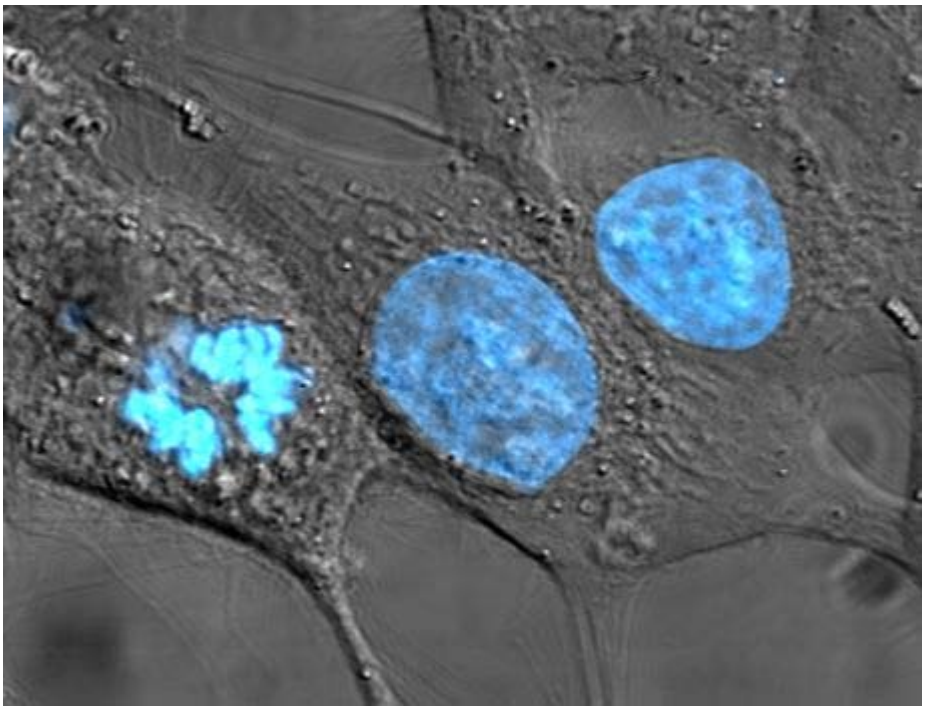
Passaging cells

Passaging (also known as subculture or splitting cells) involves transferring a small number of cells into a new vessel. Cells can be cultured for a longer time if they are split regularly, as it avoids the senescence associated with prolonged high cell density. Suspension cultures are easily passaged with a small amount of culture containing a few cells diluted in a larger volume of fresh media. For adherent cultures, cells first need to be detached; this is commonly done with a mixture of trypsin-EDTA, however other enzyme mixes are now available for this purpose. A small number of detached cells can then be used to seed a new culture.

Transfection and transduction

Another common method for manipulating cells involves the introduction of foreign DNA by transfection. This is often performed to cause cells to express a protein of interest. More recently, the transfection of RNAi constructs have been realized as a convenient mechanism for suppressing the expression of a particular gene/protein. DNA can also be inserted into cells using viruses, in methods referred to as transduction, infection or transformation. Viruses, as parasitic agents, are well suited to introducing DNA into cells, as this is a part of their normal course of reproduction.

Established human cell lines



Cultured HeLa cells have been stained with Hoechst turning their nuclei blue, and are one of the earliest human cell lines descended from Henrietta Lacks, who died of cervical cancer from which these cells originated.

Cell lines that originate with humans have been somewhat controversial in bioethics, as they may outlive their parent organism and later be used in the discovery of lucrative medical treatments. In the pioneering decision in this area, the Supreme Court of California held in *Moore v. Regents of the University of California* that human patients have no property rights in cell lines derived from organs removed with their consent.

Generation of hybridomas

It is possible to fuse normal cells with an immortalised cell line. This method is used to produce monoclonal antibodies. In brief, lymphocytes isolated from the spleen (or possibly blood) of an immunised animal are combined with an immortal myeloma cell line (B cell lineage) to produce a hybridoma which has the antibody specificity of the primary lymphoctye and the immortality of the myeloma. Selective growth medium (HA or HAT) is used to select against unfused myeloma cells; primary lymphoctyes die quickly in culture and only the fused cells survive. These are screened for production of the required antibody, generally in pools to start with and then after single cloning.

Applications of cell culture

Mass culture of animal cell lines is fundamental to the manufacture of viral vaccines and other products of biotechnology

Biological products produced by recombinant DNA (rDNA) technology in animal cell cultures include enzymes, synthetic hormones, immunobiologicals (monoclonal antibodies, interleukins, lymphokines), and anticancer agents. Although many simpler proteins can be produced using rDNA in bacterial cultures, more complex proteins that are glycosylated (carbohydrate-modified) currently must be made in animal cells. An important example of such a complex protein is the hormone erythropoietin. The cost of growing mammalian cell cultures is high, so research is underway to produce such complex proteins in insect cells or in higher plants, use of single embryonic cell and somatic embryos as a source for direct gene transfer via particle bombardment, transit gene expression and confocal microscopy observation is one of its applications. It also offers to confirm single cell origin of somatic embryos and the asymmetry of the first cell division, which starts the process. -

Tissue culture and engineering

Cell culture is a fundamental component of tissue culture and tissue engineering, as it establishes the basics of growing and maintaining cells *ex vivo*. The major application of human cell culture is in stem cell industry where mesenchymal stem cells can be cultured and cryopreserved for future use.

Vaccines

Vaccines for polio, measles, mumps, rubella, and chickenpox are currently made in cell cultures. Due to the H5N1 pandemic threat, research into using cell culture for influenza vaccines is being funded by the United States government. Novel ideas in the field include recombinant DNA-based vaccines, such as one made using human adenovirus (a common cold virus) as a vector, such as adjuvants.

Culture of non-mammalian cells

Plant cell culture methods

Plant cell cultures are typically grown as cell suspension cultures in liquid medium or as callus cultures on solid medium. The culturing of undifferentiated plant cells and calli requires the proper balance of the plant growth hormones auxin and cytokinin.

Bacterial and yeast culture methods

For bacteria and yeast, small quantities of cells are usually grown on a solid support that contains nutrients embedded in it, usually a gel such as agar, while large-scale cultures are grown with the cells suspended in a nutrient broth.

Viral culture methods

The culture of viruses requires the culture of cells of mammalian, plant, fungal or bacterial origin as hosts for the growth and replication of the virus. Whole wild type

viruses, recombinant viruses or viral products may be generated in cell types other than their natural hosts under the right conditions. Depending on the species of the virus, infection and viral replication may result in host cell lysis and formation of a viral plaque.

Common cell lines

Human cell lines

- National Cancer Institute's 60 cancer cell lines

- DU145 (Prostate cancer)
- Lncap (Prostate cancer)
- MCF-7 (breast cancer)
- MDA-MB-438 (breast cancer)
- PC3 (Prostate cancer)
- T47D (breast cancer)
- THP-1 (acute myeloid leukemia)
- U87 (glioblastoma)
- SHSY5Y Human neuroblastoma cells, cloned from a myeloma
- Saos-2 cells (bone cancer)

Primate cell lines

- Vero (African green monkey *Chlorocebus* kidney epithelial cell line initiated 1962)

Rat tumor cell lines

- GH3 (pituitary tumor)
- PC12 (pheochromocytoma)

Mouse cell lines

- MC3T3 (embryonic calvarial)

Plant cell lines

- Tobacco BY-2 cells (kept as cell suspension culture, they are model system of plant cell)

Other species cell lines

- zebrafish ZF4 and AB9 cells.
- *Madin-Darby Canine Kidney (MDCK)* epithelial cell line
- Xenopus A6 kidney epithelial cells.

List of cell lines

Cell line	Meaning	Organism	Origin tissue	Morphology	Link
293-T		Human	Kidney (embryonic)		Derivative of HEK 293ECACC
3T3 cells	"3-day transfer, inoculum 3 x 10^5 cells"	Mouse	Embryonic fibroblast		Also known as NIH 3T3 ECACC
721		Human	Melanoma		
9L		Rat	Glioblastoma		
A2780		Human	Ovary	Ovarian Cancer	ECACC
A2780ADR		Human	Ovary	Adriamycin-resistant derivative	ECACC
A2780cis		Human	Ovary	Cisplatin-resistant derivative	ECACC
A172		Human	glioblastoma	malignant glioma	ECACC
A20		Murine	B lymphoma	B lymphocyte	
A253		Human	Head and neck carcinoma	submandibular duct	
A431		Human	Skin epithelium	squamous carcinoma	ECACCCell Line Data Base
A-549		Human	Lungcarcinoma	Epithelium	DSMZECACC
ALC		Murine	bone marrow	Stroma	PubMed
B16		Murine	Melanoma		ECCAC
B35		Rat	Neuroblastoma		ATCC
BCP-1 cells		Human	PBMC	HIV+ Lymphoma	ATCC
BEAS-2B	Bronchial epithelium + Adenovirus 12-SV40 virus hybrid (Ad12SV40)	Human	Lung	Epithelial	ATCC
bEnd.3	*Brain endothelial*	Mouse	Brain / Cerebral cortex	Endothelium	ATCC
BHK-21	"Baby Hamster Kidney Fibroblast cells"	Hamster	Kidney	fibroblast	ECACCOlympus
BR 293		Human	Breast	Breast cancer	
BxPC3	Biopsy xenograph of pancreatic carcinoma line 3	Human	pancreatic adenocarcinoma	Epithelial	ATCC
C3H-10T1/2		Mouse	Embryonic mesenchymal cell line		ECACC
C6/36		Asian tiger	larval tissue		ECACC

		mosquito			
Cal-27		Human	Tongue	squamous cell carcinoma	
CHO	*Chinese hamster ovary*	hamster	Ovary	Epithelium	ECACCICLC
COR-L23		Human	Lung		ECACC
COR-L23/CPR		Human	Lung		ECACC
COR-L23/5010		Human	Lung		ECACC
COR-L23/R23		Human	Lung	Epithelial	ECACC
COS-7	*Cercopithecus aethiops, origin-defective SV-40*	Ape - *Cercopithecus aethiops* (Chlorocebus)	Kidney	fibroblast	ECACCATCC
COV-434		Human	Ovary	Metastatic granulosa cell carcinoma	ECACC
CML T1	*Chronic Myelod Leukaemia T-lymphocyte 1*	Human	CML acute phase	T cell leukaemia	Blood
CMT	*canine mammary tumor*	Dog	Mammary gland	Epithelium	
CT26		Murine	Colorectal Carcinoma	Colon	
D17		canine	osteosarcoma		ECACC
DH82		canine	histiocytosis	monocyte/macrophage	ECACC
DU145		Human	Androgen insensitive carcinoma	Prostate	J Vir Meth PubMed
DuCaP	Dura mater Cancer of the Prostate	Human	Metastatic Prostate Cancer	Epithelial	EAC { Ehrlich Ascites Carcinoma } mice
EL4		Mouse		T cell leukaemia	ECACC
EM2		Human	CML blast crisis	Ph+ CML line	Cell Line Data Base
EM3		Human	CML blast crisis	Ph+ CML line	Cell Line Data Base
EMT6/AR1		Mouse	Breast	Epithelial-like	ECACC
EMT6/AR10.0		Mouse	Breast	Epithelial-like	ECACC
FM3		Human	Metastatic lymph node	melanoma	

H1299		Human	Lung	Lung cancer	
H69		Human	Lung		ECACC
HB54		hybridoma	hybridoma	secretes L243 mAb (against HLA-DR)	Human Immunology
HB55		hybridoma	hybridoma	secretes MA2.1 mAb (against HLA-A2 and HLA-B17)	Journal of Immunology
HCA2		Human	fibroblast		Journal of General Virology
HEK-293	*Human embryonic kidney*	Human	Kidney (embryonic)	Epithelium	ATCC
HeLa	*Henrietta Lacks*	Human	Cervical cancer	Epithelium	DSMZECACC
Hepa1c1c7	clone 7 of clone 1 hepatoma line 1	Mouse	Hepatoma	Epithelial	ECACC ATCC
HL-60	*Human leukemia*	Human	Myeloblast	bloodcells	ECACCDSMZ
HMEC	*Human mammary epithelial cell*	Human		Epithelium	ECACC
HT-29		Human	Colon epithelium	Adenocarcinoma	ECACC Cell Line Data Base
Jurkat		Human	T-Cell-Leukemia	white blood cells	ECACC DSMZ
JY cells		Human	Lymphoblastoid	EBV immortalised B cell	
K562 cells		Human	Lymphoblastoid	CML blast crisis	ECACC
Ku812		Human	Lymphoblastoid	erythroleukemia	ECACC LGCstandards
KCL22		Human	Lymphoblastoid	CML	
KG1		Human	Lymphoblastoid	AML	
KYO1	Kyoto 1	Human	Lymphoblastoid	CML	DSMZ
LNCap	Lymph node Cancer of the Prostate	Human	prostatic adenocarcinoma	Epithelial	ECACCATCC
Ma-Mel 1, 2, 3....48		Human		a range of melanoma cell lines	
MC-38		Mouse		Adenocarcinoma	
MCF-7	*Michigan Cancer Foundation-7*	Human	Mammary gland	Invasive breast ductal carcinoma	ER+, PR+
MCF-10A	*Michigan*	Human	mammary gland	Epithelium	ATCC

	Cancer Foundation				
MDA-MB-231	M.D. Anderson - Metastatic Breast	Human	Breast	Cancer	ECACC
MDA-MB-468	M.D. Anderson - Metastatic Breast	Human	Breast	Cancer	ECACC
MDA-MB-435	M.D. Anderson - Metastatic Breast	Human	Breast	melanoma or carcinoma (disputed)	Cambridge Pathology ECACC
MDCK II	*Madin Darby canine kidney*	Dog	Kidney	Epithelium	ECACC ATCC
MDCK II	*Madin Darby canine kidney*	Dog	Kidney	Epithelium	ATCC
MOR/0.2R		Human	Lung		ECACC
MONO-MAC 6		Human	WBC	myeloid metaplasic AML	Cell Line Data Base
MTD-1A		Mouse		Epithelium	
MyEnd	*Myocardial endothelial*	Mouse		Endothelium	
NCI-H69/CPR		Human	Lung		ECACC
NCI-H69/LX10		Human	Lung		ECACC
NCI-H69/LX20		Human	Lung		ECACC
NCI-H69/LX4		Human	Lung		ECACC
NIH-3T3	*NIH, 3-day transfer, inoculum 3 x 10^5 cells*	Mouse	embryo	fibroblast	ECACCATCC
NALM-1			peripheral blood	blast-crisis CML	Cancer Genetics and Cytogenetics
NW-145				Melanoma	ESTDAB
OPCN / OPCT cell lines	Onyvax Prostate Cancer....			Range of prostate tumour lines	Asterand
Peer		Human	T cell leukemia		DSMZ
PNT-1A / PNT 2				Prostate tumour lines	ECACC

RenCa	Renal Carcinoma	Mouse		renal carcinoma	
RIN-5F		Mouse	Pancreas		
RMA/RMAS		Mouse		T cell tumour	
Saos-2 cells		Human		Osteosarcoma	ECACC
Sf-9	*Spodoptera frugiperda*	insect - *Spodoptera frugiperda* (moth)	Ovary		DSMZECACC
SkBr3		Human		Breast carcinoma	
T2		Human		T cell leukemia/B cell line hybridoma	DSMZ
T-47D		Human	Mammary gland	ductal carcinoma	
T84		Human	colorectal Carcinoma / Lungmetastasis	Epithelium	ECACCATCC
THP1 cell line		Human	Monocyte	AML	ECACC
U373		Human	Glioblastoma-astrocytoma	Epithelium	
U87		Human	glioblastoma-astrocytoma	Epithelial-like	Abcam
U937		Human	Leukaemic monocytic lymphoma		ECACC
VCaP	Vertebra Prostate Cancer	Human	Metastatic prostate cancer	Epithelial	ECACC ATCC

Vero cells	'Vera Reno' ('Green kidney') / 'Vero' ('truth')	African Green Monkey	Kidney epithelium		ECACC
WM39		Human	skin	Primary melanoma	
WT-49		Human	Lymphoblastoid		
X63		Mouse	Melanoma		
YAC-1		Mouse	Lymphoma		Cell Line Data Base ECACC
YAR		Human	B-cell	EBV transofrmed	Human Immunology

Note: this list is a sample of available cell lines, and is not comprehensive

Chapter- 2

ChIA-PET

Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) is a technique that incorporates chromatin immunoprecipitation (ChIP)-based enrichment, chromatin proximity ligation, Paired-End Tags, and ultra-high-throughput sequencing to determine *de novo* long-range chromatin interactions genome-wide (Fullwood & Yijun, 2009). Genes can be regulated by regions far from the promoter such as regulatory elements, insulators and boundary elements, and transcription-factor binding sites (TFBS). Uncovering the interplay between regulatory regions and gene coding regions is essential for understanding the mechanisms governing gene regulation in health and disease (Maston et al., 2006). ChIA-PET can be used to identify unique, functional chromatin interactions between distal and proximal regulatory TFBS and the promoters of the genes they interact with.

ChIA-PET can also be used to unravel the mechanisms of genome control during processes such as cell differentiation, proliferation, and development. By creating ChIA-PET interactome maps for DNA-binding regulatory proteins and promoter regions, we can better identify unique targets for therapeutic intervention (Fullwood & Yijun, 2009).

Methodology

The ChIA-PET method combines ChIP-based methods (Kuo & Allis, 1999), and Chromosome conformation capture (3C), to extend the capabilities of both approaches. ChIP-Sequencing (ChIP-Seq) is a popular method used to identify TFBS while 3C has been used to identify long-range chromatin interactions (Dekker et al., 2002). However, both suffer from limitations when used independently to identify de-novo long-range interactions genome wide. While ChIP-Seq is typically used for genome-wide identification of TFBS (Barski et al., 2007; Wei et al., 2006), it provides only linear information of protein binding sites along the chromosomes (but not interactions between them), and suffers from high genomic background noise (false positives). Additionally, only a small amount of sequences generated by ChIP-Seq uniquely map to the genome, and an even smaller amount are functional TFBS.

While 3C is capable of analyzing long-range chromatin interactions, it cannot be used genome wide and, like ChIP-Seq, also suffers from high levels of background noise. Since the noise increases in relation to the distance between interacting regions (max 100kb), laborious and tedious controls are required for accurate characterization of chromatin interactions (Dekker et al., 2006).

The ChIA-PET method successfully resolves the issues of non-specific interaction noise found in ChIP-Seq by sonicating the chip fragments in order to separate random attachments from specific interaction complexes. The next step, which is referred to as enrichment, reduces complexity for genome-wide analysis and adds specificity to chromatin interactions bound by pre-determined TFs (transcription factors). The ability of 3C approaches to identify long-range interactions is based on the theory of proximity ligation. In regards to DNA inter-ligation, fragments that are tethered by common protein complexes have greater kinetic advantages under dilute conditions, than those freely diffusing in solution or anchored in different complexes. ChIA-PET takes advantage of this concept by incorporating linker sequences onto the free ends of the DNA fragments tethered to the protein complexes. In order to build connectivity of the fragments tethered by regulatory complexes, the linker sequences are ligated during nuclear proximity ligation. Therefore, the products of linker-connected ligation can be analyzed by ultra-high-throughput PET sequencing and mapped to the reference genome. Since ChIA-PET is not dependent on specific sites for detection as 3C and 4C are, it allows unbiased, genome-wide de-novo detection of chromatin interactions.

Workflow

Wet-lab portion of the workflow:

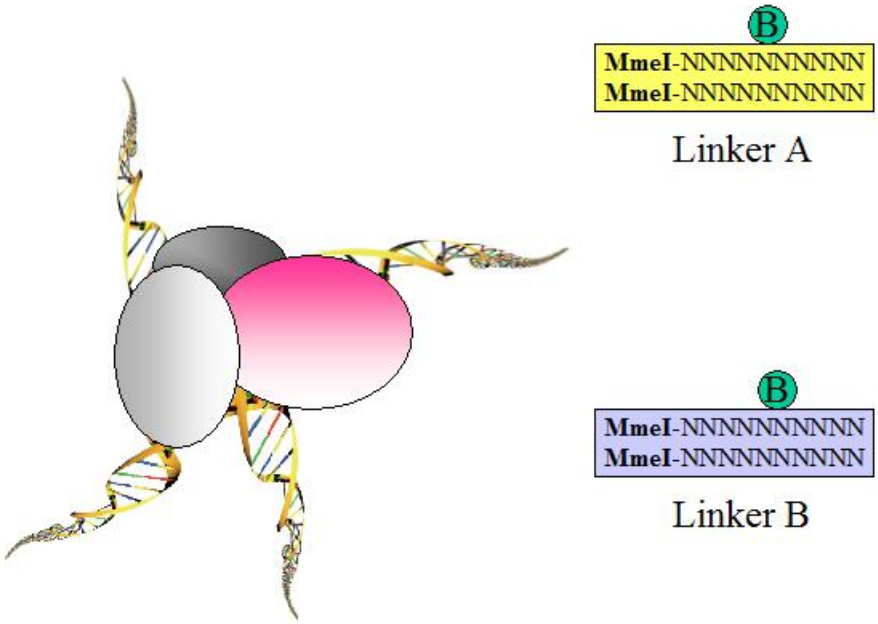


Figure 3. Biotinylated universal linkers with Mme1 restriction endonuclease sites are introduced.

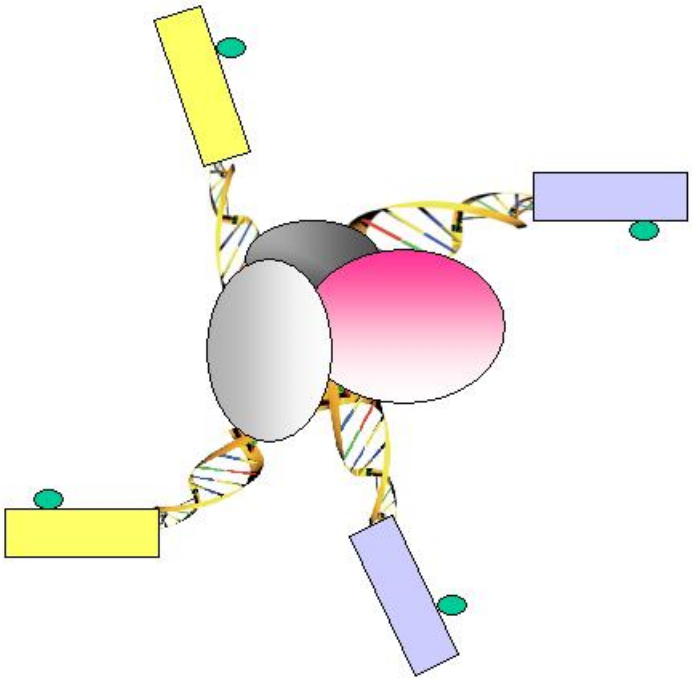


Figure 4. Biotinylated universal linkers are ligated to the free DNA ends.

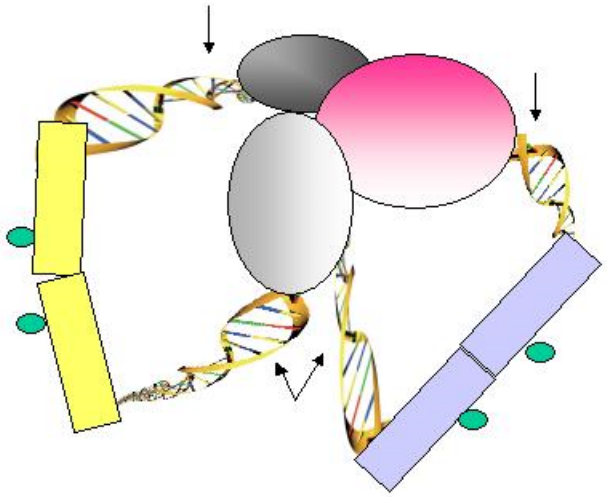


Figure 5. Ligation of linkers during proximity ligation.

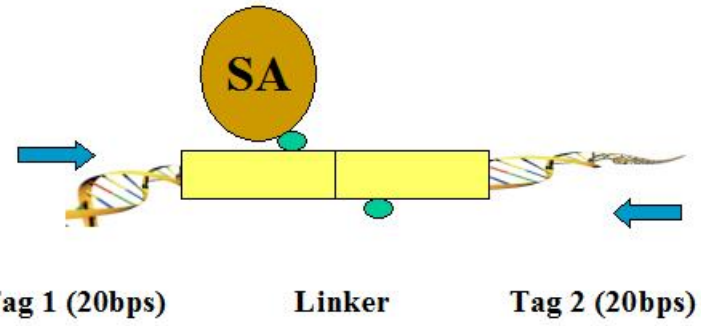


Figure 6. Pull down of biotinylated linkers by streptavidin-beads, and amplification of DNA tags.

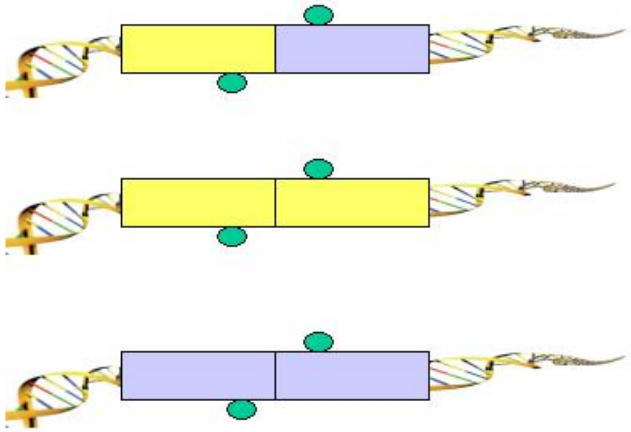


Figure 7. Conformations of universal linkers.

- Figure 1. Formaldehyde is used to cross-link the DNA-protein complexes. Sonication is used to break-up the chromatin and also to reduce non-specific interactions.

- Figure 2. A specific antibody of choice is used to enrich protein of interest bound chromatin fragments. ChIP material bound by the antibody are used to construct the ChIA-PET.

- Figure 3. Biotinylated oligonucleotide half-linkers containing flanking MmeI sites are used to connect proximity ligated DNA fragments. Two different linkers are designed (A and B) with specific nucleotide barcodes (CG or AT) for each of the two linker sequences.

- Figure 4. The linkers are ligated to the tethered DNA fragments.

- Figure 5. The linker fragments are ligated on the ChIP beads under dilute conditions. The purified DNA is then digested by MmeI, which cuts at a distance from its recognition site to release the tag-linker-tag structure.

- Figure 6. The biotinylated PETs are then immobilized on streptavidin-conjugated magnetic beads.

- Figure 7. PET sequences with AA (CG/CG) and BB (AT/AT) linker barcode composition are considered to be possible intra-complex ligation products, while the PET sequences with AB (CG/AT) linker composition are considered to be derived from chimeric ligation products between DNA fragments bounded in different chromatin complexes.

Dry-lab portion of the workflow:

PET extraction, mapping, and statistical analyses

The PET tags are extracted and mapped to the reference human genome in-silico.

Identification of ChIP enriched peaks (binding sites)

Self-ligated PET are used for identifying ChIP enriched sites because they provide the most reliable mapping (20 + 20 bps) to the reference genome.

ChIP enrichment peak-finding algorithm

A called peak is considered a binding site if there are multiple overlapping self-ligated PETs. The false discovery rate (FDR) is determined using statistical simulations to estimate the random background of PET-derived virtual DNA overlaps, and the estimated background noise.

Filtering of repetitive DNA (affects non-specific binding)

Satellite regions and binding sites present in regions with severe structural variations are removed.

ChIP enrichment count

The numbers of self-ligation and inter-ligation PETs (within + 250 bp window) are reported at each site. The total number of self-ligated and inter-ligated PETs at a specific site is called the ChIP enrichment count.

Figure 8. PET Classification: Uniquely aligned PET sequences can be classified by whether they are derived from one DNA fragment or two DNA fragments.

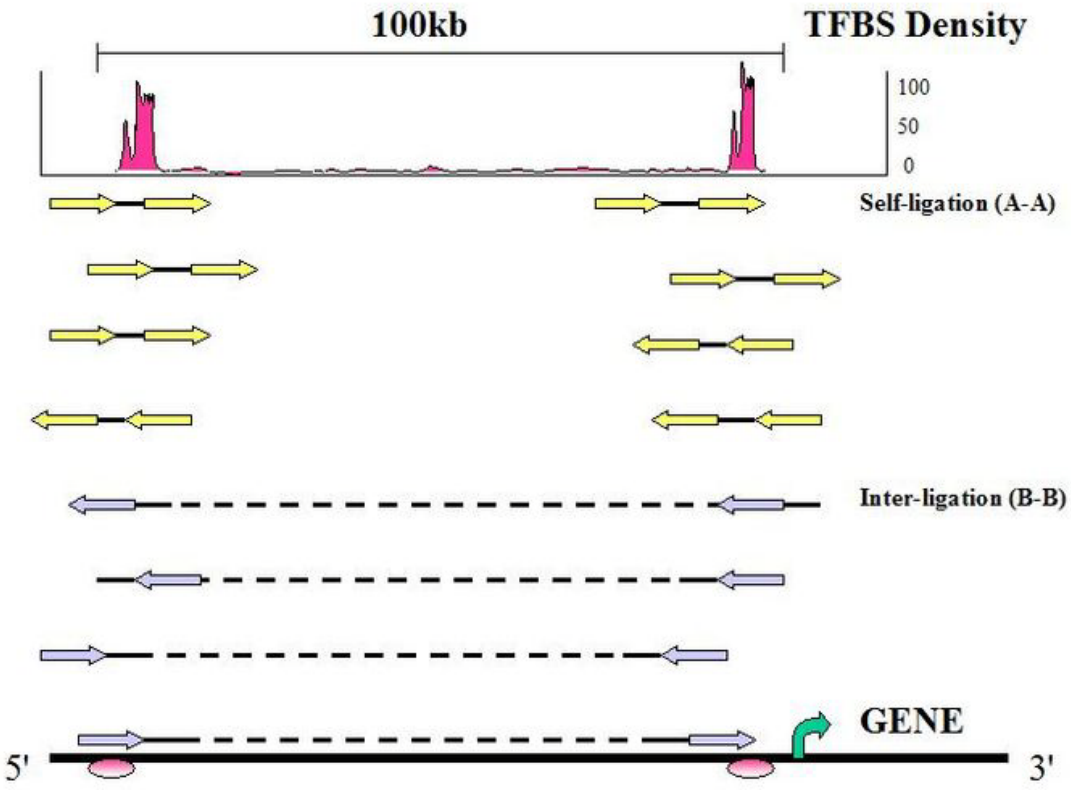


Figure 8. Intra and inter-ligated PETs are clustered around TFBS when mapped to the reference human genome.

- **Self-ligation PETs**

If the two tags of a PET are mapped on the same chromosome with the genomic span in the range of ChIP DNA fragments (less than 3 Kb), with expected self-ligation orientation and on the same strand, they are considered to be derived from a self-ligation of a single ChIP DNA fragment, and considered a self-ligation PET.

- **Inter-ligation PETs**

If a PET does not fit into these criteria, then the PET most likely resulted from a ligation product between two DNA fragments and referred to as an inter-ligation PET. The two tags of an inter-ligation PETs do not have fixed tag orientations, might not be found on the same strands, might have any genomic span, and might not map to the same chromosome.

- **Intrachromosomal inter-ligation PETs**

If the two tags of an inter-ligation PET are mapped in the same chromosome but with a span > 3 Kb in any orientation, then these PETs are called intrachromosomal inter-ligation PETs.

- **Interchromosomal inter-ligation PETs**

PETs which are mapped to different chromosomes are called interchromosomal inter-ligation PETs.

Figure 9. Proposed mechanism showing how distal regulatory elements can initiate long-range chromatin interactions involving promoter regions of target genes.

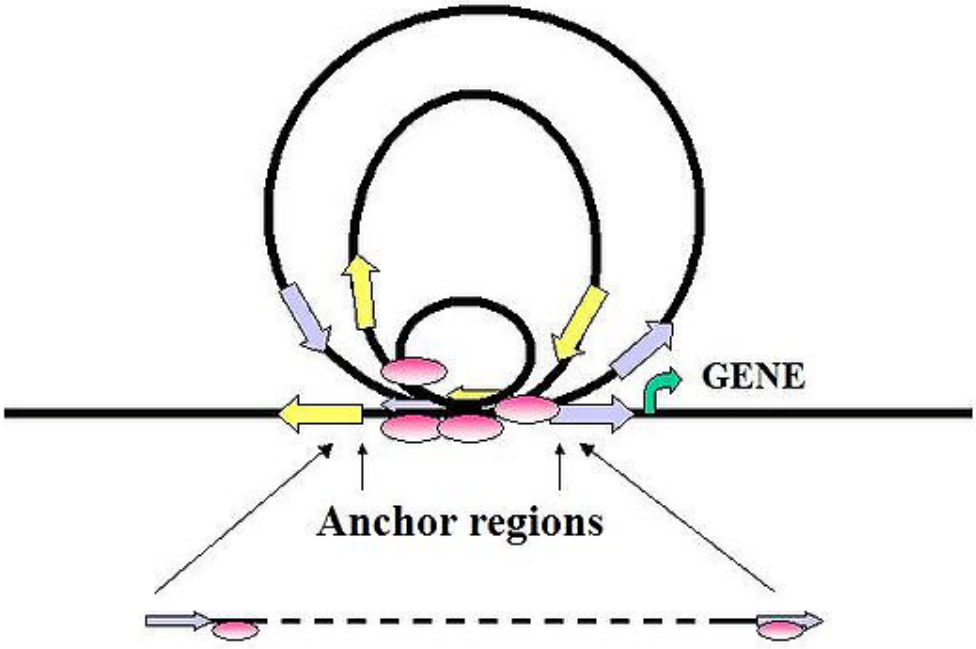


Figure 9. Proposed DNA looping mechanism between distal regulatory proteins and the promoter region

The interactions form DNA loop structures with multiple TFBS at the anchoring center. Small loops might package genes near the anchoring center in a tight sub-compartment, which could increase the local concentration of regulatory proteins for enhanced transcriptional activation. This mechanism might also enhance transcription efficiency, allowing RNA pol II to cycle the tight circular gene templates. The large interaction loops are more likely to link together distant genes at either end of the loop residing near

anchor sites for coordinated regulation, or could separate genes in long loops to prevent their activation. Adapted from Fullwood et al. (2009).

Strengths and Weaknesses

Advances of the ChIA-PET method

- ChIA-PET is an unbiased, whole-genome and de-novo approach for long-range chromatin interaction analysis.
- A ChIA-PET experiment is capable of providing two global datasets: The protein factor binding sites (self-ligated PETs); and The interactions between the binding sites (inter-ligated PETs).
- ChIA-PET involves ChIP to reduce the complexity for genome-wide analysis and adds specificity to chromatin interactions bound by specific factors of interest.
- ChIA-PET is compatible with tag-based next-generation sequencing approaches such as Roche 454 pyrosequencing, Illumina GA, ABI SOLiD, and Helicos.
- ChIA-PET is applicable to many different protein factors involved in transcriptional regulation or chromatin structural conformation.
- ChIA-PET analysis can be applied to chromatin interactions involved in a particular nuclear process. By using general TFs such as RNA Polymerase II, it may be possible to identify all chromatin interactions involved in transcription regulation. Further, the use of protein factors involved in DNA replication or chromatin structure would allow identification of all interactions due to DNA replication and chromatin structural modification (Fullwood et al., 2009).

Weaknesses

- It is well established that cis and trans-regulatory complexes contain unique combinations of proteins based on cell and tissue specific conditions (Dekker et al., 2006). While identification of single, functional TFBS is a significant advancement, the use of ChIA-PET to identify individual proteins in a complex would require guess work and multiple experiments to identify each interacting protein. This would be a costly and time consuming endeavour.
- ChIA-PET is limited by the quality, purity, and specificity of the antibodies used (Fullwood et al., 2009).
- ChIA-PET is dependent on identification of sequences that can be mapped to the reference sequence (ref).

- ChIA-PET requires the use of peak-calling computer algorithms to organize and map PET reads to the reference genome. Because of variations between software platforms, results can vary depending on which program is used.

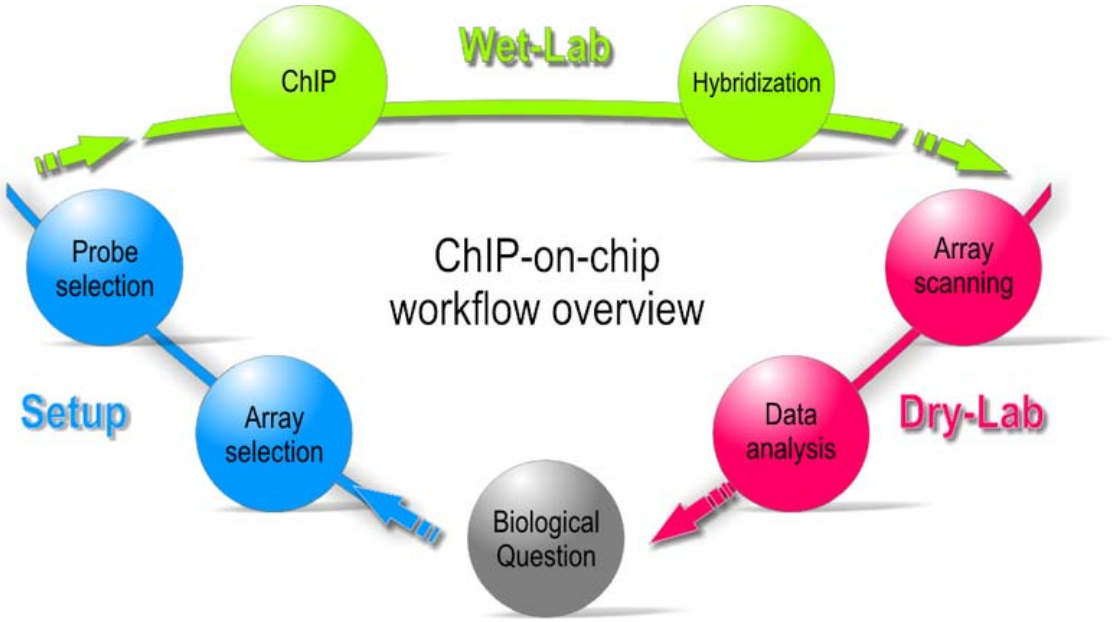
- Although repetitive DNA regions can be associated with gene regulation (Polak & Domany, 200), they need to be removed as they can affect the data (Fullwood et al., 2009).

History

Fullwood et al. (2009), used ChIA-PET to detect and map the chromatin interaction network mediated by oestrogen receptor alpha (ER-alpha) in human cancer cells. The resulting global chromatin interactome map revealed that remote ER-alpha-binding sites were also anchored to gene promoters through long-range chromatin interactions suggesting that ER-alpha functions by extensive chromatin looping in order to bring genes together for coordinated transcriptional regulation.

Chapter- 3

ChIP-on-Chip



Workflow overview of a ChIP-on-chip experiment.

ChIP-on-chip (also known as **ChIP-chip**) is a technique that combines chromatin immunoprecipitation (*"ChIP"*) with microarray technology (*"chip"*). Like regular ChIP, ChIP-on-chip is used to investigate interactions between proteins and DNA *in vivo*. Specifically, it allows the identification of the cistrome, sum of binding sites, for DNA-binding proteins on a genome-wide basis. Whole-genome analysis can be performed to determine the locations of binding sites for almost any protein of interest. As the name of the technique suggests, such proteins are generally those operating in the context of chromatin. The most prominent representatives of this class are transcription factors, replication-related proteins, like ORC, histones, their variants, and histone modifications. The goal of ChIP-on-chip is to localize protein binding sites that may help identify functional elements in the genome. For example, in the case of a transcription factor as a protein of interest, one can determine its transcription factor binding sites throughout the genome. Other proteins allow the identification of promoter regions, enhancers,

repressors and silencing elements, insulators, boundary elements, and sequences that control DNA replication. If histones are subject of interest, it is believed that the distribution of modifications and their localizations may offer new insights into the mechanisms of regulation. One of the long-term goals ChIP-on-chip was designed for is to establish a catalogue of (selected) organisms that lists all protein-DNA interactions under various physiological conditions. This knowledge would ultimately help in the understanding of the machinery behind gene regulation, cell proliferation, and disease progression. Hence, ChIP-on-chip offers not only huge potential to complement our knowledge about the orchestration of the genome on the nucleotide level, but also on higher levels of information and regulation as it is propagated by research on epigenetics.

Technological platforms

The technical platforms to conduct ChIP-on-chip experiments are DNA microarrays, or *"chips"*. They can be classified and distinguished according to various characteristics:

- *Probe type*: DNA arrays can comprise either mechanically spotted cDNAs or PCR-products, mechanically spotted oligonucleotides, or oligonucleotides that are synthesized *in situ*. The early versions of microarrays were designed to detect RNAs from expressed genomic regions (open reading frames). Although such arrays are perfectly suited to study gene expression profiles, they have limited importance in ChIP experiments since most "interesting" proteins with respect to this technique bind in intergenic regions. Nowadays, even custom-made arrays can be designed and fine-tuned to match the requirements of an experiment. Also, any sequence of nucleotides can be synthesized to cover genic as well as intergenic regions.
- *Probe size*: Early version of cDNA arrays had a probe length of about 200bp. Latest array versions use oligos as short as 70- (Microarrays, Inc.) to 25-mers (Affymetrix). (Feb 2007)
- *Probe composition*: There are tiled and non-tiled DNA arrays. Non-tiled arrays use probes selected according to non-spatial criteria, i.e., the DNA sequences used as probes have no fixed distances in the genome. Tiled arrays, however, select a genomic region (or even a whole genome) and divide it into equal chunks. Such a region is called tiled path. The average distance between each pair of neighboring chunks (measured from the center of each chunk) gives the resolution of the tiled path. A path can be overlapping, end-to-end or spaced .
- *Array size*: The first microarrays used for ChIP-on-Chip contained about 13,000 spotted DNA segments representing all ORFs and intergenic regions from the yeast genome. Nowadays, Affymetrix offers whole-genome tiled yeast arrays with a resolution of 5bp (all in all 3.2 million probes). Tiled arrays for the human genome become more and more powerful, too. Just to name example, Affymetrix offers a set of seven arrays with about 90 million probes, spanning the complete non-repetitive part of the human genome with about 35bp spacing. (Feb 2007)

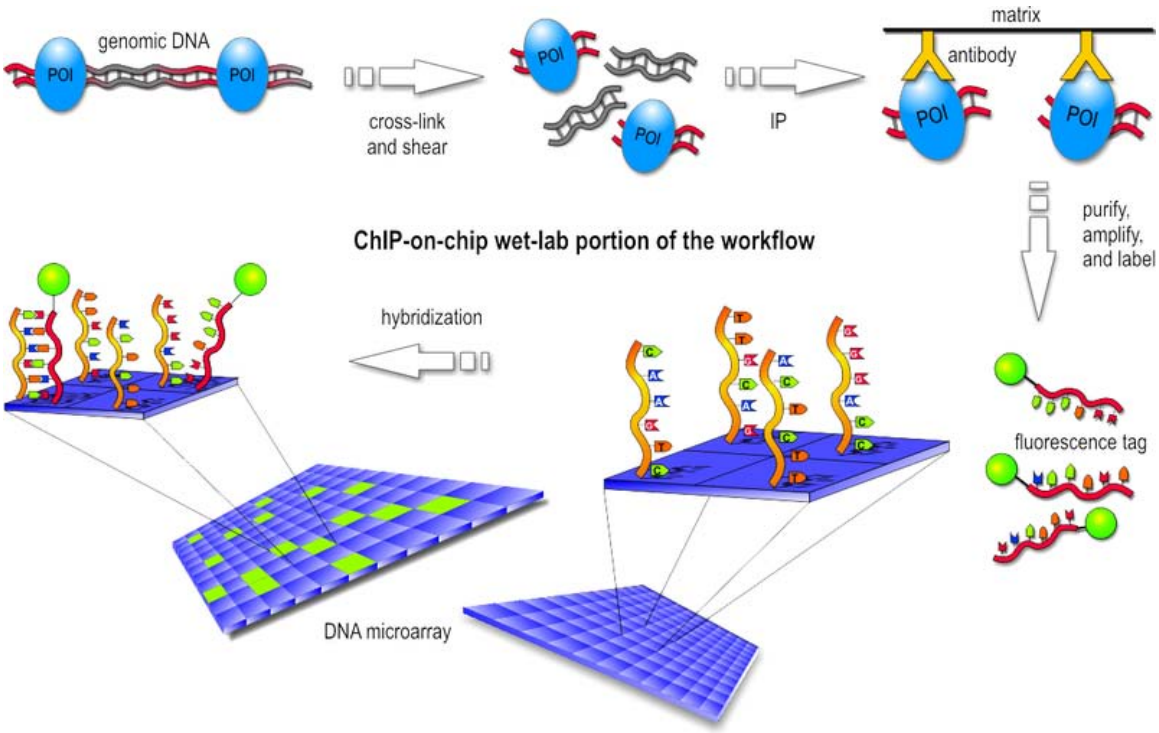
Besides the actual microarray, other hard- and software equipment is necessary to run ChIP-on-chip experiments. It is generally the case that one company's microarrays can

not be analyzed by another company's processing hardware. Hence, buying an array requires also buying the associated workflow equipment. The most important elements are, among others, hybridization ovens, chip scanners, and software packages for subsequent numerical analysis of the raw data.

Workflow of a ChIP-on-chip experiment

Starting with a biological question, a ChIP-on-chip experiment can be divided into three major steps: The first is to set up and design the experiment by selecting the appropriate array and probe type. Second, the actual experiment is performed in the wet-lab. Last, during the dry-lab portion of the cycle, gathered data are analyzed to either answer the initial question or lead to new questions so that the cycle can start again.

Wet-lab portion of the workflow

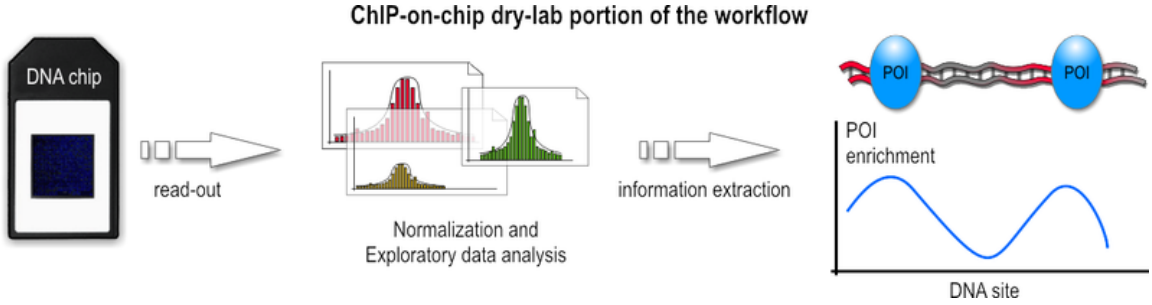


Workflow overview of the wet-lab portion of a ChIP-on-chip experiment.

- In the first step, the protein of interest (POI) is cross-linked with the DNA site it binds to in an *in vivo* environment. Usually this is done by a gentle formaldehyde fixation that is reversible with heat.
- Then, the cells are lysed and the DNA is sheared by sonication or using micrococcal nuclease. This results in double-stranded chunks of DNA fragments, normally 1 kb or less in length. Those that were cross-linked to the POI form a POI-DNA complex.

- In the next step, only these complexes are filtered out of the set of DNA fragments, using an antibody specific to the POI. The antibodies may be attached to a solid surface, may have a magnetic bead, or some other physical property that allow distributing cross-linked complexes and unbound fragments. This procedure is essentially an immunoprecipitation (IP). There are two alternative ways to implement this filtering step:
    - immunoprecipitation of the tagged protein with an antibody against the tag (ex. FLAG, HA, c-myc)
    - affinity purification that does not require antibodies, such as the Tandem Affinity Purification (TAP)
- The cross-linking of POI-DNA complexes is reversed (usually by heating) and the DNA strands are purified. For the rest of the workflow, the POI is no longer necessary.
- After an amplification and denaturation step, the single-stranded DNA fragments are labeled with a fluorescent tag such as Cy5 or Alexa 647.
- Finally, the fragments are poured over the surface of the DNA microarray, which is spotted with short, single-stranded sequences that cover the genomic portion of interest. Whenever a labeled fragment "finds" a complementary fragment on the array, they will hybridize and form again a double-stranded DNA fragment.

Dry-lab portion of the workflow



Workflow overview of the dry-lab portion of a ChIP-on-chip experiment.

- After a sufficiently large time frame to allow hybridization, the array is illuminated with fluorescence light. Those probes on the array that are hybridized to one of the labeled fragments emit a light signal that is captured by a camera. This image contains all raw data for the remaining part of the workflow.
- This raw data, encoded as false-color image, needs to be converted to numerical values before the actual analysis can be done. The analysis and information extraction of the raw data often remains the most challenging part for ChIP-on-chip experiments. Problems arise throughout this portion of the workflow, ranging from the initial chip read-out, to suitable methods to subtract background noise, and finally to appropriate algorithms that normalize the data and make it available for subsequent statistical analysis, which then hopefully lead to a better understanding of the biological question sought to answer. Furthermore, due to the different array platforms and missing standardization between them, data

storage and exchange is a huge problem, too. Generally speaking, the data analysis can be divided into three major steps:

1. During the first step, the captured fluorescence signals from the array are normalized, using control signals derived from the same or a second chip. Such control signals tell which probes on the array were hybridized correctly and which bound nonspecifically.
2. In the second step, numerical and statistical tests are applied to control data and IP fraction data to identify POI-enriched regions along the genome. The following three methods are used widely: Median percentile rank, Single-array error, and Sliding-window. These methods generally differ in a way how low-intensity signals are handled, how much background noise is accepted, and which trait for the data is emphasized during the computation. In the recent past, the sliding-window approach seems to be favored and is often described as most powerful.
3. In the third step, these regions are analyzed further. If, for example, the POI was a transcription factor, such regions would represent its binding sites. Subsequent analysis then may want to infer nucleotide motifs and other patterns to allow functional annotation of the genome.

Strengths and Weaknesses

Using tiled arrays, ChIP-on-chip allows for high resolution of genome-wide maps. These maps can determine the binding sites of many DNA-binding proteins like transcription factors and also chromatin modifications.

Although ChIP-on-chip can be a powerful technique in the area of genomics, it is very expensive. Most published studies using ChIP-on-chip repeat their experiments at least three times to ensure biologically meaningful maps. The cost of the DNA microarrays is often a limiting factor to whether a laboratory should proceed with a ChIP-on-chip experiment. Another limitation is the size of DNA fragments that can be achieved. Most ChIP-on-chip protocols utilize sonication as a method of breaking up DNA into small pieces. However, sonication is limited to a minimal fragment size of 200 bp. For higher resolution maps, this limitation should be overcome to achieve smaller fragments, preferably to single nucleosome resolution. As mentioned previously, the statistical analysis of the huge amount of data generated from arrays is a challenge and normalization procedures should aim to minimize artifacts and determine what is really biologically significant. So far, application to mammalian genomes has been a major limitation, for example, due to a significant percentage of the genome that is occupied by repeats. However, as ChIP-on-chip technology advances, high resolution whole mammalian genome maps should become achievable.

Antibodies used for ChIP-on-chip can be an important limiting factor. ChIP-on-chip requires highly specific antibodies that must recognize its epitope in free solution and also under fixed conditions. If it is demonstrated to successfully immunoprecipitate cross-linked chromatin, it is termed "ChIP-grade". Companies that provide ChIP-grade antibodies include Abcam, Cell Signaling Technology, Santa Cruz, and Upstate. To

overcome the problem of specificity, the protein of interest can be fused to a tag like FLAG or HA that are recognized by antibodies. An alternative to ChIP-on-chip that does not require antibodies is DamID.

Also available are antibodies against a specific histone modification like H3 tri methyl K4. As mentioned before, the combination of these antibodies and ChIP-on-chip has become extremely powerful in determining whole genome analysis of histone modification patterns and will contribute tremendously to our understanding of the histone code and epigenetics.

A study demonstrating the non-specific nature of DNA binding proteins has been published in PLoS Biology. This indicates that alternate confirmation of functional relevancy is a necessary step in any ChIP-chip experiment.

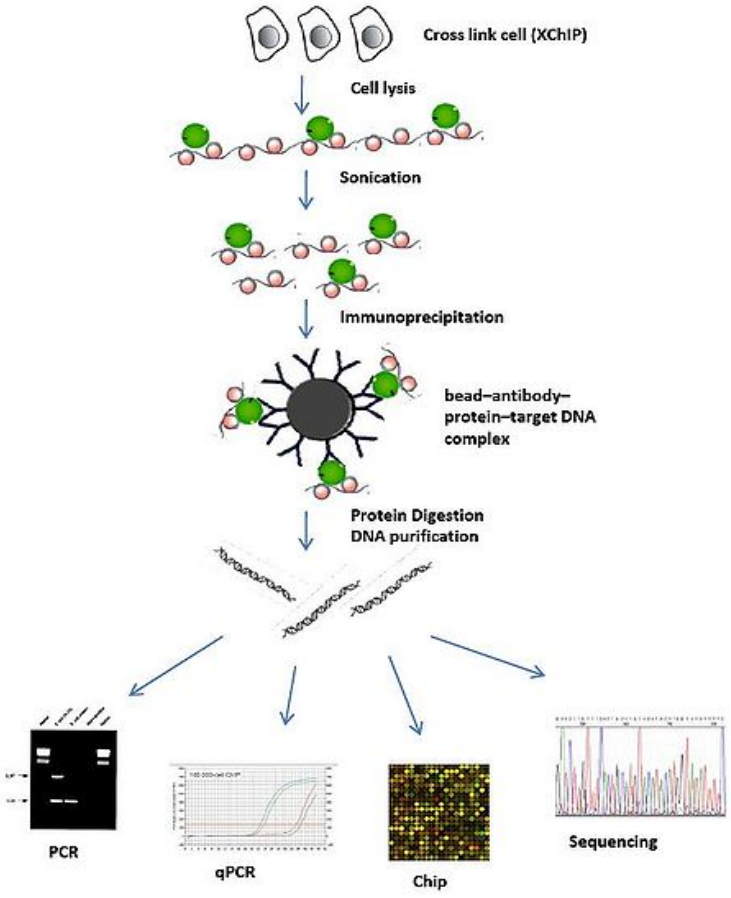
History

The ChIP-on-chip technique was first applied successfully in three papers published in 2000 and 2001 . The authors identified binding sites for individual transcription factors in the budding yeast *Saccharomyces cerevisiae*. In 2002, Richard Young's group determined the genome-wide positions of 106 transcription factors using a c-Myc tagging system in yeast. Other applications for ChIP-on-chip include DNA replication, recombination, and chromatin structure. Since then, ChIP-on-chip has become a powerful tool in determining genome-wide maps of histone modifications and many more transcription factors. ChIP-on-chip in mammalian systems has been difficult due to the large and repetitive genomes. Thus, many studies in mammalian cells have focused on select promoter regions that are predicted to bind transcription factors and have not analyzed the entire genome. However, whole mammalian genome arrays have recently become commercially available from companies like Nimblegen. In the future, as ChIP-on-chip arrays become more and more advanced, high resolution whole genome maps of DNA-binding proteins and chromatin components for mammals will be analyzed in more detail.

Chapter- 4

Chromatin Immunoprecipitation and Chip-Sequencing

Chromatin immunoprecipitation



The procedure of chromatin immunoprecipitation (ChIP) assay and methods of analysis

Chromatin Immunoprecipitation (ChIP) is a type of immunoprecipitation experimental technique used to investigate the interaction between proteins and DNA in the cell. It aims to determine whether specific proteins are associated with specific genomic regions, such as transcription factors on promoters or other DNA binding sites, and possibly defining cistromes. ChIP also aims to determine the specific location in the genome that various histone modifications are associated with, indicating the target of the histone modifiers.

Briefly, the method is as follows: protein and associated chromatin in a cell lysate are temporarily bonded, the DNA-protein complexes (chromatin-protein) are then sheared and DNA fragments associated with the protein(s) of interest are selectively immunoprecipitated, then the associated DNA fragments are purified and their sequence is determined. These DNA sequences are supposed to be associated with the protein of interest *in vivo*.

Typical ChIP

Technically, there are mainly two types of ChIP, primarily differing in the starting chromatin preparation. The first uses reversibly cross-linked chromatin sheared by sonication called cross-linked ChIP (XChIP). Native ChIP (NChIP) uses native chromatin sheared by micrococcal nuclease digestion,

Cross-linked ChIP (XChIP)

Cross-linked ChIP is mainly suited for mapping DNA target of transcription factors or other chromatin-associated proteins, by using reversibly cross-linked chromatin as starting material. The agent for reversible cross-link could be formaldehyde or UV light. Then the cross-linked chromatin are usually sheared by sonication, providing fragments of 300-1000 base pairs (bp) in length. Mild formaldehyde crosslinking followed by nuclease digestion has been used to shear the chromatin. Chromatin fragments of 400-500bp have proven to be suitable for ChIP assays as they cover two to three nucleosomes.

Cell debris in the sheared lysate is then cleared by sedimentation and protein–DNA complexes are selectively immunoprecipitated using specific antibodies to the protein(s) of interest. The antibodies are commonly coupled to agarose, sepharose or magnetic beads. The immunoprecipitated complexes (i.e., the bead–antibody–protein–target DNA sequence complex) are then collected and washed to remove non-specifically bound chromatin, the protein-DNA cross-link is reversed and proteins are removed by digestion with proteinase K.

The DNA associated with the complex is then purified and identified by polymerase chain reaction (PCR), microarrays (ChIP-on-chip), molecular cloning and sequencing, or direct high-throughput sequencing (ChIP-seq).

Native ChIP (NChIP)

Native ChIP is mainly suited for mapping the DNA target of histone modifiers. Generally, native chromatin is used as starting chromatin. As histones wrap around DNA to form nucleosomes, they are naturally linked. Then the chromatin is sheared by micrococcal nuclease digestion, which cuts DNA at the length of the linker, leaving nucleosomes intact and providing DNA fragments of one nucleosome (200bp) to five nucleosomes (1000bp) in length.

Thereafter, methods similar to XChIP are used for clearing the cell debris, immunoprecipitating the protein of interest, removing protein from the immunoprecipated complex, and purifying and analyzing the complex-associated DNA.

Comparison of XChIP and NChIP

The major advantage for NChIP is antibody specificity. It is important to note that most antibodies to modified histones are raised against unfixed, synthetic peptide antigens and that the epitopes they need to recognize in the XChIP may be disrupted or destroyed by formaldehyde cross-linking, particularly as the cross-links are likely to involve lysine e-amino groups in the N-terminals, disrupting the epitopes. This is likely to explain the consistently low efficiency of XChIP protocols compare to NChIP.

But XChIP and NChIP have different aims and advantage against each other, XChIP is for mapping target site of transcription factors and other chromatin associated proteins, NChIP is for mapping the target site of histone modifiers (see Table 1).

Table 1 Advantages and disadvantages of NChIP and XChIP

	XChIP	NChIP
Advantages	Suitable for transcriptional factors, or any other weakly binding chromatin associated proteins. Applicable to any organisms where native protein is hard to prepare	Testable antibody specificity Better antibody specificity as target protein naturally intact Better chromatin and protein recovery efficiency due to Better antibody specificity
Disadvantages	Inefficient chromatin recovery due to antibody target protein epitope disruption May cause false positive result due to fixation of transient proteins to chromatin Wide range of chromatin shearing size due to random cut by sonication.	Usually not suitable for non-histone proteins Nucleosomes may rearrange during digestion

History and New ChIP methods

XChIP was pioneered by Alexander Varshavsky and co-workers in the 1980s, and has been extensively developed and refined. NChIP approach was first described by Hebbes *et al*., 1988, and also been developed and refined quickly. The typical ChIP assay usually take 4–5 days, and require 10^6~ 10^7 cells at least. Now new techniques on ChIP could be achieved as few as 100~1000 cells and complete within one day.

- **Carrier ChIP** (CChIP): This approach could use as few as 100 cells by adding *Drosophila* cells as carrier chromatin to reduce loss and facilitate precipitation of the target chromatin. However, it demands highly specific primers for detection of the target cell chromatin from the foreign carrier chromatin background, and it takes two to three days.

- **Fast ChIP** (qChIP): The fast ChIP assay reduced the time by shortening two steps in a typical ChIP assay: *(i)* an ultrasonic bath accelerates the rate of antibody binding to target proteins—and thereby reduces immunoprecipitation time *(ii)* a resin-based (Chelex-100) DNA isolation procedure reduces the time of cross-link reversal and DNA isolation. However, the fast protocol is suitable only for large cell samples (in the range of 10^6~10^7). Up to 24 sheared chromatin samples can be processed to yield PCR-ready DNA in 5 hours, allowing multiple chromatin factors be probed simultaneously and/or looking at genomic events over several time points.

- **Quick and quantitative ChIP** (Q^2ChIP) : The assay uses 100,000 cells as starting material and is suitable for up to 1,000 histone ChIPs or 100 transcription factor ChIPs. Thus many chromatin samples can be prepared in parallel and stored, and Q^2ChIP can be undertaken in a day.

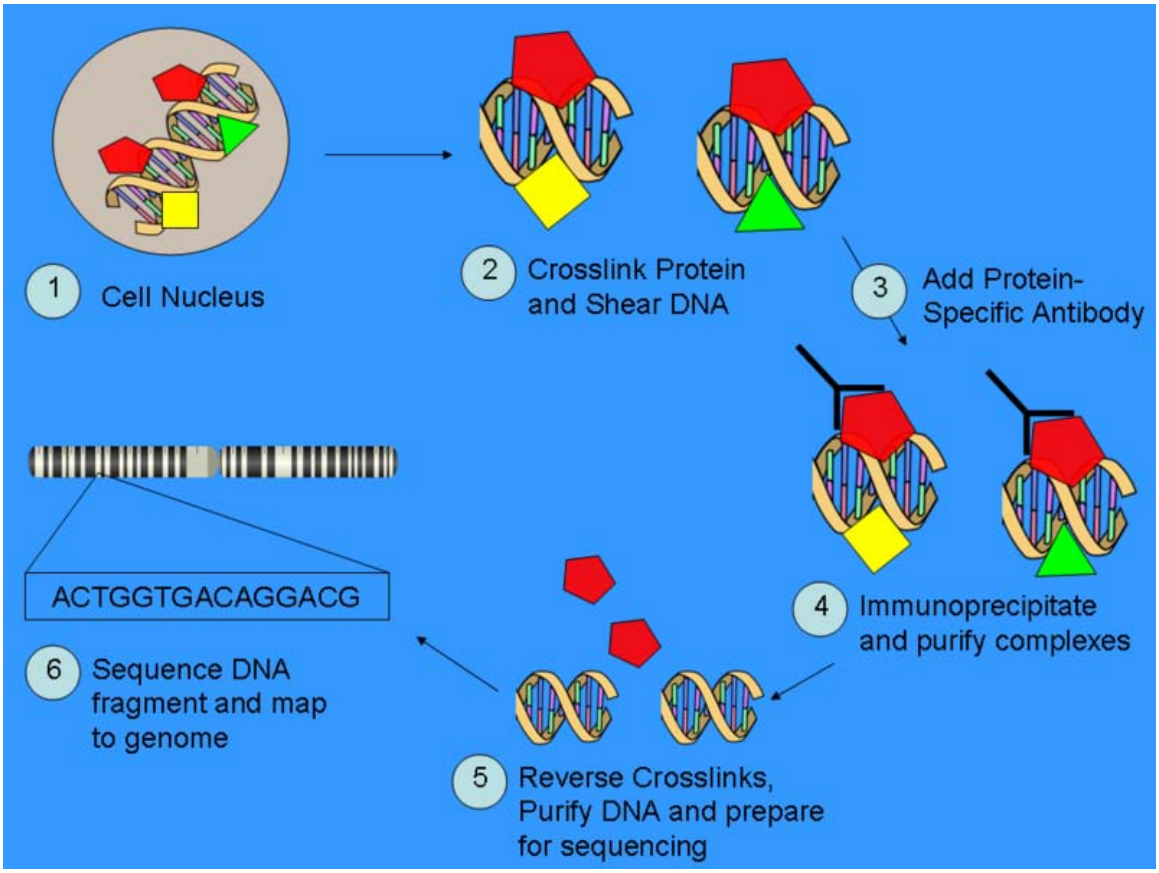
- **MicroChIP** (μChIP): chromatin is usually prepared from 1,000 cells and up to 8 ChIPs can be done in parallel without carriers. The assay can also start with 100 cells, but only suit for one ChIP. It can also use small (1 mm^3) tissue biopsies and microChIP can be done within one day.

- **Matrix ChIP**: This is a microplate-based ChIP assay with increased throughput and simplified the procedure. All steps are done in microplate wells without sample transfers, enabling a potential for automation. It enables 96 ChIP assays for histone and various DNA-bound proteins in a single day.

ChIP could also been applied for genome wide analysis when combined with microarray technology (ChIP-on-chip) and second generation DNA-sequencing technology (Chip-Sequencing). ChIP can also combine with paired-end tags sequencing in Chromatin Interaction Analysis using Paired End Tag sequencing (ChIA-PET), a technique developed for large-scale, de novo analysis of higher-order chromatin structures.

Chip-Sequencing

ChIP-Sequencing, also known as **ChIP-Seq**, is used to analyze protein interactions with DNA. ChIP-Seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the cistrome of DNA-associated proteins. It can be used to precisely map global binding sites for any protein of interest. Previously, ChIP-on-chip was the most common technique utilized to study these protein-DNA relations.



ChIP-Sequencing Workflow

Uses of ChIP-Seq

Chip-Seq is used primarily to determine how transcription factors and other chromatin-associated proteins influence phenotype-affecting mechanisms. Determining how proteins interact with DNA to regulate gene expression is essential for fully understanding many biological processes and disease states. This epigenetic information is complementary to genotype and expression analysis. ChIP-Seq technology is currently seen primarily as an alternative to ChIP-chip which requires a hybridization array. This necessarily introduces some bias, as an array is restricted to a fixed number of probes.

Sequencing, by contrast, is thought to have less bias, although the sequencing bias of different sequencing technologies is not yet fully understood.

Specific DNA sites in direct physical interaction with transcription factors and other proteins can be isolated by chromatin immunoprecipitation. ChIP produces a library of target DNA sites bound to a target *in vivo*. Massively parallel sequence analyses are used in conjunction with whole-genome sequence databases to analyze the interaction pattern of any protein with DNA, or the pattern of any epigenetic chromatin modifications. This can be applied to the set of ChIP-able proteins and modifications, such as transcription factors, polymerases and transcriptional machinery, structural proteins, protein modifications, and DNA modifications.

Workflow of ChIP-Sequencing

Part 1: ChIP

ChIP is a powerful method to selectively enrich for DNA sequences bound by a particular protein in living cells. However, the widespread use of this method has been limited by the lack of a sufficiently robust method to identify all of the enriched DNA sequences. The ChIP process enriches specific crosslinked DNA-protein complexes using an antibody against a protein of interest. Oligonucleotide adapters are then added to the small stretches of DNA that were bound to the protein of interest to enable massively parallel sequencing.

Part 2: Sequencing

After size selection, all the resulting ChIP-DNA fragments are sequenced simultaneously using a genome sequencer. A single sequencing run can scan for genome-wide associations with high resolution, meaning that features can be located precisely on the chromosomes. ChIP-chip, by contrast, requires large sets of tiling arrays for lower resolution.

There are many new sequencing methods used in this sequencing step. Some technologies that analyze the sequences can use cluster amplification of adapter-ligated ChIP DNA fragments on a solid flow cell substrate to create clusters of approximately 1000 clonal copies each. The resulting high density array of template clusters on the flow cell surface is sequenced by a Genome analyzing program. Each template cluster undergoes sequencing-by-synthesis in parallel using novel fluorescently labelled reversible terminator nucleotides. Templates are sequenced base-by-base during each read. Then, the data collection and analysis software aligns sample sequences to a known genomic sequence to identify the ChIP-DNA fragments.

Sensitivity

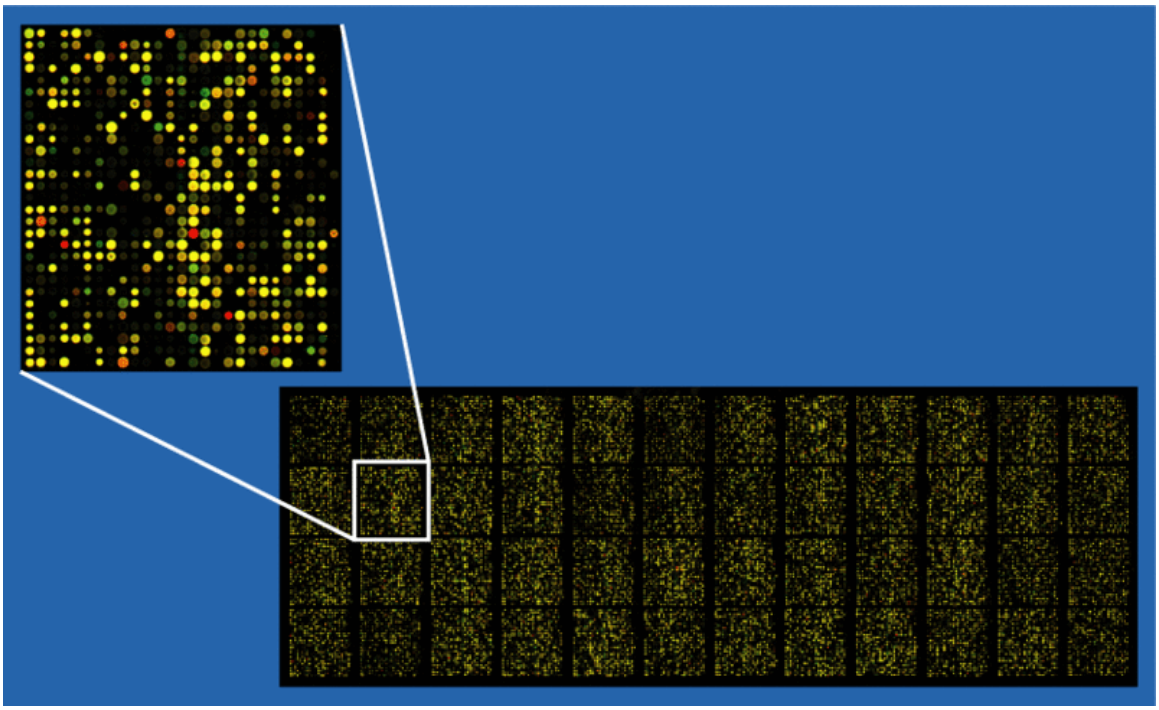
Sensitivity of this technology depends on the depth of the sequencing run (i.e. the number of mapped sequence tags), the size of the genome and the distribution of the target factor.

The sequencing depth is directly correlated with cost. If abundant binders in large genomes have to be mapped with high sensitivity, costs are high as an enormously high number of sequence tags will be required. This is in contrast to ChIP-chip in which the costs are not correlated with sensitivity.

Unlike microarray-based ChIP methods, the precision of the ChIP-Seq assay is not limited by the spacing of predetermined probes. By integrating a large number of short reads, highly precise binding site localization is obtained. Compared to ChIP-chip, ChIP-Seq data can be used to locate the binding site within few tens of base pairs of the actual protein binding site. Tag densities at the binding sites are a good indicator of protein–DNA binding affinity, which makes it easier to quantify and compare binding affinities of a protein to different DNA sites.

Chapter- 5

DNA Microarray



Example of an approximately 40,000 probe spotted oligo microarray with enlarged inset to show detail.

A **DNA microarray** is a multiplex technology used in molecular biology. It consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides, called features, each containing picomoles (10^{-12} moles) of a specific DNA sequence, known as *probes* (or *reporters*). These can be a short section of a gene or other DNA element that are used to hybridize a cDNA or cRNA sample (called *target*) under high-stringency conditions. Probe-target hybridization is usually detected and quantified by detection of fluorophore-, silver-, or chemiluminescence-labeled targets to determine relative abundance of nucleic acid sequences in the target. Since an array can contain tens of thousands of probes, a microarray experiment can accomplish many genetic tests in parallel. Therefore arrays have dramatically accelerated many types of investigation.

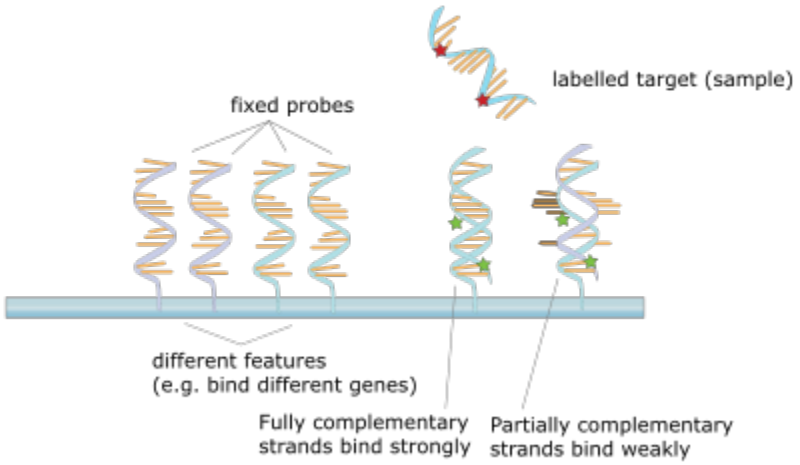
In standard microarrays, the probes are attached via surface engineering to a solid surface by a covalent bond to a chemical matrix (via epoxy-silane, amino-silane, lysine, polyacrylamide or others). The solid surface can be glass or a silicon chip, in which case they are colloquially known as an *Affy chip* when an Affymetrix chip is used. Other microarray platforms, such as Illumina, use microscopic beads, instead of the large solid support. DNA arrays are different from other types of microarray only in that they either measure DNA or use DNA as part of its detection system.

DNA microarrays can be used to measure changes in expression levels, to detect single nucleotide polymorphisms (SNPs), or to genotype or resequence mutant genomes. Microarrays also differ in fabrication, workings, accuracy, efficiency, and cost. Additional factors for microarray experiments are the experimental design and the methods of analyzing the data.

History

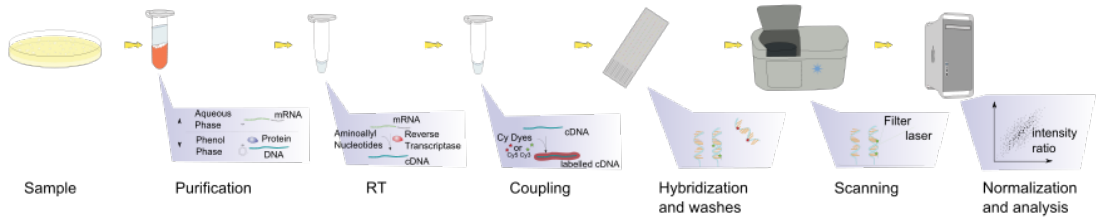
Microarray technology evolved from Southern blotting, where fragmented DNA is attached to a substrate and then probed with a known gene or fragment. Nucleic Acids Res. 1992 Apr 11;20(7):1679-84. Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ. Maskos U, Southern EM. The first reported use of this approach was the analysis of 378 arrayed lysed bacterial colonies each harboring a different sequence which were assayed in multiple replicas for expression of the genes in multiple normal and tumor tissue (Augenlicht and Kobrin, Cancer Research, 42, 1088–1093, 1982). This was expanded to analysis of more than 4000 human sequences with computer driven scanning and image processing for quantitative analysis of the sequences in human colonic tumors and normal tissue (Augenlicht *et al.*, Cancer Research, 47, 6017-6021, 1987) and then to comparison of colonic tissues at different genetic risk (Augenlicht *et al.*, Proceedings National Academy of Sciences, USA, 88, 3286-3289, 1991). The use of a collection of distinct DNAs in arrays for expression profiling was also described in 1987, and the arrayed DNAs were used to identify genes whose expression is modulated by interferon. These early gene arrays were made by spotting cDNAs onto filter paper with a pin-spotting device. The use of miniaturized microarrays for gene expression profiling was first reported in 1995, and a complete eukaryotic genome (*Saccharomyces cerevisiae*) on a microarray was published in 1997.

Principle



Hybridization of the target to the probe

The core principle behind microarrays is hybridization between two DNA strands, the property of complementary nucleic acid sequences to specifically pair with each other by forming hydrogen bonds between complementary nucleotide base pairs. A high number of complementary base pairs in a nucleotide sequence means tighter non-covalent bonding between the two strands. After washing off of non-specific bonding sequences, only strongly paired strands will remain hybridized. So fluorescently labeled target sequences that bind to a probe sequence generate a signal that depends on the strength of the hybridization determined by the number of paired bases, the hybridization conditions (such as temperature), and washing after hybridization. Total strength of the signal, from a spot (feature), depends upon the amount of target sample binding to the probes present on that spot. Microarrays use relative quantization in which the intensity of a feature is compared to the intensity of the same feature under a different condition, and the identity of the feature is known by its position. An alternative to microarrays is serial analysis of gene expression, where the transcriptome is sequenced allowing an absolute measurement.



The step required in a microarray experiment

Uses and types

Two Affymetrix chips

Many types of array exist and the broadest distinction is whether they are spatially arranged on a surface or on coded beads:

- The traditional solid-phase array is a collection of orderly microscopic "spots", called features, each with a specific probe attached to a solid surface, such as glass, plastic or silicon biochip (commonly known as a *genome chip*, *DNA chip* or *gene array*). Thousands of them can be placed in known locations on a single DNA microarray.
- The alternative bead array is a collection of microscopic polystyrene beads, each with a specific probe and a ratio of two or more dyes, which do not interfere with the fluorescent dyes used on the target sequence.

DNA microarrays can be used to detect DNA (as in comparative genomic hybridization), or detect RNA (most commonly as cDNA after reverse transcription) that may or may not be translated into proteins. The process of measuring gene expression via cDNA is called expression analysis or expression profiling.
Applications include:

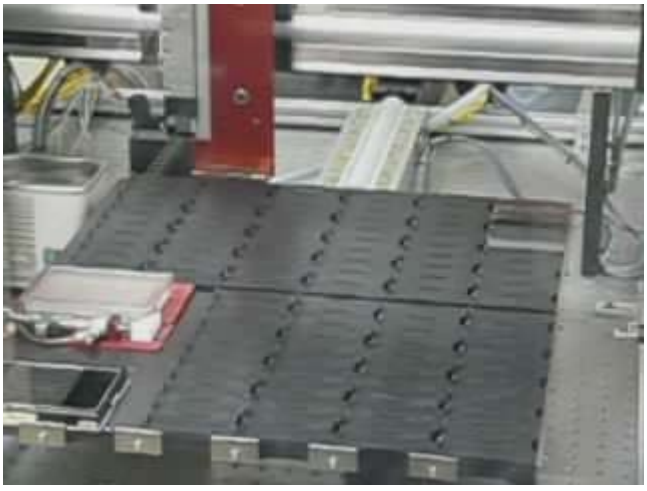
Application or technology	Synopsis
Gene expression profiling	In an mRNA or gene expression profiling experiment the expression levels of thousands of genes are simultaneously monitored to study the effects of certain treatments, diseases, and developmental stages on gene expression. For example, microarray-based gene expression profiling can be used to identify genes whose expression is changed in response to pathogens or other organisms by comparing gene expression in infected to that in uninfected cells or tissues.
Comparative genomic hybridization	Assessing genome content in different cells or closely related organisms.
GeneID	Small microarrays to check IDs of organisms in food and feed (like GMO), mycoplasms in cell culture, or pathogens for disease detection, mostly combining PCR and microarray technology.
Chromatin immunoprecipitation on Chip	DNA sequences bound to a particular protein can be isolated by immunoprecipitating that protein (ChIP), these fragments can be then hybridized to a microarray (such as a tiling array) allowing the determination of protein binding site occupancy throughout the genome. Example protein to immunoprecipitate are histone modifications (H3K27me3, H3K4me2, H3K9me3, etc.), Polycomb-group protein (PRC2:Suz12, PRC1:YY1) and trithorax-group protein (Ash1) to study the epigenetic landscape or RNA Polymerase II to study the transcription landscape.
DamID	Analogously to ChIP, genomic regions bound by a protein of interest can be isolated and used to probe a microarray to determine binding site occupancy. Unlike ChIP, DamID does not require antibodies but makes use of adenine methylation near the protein's binding sites to selectively amplify those regions, introduced by expressing minute amounts of protein of interest fused to bacterial DNA adenine methyltransferase.
SNP detection	Identifying single nucleotide polymorphism among alleles within or between populations. Several applications of microarrays make use of SNP detection, including Genotyping, forensic analysis, measuring predisposition to disease, identifying drug-candidates, evaluating germline mutations in individuals or somatic mutations in cancers, assessing loss of heterozygosity, or genetic linkage analysis.

Alternative splicing detection	An '*exon junction array* design uses probes specific to the expected or potential splice sites of predicted exons for a gene. It is of intermediate density, or coverage, to a typical gene expression array (with 1-3 probes per gene) and a genomic tiling array (with hundreds or thousands of probes per gene). It is used to assay the expression of alternative splice forms of a gene. Exon arrays have a different design, employing probes designed to detect each individual exon for known or predicted genes, and can be used for detecting different splicing isoforms.
Fusion genes microarray	A Fusion gene microarray can detect fusion transcripts, *e.g.* from cancer specimens. The principle behind this is building on the alternative splicing microarrays. The oligo design strategy enables combined measurements of chimeric transcript junctions with exon-wise measurements of individual fusion partners.
Tiling array	Genome tiling arrays consist of overlapping probes designed to densely represent a genomic region of interest, sometimes as large as an entire human chromosome. The purpose is to empirically detect expression of transcripts or alternatively splice forms which may not have been previously known or predicted.

Fabrication

Microarrays can be manufactured in different ways, depending on the number of probes under examination, costs, customization requirements, and the type of scientific question being asked. Arrays may have as few as 10 probes or up to 2.1 million micrometre-scale probes from commercial vendors.

Spotted vs. in situ synthesised arrays



A DNA microarray being printed by a robot at the University of Delaware

Microarrays can be fabricated using a variety of technologies, including printing with fine-pointed pins onto glass slides, photolithography using pre-made masks, photolithography using dynamic micromirror devices, ink-jet printing, or electrochemistry on microelectrode arrays.

In *spotted microarrays*, the probes are oligonucleotides, cDNA or small fragments of PCR products that correspond to mRNAs. The probes are synthesized prior to deposition on the array surface and are then "spotted" onto glass. A common approach utilizes an array of fine pins or needles controlled by a robotic arm that is dipped into wells containing DNA probes and then depositing each probe at designated locations on the array surface. The resulting "grid" of probes represents the nucleic acid profiles of the prepared probes and is ready to receive complementary cDNA or cRNA "targets" derived from experimental or clinical samples. This technique is used by research scientists around the world to produce "in-house" printed microarrays from their own labs. These arrays may be easily customized for each experiment, because researchers can choose the probes and printing locations on the arrays, synthesize the probes in their own lab (or collaborating facility), and spot the arrays. They can then generate their own labeled samples for hybridization, hybridize the samples to the array, and finally scan the arrays with their own equipment. This provides a relatively low-cost microarray that may be customized for each study, and avoids the costs of purchasing often more expensive commercial arrays that may represent vast numbers of genes that are not of interest to the investigator. Publications exist which indicate in-house spotted microarrays may not provide the same level of sensitivity compared to commercial oligonucleotide arrays, possibly owing to the small batch sizes and reduced printing efficiencies when compared to industrial manufactures of oligo arrays.

In *oligonucleotide microarrays*, the probes are short sequences designed to match parts of the sequence of known or predicted open reading frames. Although oligonucleotide probes are often used in "spotted" microarrays, the term "oligonucleotide array" most often refers to a specific technique of manufacturing. Oligonucleotide arrays are produced by printing short oligonucleotide sequences designed to represent a single gene or family of gene splice-variants by synthesizing this sequence directly onto the array surface instead of depositing intact sequences. Sequences may be longer (60-mer probes such as the Agilent design) or shorter (25-mer probes produced by Affymetrix) depending on the desired purpose; longer probes are more specific to individual target genes, shorter probes may be spotted in higher density across the array and are cheaper to manufacture. One technique used to produce oligonucleotide arrays include photolithographic synthesis (Agilent and Affymetrix) on a silica substrate where light and light-sensitive masking agents are used to "build" a sequence one nucleotide at a time across the entire array. Each applicable probe is selectively "unmasked" prior to bathing the array in a solution of a single nucleotide, then a masking reaction takes place and the next set of probes are unmasked in preparation for a different nucleotide exposure. After many repetitions, the sequences of every probe become fully constructed. More recently, Maskless Array Synthesis from NimbleGen Systems has combined flexibility with large numbers of probes.

Two-channel vs. one-channel detection

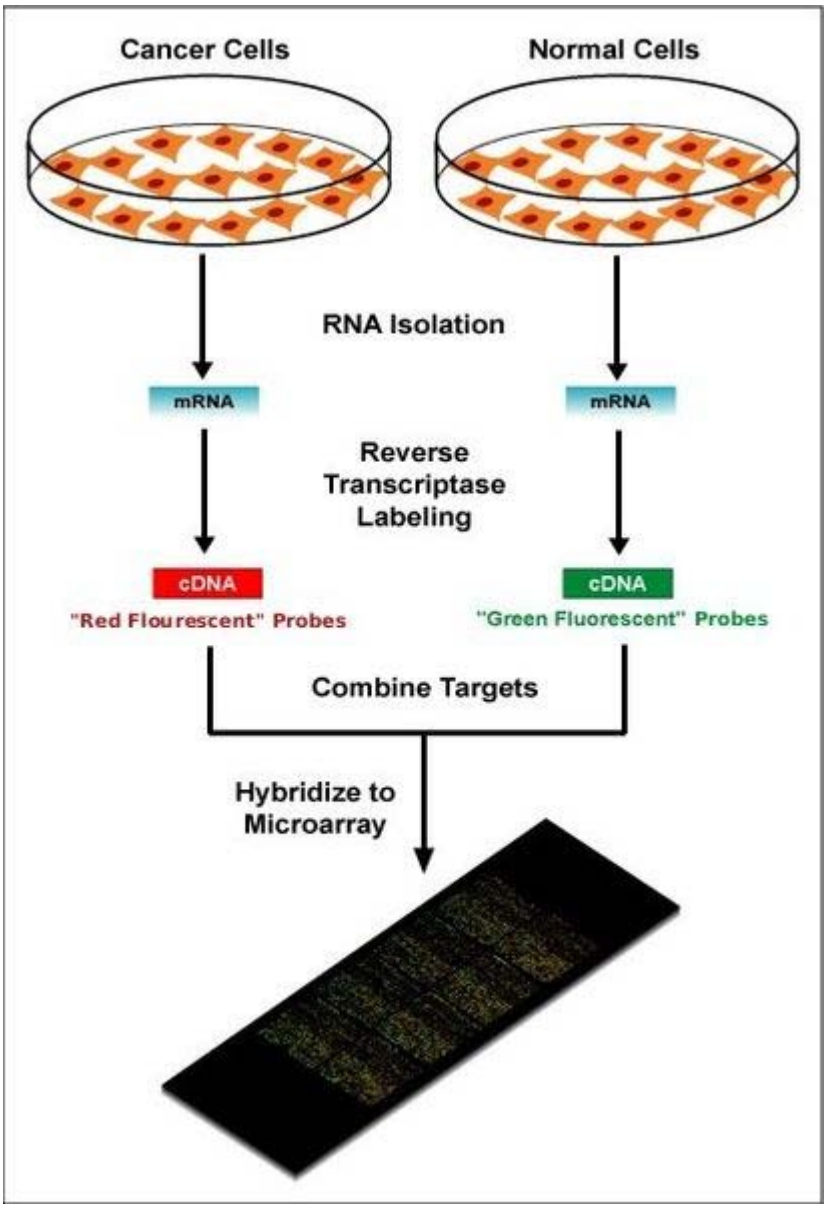


Diagram of typical dual-colour microarray experiment.

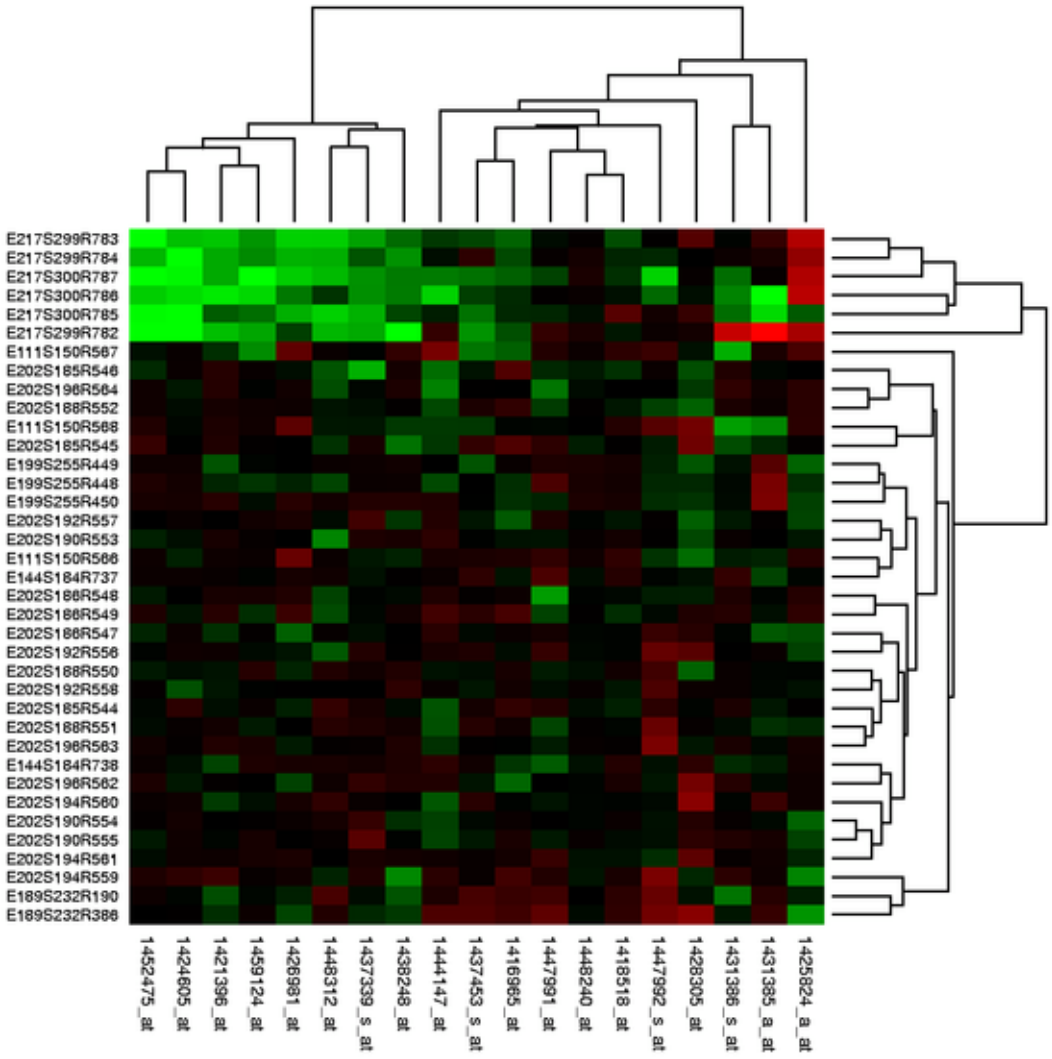
Two-color microarrays or *two-channel microarrays* are typically hybridized with cDNA prepared from two samples to be compared (e.g. diseased tissue versus healthy tissue) and that are labeled with two different fluorophores. Fluorescent dyes commonly used for cDNA labeling include Cy3, which has a fluorescence emission wavelength of 570 nm (corresponding to the green part of the light spectrum), and Cy5 with a fluorescence emission wavelength of 670 nm (corresponding to the red part of the light spectrum). The two Cy-labeled cDNA samples are mixed and hybridized to a single microarray that is then scanned in a microarray scanner to visualize fluorescence of the two fluorophores after excitation with a laser beam of a defined wavelength. Relative intensities of each

fluorophore may then be used in ratio-based analysis to identify up-regulated and down-regulated genes.

Oligonucleotide microarrays often carry control probes designed to hybridize with RNA spike-ins. The degree of hybridization between the spike-ins and the control probes is used to normalize the hybridization measurements for the target probes. Although absolute levels of gene expression may be determined in the two-color array in rare instances, the relative differences in expression among different spots within a sample and between samples is the preferred method of data analysis for the two-color system. Examples of providers for such microarrays includes Agilent with their Dual-Mode platform, Eppendorf with their DualChip platform for colorimetric Silverquant labeling, and TeleChem International with Arrayit.

In *single-channel microarrays* or *one-color microarrays*, the arrays provide intensity data for each probe or probe set indicating a relative level of hybridization with the labeled target. However, they do not truly indicate abundance levels of a gene but rather relative abundance when compared to other samples or conditions when processed in the same experiment. Each RNA molecule encounters protocol and batch-specific bias during amplification, labeling, and hybridization phases of the experiment making comparisons between genes for the same microarray uninformative. The comparison of two conditions for the same gene requires two separate single-dye hybridizations. Several popular single-channel systems are the Affymetrix "Gene Chip", Illumina "Bead Chip", Agilent single-channel arrays, the Applied Microarrays "CodeLink" arrays, and the Eppendorf "DualChip & Silverquant". One strength of the single-dye system lies in the fact that an aberrant sample cannot affect the raw data derived from other samples, because each array chip is exposed to only one sample (as opposed to a two-color system in which a single low-quality sample may drastically impinge on overall data precision even if the other sample was of high quality). Another benefit is that data are more easily compared to arrays from different experiments so long as batch effects have been accounted for. A drawback to the one-color system is that, when compared to the two-color system, twice as many microarrays are needed to compare samples within an experiment.

Microarrays and bioinformatics



Gene expression values from microarray experiments can be represented as heat maps to visualize the result of data analysis.

The advent of inexpensive microarray experiments created several specific bioinformatics challenges:

- the multiple levels of replication in experimental design (Experimental design)
- the number of platforms and independent groups and data format (Standardization)
- the treatment of the data (Statistical analysis)
- accuracy and precision (Relation between probe and gene)
- the sheer volume of data and the ability to share it (Data warehousing)

Experimental design

Due to the biological complexity of gene expression, the considerations of experimental design that are discussed in the expression profiling article are of critical importance if statistically and biologically valid conclusions are to be drawn from the data.

There are three main elements to consider when designing a microarray experiment. First, replication of the biological samples is essential for drawing conclusions from the experiment. Second, technical replicates (two RNA samples obtained from each experimental unit) help to ensure precision and allow for testing differences within treatment groups. The biological replicates include independent RNA extractions and technical replicates may be two aliquots of the same extraction. Third, spots of each cDNA clone or oligonucleotide are present as replicates (at least duplicates) on the microarray slide, to provide a measure of technical precision in each hybridization. It is critical that information about the sample preparation and handling is discussed, in order to help identify the independent units in the experiment and to avoid inflated estimates of statistical significance.

Standardization

Microarray data is difficult to exchange due to the lack of standardization in platform fabrication, assay protocols, and analysis methods. This presents an interoperability problem in bioinformatics. Various grass-roots open-source projects are trying to ease the exchange and analysis of data produced with non-proprietary chips:

- For example, the "Minimum Information About a Microarray Experiment" (MIAME) checklist helps define the level of detail that should exist and is being adopted by many journals as a requirement for the submission of papers incorporating microarray results. But MIAME does not describe the format for the information, so while many formats can support the MIAME requirements, as of 2007 no format permits verification of complete semantic compliance.
- The "MicroArray Quality Control (MAQC) Project" is being conducted by the US Food and Drug Administration (FDA) to develop standards and quality control metrics which will eventually allow the use of MicroArray data in drug discovery, clinical practice and regulatory decision-making.
- The MGED Society has developed standards for the representation of gene expression experiment results and relevant annotations.

Statistical analysis

Microarray data sets are commonly very large, and analytical precision is influenced by a number of variables. Statistical challenges include taking into account effects of background noise and appropriate normalization of the data. Normalization methods may be suited to specific platforms and, in the case of commercial platforms, the analysis may be proprietary. Algorithms that affect statistical analysis include:

- Image analysis: gridding, spot recognition of the scanned image (segmentation algorithm), removal or marking of poor-quality and low-intensity features (called *flagging*).
- Data processing: background subtraction (based on global or local background), determination of spot intensities and intensity ratios, visualisation of data, and log-transformation of ratios, global or local normalization of intensity ratios.
- Identification of statistically significant changes: t-test, ANOVA, Bayesian method Mann–Whitney test methods tailored to microarray data sets, which take into account multiple comparisons or cluster analysis. These methods assess statistical power based on the variation present in the data and the number of experimental replicates, and can help minimize Type I and type II errors in the analyses.
- Network-based methods: Statistical methods that take the underlying structure of gene networks into account, representing either associative or causative interactions or dependencies among gene products.

Microarray data may require further processing aimed at reducing the dimensionality of the data to aid comprehension and more focused analysis. Other methods permit analysis of data consisting of a low number of biological or technical replicates; for example, the Local Pooled Error (LPE) test pools standard deviations of genes with similar expression levels in an effort to compensate for insufficient replication.

Relation between probe and gene

The relation between a probe and the mRNA that it is expected to detect is not trivial. Some mRNAs may cross-hybridize probes in the array that are supposed to detect another mRNA. In addition, mRNAs may experience amplification bias that is sequence or molecule-specific. Thirdly, probes that are designed to detect the mRNA of a particular gene may be relying on genomic EST information that is incorrectly associated with that gene.

Data warehousing

Microarray data was found to be more useful when compared to other similar datasets. The sheer volume (in bytes), specialized formats (such as MIAME), and curation efforts associated with the datasets require specialized databases to store the data.

Chapter- 6

DNA Sequencing

The term **DNA sequencing** is the use of sequencing for determining the order of the nucleotide bases—adenine, guanine, cytosine, and thymine—in a molecule of DNA.

Knowledge of DNA sequences has become indispensable for basic biological research, other research branches utilizing DNA sequencing, and in numerous applied fields such as diagnostic, biotechnology, forensic biology and biological systematics. The advent of DNA sequencing has significantly accelerated biological research and discovery. The rapid speed of sequencing attained with modern DNA sequencing technology has been instrumental in the sequencing of the human genome, in the Human Genome Project. Related projects, often by scientific collaboration across continents, have generated the complete DNA sequences of many animal, plant, and microbial genomes.

The first DNA sequences were obtained in the early 1970s by academic researchers using laborious methods based on two-dimensional chromatography. Following the development of dye-based sequencing methods with automated analysis, DNA sequencing has become easier and orders of magnitude faster.

History

RNA sequencing was one of the earliest forms of nucleotide sequencing. The major landmark of RNA sequencing is the sequence of the first complete gene and the complete genome of Bacteriophage MS2, identified and published by Walter Fiers and his coworkers at the University of Ghent (Ghent, Belgium), between 1972 and 1976.

Prior to the development of rapid DNA sequencing methods in the early 1970s by Frederick Sanger at the University of Cambridge, in England and Walter Gilbert and Allan Maxam at Harvard, a number of laborious methods were used. For instance, in 1973, Gilbert and Maxam reported the sequence of 24 basepairs using a method known as wandering-spot analysis.

The chain-termination method developed by Sanger and coworkers in 1975 soon became the method of choice, owing to its relative ease and reliability.

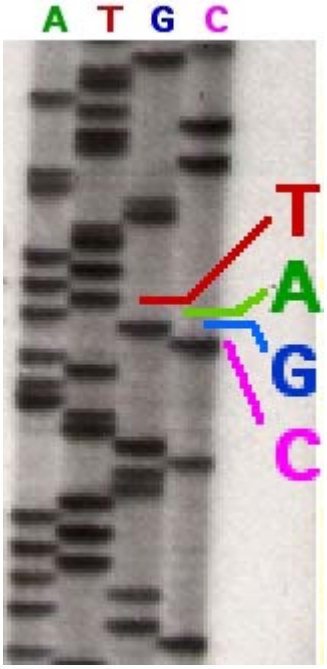
Maxam–Gilbert sequencing

In 1976–1977, Allan Maxam and Walter Gilbert developed a DNA sequencing method based on chemical modification of DNA and subsequent cleavage at specific bases. Although Maxam and Gilbert published their chemical sequencing method two years after the ground-breaking paper of Sanger and Coulson on plus-minus sequencing, Maxam–Gilbert sequencing rapidly became more popular, since purified DNA could be used directly, while the initial Sanger method required that each read start be cloned for production of single-stranded DNA. However, with the improvement of the chain-termination method, Maxam-Gilbert sequencing has fallen out of favour due to its technical complexity prohibiting its use in standard molecular biology kits, extensive use of hazardous chemicals, and difficulties with scale-up.

The method requires radioactive labeling at one 5' end of the DNA (typically by a kinase reaction using gamma-^{32}P ATP) and purification of the DNA fragment to be sequenced. Chemical treatment generates breaks at a small proportion of one or two of the four nucleotide bases in each of four reactions (G, A+G, C, C+T). For example, the purines (A+G) are depurinated using formic acid, the guanines (and to some extent the adenines) are methylated by dimethyl sulfate, and the pyrimidines (C+T) are methylated using hydrazine. The addition of salt (sodium chloride) to the hydrazine reaction inhibits the methylation of thymine for the C-only reaction. The modified DNAs are then cleaved by hot piperidine at the position of the modified base. The concentration of the modifying chemicals is controlled to introduce on average one modification per DNA molecule. Thus a series of labeled fragments is generated, from the radiolabeled end to the first "cut" site in each molecule. The fragments in the four reactions are electrophoresed side by side in denaturing acrylamide gels for size separation. To visualize the fragments, the gel is exposed to X-ray film for autoradiography, yielding a series of dark bands each corresponding to a radiolabeled DNA fragment, from which the sequence may be inferred.

Also sometimes known as "chemical sequencing", this method led to the Methylation Interference Assay used to map DNA-binding sites for DNA-binding proteins.

Chain-termination methods



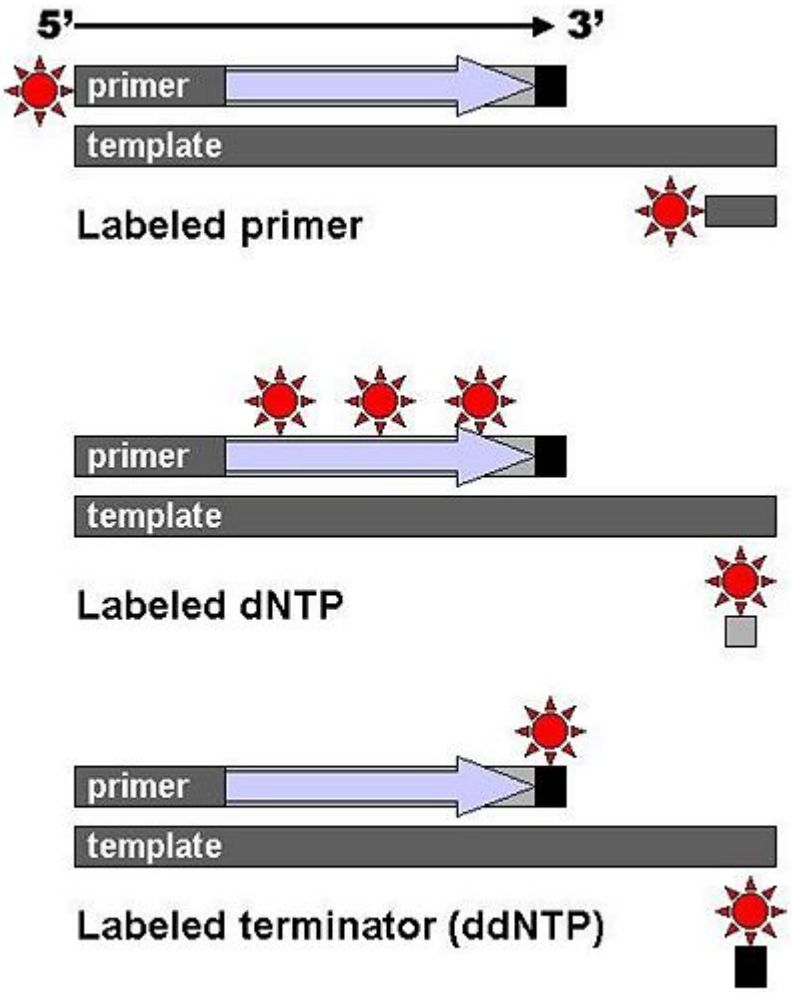
Part of a radioactively labelled sequencing gel

Because the chain-terminator method (or Sanger method after its developer Frederick Sanger) is more efficient and uses fewer toxic chemicals and lower amounts of radioactivity than the method of Maxam and Gilbert, it rapidly became the method of choice. The key principle of the Sanger method was the use of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators.

The classical chain-termination method requires a single-stranded DNA template, a DNA primer, a DNA polymerase, normal deoxynucleotidephosphates (dNTPs), and modified nucleotides (dideoxyNTPs) that terminate DNA strand elongation. These ddNTPs will also be radioactively or fluorescently labelled for detection in automated sequencing machines. The DNA sample is divided into four separate sequencing reactions, containing all four of the standard deoxynucleotides (dATP, dGTP, dCTP and dTTP) and the DNA polymerase. To each reaction is added only one of the four dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP) which are the chain-terminating nucleotides, lacking a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides, thus terminating DNA strand extension and resulting in DNA fragments of varying length.

The newly synthesized and labelled DNA fragments are heat denatured, and separated by size (with a resolution of just one nucleotide) by gel electrophoresis on a denaturing polyacrylamide-urea gel with each of the four reactions run in one of four individual lanes (lanes A, T, G, C); the DNA bands are then visualized by autoradiography or UV light, and the DNA sequence can be directly read off the X-ray film or gel image. In the image on the right, X-ray film was exposed to the gel, and the dark bands correspond to

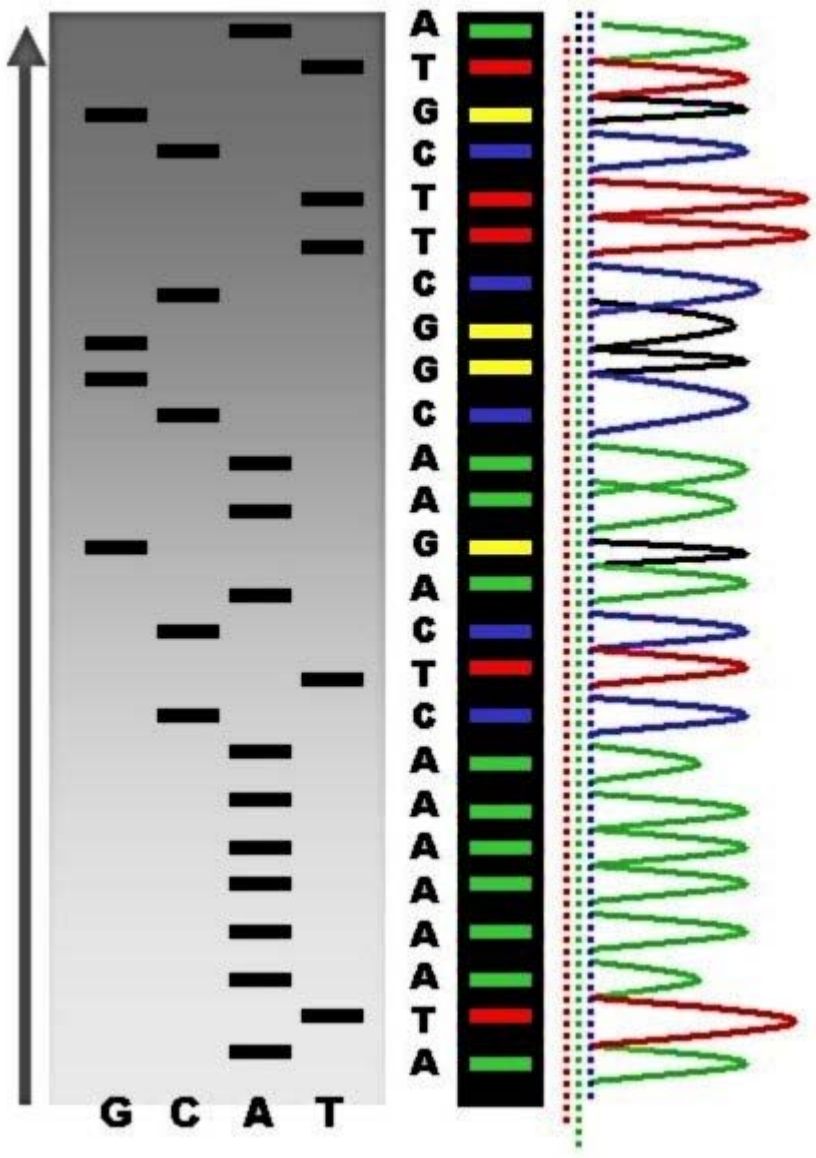
DNA fragments of different lengths. A dark band in a lane indicates a DNA fragment that is the result of chain termination after incorporation of a dideoxynucleotide (ddATP, ddGTP, ddCTP, or ddTTP). The relative positions of the different bands among the four lanes are then used to read (from bottom to top) the DNA sequence.



DNA fragments are labelled with a radioactive or fluorescent tag on the primer (1), in the new DNA strand with a labeled dNTP, or with a labeled ddNTP.

Technical variations of chain-termination sequencing include tagging with nucleotides containing radioactive phosphorus for radiolabelling, or using a primer labeled at the 5' end with a fluorescent dye. Dye-primer sequencing facilitates reading in an optical system for faster and more economical analysis and automation. The later development

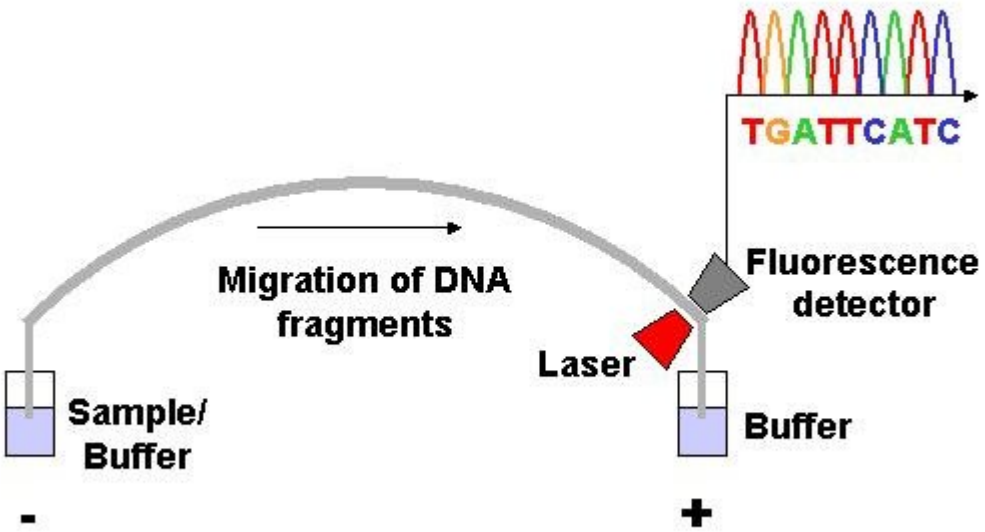
by Leroy Hood and coworkers of fluorescently labeled ddNTPs and primers set the stage for automated, high-throughput DNA sequencing.



Sequence ladder by radioactive sequencing compared to fluorescent peaks

Chain-termination methods have greatly simplified DNA sequencing. For example, chain-termination-based kits are commercially available that contain the reagents needed for sequencing, pre-aliquoted and ready to use. Limitations include non-specific binding of the primer to the DNA, affecting accurate read-out of the DNA sequence, and DNA secondary structures affecting the fidelity of the sequence.

Dye-terminator sequencing



Capillary electrophoresis

Dye-terminator sequencing utilizes labelling of the chain terminator ddNTPs, which permits sequencing in a single reaction, rather than four reactions as in the labelled-primer method. In dye-terminator sequencing, each of the four dideoxynucleotide chain terminators is labelled with fluorescent dyes, each of which emit light at different wavelengths.

Owing to its greater expediency and speed, dye-terminator sequencing is now the mainstay in automated sequencing. Its limitations include dye effects due to differences in the incorporation of the dye-labelled chain terminators into the DNA fragment, resulting in unequal peak heights and shapes in the electronic DNA sequence trace chromatogram after capillary electrophoresis.

This problem has been addressed with the use of modified DNA polymerase enzyme systems and dyes that minimize incorporation variability, as well as methods for eliminating "dye blobs". The dye-terminator sequencing method, along with automated high-throughput DNA sequence analyzers, is now being used for the vast majority of sequencing projects.

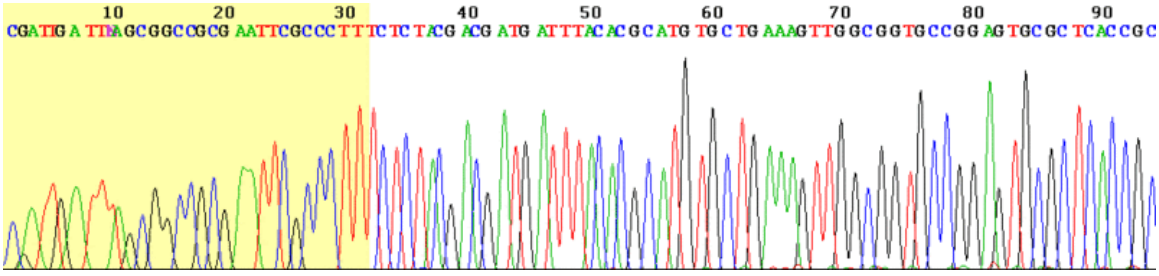
Challenges

Common challenges of DNA sequencing include poor quality in the first 15–40 bases of the sequence and deteriorating quality of sequencing traces after 700–900 bases. Base calling software typically gives an estimate of quality to aid in quality trimming.

In cases where DNA fragments are cloned before sequencing, the resulting sequence may contain parts of the cloning vector. In contrast, PCR-based cloning and emerging sequencing technologies based on pyrosequencing often avoid using cloning vectors. Recently, one-step Sanger sequencing (combined amplification and sequencing) methods such as Ampliseq and SeqSharp have been developed that allow rapid sequencing of target genes without cloning or prior amplification.

Current methods can directly sequence only relatively short (300–1000 nucleotides long) DNA fragments in a single reaction. The main obstacle to sequencing DNA fragments above this size limit is insufficient power of separation for resolving large DNA fragments that differ in length by only one nucleotide. In all cases the use of a primer with a free 5' end is essential.

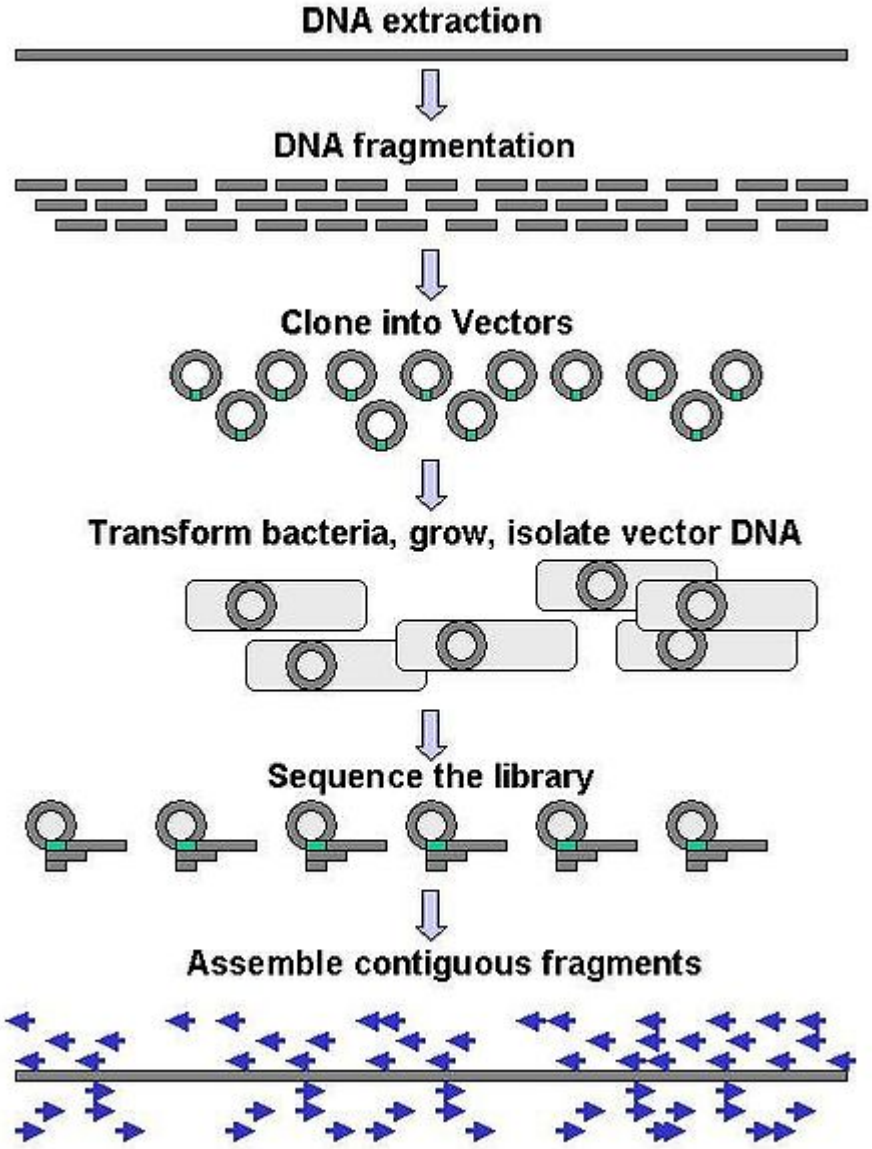
Automation and sample preparation



View of the start of an example dye-terminator read

Automated DNA-sequencing instruments (DNA sequencers) can sequence up to 384 DNA samples in a single batch (run) in up to 24 runs a day. DNA sequencers carry out capillary electrophoresis for size separation, detection and recording of dye fluorescence, and data output as fluorescent peak trace chromatograms. Sequencing reactions by thermocycling, cleanup and re-suspension in a buffer solution before loading onto the sequencer are performed separately. A number of commercial and non-commercial software packages can trim low-quality DNA traces automatically. These programs score the quality of each peak and remove low-quality base peaks (generally located at the ends of the sequence). The accuracy of such algorithms is below visual examination by a human operator, but sufficient for automated processing of large sequence data sets.

Amplification and clonal selection



Genomic DNA is fragmented into random pieces and cloned as a bacterial library. DNA from individual bacterial clones is sequenced and the sequence is assembled by using overlapping DNA regions.

Large-scale sequencing aims at sequencing very long DNA pieces, such as whole chromosomes. Common approaches consist of cutting (with restriction enzymes) or shearing (with mechanical forces) large DNA fragments into shorter DNA fragments. The fragmented DNA is cloned into a DNA vector, and amplified in *Escherichia coli*. Short DNA fragments purified from individual bacterial colonies are individually sequenced and assembled electronically into one long, contiguous sequence.

This method does not require any pre-existing information about the sequence of the DNA and is referred to as *de novo* sequencing. Gaps in the assembled sequence may be filled by primer walking. The different strategies have different tradeoffs in speed and accuracy; *shotgun methods* are often used for sequencing large genomes, but its assembly is complex and difficult, particularly with sequence repeats often causing gaps in genome assembly.

Most sequencing approaches use an *in vitro* cloning step to amplify individual DNA molecules, because their molecular detection methods are not sensitive enough for single molecule sequencing. Emulsion PCR isolates individual DNA molecules along with primer-coated beads in aqueous droplets within an oil phase. Polymerase chain reaction (PCR) then coats each bead with clonal copies of the DNA molecule followed by immobilization for later sequencing. Emulsion PCR is used in the methods by Marguilis et al. (commercialized by 454 Life Sciences), Shendure and Porreca et al. (also known as "Polony sequencing") and SOLiD sequencing, (developed by Agencourt, now Applied Biosystems).

Another method for *in vitro* clonal amplification is *bridge PCR*, where fragments are amplified upon primers attached to a solid surface, used in the Illumina Genome Analyzer. Single-molecule methods, such as that developed by Stephen Quake's laboratory (later commercialized by Helicos) is an exception: it uses bright fluorophores and laser excitation to detect base addition events from individual DNA molecules fixed to a surface, eliminating the need for molecular amplification.

High-throughput sequencing

The high demand for low-cost sequencing has driven the development of high-throughput sequencing technologies that parallelize the sequencing process, producing thousands or millions of sequences at once. High-throughput sequencing technologies are intended to lower the cost of DNA sequencing beyond what is possible with standard dye-terminator methods.

Lynx Therapeutics' Massively Parallel Signature Sequencing (MPSS)

The first of the "next-generation" sequencing technologies, MPSS was developed in 1990s at Lynx Therapeutics, a company founded in 1992 by Sydney Brenner and Sam Eletr. MPSS was a bead-based method that used a complex approach of adapter ligation followed by adapter decoding, reading the sequence in increments of four nucleotides; this method made it susceptible to sequence-specific bias or loss of specific sequences. Because the technology was so complex, MPSS was only performed 'in-house' by Lynx Therapeutics and no machines were sold; when the merger with Solexa later lead to the development of sequencing-by-synthesis, a more simple approach with numerous advantages, MPSS became obsolete. However, the essential properties of the MPSS output were typical of later "next-gen" data types, including hundreds of thousands of short DNA sequences. In the case of MPSS, these were typically used for sequencing

cDNA for measurements of gene expression levels. Lynx Therapeutics merged with Solexa in 2004, and this company was later purchased by Illumina.

Polony sequencing

Polony sequencing, developed in George Church's lab at Harvard, was among the first next-generation sequencing systems used to sequence a full genome in 2005. It combined an in vitro paired-tag library with emulsion PCR, an automated microscope, and ligation-based sequencing chemistry to sequence an E. coli genome at an accuracy of > 99.9999% and a cost approximately 1/10 that of Sanger sequencing. The technology was licensed to Agencourt Biosciences, subsequently spun out into Agencourt Personal Genomics, and ultimately incorporated into the Applied Biosystems SOLiD platform.

454 pyrosequencing

A parallelized version of pyrosequencing was developed by 454 Life Sciences. The method amplifies DNA inside water droplets in an oil solution (emulsion PCR), with each droplet containing a single DNA template attached to a single primer-coated bead that then forms a clonal colony. The sequencing machine contains many picolitre-volume wells each containing a single bead and sequencing enzymes. Pyrosequencing uses luciferase to generate light for detection of the individual nucleotides added to the nascent DNA, and the combined data are used to generate sequence read-outs. This technology provides intermediate read length and price per base compared to Sanger sequencing on one end and Solexa and SOLiD on the other. 454 Life Sciences has since been acquired by Roche Diagnostics.

Illumina (Solexa) sequencing

Solexa, now part of Illumina, developed a sequencing technology based on reversible dye-terminators. DNA molecules are first attached to primers on a slide and amplified so that local clonal colonies are formed (bridge amplification). Four types of ddNTPs are added, and non-incorporated nucleotides are washed away. Unlike pyrosequencing, the DNA can only be extended one nucleotide at a time. A camera takes images of the fluorescently labeled nucleotides then the dye along with the terminal 3' blocker is chemically removed from the DNA, allowing a next cycle.

SOLiD sequencing

Applied Biosystems' SOLiD technology employs sequencing by ligation. Here, a pool of all possible oligonucleotides of a fixed length are labeled according to the sequenced position. Oligonucleotides are annealed and ligated; the preferential ligation by DNA ligase for matching sequences results in a signal informative of the nucleotide at that position. Before sequencing, the DNA is amplified by emulsion PCR. The resulting bead, each containing only copies of the same DNA molecule, are deposited on a glass slide. The result is sequences of quantities and lengths comparable to Illumina sequencing.

Ion semiconductor sequencing

Ion Torrent Systems Inc. developed a system based on ion semiconductor sequencing. This method of sequencing is based on the detection of hydrogen ions that are released during the polymerisation of DNA. A microwell containing a template DNA strand to be sequenced is flooded with a single type of nucleotide. If the introduced nucleotide is complementary to the leading template nucleotide it is incorporated into the growing complementary strand. This causes the release of a hydrogen ion that triggers a hypersensitive ion sensor, which indicates that a reaction has occurred. If homopolymer repeats are present in the template sequence multiple nucleotides will be incorporated in a single cycle. This leads to a corresponding number of released hydrogens and a proportionally higher electronic signal.

DNA nanoball sequencing

DNA nanoball sequencing is a type of high throughput sequencing technology used to determine the entire genomic sequence of an organism. The company Complete Genomics uses this technology to sequence samples that researchers submit from several projects. The method uses rolling circle replication to amplify small fragments of genomic DNA into DNA nanoballs. Unchained sequencing by ligation is then used to determine the nucleotide sequence. This method of DNA sequencing allows large numbers of DNA nanoballs to be sequenced per run and at low reagent costs compared to other next generation sequencing platforms. However, only short sequences of DNA are determined from each DNA nanoball which makes mapping the short reads to a reference genome difficult. This technology has been used for multiple genome sequencing projects and is scheduled to be used for many more.

Future methods

Sequencing by hybridization is a non-enzymatic method that uses a DNA microarray. A single pool of DNA whose sequence is to be determined is fluorescently labeled and hybridized to an array containing known sequences. Strong hybridization signals from a given spot on the array identifies its sequence in the DNA being sequenced. Mass spectrometry may be used to determine mass differences between DNA fragments produced in chain-termination reactions.

DNA sequencing methods currently under development include labeling the DNA polymerase, reading the sequence as a DNA strand transits through nanopores, and microscopy-based techniques, such as AFM or transmission electron microscopy that are used to identify the positions of individual nucleotides within long DNA fragments (>5,000 bp) by nucleotide labeling with heavier elements (e.g., halogens) for visual detection and recording.

In microfluidic Sanger sequencing the entire thermocycling amplification of DNA fragments as well as their separation by electrophoresis is done on a single glass wafer (approximately 10 cm in diameter) thus reducing the reagent usage as well as cost. In

some instances researchers have shown that they can increase the throughput of conventional sequencing through the use of microchips. Research will still need to be done in order to make this use of technology effective.

In October 2006, the X Prize Foundation established an initiative to promote the development of full genome sequencing technologies, called the Archon X Prize, intending to award $10 million to "the first Team that can build a device and use it to sequence 100 human genomes within 10 days or less, with an accuracy of no more than one error in every 100,000 bases sequenced, with sequences accurately covering at least 98% of the genome, and at a recurring cost of no more than $10,000 (US) per genome."

Each year NHGRI promotes grants for new research and developments in genomics. 2010 grants and 2011 candidates include continuing work in microfluidic, polony and base-heavy sequencing methodologies

Chapter- 7

Eastern Blotting and Combined Bisulfite Restriction Analysis

Eastern blotting

Eastern blotting is a biochemical technique used to analyze protein post translational modifications (PTM) such as lipids and glycoconjugates. It is most often used to detect carbohydrate epitopes. Thus, Eastern blotting can be considered an extension of the biochemical technique of Western blotting. Multiple techniques have been described by the term Eastern blotting, most use proteins or lipids blotted from SDS-PAGE gel on to a PVDF or nitrocellulose membrane. Transferred proteins are analyzed for post-translational modifications using probes that may detect lipids, carbohydrate, phosphorylation or any other protein modification. Eastern blotting should be used to refer to methods that detect their targets through specific interaction of the PTM and the probe, distinguishing them from a standard Far-western blot. In principle, Eastern blotting is similar to lectin blotting (i.e. detection of carbohydrate epitopes on proteins or lipids); however, the term *lectin blotting* is more prevalent in the literature.

History and multiple definitions

Definition of the term *Eastern blotting* is somewhat confused due to multiple sets of authors dubbing a new method as *Eastern blotting*, or a derivative therof. All of the definitions are a derivative of the technique of Western blotting developed by Towbin in 1979. The current definitions are summarized below in order of the first use of the name; however, all are based on some earlier works. In some cases, the technique had been in practice for some time before the introduction of the term.

- (1982) The term *Eastern blotting* was specifically rejected by two separate groups: Reinhart and Malamud referred to a protein blot of a native gel as a *native blot*; Peferoen et al., opted to refer to their method of drawing SDS-gel separated proteins onto nitrocellulose using a vacuum as *Vacuum blotting*.

- (1984) *Middle Eastern blotting* has been described as a blot of polyA RNA (resolved by agarose) which is then immobilized. The immobilized RNA is then probed using DNA.
- (1996) *Eastern-Western blot* was first used by Bogdanov et al. The method involved blotting of phospholipids on PVDF or nitrocellulose membrane prior to transfer of proteins onto the same nitrocellulose membrane by conventional Western blotting and probing with conformation specific antibodies. This method is based on earlier work by Taki et al. in 1994, which they originally dubbed *TLC blotting*, and was based on a similar method introduced by Towbin in 1984.
- (2000) *Far-Eastern blotting* seems to have been first named in 2000 by Ishikawa & Taki. The method is described more fully in Far-Eastern blotting, but is based on antibody or lectin staining of lipids transferred to PVDF membranes.
- (2001) *Eastern blotting* was described as a technique for detecting glycoconjugates generated by blotting BSA onto PVDF membranes, followed by periodate treatment. The oxidized protein is then treated with a complex mixture, generating a new conjugate on the membrane. The membrane is then probed with antibodies for epitopes of interest. This method has also been discussed in later work by the same group. The method is essentially Far-Eastern blotting.
- (2002) *Eastern blot* has also been used to describe an immunoblot performed on proteins blotted to a PVDF membrane from a PAGE gel run with opposite polarity. Since this is essentially a Western blot, the charge reversal was used to dub this method an *Eastern blot*.
- (2005) *Eastern blot* has been used to describe a blot of proteins on PVDF membrane where the probe is an aptamer rather than an antibody. This could be seen as similar to a Southern blot, however the interaction is between a DNA molecule(the aptamer) and a protein, rather than two DNA molecules.
- (2006) *Eastern blotting* has been used to refer to the detection of fusion proteins through complementation. The name is based on the use of an enzyme activator (EA) as part of the detection.
- (2009) *Eastern blotting* has most recently been re-dubbed by Thomas et al. as a technique which probes proteins blotted to PVDF membrane with lectins, cholera toxin and chemical stains to detect glycosylated, lipoylated or phosphorylated proteins. These authors distinguish the method from the *Far-eastern blot* named by Taki et al. in that they use lectin probes and other staining reagents.

There is clearly no single accepted definition of the term. A recent highlight article has interviewed Ed Southern, originator of the Southern blot, regarding a re-christening of *Eastern blotting* from Tanaka et al. The article likens the *Eastern blot* to "fairies, unicorns, and a free lunch" and states that Eastern blots "don't exist." The *Eastern blot* is mentioned in an Immunology textbook which compares the common blotting methods (Southern, Northern, and Western), and states that "the Eastern blot, however, exists only in test questions."

The principles used for Eastern blotting to detect glycans can be traced back to the use of lectins to detect protein glycosylation. The earliest example for this mode of detection is Tanner and Anstee in 1976, where lectins were used to detect glycosylated proteins

isolated from human erythrocytes. The specific detection of glycosylation through blotting is usually referred to as *lectin blotting*. A summary of more recent improvements of the protocol has been provided by H. Freeze.

Applications

One application of the technique includes detection of protein modifications in two bacterial species *Ehrlichia- E. muris* and IOE. Cholera toxin B subunit (which binds to gangliosides), Concanavalin A (which detects mannose-containing glycans) and nitrophospho molybdate-methyl green (which detects phosphoproteins) were used to detect protein modifications. The technique showed that the antigenic proteins of the non-virulent *E.muris* is more post-translationally modified than the highly virulent IOE.

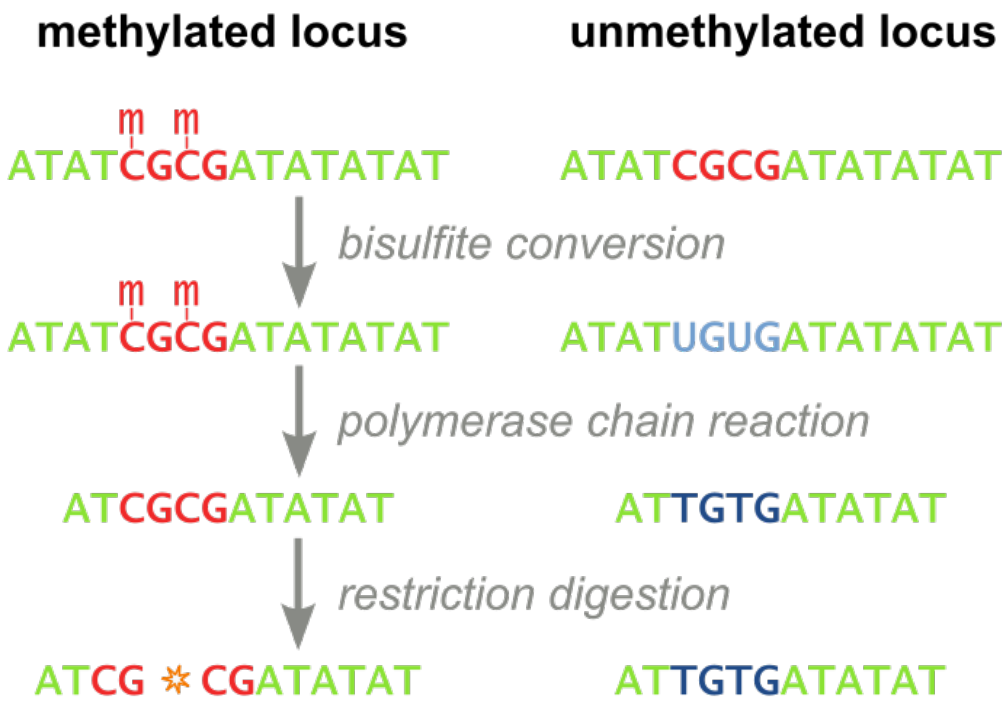
Significance

Most proteins that are translated from mRNA undergo modifications before becoming functional in cells. These modifications are collectively known as post-translational modifications (PTMs). The nascent or folded proteins, which are stable under physiological conditions, are then subjected to a battery of specific enzyme-catalyzed modifications on the side chains or backbones.

Post-translational protein modifications can include: acetylation, acylation (myristoylation, palmitoylation), alkylation, arginylation, biotinylation, formylation, geranylgeranylation, glutamylation, glycosylation, glycylation, hydroxylation, isoprenylation, lipoylation, methylation, nitroalkylation, phosphopantetheinylation, phosphorylation, prenylation, selenation, S-nitrosylation, sulfation, transglutamination and ubiquitination (sumoylation).

Post-translational modifications occurring at the N-terminus of the amino acid chain play an important role in translocation across biological membranes. These include secretory proteins in prokaryotes and eukaryotes and also proteins that are intended to be incorporated in various cellular and organelle membranes such as lysosomes, chloroplast, mitochondria and plasma membrane. Expression of posttranslated proteins is important in several diseases.

Combined bisulfite restriction analysis



The first few steps of COBRA, and the molecular changes caused by each step to methylated and unmethylated CpG sites.

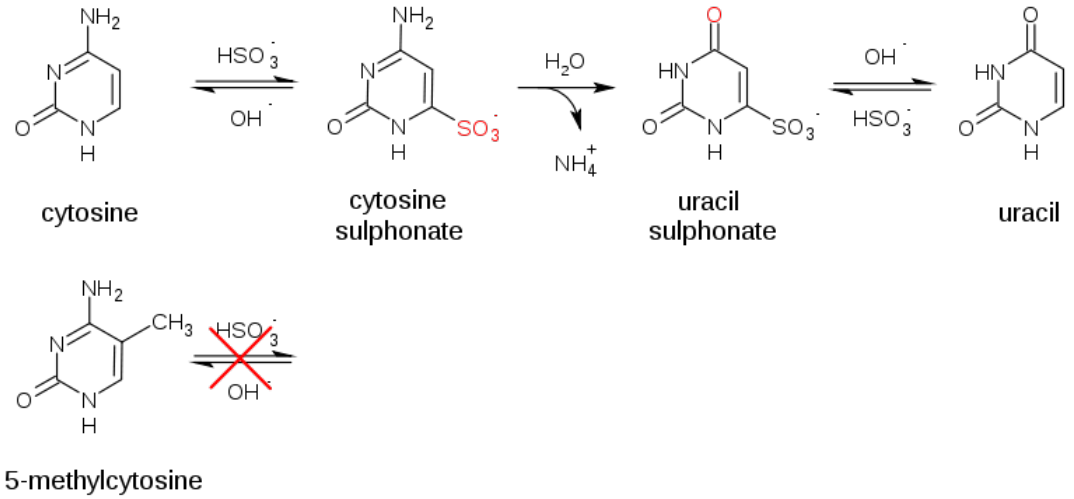
Combined Bisulfite Restriction Analysis (or **COBRA**) is a molecular biology technique that allows for the sensitive quantification of DNA methylation levels at a specific genomic loci on a DNA sequence in a small sample of genomic DNA. The technique is a variation of bisulfite sequencing, and combines bisulfite conversion based polymerase chain reaction with restriction digestion. Originally developed to reliably handle minute amounts of genomic DNA from microdissected paraffin-embedded tissue samples, the technique has since seen widespread usage in cancer research and epigenetics studies.

Procedure

Bisulfite Treatment

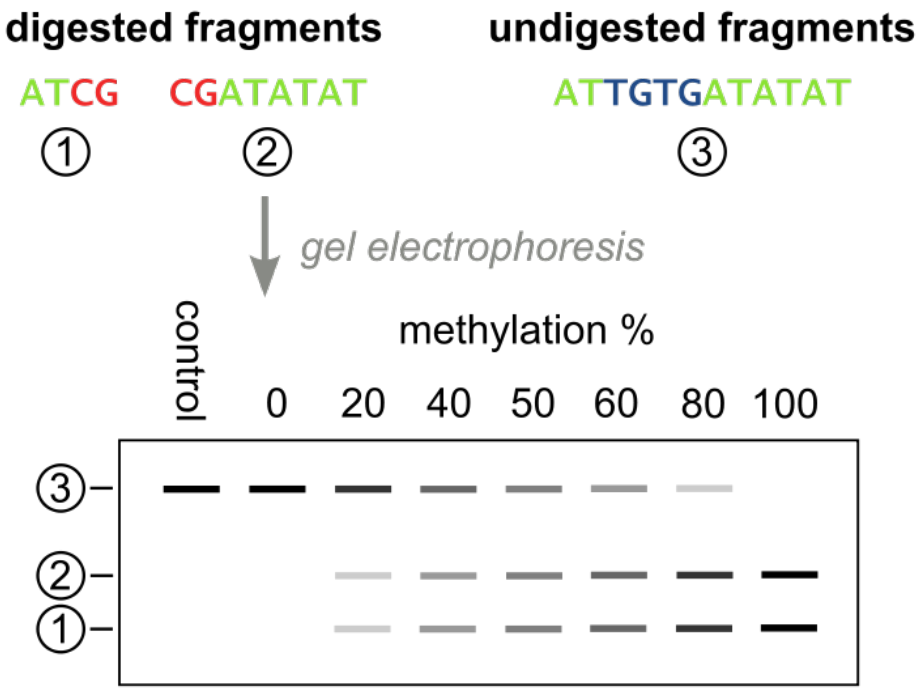
Genomic DNA of interest is treated with sodium bisulfite, which introduces methylation-dependent sequence differences. During sodium bisulfite treatment, unmethylated

cytosine residues are converted to uracil, while methylated cytosine residues are unaffected.



Bisulfite conversion, changing unmethylated cytosine to uracil, while 5-methylcytosine remains unaffected by the treatment.

PCR Amplificiation



The final quantification step of COBRA, where the DNA methylation levels of the original input sample can be determined by comparing and quantifying the number of digested and undigested fragments.

Bisulfite treated DNA is then PCR amplified, resulting in cytosine residues at originally methylated positions, and thyamine residues at originally unmethylated position (that were converted to uracil). Primers used during this step do not contain CpG sites (the common target of cytosine methylation), so the amplificiation process does not discriminate between templates based on methylation status. PCR products are purified to ensure complete digestion in the following step.

Restriction Digest

The above steps lead to the methylation dependent retention or loss of CpG-containing restriction enzyme sites, such as those for TaqI (TCGA) and BstUI (CGCG), depending on whether the cytosine residue was originally methylated or not, respectively. Due to the methylation-independent amplification in the above step, the resulting PCR products will be a mixed population of fragments that have lost or retained CpG-containing restriction enzyme sites, whose respective percentages will be directly correlated to the original level of DNA methylation in the sample DNA.

PCR products are then treated with an restriction enzyme (*e.g.* BstUI), which will only cleave sites that were originally methylated (CGCG), while leaving sites that were originally unmethylated (TGTG). To ensure that all CpG sites are retained due to originally being methylated, and not a remnant of incomplete bisulfite conversion, a control digestion is performed, with enzymes such as Hsp92II which recognizes the sequence CATG, none of which should be remaining after bisulfite conversion (with the rare exception of non-CpG methylation) and thus no cleavage should occur if bisulfite conversion was complete.

Quantification

The digested fragments are then separated by polyacrylamide gel electrophoresis with the expected appearance of bands corresponding to a single large undigested fragment, and multiple smaller bands corresponding to digested fragments. Quantitative amount of DNA in these bands can be determined with a device such as a phosphoimager, after which the methylation percentage of the original sample can be calculated by:

$$\% \text{ Methylation} = 100 \times \frac{\text{Digested Fragments}}{\text{Undigested Fragments} + \text{Digested Fragments}}$$

Usage and Applications

COBRA has been used extensively in many research-based applications such as screening for DNA methylation changes at gene promoters in cancer studies, detecting

altered methylation patterns at imprinted genes, and characterizing methylation patterns in the genome during development in mammals.

In medicine, COBRA has been used as a tool to help diagnose human disease involving aberrant DNA methylation. Researchers utilized COBRA in conjunction with denaturing high performance liquid chromatography in the diagnosis of the genetic imprinting disorder Russell-Silver syndrome where hypomethylation of the imprinted gene H19 is responsible for the disorder in up to 50% of patients.

Strengths

- **Simple, fast and inexpensive**: In COBRA, DNA methylation levels are easily and quickly measured without the need for laborious sub-cloning and sequencing, as with bisulfite sequencing. The assay is straight-forward and can be done with standard inexpensive molecular biology reagents.

- **High compatibility**: Due to the PCR and purification steps, the method not only works with very small amounts of genomic DNA, but also samples that have been treated with paraffin, both of which can be problems in other DNA methylation quantification protocols such as Southern blotting and methylation-sensitive restriction enzyme digestion followed by PCR.

- **Quantitative**: This is in contrast to methylation-specific PCR, which is qualitative. With COBRA, DNA methylation levels can be directly quantified at a given locus, yielding more information per assay.

- **Scalability for high-throughput sample processing**: With COBRA, many regions of interest can be processed in parallel in separate samples digested with the same restriction enzyme. This is in contrast to bisulfite sequencing analysis, where each region needs to be examined rigorously by sequencing many clones per locus, costing more time.

- **Multiple queries per assay**: Methylation status can be interrogated at multiple CpG-containing restriction sites in a single digestion assay.

Weaknesses

- The assay is **limited to using existing restriction sites** in the region of interest, and methylation that does not occur in the context of a specific restriction site will not be assayed.

- **Incomplete digestion** by restriction enzymes after PCR can confound the analysis: incomplete digestion would suggest lack of DNA methylation (if cutting with a methylation-sensitive enzyme such as HpaII). It is also known that BstUI can cut at uncoverted sites, leading to overestimation of methylation levels and so the use of HpaII is often needed.

- In complex samples, **cell-type heterogeneity can confound the analysis** since the DNA is not being sequenced, heterogeneity in sequences from different cells in the sample (*i.e.* different cell populations within a tumor) that have acquired mutations in the interrogated region, such as changing the CG dinucleotide to CA or CT, would result in loss of the restriction site giving rise to an apparently methylated region due to lack of digestion. This would skew the quantification of DNA methylation levels in a given sample.

Chapter- 8

Immunoprecipitation

Immunoprecipitation (IP) is the technique of precipitating a protein antigen out of solution using an antibody that specifically binds to that particular protein. This process can be used to isolate and concentrate a particular protein from a sample containing many thousands of different proteins. Immunoprecipitation requires that the antibody be coupled to a solid substrate at some point in the procedure.

Types of immunoprecipitation

Individual protein Immunoprecipitation (IP)

Involves using an antibody that is specific for a known protein to isolate that particular protein out of a solution containing many different proteins. These solutions will often be in the form of a crude lysate of a plant or animal tissue. Other sample types could be bodily fluids or other samples of biological origin.

Protein complex immunoprecipitation (Co-IP)

Immunoprecipitation of intact protein complexes (i.e.: antigen along with any proteins or ligands that are bound to it) is known as co-immunoprecipitation (Co-IP). Co-IP works by selecting an antibody that targets a known protein that is believed to be a member of a larger complex of proteins. By targeting this *known* member with an antibody it may become possible to pull the entire protein complex out of solution and thereby identify *unknown* members of the complex.

This works when the proteins involved in the complex bind to each other tightly, making it possible to pull multiple members of the complex out of solution by latching onto one member with an antibody. This concept of pulling protein complexes out of solution is sometimes referred to as a "pull-down". Co-IP is a powerful technique that is used regularly by molecular biologists to analyze protein-protein interactions.

Identifying the members of protein complexes may require several rounds of precipitation with different antibodies for a number of reasons:

- A particular antibody often selects for a subpopulation of its target protein that has the epitope exposed, thus failing to identify any proteins in complexes that hide the epitope. This can be seen in that it is rarely possible to precipitate even half of a given protein from a sample with a single antibody, even when a large excess of antibody is used.

- The first round of IP will often result in the identification of many new proteins that are putative members of the complex being studied. The researcher will then obtain antibodies that specifically target one of the newly identified proteins and repeat the entire immunoprecipitation experiment. This second round of precipitation may result in the recovery of additional new members of a complex that were not identified in the previous experiment. As successive rounds of targeting and immunoprecipitations take place, the number of identified proteins may continue to grow. The identified proteins may not ever exist in a single complex at a given time, but may instead represent a network of proteins interacting with one another at different times for different purposes.

- Repeating the experiment by targeting different members of the protein complex allows the researcher to double-check the result. Each round of pull-downs should result in the recovery of both the original known protein as well as other previously identified members of the complex (and even new additional members). By repeating the immunoprecipitation in this way, the researcher verifies that each identified member of the protein complex was a valid identification. If a particular protein can only be recovered by targeting one of the known members but not by targeting other of the known members then that protein's status as a member of the complex may be subject to question.

Chromatin immunoprecipitation (ChIP)

Chromatin immunoprecipitation (ChIP) is a method used to determine the location of DNA binding sites on the genome for a particular protein of interest. This technique gives a picture of the protein-DNA interactions that occur inside the nucleus of living cells or tissues. The *in vivo* nature of this method is in contrast to other approaches traditionally employed to answer the same questions.

The principle underpinning this assay is that DNA-binding proteins (including transcription factors and histones) in living cells can be cross-linked to the DNA that they are binding. By using an antibody that is specific to a putative DNA binding protein, one can immunoprecipitate the protein-DNA complex out of cellular lysates. The crosslinking is often accomplished by applying formaldehyde to the cells (or tissue), although it is sometimes advantageous to use a more defined and consistent crosslinker such as DTBP. Following crosslinking, the cells are lysed and the DNA is broken into pieces 0.2–1 kb in length by sonication. At this point the immunoprecipitation is performed resulting in the purification of protein-DNA complexes. The purified protein-DNA complexes are then heated to reverse the formaldehyde cross-linking of the protein and DNA complexes, allowing the DNA to be separated from the proteins. The identity and quantity of the

DNA fragments isolated can then be determined by PCR. The limitation of performing PCR on the isolated fragments is that one must have an idea which genomic region is being targeted in order to generate the correct PCR primers. This limitation is very easily circumvented simply by cloning the isolated genomic DNA into a plasmid vector and then using primers that are specific to the cloning region of that vector. Alternatively, when one wants to find where the protein binds on a genome-wide scale, a DNA microarray can be used (ChIP-on-chip or ChIP-chip) allowing for the characterization of the cistrome. As well, ChIP-Sequencing has recently emerged as a new technology that can localize protein binding sites in a high-throughput, cost-effective fashion.

RNA immunoprecipitation (RIP)

Similar to chromatin immunoprecipitation (ChIP) outlined above, but rather than targeting DNA binding proteins as in ChIP, RNA immunoprecipitation targets RNA binding proteins. RIP is also an *in vivo* method in that live cells are exposed to formaldehyde in order to create cross-links between RNA and RNA-binding proteins. Cells are then lysed and the immunoprecipitation is performed with an antibody that targets the protein of interest. By isolating the protein, the RNA will also be isolated as it is cross-linked to the protein. The purified RNA-protein complexes can be separated by reversing the cross-link and the identity of the RNA can be determined by cDNA sequencing or RT-PCR.

Tagged proteins

One of the major technical hurdles with immunoprecipitation is the great difficulty in generating an antibody that specifically targets a single known protein. To get around this obstacle, many groups will engineer **tags** onto either the C- or N- terminal end of the protein of interest. The advantage here is that the same tag can be used time and again on many different proteins and the researcher can use the same antibody each time. The advantages with using tagged proteins are so great that this technique has become commonplace for all types of immunoprecipitation including all of the types of IP detailed above. Examples of tags in use are the Green Fluorescent Protein (GFP) tag, Glutathione-S-transferase (GST) tag and the FLAG-tag tag. While the use of a tag to enable pull-downs is convenient, it raises some concerns regarding biological relevance because the tag itself may either obscure native interactions or introduce new and unnatural interactions.

Methods

The two general methods for immunoprecipitation are the direct capture method and the indirect capture method.

Direct

Antibodies that are specific for a particular protein (or group of proteins) are immobilized on a solid-phase substrate such as superparamagnetic microbeads or on microscopic

agarose (non-magnetic) beads. The beads with bound antibodies are then added to the protein mixture and the proteins that are targeted by the antibodies are captured onto the beads via the antibodies, in other words, they become immunoprecipitated.

Indirect

Antibodies that are specific for a particular protein, or a group of proteins, are added directly to the mixture of protein. The antibodies have not been attached to a solid-phase support yet. The antibodies are free to float around the protein mixture and bind their targets. As time passes, the beads coated in protein A/G are added to the mixture of antibody and protein. At this point, the antibodies, which are now bound to their targets, will stick to the beads.

From this point on, the direct and indirect protocols converge because the samples now have the same ingredients. Both methods gives the same end-result with the protein or protein complexes bound to the antibodies which themselves are immobilized onto the beads.

Selection

An indirect approach is sometimes preferred when the concentration of the protein target is low or when the specific affinity of the antibody for the protein is weak. The indirect method is also used when the binding kinetics of the antibody to the protein is slow for a variety of reasons. In most situations, the direct method is the default, and the preferred, choice.

Technological advances

Agarose

Historically the solid-phase support for immunoprecipitation used by the majority of scientists has been highly-porous **agarose beads** (also known as agarose resins or slurries). The advantage of this technology is a very high potential binding capacity, as virtually the entire sponge-like structure of the agarose particle (50 to 150µm in size) is available for binding antibodies (which will in turn bind the target proteins) and the use of standard laboratory equipment for all aspects of the IP protocol without the need for any specialized equipment. The advantage of an extremely high binding capacity must be carefully balanced with the quantity of antibody that the researcher is prepared to use to coat the agarose beads. Because antibodies can be a cost-limiting factor, it is best to calculate backward ***from*** the amount of protein that needs to be captured (depending upon the analysis to be performed downstream), ***to*** the amount of antibody that is required to bind that quantity of protein (with a small excess added in to account for inefficiencies of the system), and back still further ***to*** the quantity of agarose that is needed to bind that particular quantity of antibody. In cases where antibody saturation is not required, this technology is unmatched in its ability to capture extremely large quantities of captured target proteins. The caveat here is that the ***"high capacity***

advantage" can become a ***"high capacity disadvantage"*** that is manifested when the enormous binding capacity of the sepharose/agarose beads is not completely saturated with antibodies. It often happens that the amount of antibody available to the researcher for the their immunoprecipitation experiment is less than sufficient to saturate the agarose beads to be used in the immunoprecipitation. In these cases the researcher can end up with agarose particles that are only partially coated with antibodies, and the portion of the binding capacity of the agarose beads that is not coated with antibody is then free to bind anything that will stick, resulting in an elevated background signal due to non-specific binding of lysate components to the beads, which can make data interpretation difficult. While some may argue that for these reasons it is prudent to match the quantity of agarose (in terms of binding capacity) to the quantity of antibody that one wishes to be bound for the immunoprecipitation, a simple way to reduce the issue of non-specific binding to agarose beads and increase specificity is to preclear the lysate, which for any immunoprecipitation is highly recommended .

Preclearing

Lysates are complex mixtures of proteins, lipids, carbohydrates and nucleic acids, and one must assume that some amount of non-specific binding to the IP antibody, Protein A/G or the beaded support will occur and negatively affect the detection of the immunoprecipitated target(s). In most cases, ***preclearing*** the lysate at the start of each immunoprecipitation experiment is a way to remove potentially reactive components from the cell lysate prior to the immunoprecipitation to prevent the non-specific binding of these components to the IP beads or antibody. The basic preclearing procedure is described below, wherein the lysate is incubated with beads alone, which are then removed and discarded prior to the immunoprecipitation This approach, though, does not account for non-specific binding to the IP antibody, which can be considerable. Therefore, an alternative method of preclearing is to incubate the protein mixture with exactly the same components that will be used in the immunoprecipitation, except that a non-target, irrelevant antibody of the same antibody subclass as the IP antibody is used instead of the IP antibody itself . This approach attempts to use as close to the exact IP conditions and components as the actual immunoprecipitation to remove any non-specific cell constituent without capturing the target protein (unless, of course, the target protein non-specifically binds to some other IP component, which should be properly controlled for by analyzing the discarded beads used to preclear the lysate). The target protein can then be immunoprecipitated with the reduced risk of non-specific binding interfering with data interpretation.

Superparamagnetic beads

While the vast majority of immunoprecipitations are performed with agarose beads, the use of superparamagnetic beads for immunoprecipitaion is a much newer approach that is only recently gaining in popularity as an alternative to agarose beads for IP applications. Unlike agarose, magnetic beads are solid and can be spherical, depending on the type of bead, and antibody binding is limited to the surface of each bead. While these beads do not have the advantage of a porous center to increase the binding capacity, magnetic

beads are significantly smaller than agarose beads (1 to 4μm), and the greater number of magnetic beads per volume than agarose beads collectively gives magnetic beads an effective surface area-to-volume ratio for optimum antibody binding.

Commercially available magnetic beads can be separated based by size uniformity into monodisperse and polydisperse beads. Monodisperse beads, also called microbeads, exhibit exact uniformity, and therefore all beads exhibit identical physical characteristics, including the binding capacity and the level of attraction to magnets. Polydisperse beads, while similar in size to monodisperse beads, show a wide range in size variability (1 to 4μm) that can influence their binding capacity and magnetic capture. Although both types of beads are commercially available for immunoprecipitation applications, the higher quality monodisperse superparamagnetic beads are more ideal for automatic protocols because of their consistent size, shape and performance. Monodisperse and polydisperse superparamagnetic beads are offered by many companies, including Invitrogen, Thermo Scientific, and Millipore.

Agarose vs. Magnetic Beads

Proponents of magnetic beads claim that the beads exhibit a faster rate of protein binding over agarose beads for immunoprecipitation applications, although standard agarose bead-based immunoprecipitations have been performed in 1 hour . Claims have also been made that magnetic beads are better for immunoprecipitating extremely large protein complexes because of the complete lack of an upper size limit for such complexes, although there is no unbiased evidence stating this claim. The nature of magnetic bead technology does results in less sample handling due to the reduced physical stress on samples of magnetic separation versus repeated centrifugation when using agarose, which may contribute greatly to increasing the yield of labile (fragile) protein complexes. Additional factors, though, such as the binding capacity, cost of the reagent, the requirement of extra equipment and the capability to automate IP processes should be considered in the selection of an immunoprecipitation support.

Binding Capacity

Proponents of both agarose and magnetic beads can argue whether the vast difference in the binding capacities of the two beads favors one particular type of bead. In a bead-to-bead comparison, agarose beads have significantly greater surface area and therefore a greater binding capacity than magnetic beads due to the large bead size and sponge-like structure. But the variable pore size of the agarose causes a potential upper size limit that may affect the binding of extremely large proteins or protein complexes to internal binding sites, and therefore magnetic beads may be better suited for immunoprecipitating large proteins or protein complexes than agarose beads, although there is a lack of independent comparative evidence that proves either case.

Some argue that the significantly greater binding capacity of agarose beads may be a disadvantage because of the larger capacity of non-specific binding. Others may argue for the use of magnetic beads because of the greater quantity of antibody required to saturate

the total binding capacity of agarose beads, which would obviously be an economical disadvantage of using agarose. While these arguments are correct outside the context of their practical use, these lines of reasoning ignore two key aspects of the principle of immunoprecipitation that demonstrates that the decision to use agarose or magnetic beads is not simply determined by binding capacity.

First, non-specific binding is not limited to the antibody-binding sites on the immobilized support; any surface of the antibody or component of the immunoprecipitation reaction can bind to nonspecific lysate constituents, and therefore nonspecific binding will still occur even when completely saturated beads are used. This is why it is important to preclear the sample before the immunoprecipitation is performed.

Second, the ability to capture the target protein is directly dependent upon the amount of immobilized antibody used, and therefore, in a side-by-side comparison of agarose and magnetic bead immunoprecipitation, the most protein that either support can capture is limited by the amount of antibody added. So the decision to saturate any type of support depends on the amount of protein required, as described above in the Agarose section of this page.

Cost

The price of using either type of support is a key determining factor in using agarose or magnetic beads for immunoprecipitation applications. A typical first-glance calculation on the cost of magnetic beads compared to sepharose beads may make the sepharose beads appear less expensive. But magnetic beads may be competitively priced compared to agarose for analytical-scale immunoprecipitations depending on the IP method used and the volume of beads required per IP reaction.

Using the traditional batch method of immunoprecipitation as listed below, where all components are added to a tube during the IP reaction, the physical handling characteristics of agarose beads necessitate a minimum quantity of beads for each IP experiment (typically in the range of 25 to 50µl beads per IP). This is because sepharose beads must be concentrated at the bottom of the tube by centrifugation and the supernatant removed after each incubation, wash, etc. This imposes absolute physical limitations on the process, as pellets of agarose beads less than 25 to 50µl are difficult if not impossible to visually identify at the bottom of the tube. With magnetic beads, there is no minimum quantity of beads required due to magnetic handling, and therefore, depending on the target antigen and IP antibody, it is possible to use considerably less magnetic beads.

Conversely, spin columns may be employed instead of normal microfuge tubes to significantly reduce the amount of agarose beads required per reaction. Spin columns contain a filter that allows all IP components except the beads to flow through using a brief centrifugation and therefore provide a method to use significantly less agarose beads with minimal loss.

Equipment

As mentioned above, only standard laboratory equipment is required for the use of agarose beads in immunoprecipitation applications, while high-power magnets are required for magnetic bead-based IP reactions. While the magnetic capture equipment may be cost-prohibitive, the rapid completion of immunoprecipitations using magnetic beads may be a financially beneficial approach when grants are due, because a 30 minute protocol with magnetic beads compared to overnight incubation at 4°C with agarose beads may result in more data generated in a shorter length of time.

Automation

An added benefit of using magnetic beads is that automated immunoprecipitation devices are becoming more readily available. These devices not only reduce the amount of work and time to perform an IP, but they can also be used for high-throughput applications.

Summary

While clear benefits of using magnetic beads include the increased reaction speed, more gentle sample handling and the potential for automation, the choice of using agarose or magnetic beads based on the binding capacity of the support medium and the cost of the product may depend on the protein of interest and the IP method used. As with all assays, empirical testing is required to determine which method is optimal for a given application.

Protocol

Background

Once the solid substrate bead technology has been chosen, antibodies are coupled to the beads and the antibody-coated-beads can be added to the heterogeneous protein sample (e.g. homogenized tissue). At this point, antibodies that are immobilized to the beads will bind to the proteins that they specifically recognize. Once this has occurred the immunoprecipitation portion of the protocol is actually complete, as the specific proteins of interest are bound to the antibodies that are themselves immobilized to the beads. Separation of the immunocomplexes from the lysate is an extremely important series of steps, because the protein(s) must remain bound to each other (in the case of co-IP) AND bound to the antibody during the wash steps to remove non-bound proteins and reduce background.

When working with agarose beads, the beads must be pelleted out of the sample by briefly spinning in a centrifuge with forces between 600-3,000 x g (times the standard gravitational force). This step may be performed in a standard microcentrifuge tube, but for faster separation, greater consistency and higher recoveries, the process is often performed in small spin columns with a pore size that allows liquid, but not agarose beads, to pass through. After centrifugation, the agarose beads will form a very loose

fluffy pellet at the bottom of the tube. The supernatant containing contaminants can be carefully removed so as not to disturb the beads. The wash buffer can then be added to the beads and after mixing, the beads are again separated by centrifugation.

With superparamagnetic beads, the sample is placed in a magnetic field so that the beads can collect on the side of the tube. This procedure is generally complete in approximately 30 seconds, and the remaining (unwanted) liquid is pipetted away. Washes are accomplished by resuspending the beads (off the magnet) with the washing solution and then concentrating the beads back on the tube wall (by placing the tube back on the magnet). The washing is generally repeated several times to ensure adequate removal of contaminants. If the superparamagnetic beads are homogeneous in size and the magnet has been designed properly, the beads will concentrate uniformly on the side of the tube and the washing solution can be easily and completely removed.

After washing, the precipitated protein(s) are eluted and analyzed by gel electrophoresis, mass spectrometry, western blotting, or any number of other methods for identifying constituents in the complex. Protocol times for immunoprecipitation vary greatly due to a variety of factors, with protocol times increasing with the number of washes necessary or with the slower reaction kinetics of porous agarose beads.

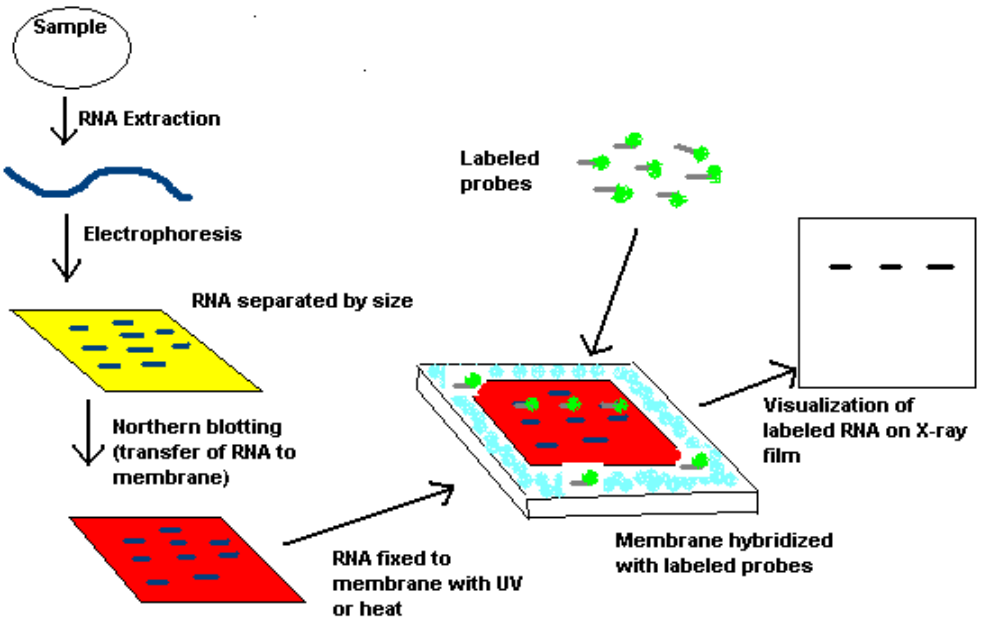
Steps

1. Lyse cells and prepare sample for immunoprecipitation.
2. Pre-clear the sample by passing the sample over beads alone or bound to an irrelevant antibody to soak up any proteins that non-specifically bind to the IP components.
3. Incubate solution with antibody against the protein of interest. Antibody can be attached to solid support before this step (direct method) or after this step (indirect method). Continue the incubation to allow antibody-antigen complexes to form.
4. Precipitate the complex of interest, removing it from bulk solution.
5. Wash precipitated complex several times. Spin each time between washes when using agarose beads or place tube on magnet when using superparamagnetic beads and then remove the supernatant. After the final wash, remove as much supernatant as possible.
6. Elute proteins from the solid support using low-pH or SDS sample loading buffer.
7. Analyze complexes or antigens of interest. This can be done in a variety of ways:
    1. SDS-PAGE (sodium dodecyl sulfate-polyacrylamide gel electrophoresis) followed by gel staining.
    2. SDS-PAGE followed by: gel staining, cutting out individual stained protein bands, and sequencing the proteins in the bands by MALDI-Mass Spectrometry
    3. Transfer and Western Blot using another antibody for proteins that were interacting with the antigen followed by chemiluminesent visualization.

Chapter- 9

Northern Blot

The **northern blot** is a technique used in molecular biology research to study gene expression by detection of RNA (or isolated mRNA) in a sample.



Flow diagram outlining the general procedure for RNA detection by northern blotting.

With northern blotting it is possible to observe cellular control over structure and function by determining the particular gene expression levels during differentiation, morphogenesis, as well as abnormal or diseased conditions. Northern blotting involves the use of electrophoresis to separate RNA samples by size, and detection with a hybridization probe complementary to part of or the entire target sequence. The term 'northern blot' actually refers specifically to the capillary transfer of RNA from the electrophoresis gel to the blotting membrane, however the entire process is commonly

referred to as northern blotting. The northern blot technique was developed in 1977 by James Alwine, David Kemp, and George Stark at Stanford University. Northern blotting takes its name from its similarity to the first blotting technique, the Southern blot, named for biologist Edwin Southern. The major difference is that RNA, rather than DNA, is analyzed in the Northern blot.

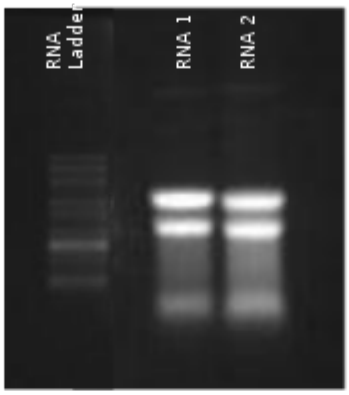
Procedure

A general blotting procedure starts with extraction of total RNA from a homogenized tissue sample. The mRNA can then be isolated through the use of oligo (dT) cellulose chromatography to maintain only those RNAs with a poly(A) tail. RNA samples are then separated by gel electrophoresis. Since the gels are fragile and the probes are unable to enter the matrix, the RNA samples, now separated by size, are transferred to a nylon membrane through a capillary or vacuum blotting system.

Capillary blotting system setup for the transfer of RNA from an electrophoresis gel to a blotting membrane.

A nylon membrane with a positive charge is the most effective for use in northern blotting since the negatively charged nucleic acids have a high affinity for them. The transfer buffer used for the blotting usually contains formamide because it lowers the annealing temperature of the probe-RNA interaction preventing RNA degradation by high temperatures. Once the RNA has been transferred to the membrane it is immobilized through covalent linkage to the membrane by UV light or heat. After a probe has been labeled, it is hybridized to the RNA on the membrane. Experimental conditions that can affect the efficiency and specificity of hybridization include ionic strength, viscosity, duplex length, mismatched base pairs, and base composition. The membrane is washed to ensure that the probe has bound specifically and to avoid background signals from arising. The hybrid signals are then detected by X-ray film and can be quantified by densitometry. To create controls for comparison in a northern blot, samples not displaying the gene product of interest can be used after determination by microarrays or RT-PCR.

Gels



Formaldehyde gel (1%) with RNA samples run at 100V for 1 hour in 1x MOPS buffe

RNA run on a formaldehyde agarose gel to highlight the 28S (top band) and 18S (lower band) ribosomal subunits.

The RNA samples are most commonly separated on agarose gels containing formaldehyde as a denaturing agent for the RNA to limit secondary structure. The gels can be stained with ethidium bromide (EtBr) and viewed under UV light to observe the quality and quantity of RNA before blotting. Polyacrylamide gel electrophoeresis with urea can also be used in RNA separation but it is most commonly used for fragmented RNA or microRNAs. An RNA ladder is often run alongside the samples on an electrophoresis gel to observe the size of fragments obtained but in total RNA samples the ribosomal subunits can act as size markers. Since the large ribosomal subunit is 28S (approximately 5kb) and the small ribosomal subunit is 18S (approximately 2kb) two

prominent bands will appear on the gel, the larger at close to twice the intensity of the smaller.

Probes

Probes for northern blotting are composed of nucleic acids with a complementary sequence to all or part of the RNA of interest, they can be DNA, RNA, or oligonucleotides with a minimum of 25 complementary bases to the target sequence. RNA probes (riboprobes) that are transcribed in vitro are able to withstand more rigorous washing steps preventing some of the background noise. Commonly cDNA is created with labelled primers for the RNA sequence of interest to act as the probe in the northern blot. The probes need to be labelled either with radioactive isotopes (^{32}P) or with chemiluminescence in which alkaline phosphatase or horseradish peroxidase breakdown chemiluminescent substrates producing a detectable emission of light. The chemiluminescent labelling can occur in two ways: either the probe is attached to the enzyme, or the probe is labelled with a ligand (e.g. biotin) for which the antibody (e.g. avidin or streptavidin) is attached to the enzyme. X-ray film can detect both the radioactive and chemiluminescent signals and many researchers prefer the chemiluminescent signals because they are faster, more sensitive, and reduce the health hazards that go along with radioactive labels. The same membrane can be probed up to five times without a significant loss of the target RNA.

Applications

Northern blotting allows one to observe a particular gene's expression pattern between tissues, organs, developmental stages, environmental stress levels, pathogen infection, and over the course of treatment. The technique has been used to show overexpression of oncogenes and downregulation of tumor-suppressor genes in cancerous cells when compared to 'normal' tissue, as well as the gene expression in the rejection of transplanted organs. If an upregulated gene is observed by an abundance of mRNA on the northern blot the sample can then be sequenced to determine if the gene is known to researchers or if it is a novel finding. The expression patterns obtained under given conditions can provide insight into the function of that gene. Since the RNA is first separated by size, if only one probe type is used variance in the level of each band on the membrane can provide insight into the size of the product, suggesting alternative splice products of the same gene or repetitive sequence motifs. The variance in size of a gene product can also indicate deletions or errors in transcript processing, by altering the probe target used along the known sequence it is possible to determine which region of the RNA is missing.

BlotBase is an online database publishing northern blots. BlotBase has over 700 published northern blots of human and mouse samples, in over 650 genes across more than 25 different tissue types. Northern blots can be searched by a blot ID, paper reference, gene identifier, or by tissue. The results of a search provide the blot ID, species, tissue, gene, expression level, blot image (if available), and links to the publication that the work originated from. This new database provides sharing of

information between members of the science community that was not previously seen in northern blotting as it was in sequence analysis, genome determination, protein structure, etc.

Disadvantages and Advantages

Analysis of gene expression can be done by several different methods including RT-PCR, RNase protection assays, microarrays, serial analysis of gene expression (SAGE), as well as northern blotting. Microarrays are quite commonly used and are usually consistent with data obtained from northern blots; however, at times northern blotting is able to detect small changes in gene expression that microarrays cannot. The advantage that microarrays have over northern blots is that thousands of genes can be visualized at a time, while northern blotting is usually looking at one or a small number of genes.

A problem in northern blotting is often sample degradation by RNases (both endogenous to the sample and through environmental contamination), which can be avoided by proper sterilization of glassware and the use of RNase inhibitors such as DEPC (diethylpyrocarbonate). The chemicals used in most northern blots can be a risk to the researcher, since formaldehyde, radioactive material, ethidium bromide, DEPC, and UV light are all harmful under certain exposures. Compared to RT-PCR, northern blotting has a low sensitivity, but it also has a high specificity which is important to reduce false positive results.

The advantages of using northern blotting include the detection of RNA size, the observation of alternate splice products, the use of probes with partial homology, the quality and quantity of RNA can be measured on the gel prior to blotting, and the membranes can be stored and reprobed for years after blotting.

Reverse northern blot

A variant of the procedure known as the reverse northern blot is occasionally used. In this procedure, the substrate nucleic acid (that is affixed to the membrane) is a collection of isolated DNA fragments, and the probe is RNA extracted from a tissue and radioactively labelled.

The use of DNA microarrays that have come into widespread use in the late 1990s and early 2000s is more akin to the reverse procedure, in that they involve the use of isolated DNA fragments affixed to a substrate, and hybridization with a probe made from cellular RNA. Thus the reverse procedure, though originally uncommon, enabled northern analysis to evolve into gene expression profiling, in which many (possibly all) of the genes in an organism may have their expression monitored.

Chapter- 10

Nucleic Acid Structure Determination

Structure probing of nucleic acids is the process by which biochemical techniques are used to determine nucleic acid structure. This analysis can be used to define the patterns which can infer the molecular structure, experimental analysis of molecular structure and function, and further understanding on development of smaller molecules for further biological research. Structure probing analysis can be done through many different methods, which include chemical probing, hydroxyl radical probing, SHAPE, nucleotide analog interference mapping (NAIM), and in-line probing.

RNA sequencing

While methods for each type of probing differ in steps, the probing of secondary structure involves certain steps in order to determine the structure. The first step involved is submitting the structural RNA to the probe of interest and incubating over a certain amount of time to allow the reaction to occur. The RNA is then transcribed using reverse transcriptase PCR, where this results in different lengths of bands due to the modification of the RNA at specific sites, which causes the reverse transcriptase to fall off. These bands are then run on a gel with sequencing data determined from RNA sequencing.

Chemical probing

RNA chemical probing can involve many different chemicals which serve to modify specific bases at certain sites to show certain locations available for specific modification type.

DMS

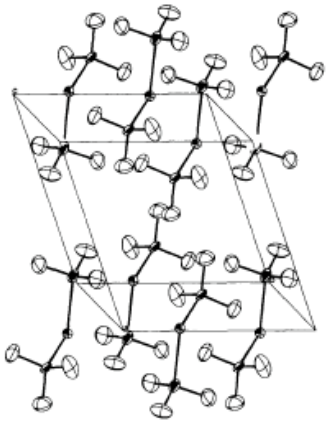


Figure 1. Structure of DMS showing the two methyl groups indicated as circles connected to sulfate.

Dimethyl sulfate, known as DMS, is a chemical that can be used to modify nucleic acids in order to determine secondary structure. DMS modifies certain bases through methylation. One set of methylation products that is used for RNA is the methylation of N1 adenosine and N3 of cytosine which prevents the natural hydrogen bonds to form between bases modified by DMS. This enables modification sites to be detected by RT-PCR, as modified sites cannot be basepaired, which results in the reverse transcriptase to fall off and different band sizes. DMS can only modify bases that are not basepaired, as single stranded nucleotides have the bases exposed so the chemical can modify. Detection of these modifications is done by examination of a gel and looking at bands present. One thing of note is that since modification prevents the addition of nucleotides at the modification site each of the bands generated is shifted by one base down on the gel. The PCR products indicate which adenine and cytosine nucleotides are double stranded, that is protected from DMS modification, or single stranded, that is accessible to DMS modification. This can be used to begin to develop a model of secondary structure for the RNA molecule. Structure probing by DMS allows for detection of secondary structure changes due to binding of RNA molecules along with detecting changes in tertiary structure, but it lacks the ability to determine the origin of these changes, as well as protection of the RNA backbone.

DMS modification can also be used for DNA, for example in footprinting DNA-protein interactions.

CMCT

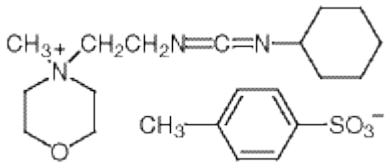


Figure 2. Structure of CMCT used in RNA structure probing

1-cyclohexyl-(2-morpholinoethyl)carbodiimide metho-p-toluene sulfonate known as CMCT is another chemical that is used in structure probing. CMCT like DMS serves to modify the exposed bases of specific nucleotides, which are uridine, and to a smaller extent guanine. CMCT reacts primarily with N3 of uridine and N1 of guanine modifying two sites responsible for hydrogen bonding on the bases. Modification using CMCT is analogous in detection to DMS as modification prevents basepairing at specific nucleotides which are then detected using rt-PCR and running of the PCR products on an agarose gel as shown in Figure 1. Structure probing of RNA by CMCT indicates the presence of uridine and guanine in single stranded regions by accessibility to modification or the presence of uridine and guanine in double stranded regions by protection from CMCT, which is the absence of a band. CMCT, like DMS can detect secondary and tertiary structure changes, but still has the same weaknesses as the method of modification is the same as DMS.

Kethoxal

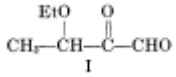


Figure 3. Structure of kethoxal that attacks guanine in structure probing.

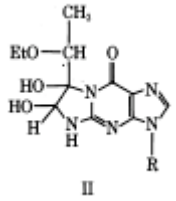


Figure 4. Binding of kethoxal to modified guanine preventing basepairing.

1,1-Dihydroxy-3-ethoxy-2-butanone, known as kethoxal, is used like DMS and CMCT, where treatment with kethoxal causes the modification of guanine, specifically altering the N1 and N2 by covalent interaction. The modification by kethoxal prevents single stranded guanine nucleotides to become modified, so that when reverse transcriptase reaches the modified guanine it falls off resulting in a band. After rt-PCR the products will be of different lengths where a band present indicates modification of a guanine base, and the absence of a band indicates that the base was not available for modification. Using this gel in combination with DMS and CMCT a model for structural RNAs can be formed through comparison between the gels which indicate protected and accessible positions in all the bases, which can be used to form a preliminary model.

SHAPE

Selective 2′-**h**ydroxyl **a**cylation analyzed by **p**rimer **e**xtension, or **SHAPE**, takes advantage of reagents that preferentially modify the backbone of RNA in structurally flexible regions.

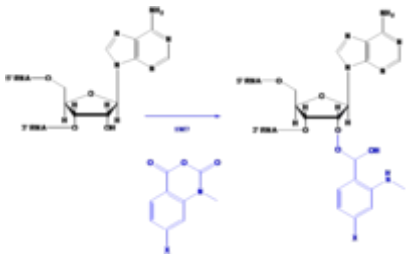


Figure 5. 1-methyl-7-nitroisatoic anhydride (1M7) undergoes hydrolysis to form adducts on the backbone of unpaired RNA nucleotides.

Reagents such as N-methylisotoic anhydride (NMIA) and 1-methyl-7-nitroisatoic anhydride (1M7) undergo hydrolysis to form adducts on the 2'-hydroxyl of the RNA backbone. Compared to the chemicals used in other RNA probing techniques, these reagents have the advantage of being largely unbiased to base identity, while remaining very sensitive to conformational dynamics. Nucleotides which are constrained (usually by base-pairing) show less adduct formation than nucleotides which are unpaired. Adduct formation is quantified for each nucleotide in a given RNA by extension of a complementary DNA primer with reverse transcriptase and comparison of the resulting fragments with those from an unmodified control. SHAPE therefore reports on RNA structure at the individual nucleotide level. This data can be used as input to generate highly accurate secondary structure models. SHAPE has been used to analyze diverse RNA structures, including that of an entire HIV-1 genome.

Hydroxyl radical probing

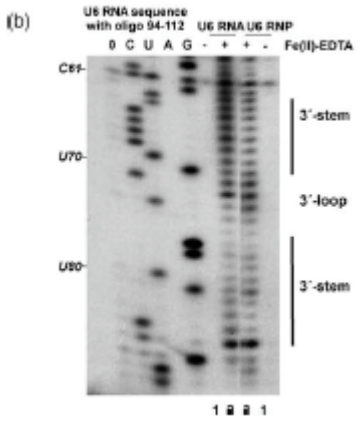


Figure 6. Hydroxyl radical probing gel showing bands at positions and dots indicating strength of protection.

Probing with hydroxyl radicals involves an additional step, as hydroxyl radicals are short lived in solution they need to be generated. This can be done using H_2O_2, ascorbic acid, and Fe(II)-EDTA complex which is attached to the backbone through EDTA. The Fe(II) along with ascorbic acid generates hydroxyl radicals which then can react with the nucleic acid molecules. Hydroxyl radical probing is often used in conjunction with chemical probing of nucleic acid molecules that are thought to associate with proteins,

this is due to the modification done by hydroxyl radicals. Hydroxyl radicals attack the ribose/deoxyribose ring and this results in breaking of the phosphate backbone, which is independent of secondary structure, as all backbone is accessible, but is instead resultant of protein or tertiary structure protection. Probing with hydroxyl radicals shows the protection of structured nucleic acids by the proteins thought to be associated or folding on itself, where cleavage again results in a band formed through gel electrophoresis (after rt-PCR in the case of RNA) that is shorter than the full nucleic acid depending upon where it is cleaved. In this case since the last nucleotide is not modified the band length is indicative of the base that was cleaved. When examining the gel produced by running the gels on a band the areas of various strength of protection where areas of stronger protection for hydroxyl radicals can be said to have tighter association with a protein, or if no protein associates with the nucleic acid it can be caused by the tertiary fold.

In-line probing

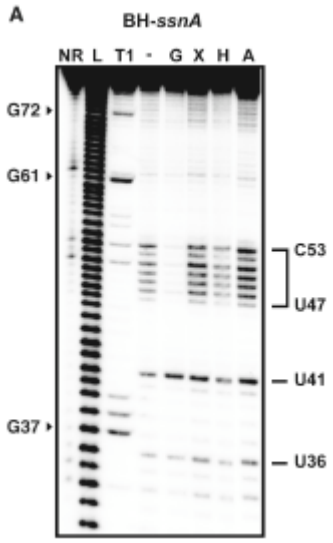


Figure 7. In-line probing assay of guanine riboswitches showing change in flexibility in response to various nucleotide ligands

In-line probing does not involve treatment with any type of chemicals or reagents to modify exposed RNA structures. This type of probing assay uses the structure dependent cleavage of RNA, as areas that are single stranded are more flexible and over time will degrade, as RNA structure is not always stable. The process of in-line probing is often used to determine changes in structure due to ligand binding as this can result in different cleavage patterns. The process of in-line probing involves incubation of structural or functional RNAs over a long period of time, can be several days, but varies in each experiment, and then running the incubated products on a gel to visualize the bands. This experiment is often done using two different conditions: 1) with ligand and 2) in the absence of ligand. Cleavage results in shorter band lengths and is indicative of areas that are not basepaired, as those that are tend to be less sensitive to spontaneous cleavage. In line probing is a functional assay in ligand binding changes of RNA because it can show directly the change in flexibility and binding of regions of DNA in response to a ligand,

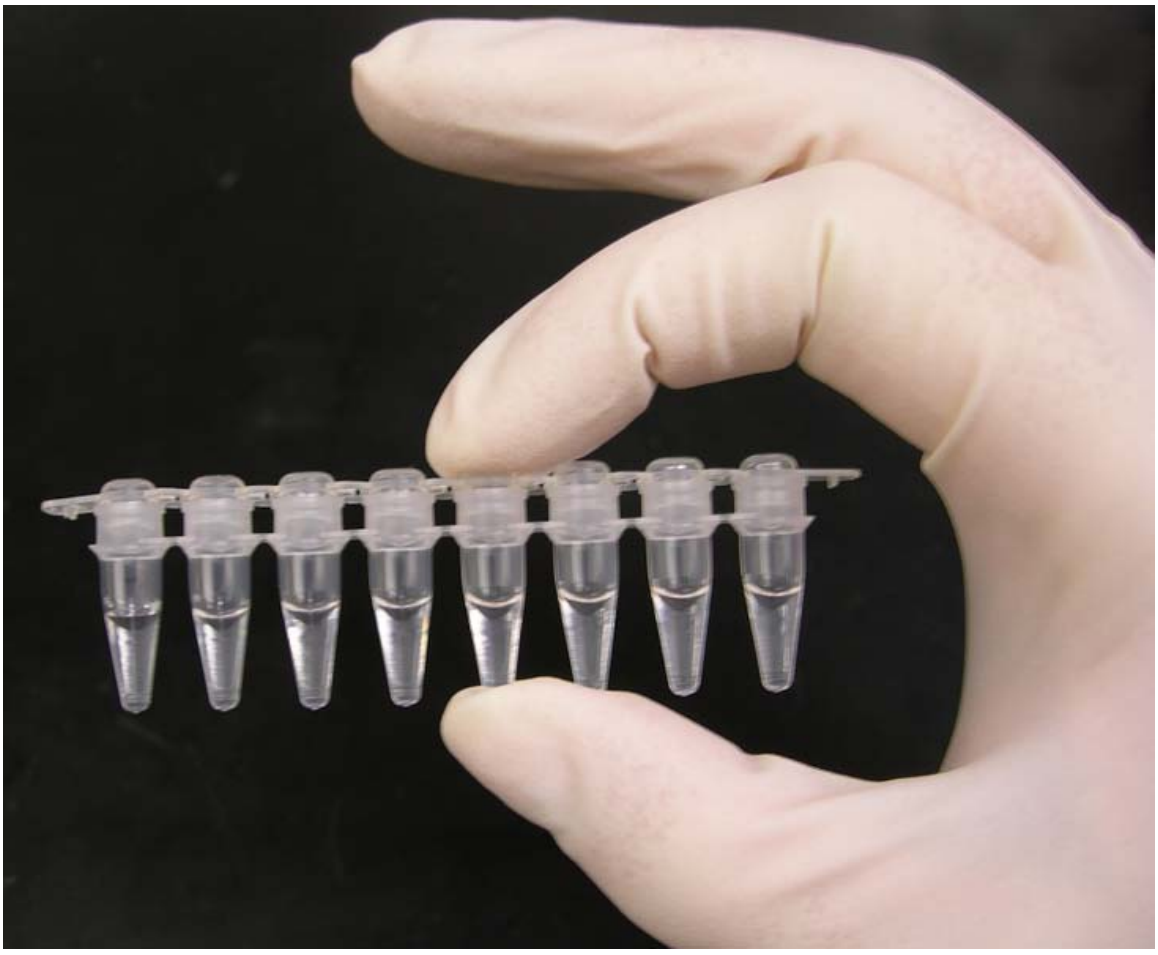
as well as compare that response to analogous ligands. An in-line probing assay is then commonly used in dynamic studies, specifically when examining riboswitches

Nucleotide analog interference mapping

Nucleotide analog interference mapping (NAIM) is the process of using nucleotide analogs, molecules that are similar in some ways to nucleotides but lack function, to determine the importance of a functional group at each location of an RNA molecule. The process of NAIM is to insert a single nucleotide analog into a unique site. This can be done by transcribing a short RNA using T7 RNA polymerase, then synthesizing a short oligonucleotide containing the analog in a specific position, then ligating them together on the DNA template using a ligase. The nucleotide analogs are tagged with a phosphorothioate, the active members of the RNA population are then distinguished from the inactive members, the inactive members then have the phophorothioate tag removed and the analog sites are identified using gel electrophoresis and autoradiography. This indicates a functionally important nucleotides, as cleavage of the phosphorothioate by iodine results in an RNA that is cleaved at the site of the nucleotide analog insert. By running these truncated RNA molecules on a gel, the nucleotide of interest can be identified against a sequencing experiment Site directed incorporation results indicate positions of importance where when running on a gel, functional RNAs that have the analog incorporated at that position will have a band present, but if the analog results in non-functionality, when the functional RNA molecules are run on a gel there will be no band corresponding to that position on the gel. This process can be used to evaluate an entire area, where analogs are placed in site specific locations, differing by a single nucleotide, then when functional RNAs are isolated and run on a gel, all areas where bands are produced indicate non-essential nucleotides, but areas where bands are absent from the functional RNA indicate that inserting a nucleotide analog in that position caused the RNA molecule to become non-functional

Chapter- 11

Polymerase Chain Reaction



A strip of eight PCR tubes, each containing a 100 µl reaction mixture

The **polymerase chain reaction** (**PCR**) is a scientific technique in molecular biology to amplify a single or a few copies of a piece of DNA across several orders of magnitude, generating thousands to millions of copies of a particular DNA sequence.

Developed in 1983 by Kary Mullis, PCR is now a common and often indispensable technique used in medical and biological research labs for a variety of applications. These include DNA cloning for sequencing, DNA-based phylogeny, or functional analysis of genes; the diagnosis of hereditary diseases; the identification of genetic fingerprints (used in forensic sciences and paternity testing); and the detection and diagnosis of infectious diseases. In 1993, Mullis was awarded the Nobel Prize in Chemistry for his work on PCR.

The method relies on thermal cycling, consisting of cycles of repeated heating and cooling of the reaction for DNA melting and enzymatic replication of the DNA. Primers (short DNA fragments) containing sequences complementary to the target region along with a DNA polymerase (after which the method is named) are key components to enable selective and repeated amplification. As PCR progresses, the DNA generated is itself used as a template for replication, setting in motion a chain reaction in which the DNA template is exponentially amplified. PCR can be extensively modified to perform a wide array of genetic manipulations.

Almost all PCR applications employ a heat-stable DNA polymerase, such as Taq polymerase, an enzyme originally isolated from the bacterium *Thermus aquaticus*. This DNA polymerase enzymatically assembles a new DNA strand from DNA building blocks, the nucleotides, by using single-stranded DNA as a template and DNA oligonucleotides (also called DNA primers), which are required for initiation of DNA synthesis. The vast majority of PCR methods use thermal cycling, i.e., alternately heating and cooling the PCR sample to a defined series of temperature steps. These thermal cycling steps are necessary first to physically separate the two strands in a DNA double helix at a high temperature in a process called DNA melting. At a lower temperature, each strand is then used as the template in DNA synthesis by the DNA polymerase to selectively amplify the target DNA. The selectivity of PCR results from the use of primers that are complementary to the DNA region targeted for amplification under specific thermal cycling conditions.

PCR principles and procedure

Figure 1a: A thermal cycler for PCR

Figure 1b: An older model three-temperature thermal cycler for PCR

PCR is used to amplify a specific region of a DNA strand (the DNA target). Most PCR methods typically amplify DNA fragments of up to ~10 kilo base pairs (kb), although some techniques allow for amplification of fragments up to 40 kb in size.

A basic PCR set up requires several components and reagents. These components include:

- *DNA template* that contains the DNA region (target) to be amplified.
- Two *primers* that are complementary to the 3' (three prime) ends of each of the sense and anti-sense strand of the DNA target.

- *Taq polymerase* or another DNA polymerase with a temperature optimum at around 70 °C.
- *Deoxynucleotide triphosphates* (dNTPs), the building blocks from which the DNA polymerase synthesizes a new DNA strand.
- *Buffer solution,* providing a suitable chemical environment for optimum activity and stability of the DNA polymerase.
- *Divalent cations,* magnesium or manganese ions; generally Mg^{2+} is used, but Mn^{2+} can be utilized for PCR-mediated DNA mutagenesis, as higher Mn^{2+} concentration increases the error rate during DNA synthesis
- *Monovalent cation* potassium ions.

The PCR is commonly carried out in a reaction volume of 10–200 µl in small reaction tubes (0.2–0.5 ml volumes) in a thermal cycler. The thermal cycler heats and cools the reaction tubes to achieve the temperatures required at each step of the reaction. Many modern thermal cyclers make use of the Peltier effect which permits both heating and cooling of the block holding the PCR tubes simply by reversing the electric current. Thin-walled reaction tubes permit favorable thermal conductivity to allow for rapid thermal equilibration. Most thermal cyclers have heated lids to prevent condensation at the top of the reaction tube. Older thermocyclers lacking a heated lid require a layer of oil on top of the reaction mixture or a ball of wax inside the tube.

Procedure

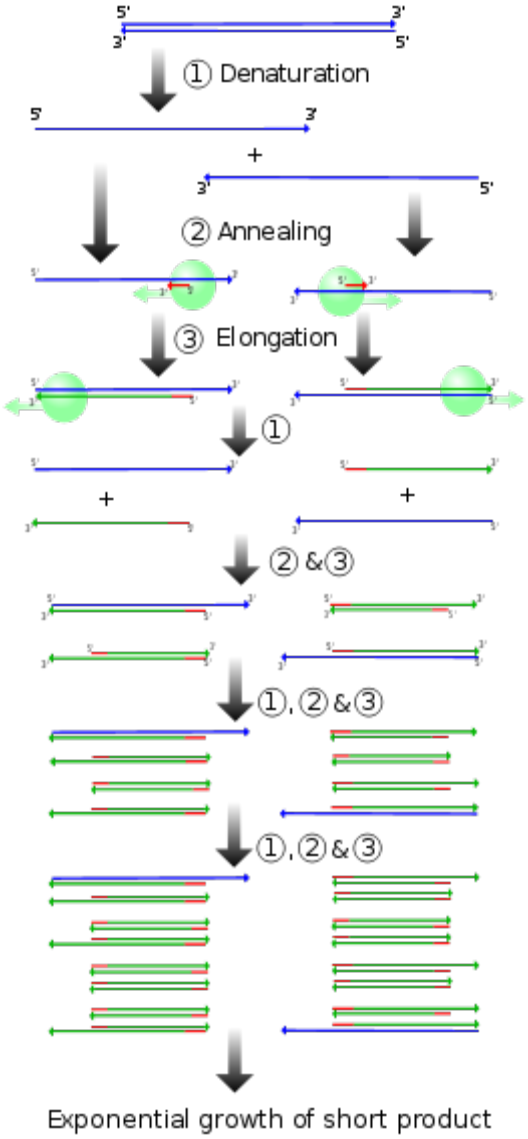


Figure 2: Schematic drawing of the PCR cycle. **(1) Denaturing at 94–96 °C. (2) Annealing at ~65 °C (3) Elongation at 72 °C**. Four cycles are shown here. The blue lines represent the DNA template to which primers (red arrows) anneal that are extended by the DNA polymerase (light green circles), to give shorter DNA products (green lines), which themselves are used as templates as PCR progresses.

Typically, PCR consists of a series of 20-40 repeated temperature changes, called cycles, with each cycle commonly consisting of 2-3 discrete temperature steps, usually three (Fig. 2). The cycling is often preceded by a single temperature step (called *hold*) at a high temperature (>90°C), and followed by one hold at the end for final product extension or brief storage. The temperatures used and the length of time they are applied in each cycle depend on a variety of parameters. These include the enzyme used for DNA synthesis,

the concentration of divalent ions and dNTPs in the reaction, and the melting temperature (Tm) of the primers.

- *Initialization step*: This step consists of heating the reaction to a temperature of 94–96 °C (or 98 °C if extremely thermostable polymerases are used), which is held for 1–9 minutes. It is only required for DNA polymerases that require heat activation by hot-start PCR.

- *Denaturation step*: This step is the first regular cycling event and consists of heating the reaction to 94–98 °C for 20–30 seconds. It causes DNA melting of the DNA template by disrupting the hydrogen bonds between complementary bases, yielding single-stranded DNA molecules.

- *Annealing step*: The reaction temperature is lowered to 50–65 °C for 20–40 seconds allowing annealing of the primers to the single-stranded DNA template. Typically the annealing temperature is about 3-5 degrees Celsius below the Tm of the primers used. Stable DNA-DNA hydrogen bonds are only formed when the primer sequence very closely matches the template sequence. The polymerase binds to the primer-template hybrid and begins DNA synthesis.

- *Extension/elongation step*: The temperature at this step depends on the DNA polymerase used; Taq polymerase has its optimum activity temperature at 75–80 °C, and commonly a temperature of 72 °C is used with this enzyme. At this step the DNA polymerase synthesizes a new DNA strand complementary to the DNA template strand by adding dNTPs that are complementary to the template in 5' to 3' direction, condensing the 5'-phosphate group of the dNTPs with the 3'-hydroxyl group at the end of the nascent (extending) DNA strand. The extension time depends both on the DNA polymerase used and on the length of the DNA fragment to be amplified. As a rule-of-thumb, at its optimum temperature, the DNA polymerase will polymerize a thousand bases per minute. Under optimum conditions, i.e., if there are no limitations due to limiting substrates or reagents, at each extension step, the amount of DNA target is doubled, leading to exponential (geometric) amplification of the specific DNA fragment.

- *Final elongation*: This single step is occasionally performed at a temperature of 70–74 °C for 5–15 minutes after the last PCR cycle to ensure that any remaining single-stranded DNA is fully extended.

- *Final hold*: This step at 4–15 °C for an indefinite time may be employed for short-term storage of the reaction.

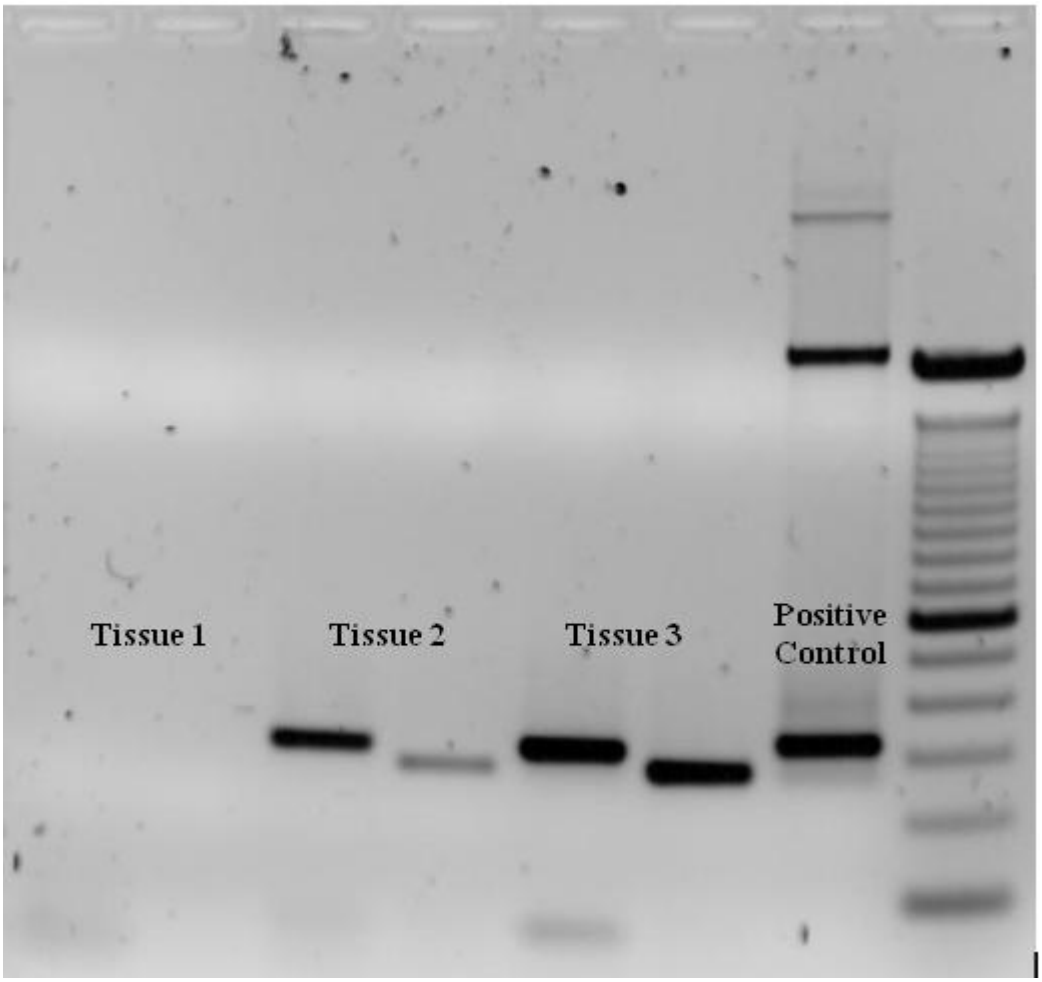


Figure 3: Ethidium bromide-stained PCR products after gel electrophoresis. Two sets of primers were used to amplify a target sequence from three different tissue samples. No amplification is present in sample #1; DNA bands in sample #2 and #3 indicate successful amplification of the target sequence. The gel also shows a positive control, and a DNA ladder containing DNA fragments of defined length for sizing the bands in the experimental PCRs.

To check whether the PCR generated the anticipated DNA fragment (also sometimes referred to as the amplimer or amplicon), agarose gel electrophoresis is employed for size separation of the PCR products. The size(s) of PCR products is determined by comparison with a DNA ladder (a molecular weight marker), which contains DNA fragments of known size, run on the gel alongside the PCR products.

PCR stages

The PCR process can be divided into three stages:

Exponential amplification: At every cycle, the amount of product is doubled (assuming 100% reaction efficiency). The reaction is very sensitive: only minute quantities of DNA need to be present.

Levelling off stage: The reaction slows as the DNA polymerase loses activity and as consumption of reagents such as dNTPs and primers causes them to become limiting.

Plateau: No more product accumulates due to exhaustion of reagents and enzyme.

PCR optimization

In practice, PCR can fail for various reasons, in part due to its sensitivity to contamination causing amplification of spurious DNA products. Because of this, a number of techniques and procedures have been developed for optimizing PCR conditions. Contamination with extraneous DNA is addressed with lab protocols and procedures that separate pre-PCR mixtures from potential DNA contaminants. This usually involves spatial separation of PCR-setup areas from areas for analysis or purification of PCR products, use of disposable plasticware, and thoroughly cleaning the work surface between reaction setups. Primer-design techniques are important in improving PCR product yield and in avoiding the formation of spurious products, and the usage of alternate buffer components or polymerase enzymes can help with amplification of long or otherwise problematic regions of DNA. Addition of reagents, such as formamide, in buffer systems may increase the specificity and yield of PCR.

Application of PCR

Selective DNA isolation

PCR allows isolation of DNA fragments from genomic DNA by selective amplification of a specific region of DNA. This use of PCR augments many methods, such as generating hybridization probes for Southern or northern hybridization and DNA cloning, which require larger amounts of DNA, representing a specific DNA region. PCR supplies these techniques with high amounts of pure DNA, enabling analysis of DNA samples even from very small amounts of starting material.

Other applications of PCR include DNA sequencing to determine unknown PCR-amplified sequences in which one of the amplification primers may be used in Sanger sequencing, isolation of a DNA sequence to expedite recombinant DNA technologies involving the insertion of a DNA sequence into a plasmid or the genetic material of another organism. Bacterial colonies (E. coli) can be rapidly screened by PCR for correct DNA vector constructs. PCR may also be used for genetic fingerprinting; a forensic technique used to identify a person or organism by comparing experimental DNAs through different PCR-based methods.

Some PCR 'fingerprints' methods have high discriminative power and can be used to identify genetic relationships between individuals, such as parent-child or between

siblings, and are used in paternity testing (Fig. 4). This technique may also be used to determine evolutionary relationships among organisms.

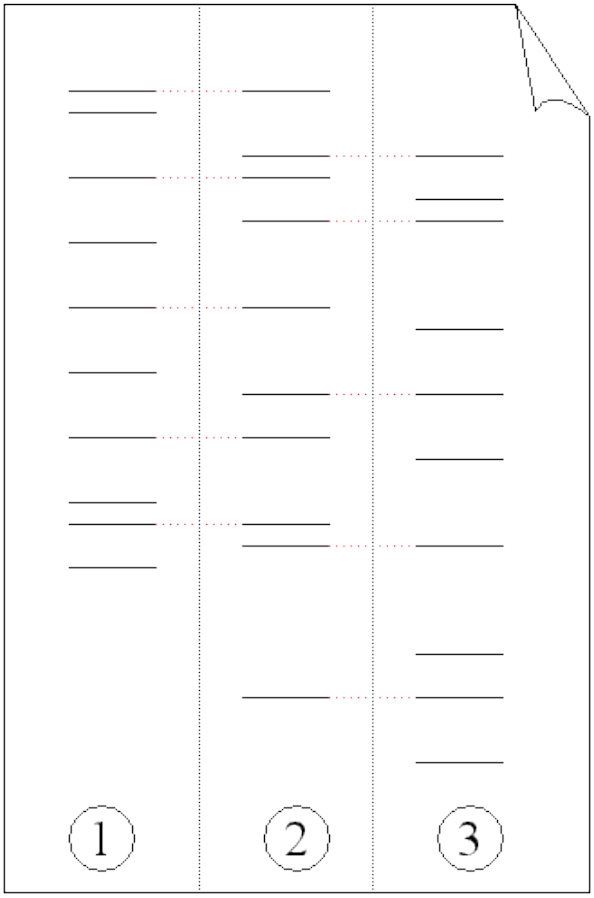


Figure 4: Electrophoresis of PCR-amplified DNA fragments. (1) Father. (2) Child. (3) Mother. The child has inherited some, but not all of the fingerprint of each of its parents, giving it a new, unique fingerprint.

Amplification and quantification of DNA

Because PCR amplifies the regions of DNA that it targets, PCR can be used to analyze extremely small amounts of sample. This is often critical for forensic analysis, when only a trace amount of DNA is available as evidence. PCR may also be used in the analysis of ancient DNA that is tens of thousands of years old. These PCR-based techniques have been successfully used on animals, such as a forty-thousand-year-old mammoth, and also on human DNA, in applications ranging from the analysis of Egyptian mummies to the identification of a Russian tsar.

Quantitative PCR methods allow the estimation of the amount of a given sequence present in a sample—a technique often applied to quantitatively determine levels of gene expression. Real-time PCR is an established tool for DNA quantification that measures the accumulation of DNA product after each round of PCR amplification.

PCR in diagnosis of diseases

PCR permits early diagnosis of malignant diseases such as leukemia and lymphomas, which is currently the highest developed in cancer research and is already being used routinely. PCR assays can be performed directly on genomic DNA samples to detect translocation-specific malignant cells at a sensitivity which is at least 10,000 fold higher than other methods.

PCR also permits identification of non-cultivatable or slow-growing microorganisms such as mycobacteria, anaerobic bacteria, or viruses from tissue culture assays and animal models. The basis for PCR diagnostic applications in microbiology is the detection of infectious agents and the discrimination of non-pathogenic from pathogenic strains by virtue of specific genes.

Viral DNA can likewise be detected by PCR. The primers used need to be specific to the targeted sequences in the DNA of a virus, and the PCR can be used for diagnostic analyses or DNA sequencing of the viral genome. The high sensitivity of PCR permits virus detection soon after infection and even before the onset of disease. Such early detection may give physicians a significant lead in treatment. The amount of virus ("viral load") in a patient can also be quantified by PCR-based DNA quantitation techniques.

Variations on the basic PCR technique

- *Allele-specific PCR*: a diagnostic or cloning technique which is based on single-nucleotide polymorphisms (SNPs) (single-base differences in DNA). It requires prior knowledge of a DNA sequence, including differences between alleles, and uses primers whose 3' ends encompass the SNP. PCR amplification under stringent conditions is much less efficient in the presence of a mismatch between template and primer, so successful amplification with an SNP-specific primer signals presence of the specific SNP in a sequence.

- *Assembly PCR* or *Polymerase Cycling Assembly (PCA)*: artificial synthesis of long DNA sequences by performing PCR on a pool of long oligonucleotides with short overlapping segments. The oligonucleotides alternate between sense and antisense directions, and the overlapping segments determine the order of the PCR fragments, thereby selectively producing the final long DNA product.

- *Asymmetric PCR*: preferentially amplifies one DNA strand in a double-stranded DNA template. It is used in sequencing and hybridization probing where amplification of only one of the two complementary strands is required. PCR is carried out as usual, but with a great excess of the primer for the strand targeted for amplification. Because of the slow (arithmetic) amplification later in the reaction after the limiting primer has been used up, extra cycles of PCR are required. A recent modification on this process, known as *L*inear-*A*fter-*T*he-*E*xponential-PCR (LATE-PCR), uses a limiting primer with a higher melting

temperature (Tm) than the excess primer to maintain reaction efficiency as the limiting primer concentration decreases mid-reaction.

- *Helicase-dependent amplification*: similar to traditional PCR, but uses a constant temperature rather than cycling through denaturation and annealing/extension cycles. DNA helicase, an enzyme that unwinds DNA, is used in place of thermal denaturation.

- *Hot-start PCR*: a technique that reduces non-specific amplification during the initial set up stages of the PCR. It may be performed manually by heating the reaction components to the melting temperature (e.g., 95°C) before adding the polymerase. Specialized enzyme systems have been developed that inhibit the polymerase's activity at ambient temperature, either by the binding of an antibody or by the presence of covalently bound inhibitors that only dissociate after a high-temperature activation step. Hot-start/cold-finish PCR is achieved with new hybrid polymerases that are inactive at ambient temperature and are instantly activated at elongation temperature.

- *Intersequence-specific PCR* (ISSR): a PCR method for DNA fingerprinting that amplifies regions between simple sequence repeats to produce a unique fingerprint of amplified fragment lengths.

- *Inverse PCR*: is commonly used to identify the flanking sequences around genomic inserts. It involves a series of DNA digestions and self ligation, resulting in known sequences at either end of the unknown sequence.

- *Ligation-mediated PCR*: uses small DNA linkers ligated to the DNA of interest and multiple primers annealing to the DNA linkers; it has been used for DNA sequencing, genome walking, and DNA footprinting.

- *Methylation-specific PCR* (MSP): developed by Stephen Baylin and Jim Herman at the Johns Hopkins School of Medicine, and is used to detect methylation of CpG islands in genomic DNA. DNA is first treated with sodium bisulfite, which converts unmethylated cytosine bases to uracil, which is recognized by PCR primers as thymine. Two PCRs are then carried out on the modified DNA, using primer sets identical except at any CpG islands within the primer sequences. At these points, one primer set recognizes DNA with cytosines to amplify methylated DNA, and one set recognizes DNA with uracil or thymine to amplify unmethylated DNA. MSP using qPCR can also be performed to obtain quantitative rather than qualitative information about methylation.

- *Miniprimer PCR*: uses a thermostable polymerase (S-Tbr) that can extend from short primers ("smalligos") as short as 9 or 10 nucleotides. This method permits PCR targeting to smaller primer binding regions, and is used to amplify conserved DNA sequences, such as the 16S (or eukaryotic 18S) rRNA gene.

- *Multiplex Ligation-dependent Probe Amplification* (*MLPA*): permits multiple targets to be amplified with only a single primer pair, thus avoiding the resolution limitations of multiplex PCR.

- *Multiplex-PCR*: consists of multiple primer sets within a single PCR mixture to produce amplicons of varying sizes that are specific to different DNA sequences. By targeting multiple genes at once, additional information may be gained from a single test run that otherwise would require several times the reagents and more time to perform. Annealing temperatures for each of the primer sets must be optimized to work correctly within a single reaction, and amplicon sizes, i.e., their base pair length, should be different enough to form distinct bands when visualized by gel electrophoresis.

- *Nested PCR*: increases the specificity of DNA amplification, by reducing background due to non-specific amplification of DNA. Two sets of primers are used in two successive PCRs. In the first reaction, one pair of primers is used to generate DNA products, which besides the intended target, may still consist of non-specifically amplified DNA fragments. The product(s) are then used in a second PCR with a set of primers whose binding sites are completely or partially different from and located 3' of each of the primers used in the first reaction. Nested PCR is often more successful in specifically amplifying long DNA fragments than conventional PCR, but it requires more detailed knowledge of the target sequences.

- *Overlap-extension PCR*: a genetic engineering technique allowing the construction of a DNA sequence with an alteration inserted beyond the limit of the longest practical primer length.

- *Quantitative PCR (Q-PCR)*: used to measure the quantity of a PCR product (commonly in real-time). It quantitatively measures starting amounts of DNA, cDNA or RNA. Q-PCR is commonly used to determine whether a DNA sequence is present in a sample and the number of its copies in the sample. *Quantitative real-time PCR* has a very high degree of precision. QRT-PCR methods use fluorescent dyes, such as Sybr Green, EvaGreen or fluorophore-containing DNA probes, such as TaqMan, to measure the amount of amplified product in real time. It is also sometimes abbreviated to RT-PCR (*R*eal *T*ime PCR) or RQ-PCR. QRT-PCR or RTQ-PCR are more appropriate contractions, since RT-PCR commonly refers to reverse transcription PCR, often used in conjunction with Q-PCR.

- *R*everse *T*ranscription PCR (RT-PCR): for amplifying DNA from RNA. Reverse transcriptase reverse transcribes RNA into cDNA, which is then amplified by PCR. RT-PCR is widely used in expression profiling, to determine the expression of a gene or to identify the sequence of an RNA transcript, including transcription start and termination sites. If the genomic DNA sequence of a gene is known, RT-PCR can be used to map the location of exons and introns in the gene. The 5' end

of a gene (corresponding to the transcription start site) is typically identified by RACE-PCR (*Rapid Amplification of cDNA Ends*).

- *Solid Phase PCR*: encompasses multiple meanings, including Polony Amplification (where PCR colonies are derived in a gel matrix, for example), Bridge PCR (primers are covalently linked to a solid-support surface), conventional Solid Phase PCR (where Asymmetric PCR is applied in the presence of solid support bearing primer with sequence matching one of the aqueous primers) and Enhanced Solid Phase PCR (where conventional Solid Phase PCR can be improved by employing high Tm and nested solid support primer with optional application of a thermal 'step' to favour solid support priming).

- *Thermal asymmetric interlaced PCR (TAIL-PCR)*: for isolation of an unknown sequence flanking a known sequence. Within the known sequence, TAIL-PCR uses a nested pair of primers with differing annealing temperatures; a degenerate primer is used to amplify in the other direction from the unknown sequence.

- *Touchdown PCR* (*Step-down PCR*): a variant of PCR that aims to reduce nonspecific background by gradually lowering the annealing temperature as PCR cycling progresses. The annealing temperature at the initial cycles is usually a few degrees (3-5°C) above the T_m of the primers used, while at the later cycles, it is a few degrees (3-5°C) below the primer T_m. The higher temperatures give greater specificity for primer binding, and the lower temperatures permit more efficient amplification from the specific products formed during the initial cycles.

- *PAN-AC*: uses isothermal conditions for amplification, and may be used in living cells.

- *Universal Fast Walking*: for genome walking and genetic fingerprinting using a more specific 'two-sided' PCR than conventional 'one-sided' approaches (using only one gene-specific primer and one general primer - which can lead to artefactual 'noise') by virtue of a mechanism involving lariat structure formation. Streamlined derivatives of UFW are LaNe RAGE (lariat-dependent nested PCR for rapid amplification of genomic DNA ends), 5'RACE LaNe and 3'RACE LaNe.

History

A 1971 paper in the Journal of Molecular Biology by Kleppe and co-workers first described a method using an enzymatic assay to replicate a short DNA template with primers *in vitro*. However, this early manifestation of the basic PCR principle did not receive much attention, and the invention of the polymerase chain reaction in 1983 is generally credited to Kary Mullis.

At the core of the PCR method is the use of a suitable DNA polymerase able to withstand the high temperatures of >90 °C (194 °F) required for separation of the two DNA strands in the DNA double helix after each replication cycle. The DNA polymerases initially

employed for in vitro experiments presaging PCR were unable to withstand these high temperatures. So the early procedures for DNA replication were very inefficient, time consuming, and required large amounts of DNA polymerase and continual handling throughout the process.

The discovery in 1976 of Taq polymerase — a DNA polymerase purified from the thermophilic bacterium, *Thermus aquaticus*, which naturally lives in hot (50 to 80 °C (122 to 176 °F)) environments such as hot springs — paved the way for dramatic improvements of the PCR method. The DNA polymerase isolated from *T. aquaticus* is stable at high temperatures remaining active even after DNA denaturation, thus obviating the need to add new DNA polymerase after each cycle. This allowed an automated thermocycler-based process for DNA amplification.

When Mullis developed the PCR in 1983, he was working in Emeryville, California for Cetus Corporation, one of the first biotechnology companies. There, he was responsible for synthesizing short chains of DNA. Mullis has written that he conceived of PCR while cruising along the Pacific Coast Highway one night in his car. He was playing in his mind with a new way of analyzing changes (mutations) in DNA when he realized that he had instead invented a method of amplifying any DNA region through repeated cycles of duplication driven by DNA polymerase. In *Scientific American*, Mullis summarized the procedure: "Beginning with a single molecule of the genetic material DNA, the PCR can generate 100 billion similar molecules in an afternoon. The reaction is easy to execute. It requires no more than a test tube, a few simple reagents, and a source of heat." He was awarded the Nobel Prize in Chemistry in 1993 for his invention, seven years after he and his colleagues at Cetus first put his proposal to practice. However, some controversies have remained about the intellectual and practical contributions of other scientists to Mullis' work, and whether he had been the sole inventor of the PCR principle.

Patent wars

The PCR technique was patented by Kary Mullis and assigned to Cetus Corporation, where Mullis worked when he invented the technique in 1983. The *Taq* polymerase enzyme was also covered by patents. There have been several high-profile lawsuits related to the technique, including an unsuccessful lawsuit brought by DuPont. The pharmaceutical company Hoffmann-La Roche purchased the rights to the patents in 1992 and currently holds those that are still protected.

A related patent battle over the Taq polymerase enzyme is still ongoing in several jurisdictions around the world between Roche and Promega. The legal arguments have extended beyond the lives of the original PCR and Taq polymerase patents, which expired on March 28, 2005.

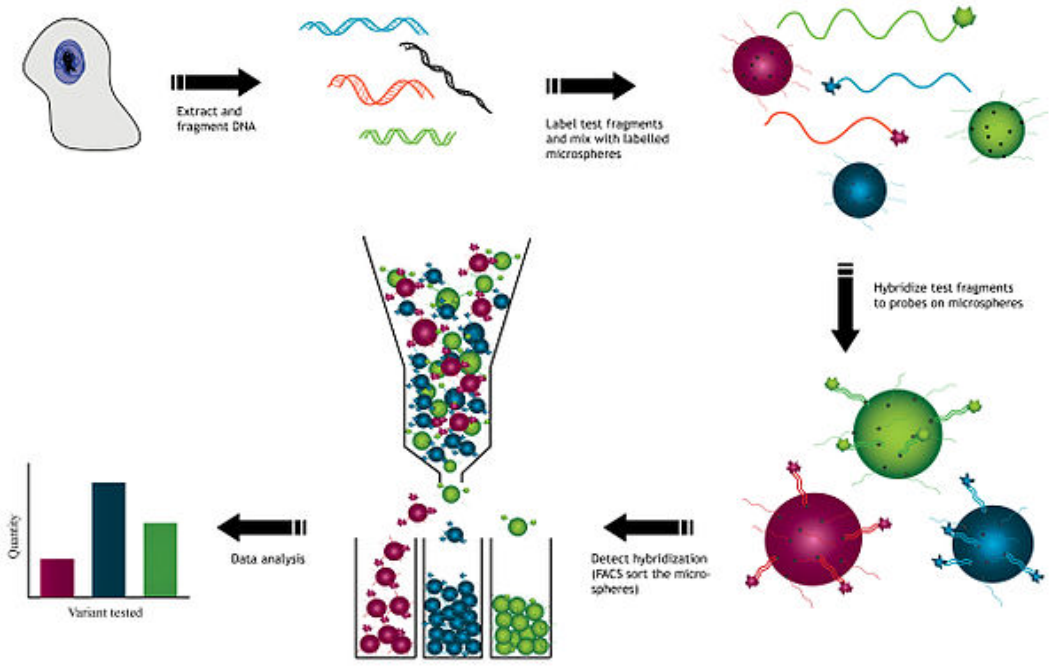
Chapter- 12

Suspension Array Technology, Southern Blot and Subcloning

Suspension array technology

Suspension Array Technology (or **SAT**) is a high throughput, large-scale, and multiplexed screening platform used in molecular biology. SAT has been widely applied to genomic and proteomic research, such as single nucleotide polymorphism (SNP) genotyping, genetic disease screening, gene expression profiling, screening drug discovery and clinical diagnosis. SAT uses microsphere beads (5.6 um in diameter) to prepare arrays. SAT allows for the simultaneous testing of multiple gene variants through the use of these microshpere beads as each type of microsphere bead has a unique identification based on variations in optical properties, most common is fluorescent colour. As each colour and intensity of colour has a unique wavelength, beads can easily be differentiated based on their wavelength intensity. Microspheres are readily suspendable in solution and exhibit favorable kinetics during an assay. Similar to flat microarrays (eg. DNA microarray), an appropriate receptor molecule, such as DNA oligonucleotide probes, antibodies, or other proteins, attach themselves to the differently labeled microspheres. This produces thousands of microsphere array elements. Probe-target hybridization is usually detected by optically labeled targets, which determines the relative abundance of each target in the sample.

Overview of SAT using DNA hybridization



Overview of the suspension array technology procedures, using DNA hybridization as a model.

DNA is extracted from cells used to create test fragments. These test fragments are added to a solution containing a variety of microsphere beads. Each type of microsphere bead contains a known DNA probe with a unique fluorescent identity. Test fragments and probes on the microsphere beads are allowed to hybridize to each other. Once hybridized, the microsphere beads are sorted, usually using flow cytometry. This allows for the detection of each of the gene variants from the original sample. The resulting data collected will indicate the relative abundance of each hybridized sample to the microsphere.

Multiplexing

Since microsphere beads are easily suspended in solution and each microsphere retains its identity when hybridized to the test sample, a typical suspension array experiment can analyze wide range of biological analysis in a single reaction, called “multiplexing”. In general, each type of microsphere used in an array is individually prepared in bulk. For example, the commercially available microsphere arrays from Luminex xMAP^{TM} technology uses 10X10 element array. This array involves beads with red and infrared dyes, each with ten different intensities, to give a 100-element array. Thus, the array size would increase exponentially if multiple dyes are used. For example, five different dyes with 10 different intensities per dye will give rise to 10,000 different array elements.

Procedure

Sample targeting

When using different types of microspheres, SAT is capable of simultaneously testing multiple variables, such as DNA and proteins, in a given sample. This allows SAT to analyze variety of molecular targets during a single reaction. The common nucleic acid detection method includes direct DNA hybridization. The direct DNA hybridization approach is the simplest suspension array assay whereby 15 to 20 bp DNA oligonucleotides attached to microspheres are amplified using PCR. This is the optimized probe length as it minimizes the melting temperature variation among different probes during probe-target hybridization. After amplifying one DNA oligoprobe of interest, it can be used to create 100 different probes on 100 different sets of microspheres, each with the capability of capturing 100 potential targets (if using a 100-plex array). Similarly, target DNA samples are usually PCR amplified and labeled. Hybridization between the capture probe and the target DNA is achieved by melting and annealing complementary target DNA sequence to their capture probes located on the microspheres. After washing to remove non-specific binding between sequences, only strongly paired probe-target will remain hybridized.

Sorting and detection with flow cytometry

Since the optical identity of each microsphere is known, the quantification of target samples hybridized to the microspheres can be achieved by comparing the relative intensity of target markers in one set of microspheres to target markers in another set of microspheres using flow cytometry. Microspheres can be sorted based using both their unique optical properties and level of hybridization to the target sequence.

Strengths

- **Rapid/high throughput:** In multiplex analysis, a 100-plex assay can be analyzed in every 30 seconds. The recent reported high-throughput flow cytometry can sample a 96-well plate in 1 minute, and theoretically, the 100-plex assay with this system can be analyzed in less than 1 second, or potentially deliver 12 million samples per day.
- **High array density/multiplex:** Compared to flat microarrays, SAT allows one to perform parallel measurements. A few microliters of microspheres could contain thousands of array elements and each array element is represented by hundreds of individual microspheres. Thus, the measurement by flow cytometry represents a replicate analysis of each array element.
- **Effective gathering of information:** One of the benefits of using SAT is that it allows you to take one sample from a patient or research organism and simultaneously test for multiple gene variants. Thus, from a single sample you can determine which virus from a series of viruses a patient has, or which base pair mutation is present in the organism with a unique phenotype.

- **Cost-effective:** Currently, commercially available suspension array kits costs $0.10-$0.25 per sequence tested.

Weaknesses

- **Relatively low array size**: Although it has the potential to use an increased amount of dyes to generate millions of different array elements, the current generation of commercially available microsphere arrays (from Luminex xMAP^{TM} technology) only uses two sets of dyes and therefore can only detect ~100 targets per experiment.
- Hybridization between different sets of probes and target sequences requires a **specific annealing temperature**, which is affected by length and sequence of the oligonucleotide probe. Therefore, for every experiment, only one possible annealing temperature can be used. Thus, all probes used in given experiment must be designed to hybridize to the target at the same temperature. Although introducing base pair mismatch in some sets of the probes could minimize annealing temperature difference between each sets of probes, the hybridization problem is still significant if more than 10-20 targets are tested in one reaction.

Southern blot

A **Southern blot** is a method routinely used in molecular biology for detection of a specific DNA sequence in DNA samples. Southern blotting combines transfer of electrophoresis-separated DNA fragments to a filter membrane and subsequent fragment detection by probe hybridization. The method is named after its inventor, the British biologist Edwin Southern. Other blotting methods (i.e., Western blot, Northern blot, Eastern blot, Southwestern blot) that employ similar principles, but using RNA or protein, have later been named in reference to Edwin Southern's name. As the technique was eponymously named, Southern blot is capitalized as is conventional for proper nouns. The names for other blotting methods may follow this convention, by analogy.

Method

1. Restriction endonucleases are used to cut high-molecular-weight DNA strands into smaller fragments.
2. The DNA fragments are then electrophoresed on an agarose gel to separate them by size.
3. If some of the DNA fragments are larger than 15 kb, then prior to blotting, the gel may be treated with an acid, such as dilute HCl, which depurinates the DNA fragments, breaking the DNA into smaller pieces, thus allowing more efficient transfer from the gel to membrane.

4. If alkaline transfer methods are used, the DNA gel is placed into an alkaline solution (typically containing sodium hydroxide) to denature the double-stranded DNA. The denaturation in an alkaline environment may improve binding of the negatively charged DNA to a positively charged membrane, separating it into single DNA strands for later hybridization to the probe, and destroys any residual RNA that may still be present in the DNA. The choice of alkaline over neutral transfer methods, however, is often empirical and may result in equivalent results.
5. A sheet of nitrocellulose (or, alternatively, nylon) membrane is placed on top of (or below, depending on the direction of the transfer) the gel. Pressure is applied evenly to the gel (either using suction, or by placing a stack of paper towels and a weight on top of the membrane and gel), to ensure good and even contact between gel and membrane. If transferring by suction 20X SSC buffer is used to ensure a seal and prevent drying of the gel. Buffer transfer by capillary action from a region of high water potential to a region of low water potential (usually filter paper and paper tissues) is then used to move the DNA from the gel on to the membrane; ion exchange interactions bind the DNA to the membrane due to the negative charge of the DNA and positive charge of the membrane.
6. The membrane is then baked in a vacuum or regular oven at 80 °C for 2 hours (standard conditions; nitrocellulose or nylon membrane) or exposed to ultraviolet radiation (nylon membrane) to permanently attach the transferred DNA to the membrane.
7. The membrane is then exposed to a hybridization probe—a single DNA fragment with a specific sequence whose presence in the target DNA is to be determined. The probe DNA is labelled so that it can be detected, usually by incorporating radioactivity or tagging the molecule with a fluorescent or chromogenic dye. In some cases, the hybridization probe may be made from RNA, rather than DNA. To ensure the specificity of the binding of the probe to the sample DNA, most common hybridization methods use salmon or herring sperm DNA for blocking of the membrane surface and target DNA, deionized formamide, and detergents such as SDS to reduce non-specific binding of the probe.
8. After hybridization, excess probe is washed from the membrane (typically using SSC buffer), and the pattern of hybridization is visualized on X-ray film by autoradiography in the case of a radioactive or fluorescent probe, or by development of color on the membrane if a chromogenic detection method is used.

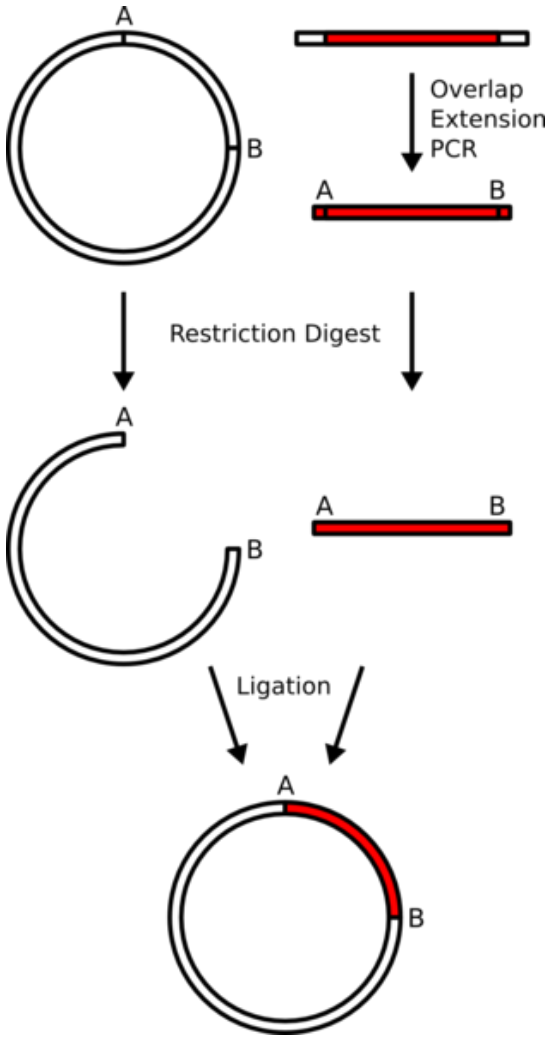
Result

Hybridization of the probe to a specific DNA fragment on the filter membrane indicates that this fragment contains DNA sequence that is complementary to the probe.

The transfer step of the DNA from the electrophoresis gel to a membrane permits easy binding of the labeled hybridization probe to the size-fractionated DNA. It also allows for the fixation of the target-probe hybrids, required for analysis by autoradiography or other detection methods.

Southern blots performed with restriction enzyme-digested genomic DNA may be used to determine the number of sequences (e.g., gene copies) in a genome. A probe that hybridizes only to a single DNA segment that has not been cut by the restriction enzyme will produce a single band on a Southern blot, whereas multiple bands will likely be observed when the probe hybridizes to several highly similar sequences (e.g., those that may be the result of sequence duplication). Modification of the hybridization conditions (for example, increasing the hybridization temperature or decreasing salt concentration) may be used to increase specificity and decrease hybridization of the probe to sequences that are less than 100% similar.

Subcloning



This image diagrams the procedure of subcloning as outlined to the left.

In molecular biology, **subcloning** is a technique used to move a particular gene of interest from a *parent vector* to a *destination vector* in order to further study its functionality.

Procedure

Restriction enzymes are used to excise the gene of interest (the *insert*) from the parent. The insert is purified in order to isolate it from background junk. A common purification method is gel isolation. The number of copies of the gene is then amplified using Polymerase Chain Reaction (PCR).

Simultaneously, the same restriction enzymes are used to digest (cut) the destination. The idea behind using the same restriction enzymes is to create complementary sticky ends, which will facilitate ligation later on. A phosphatase (commonly Calf Intestinal Alkaline Phosphatase; CIAP) is also added to prevent self-ligation of the destination vector. The digested destination vector is isolated/purified.

The insert and the destination vector are then mixed together with DNA ligase. A typical ratio of insert genes to destination vectors is 3:1 ; by increasing the insert concentration, self-ligation is further decreased. After letting the reaction mixture sit for a set amount of time at a specific temperature (dependent upon the size of the strands being ligated), the insert should become successfully incorporated into the destination plasmid.

Amplification of product plasmid

The plasmid is often transformed into a bacterium like *E. coli*. Ideally when the bacterium divides the plasmid should also be replicated. In the best case scenario, each bacteria cell should have several copies of the plasmid. After a good number of bacterial colonies have grown, they can be miniprepped to harvest the plasmid DNA.

Selection

In order to ensure growth of only transformed bacteria (which carry the desired plasmids to be harvested), a marker gene is used in the destination vector for selection. Typical marker genes are for antibiotic resistance or nutrient biosynthesis. So, for example, the "marker gene" could be for resistance to the antibiotic ampicillin. If the bacteria that were supposed to pick up the desired plasmid had picked up the desired gene then they would also contain the "marker gene". Now the bacteria that picked up the plasmid would be able to grow in ampicillin whereas the bacteria that did not pick up the desired plasmid would still be vulnerable to destruction by the ampicillin. Therefore, successfully transformed bacteria would be "selected."

Example Case: Bacterial plasmid subcloning

In this example, a gene from mammalian gene library will be subcloned into a bacterial plasmid (destination platform). The bacterial plasmid is a piece of circular DNA which

contains regulatory elements allowing for the bacteria to produce a gene product (gene expression) if it is placed in the correct place in the plasmid. The production site is flanked by two restriction enzyme cutting sites "A" and "B" with incompatible sticky ends.

The mammalian DNA does not come with these restriction sites, so they are built in by overlap extension PCR. The primers are designed to put the restriction sites carefully, so that the coding of the protein is in-frame, and a minimum of extra amino acids is implanted on either side of the protein.

Both the PCR product containing the mammalian gene with the new restriction sites and the destination plasmid are subjected to restriction digestion, and the digest products are purified by gel electrophoresis.

The digest products, now containing compatible sticky ends with each other (but incompatible sticky ends with themselves) are subjected to ligation, creating a new plasmid which contains the background elements of the original plasmid with a different insert.

The plasmid is transformed into bacteria and is checked by DNA sequencing.