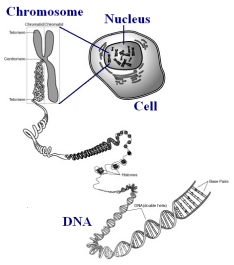
Differential gene expression

General Introduction

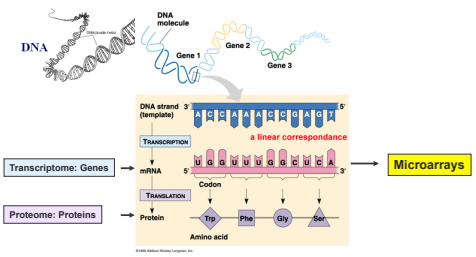
Overview (1)

- Reminder of biology
- Major steps in microarray analysis
  - Microarray preparation design, clone/probe selection
  - RNA extraction, hybridization on chip
  - Scanning, data extraction from image
  - "Low-level" Quality Control
  - Summarization of per-chip information (one number per feature)
  - "High-level" analysis
- High-throughput RNA-level technologies
  - Microarrays
  - Affymetrix Chips
  - SAGE
  - MPSS

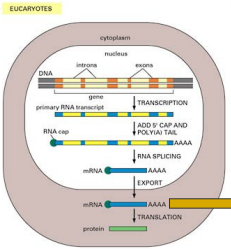
Biology Fundamentals - Genes



Biology Fundamentals - Expression



Genomics Fundamentals - Complexity

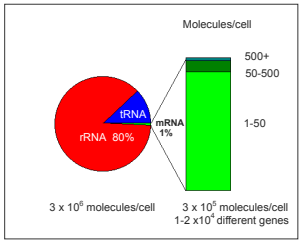


mRNA purification

Difficulties:

- Contaminations
- Alternative Splicing
- Alternative PolyAdenylation

RNA abundance in mammalian cells



Expression analysis

- Low throughput
  - Northern blot
  - Differential display
  - Quantitative PCR
- High throughput
  - DNA arrays / Chips
    - Spotted arrays (Stanford arrays)
    - Affymetrix (photolithography inspired)
    - Oligo-arrays (Agilent, NimbleGen)
  - Serial Analysis of Gene Expression (SAGE)
  - RNASeq

What are DNA Microarrays ?

Microarray analysis is a technology that allows scientists to simultaneously detect thousands of genes in a small sample and to analyze the expression of those genes.

Microarrays are simply ordered sets of DNA molecules of known sequence. Usually rectangular shaped, they can consist of a few hundred to hundreds of thousands of sets. Each individual sequence goes on the array at precisely defined location.

Potential application domains

- Identification of complex genetic diseases
- Drug discovery and toxicology studies
- Mutation/polymorphism detection (SNP's)
- Pathogen analysis
- Differing expression of genes over time, between tissues, and disease states
- Preventive medicine
- Specific genotype (population) targeted drugs
- More targeted drug treatments – AIDS
- Genetic testing and privacy

The challenge

The big revolution here is in the "micro" term. New slides will contain a survey of the human genome on a 2 cm^2 chip! The use of this large-scale method tends to create phenomenal amounts of data, that have then to be analyzed, processed and stored.

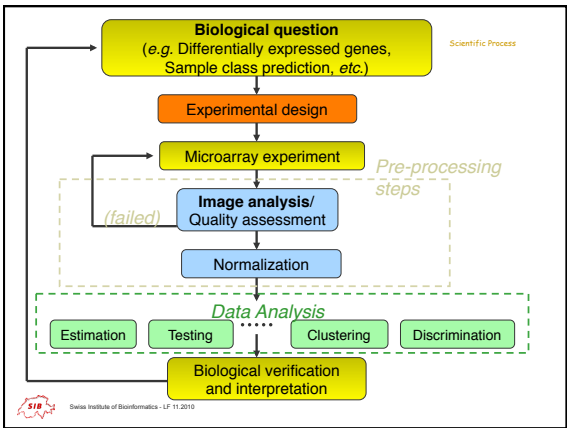
This is a job for… Bioinformatics !

General overview

- Making the chip
  - Experiment design, clone/probe selection, collection maintenance, PCR, spotting, printing, synthesis
- Sample hybridization
  - Sample purification, labelling, hybridization, washing

} wet lab

- Scanning and image treatment
  - Fluorescence correction, find spots, background
- Analysing the data
  - Filtering, normalisation
  - Clustering (hierarchical, centroid,…)
- Representation, storage
  - Graphics, databases, web public resources



Question addressed by microarrays

- What are the differences (in gene expression) between two cell lines ?
- What is the difference between knock-out and wild-type mice?
- What is the difference between a tumor and a healthy tissue ?
- Are there different tumor types ?

- Key concept: *Compare* gene expression in two (or more) cell/ tissue types ?
  - Gene expression assessed by measuring the number of RNA transcripts.
  - No absolute measurement.

THE EXPERIMENT : making the chip

1- Designing the chip : choosing genes of interest for the experiment and/or select the samples

- Selection of sequences that represent the investigated genes.

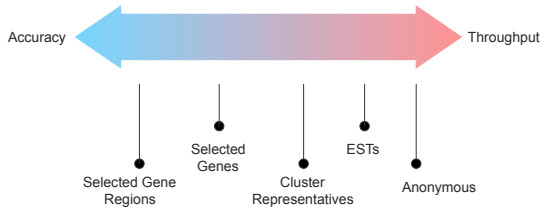
- Finding sequences, usually in the EST database.

- Problems : sequencing errors, alternative splicing, chimeric sequences, contamination…

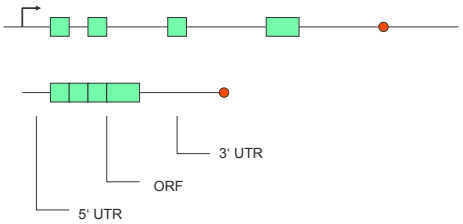
Clone/probe selection

- General
  - Not too short (sensitivity, selectivity)
  - Not too long (viscosity, surface properties)
  - Not too heterogeneous (robustness)
  - Degree of importance depends on method
- Single strand methods (Oligos, ss-cDNA)
  - Orientation must be known
  - ss-cDNA methods are not perfect
  - ds-cDNA methods don't care

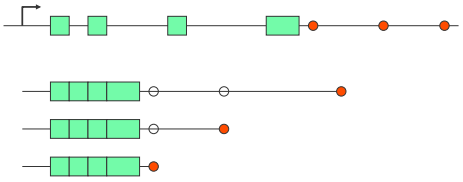
Probe selection approaches



Selection of gene regions

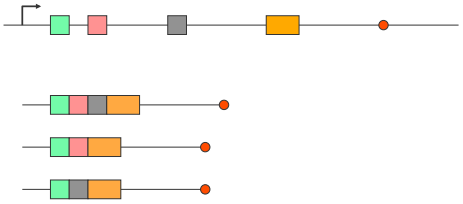


Alternative polyadenylation

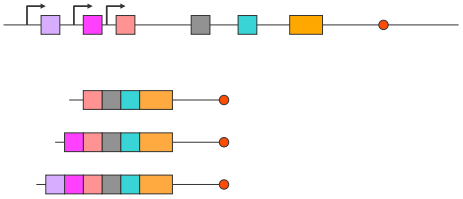


Particular problem with Affymetrix

Alternative splicing



Alternative promoter usage



Selection of gene regions - summary

- Coding region (ORF)
  - Annotation relatively safe
  - No problems with alternative polyA sites
  - No repetitive elements or other funny sequences
  - danger of close isoforms
  - danger of alternative splicing
  - might be missing in short RT products
- 3' untranslated region
  - Annotation less safe
  - danger of alternative polyA sites
  - danger of repetitive elements
  - less likely to cross-hybridize with isoforms
  - little danger of alternative splicing
- 5' untranslated region
  - close linkage to promoter
  - frequently not available

A checklist

- Pick a gene
- Try to get a complete cDNA sequence
- Verify sequence architecture (e.g. cross-species comparison)
- Mask repetitive elements (and vector!)
- If possible, discard 3'-UTR beyond first polyA signal
- Look for alternative splice events
- Use remaining region of interest for similarity searches
- Mask regions that could cross-hybridize

- Use the remaining region for probe amplification or EST selection
- When working with ESTs, use sequence-verified clones

THE EXPERIMENT : making the chip

2- Spotting the sequences on the substrate

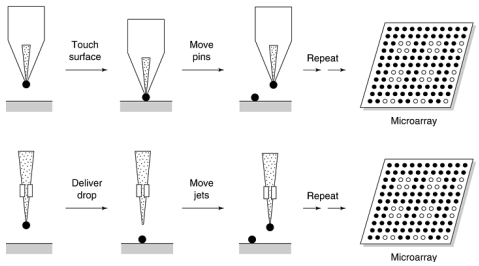
- Substrate : usually glass, but also nylon membranes, plastic, ceramic…

- Sequences : cDNA (500-5000 nucleotides), oligonucleotides (20~80-mer oligos), genomic DNA (~50' 000 bases)

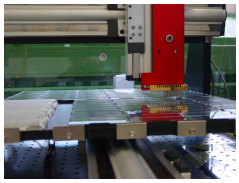
- Printing methods : microspotting, ink-jetting or in-situ printing, photolithography

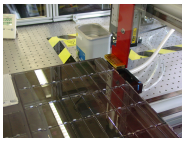
Microarrays: the making of

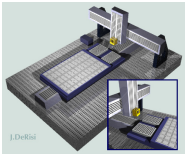
Microspotting and ink-jetting



Array Production: Spotting

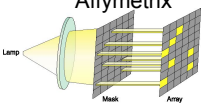




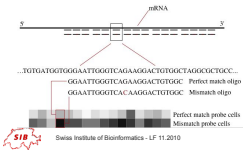


Array Production: ”photolithography"

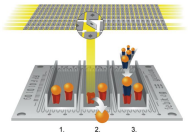
Affymetrix



- Each probe 25 bp long
- 22-40 probes per gene
- Perfect Match (PM) as well as MisMatch (MM) probes



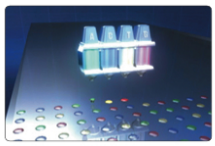
Febit/NimbleGen

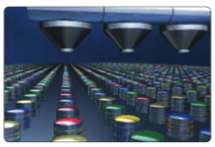


- Probe length: 24mer -70mer
- Gene/Array: Up to 38,000
- Probes/Gene: 10-25
- Only perfect match probes

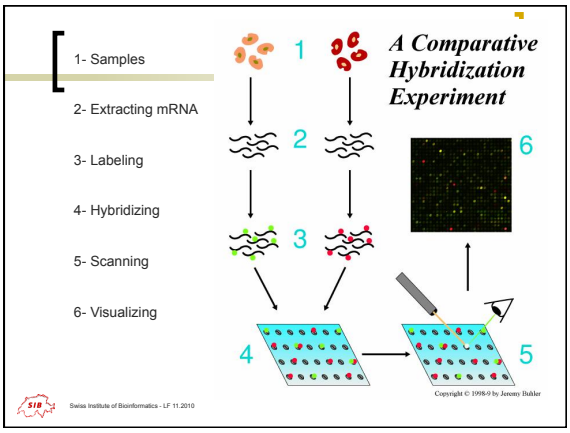
Array Production: “Inkjet”

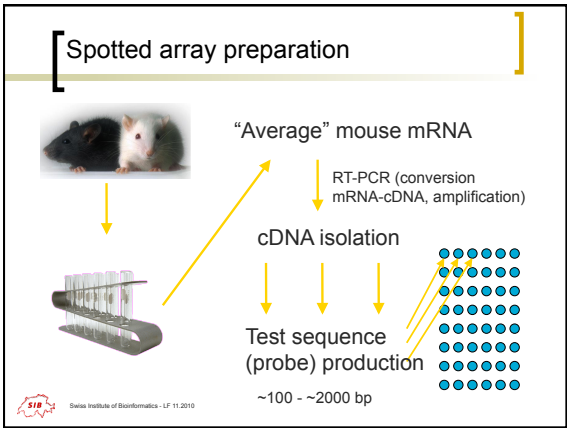
Agilent (HP SurePrint technology)

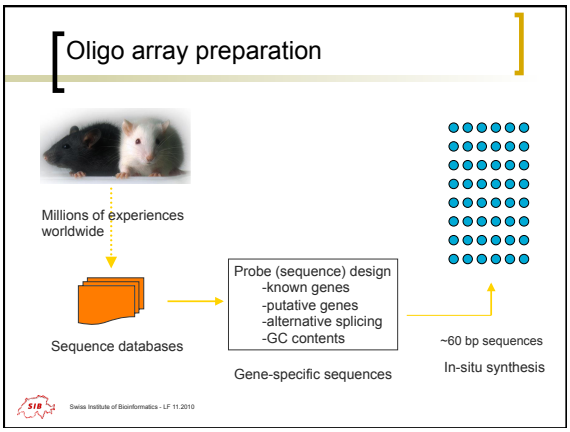


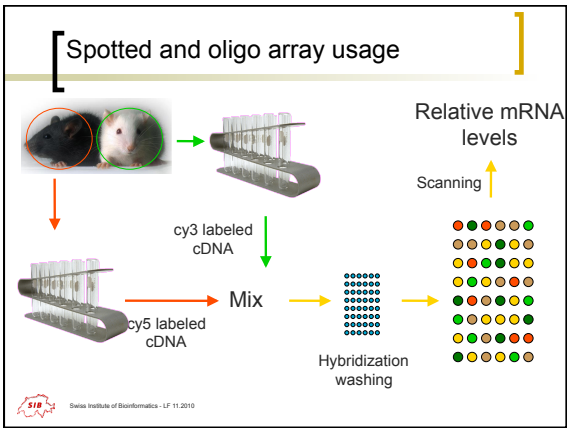


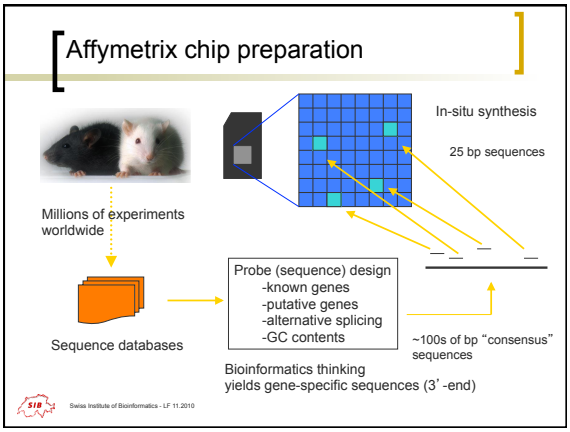
- cDNA printing
- 60bp oligo in-situ synthesis

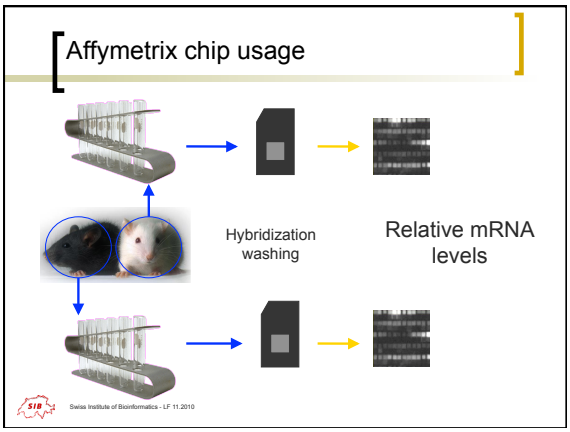

1- Samples
2- Extracting mRNA
3- Labeling
4- Hybridizing
5- Scanning
6- Visualizing
A Comparative Hybridization Experiment
1
2
3
4
5
6
Copyright © 1998-9 by Jeremy Buhler
Swiss Institute of Bioinformatics - LF 11.2010


Spotted array preparation
"Average" mouse mRNA
RT-PCR (conversion mRNA-cDNA, amplification)
cDNA isolation
Test sequence (probe) production
~100 - ~2000 bp
Swiss Institute of Bioinformatics - LF 11.2010

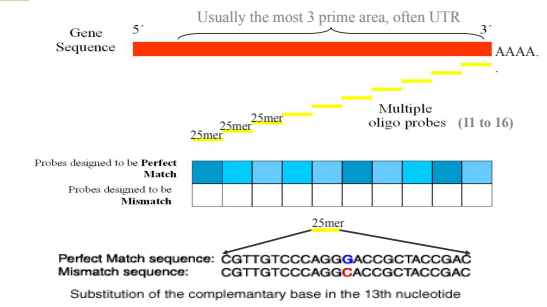

Oligo array preparation
Millions of experiences worldwide
Sequence databases
Probe (sequence) design
-known genes
-putative genes
-alternative splicing
-GC contents
Gene-specific sequences
~60 bp sequences
In-situ synthesis
Swiss Institute of Bioinformatics - LF 11.2010


Spotted and oligo array usage
Relative mRNA levels
Scanning
cy3 labeled cDNA
Mix
cy5 labeled cDNA
Hybridization washing
Swiss Institute of Bioinformatics - LF 11.2010

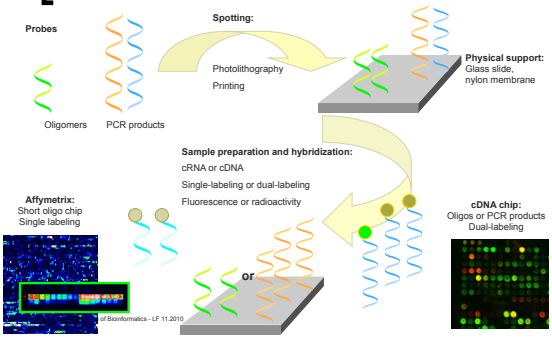

Affymetrix chip preparation
In-situ synthesis
25 bp sequences
Millions of experiments worldwide
Sequence databases
Probe (sequence) design
-known genes
-putative genes
-alternative splicing
-GC contents
~100s of bp "consensus" sequences
Bioinformatics thinking
yields gene-specific sequences (3' -end)
Swiss Institute of Bioinformatics - LF 11.2010


Affymetrix chip usage
Hybridization washing
Relative mRNA levels
Swiss Institute of Bioinformatics - LF 11.2010

Affymetrix system



Technological battle?

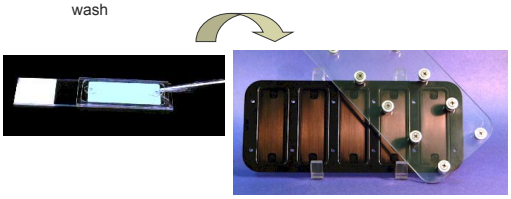


Comparison of techniques

In-situ Synthesis / Oligos	PCR Products / cDNA Probes
Advantages • No need to isolate and purify cDNAs because oligonucleotides can be synthesized. • Short oligonucleotides are less likely to have cross-reactivity with other sequences in the target DNA. • Density of chips is higher than with cDNAs.	**Advantages** • Flexibility to study cDNAs from any source. • cDNAs do not require any a priori information about the corresponding genes. • Longer sequences increase hybridization specificity, which reduces false positives.
Limitations • The sequence has to be known. • Synthesis can be expensive and time-consuming. • The short sequences are not as specific for target DNA, so appropriate controls must be added.	**Limitations** • Isolation of individual cDNAs to immobilize on each spot can be cumbersome. • Density is lower than synthesizing oligonucleotides on the surface of the chip. • cDNAs are longer sequences and are more likely to randomly contain sequences found in target DNA, which results in cross-reactivity.

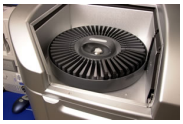
Probe preparation & hybridization

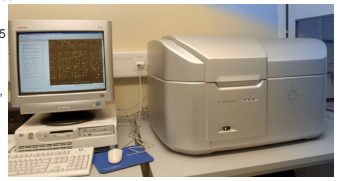
- Extract mRNA or total RNA
- RT, add 5' anchor
- PCR with labelled nucleotide (Cy3, Cy5, DIG or radiolabelling)
- Overlay probe on the chip, put in the hybridization chamber, wash



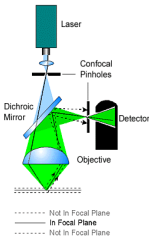
Scanner basics

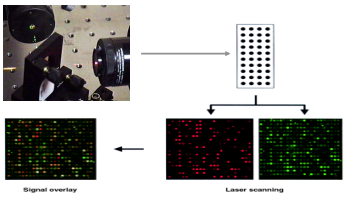
- Based on fluorescence
  - 1 or 2 lasers: cy3 cy5 (seldom more)
- Most scanners are confocal
  - Target a very limited volume of space (signal only from focal plane)
  - Need to "scan" the surface
- 16-bits ADC converters
  - Range of values: 0-65535
  - Log2 range: 0 – 16
- Scan various supports
  - Glass Slide (e.g. Agilent, PerkinElmer)
  - Affymetrix



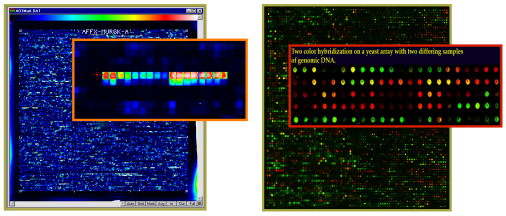


Confocal scanner





Scanner output: image(s)



Affymetrix chip

1 channel, false colors

Stanford array

dual-channel, color addition

Image analysis (scanner variability)

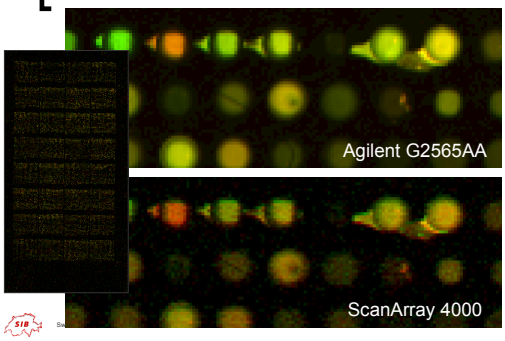
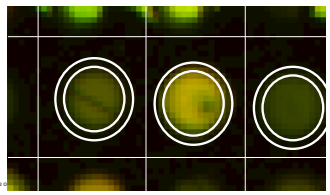


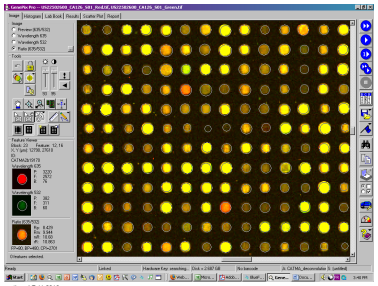
Image processing

- Align channels
- Identify spot pixels
- Identify background pixels
- Compute representative value, e.g.
  - Mean foreground value
  - Median background value

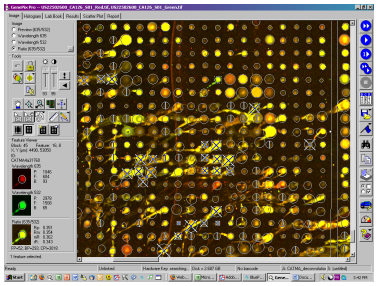


2-color Arrays Image Processing

GenePix



2-color Arrays Image Processing



A difficult case… ☺

Other high-throughput techniques: sequencing

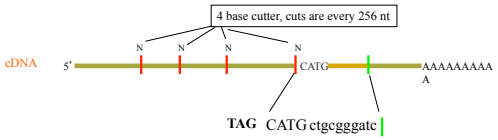
- EST counting
- SAGE
  (Serial Analysis of Gene Expression)
- MPSS
  (Massively Parallel Signature Sequencing)
- RNASeq

Comparison of techniques

Method	Microarrays	EST counts	SAGE/MPSS
Genes	Selected genes	Highly expressed	Almost all genes
Sampling	Analog	Digital	Digital
Statistics	Robust	Robust	Robust
Molecules sampled	High	Low-medium	Medium-high
Duplicates	Required	Desirable	Not required
Costs	Medium	High	High
Sharing	Variable	Easy	Easy

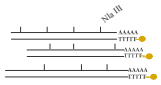
SAGE (Principle)

- A short nucleotide sequence "TAG" (9 to 10 bp) from a defined position within the transcript contains sufficient information to uniquely identify a transcript.
  - e.g. a sequence of 10 bp can distinguish 1,048,576 transcripts (4^{10}) given random nucleotide distribution at the tag site.
  - current estimates suggest the human genome only encodes ~ 100,000 transcripts (from 35,000 genes)

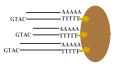


SAGE Protocol (I)

- cDNA synthesis using biotinylated oligo-dT nucleotide



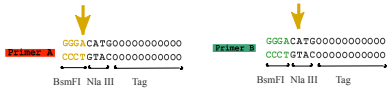
- Digestion of biotinylated cDNA with Nla III (anchoring enzyme) and binding to magnetic beads



SAGE Protocol (II)

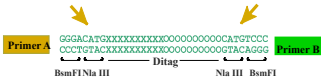
- Divide in half and ligate to linkers A and B

- Digest with BsmFI enzyme (tagging enzyme, TE), keep supernatant
- Blunt end (fill in)

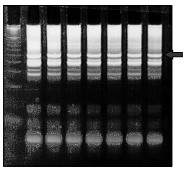


SAGE Protocol (III)

- Ligated and amplified with primers A & B



- PCR amplification using primers A & B

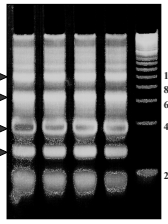


102 bp Ditag + primer A & B

- Contamination precautions!

SAGE Protocol (IV)

- Purification of 102 bp "ditag" band and Nla III digestion



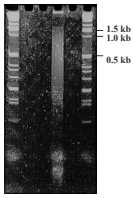
Undigested Ditag
Primer + Ditag
Primer A&B

XXXXXXXXXXOOOOOOOOOOCATG
GTACXXXXXXXXXXOOOOOOOOOO
Nla III Nla III

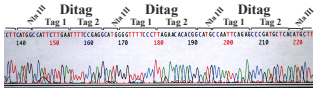
100 bp
80 bp
60 bp
40 bp
20 bp

SAGE Protocol (V)

- Purification of 26 bp "ditag" band and concatemerization

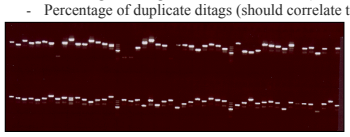


- Purification of 0.7-1.5 kb long concatemers subcloning and sequencing

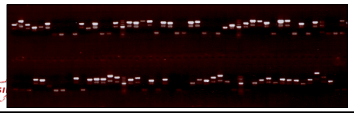


SAGE Protocol (VI)

- Library verification
  - Insert size
  - Percentage inserts
  - Percentage of ditags derived from linkers (between 2 and 15%)
  - Percentage of duplicate ditags (should correlate to TAG abundance)



- 89/96 inserts (93%)
- Ave size ~ 900bp



- 67/96 inserts (70%)
- Ave size ~ 600bp

DiTAG and TAG extraction from sequence data

SAGE300: (version 3.03, 1998)
Kinzler: kinzlk@welchlink.welch.jhu.edu

SAGE
Serial Analysis of Gene Expression

12TC7T7
NCNTTCCCGGCGCTTGGCCTCGATATG**CATG**CCCATCGTCCTAGTAACAAGAT**CATG**AAACAGTA
TGTGGGGATGCTTG**CATG**TCAGCTCCCATATTTTCCGAGG**CATG**TCCCTATTAAGCAGAGGACCA
A**CATG**CTGCAACCTATGAAGCCTTAT**CATG**AAGCAGTTACAAAACAGCCAGC**CATG**CACCTAATT
GGAGGCCTGTGAT**CATG**TACATAATTACTGGGGTTTCGA**CATG**TGAGAGACATCTANAACTTTTA
C**CATG**CTCGAGCGGCCGCCAGTGTGATGGATATCTGCAGAATTCCCGCACACTGGCGGGG

DiTAG and TAG Extraction From Sequence Data

Input File: X:\sage\XXXX\12TC7T7.SEQ

```
1 ) CCCATCGTCCTAGTAACAAGAT -  347574 -  229106
2 ) AAACAGTATGTGGGGATGCTTG -  4815 -  271574
3 ) TCAGCTCCCATATTTTCCGAGG -  862037 -  383489
4 ) TCCCTATTAAGCAGAGGACCAA -  875761 -  1027448
5 ) CTGCAACCTATGAAGCCTTAT -  495709 -  199294
6 ) AAGCAGTTACAAAACAGCCAGC -  37618 -  649712
7 ) CACCTAATTGGAGGCCTGTGAT -  285759 -  214182
8 ) TACATAATTACTGGGGTTTCGA -  805949 -  884822
9 ) TGAGAGACATCTANAACTTTTAC -  926228 -  0
```

Total Dimers: 9
Short Dimers: 0
Long Dimers: 0
Duplicate Dimers: 0
Good Tags: 17

Comparisons of SAGE Projects

Lib # 1	Lib # 2	Total	Tag Sequence
962	350	1312	GTGGCTCACA
614	687	1301	GCTGCCCTCC
288	140	428	AGCAGTCCCC
262	135	397	GCTTCGTCCA
221	136	357	TCAGGCTGCC
233	119	352	ATACTGACAT
202	102	304	AAAAAAAAAA
239	65	304	GCCTCCAAGG
212	90	302	GTGACCTGGC
233	64	297	GCGGGGTCGC
160	97	257	GCACAACTTG
180	70	250	CATCGCCAGT
163	71	234	GAGCGTTTTG

Massively Parallel Signature Sequencing

- Alternative to SAGE: generate 13-nt tags from a large (>10^5) sample of cDNAs
- Longer tags means higher specificity
- Solid-state technology gives high throughput
- Cost comparable to SAGE, but much larger number of tags obtained

Overall strategy of MPSS

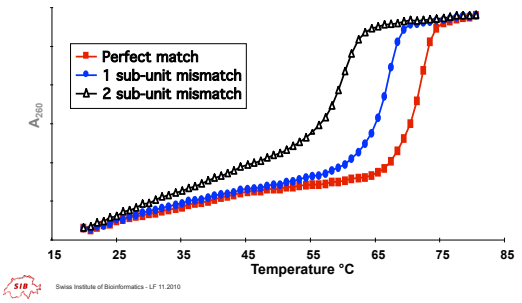
- "Megacloning"
  - Generate cDNA population where each molecule is attached to a different tag
  - Amplify this population
  - Attach to beads carrying anti-tags
  - Purify beads that have captured cDNAs
- Massively parallel sequencing
  - Use cycles of Rx enzyme cleavage, ligation and hybridization to read blocks of 4 nucleotides

"Words" for tag construction

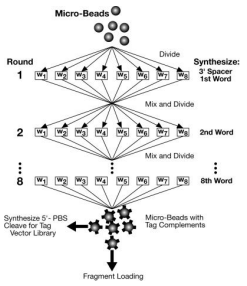
- TTAC, AATC, TACT, ATCA, ACAT, TCTA, CTTT, and CAAA
  - No restriction sites
  - Isothermal denaturation
  - Large ΔT_m for match/mismatch pairs
  - Large total repertoire (8^8 or 16,777,216)
  - Example:

```
5'-TACT.TTAC.ACAT.ATCA.CTTT.CTTT.CAAA.AATC-3'
3'-ATGA.AATG.TGTA.TAGT.GAAA.GAAA.GTTT.TTAG-5'
```


Tag Hybridization Specificity



Generation of tags from “words”

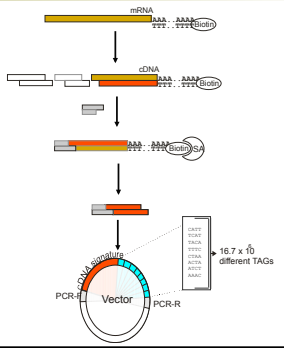


Signature Capture & Tagging

cDNA synthesis &
DpnII digestion

Signature
Capture

Signature
Tagging

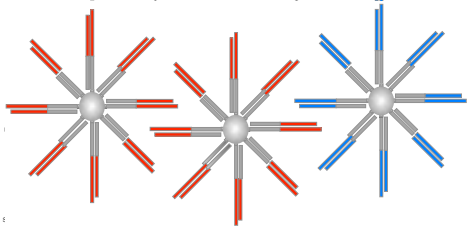


Loading DNA on Microbeads

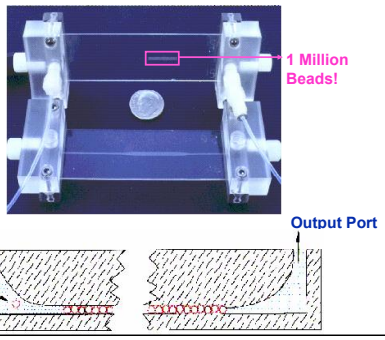
Each bead collects ~10^5 copies of a given fragment.

After loading one strand is covalently ligated to the bead

Clones of X_1 Clones of X_2 Clones of Y



Beads Immobilized in a Flow Cell

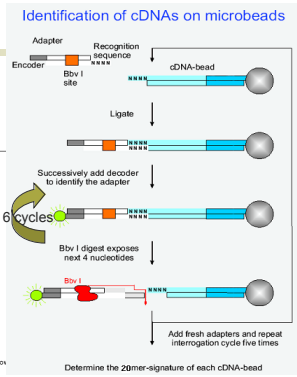


Flowchart for the sequencing reactions

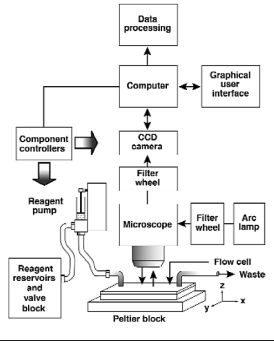
Table 1. Sequences of encoded adaptors[a]

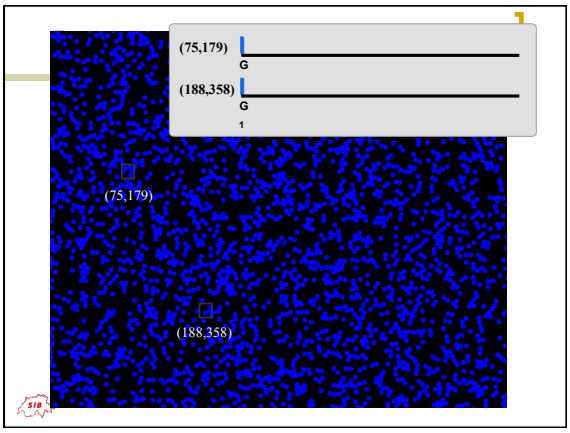
Common strand:
5'-GACTGGCAGCTCGT
Encoded adaptors for detecting base 1:
5'-**NNNN**ACGAGCTGCCAGTCcatttaggcg
5'-**NNNG**ACGAGCTGCCAGTCctgattacog
5'-**NNNC**ACGAGCTGCCAGTCaccaatacogg
5'-**NNNT**ACGAGCTGCCAGTCcogcttttgtag
Encoded adaptors for detecting base 2:
5'-**NNAN**ACGAGCTGCCAGTCggaacctgaa
5'-**NNGN**ACGAGCTGCCAGTCtgtgcgtgat
5'-**NNCN**ACGAGCTGCCAGTCaccgacattc
5'-**NNTN**ACGAGCTGCCAGTCattcctcctc
Encoded adaptors for detecting base 3:
5'-**NANN**ACGAGCTGCCAGTCcgaagaagtc
5'-**NGNN**ACGAGCTGCCAGTCtggtctctct
5'-**NCNN**ACGAGCTGCCAGTCtagcggactt
5'-**NTNN**ACGAGCTGCCAGTCcggcgataact
Encoded adaptors for detecting base 4:
5'-**ANNN**ACGAGCTGCCAGTCcgcatccatct
5'-**GNNN**ACGAGCTGCCAGTCcaactcgtca
5'-**CNNN**ACGAGCTGCCAGTCcacsgcaaca
5'-**TNNN**ACGAGCTGCCAGTCgccagtgtta

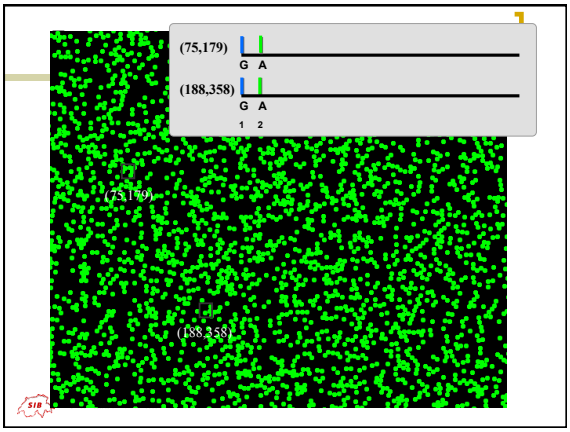
[a]Four-base overhangs in bold and decoder binding sites in low

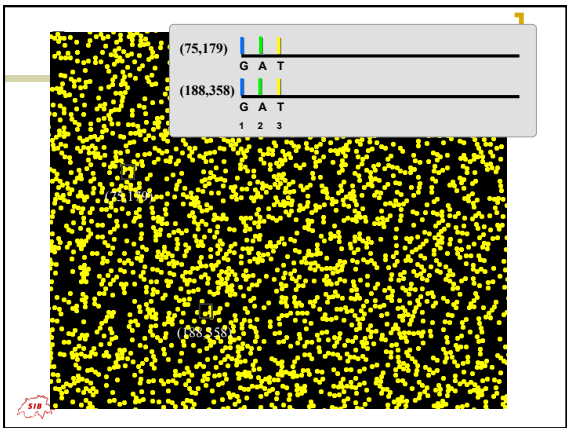


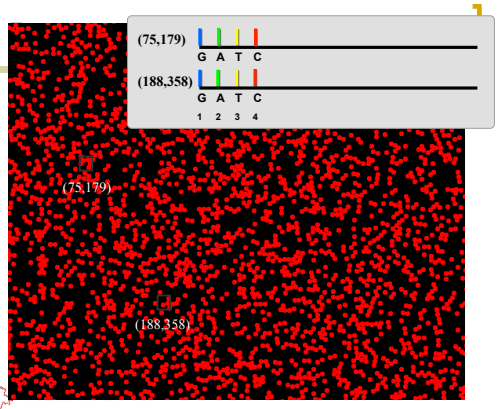
Setup for acquiring signature sequences



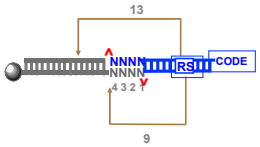

(75,179)
G
(188,358)
G
1
(75,179)
(188,358)


(75,179)
G A
(188,358)
G A
1 2
(75,179)
(188,358)


(75,179)
G A T
(188,358)
G A T
1 2 3
(75,179)
(188,358)

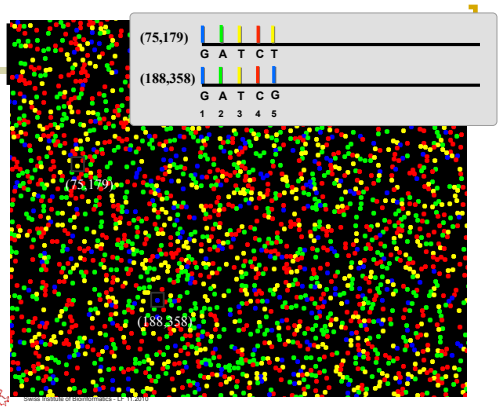


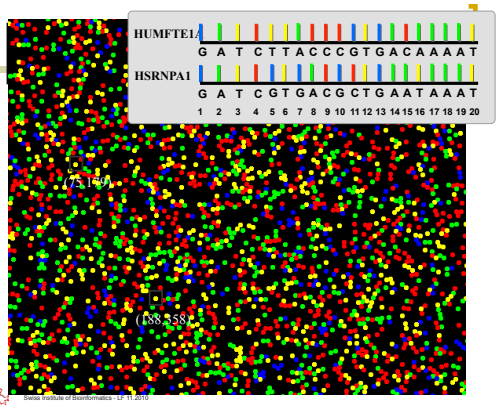
Stepping in Four Bases



Digest with Type IIs enzyme to uncover next 4 bases

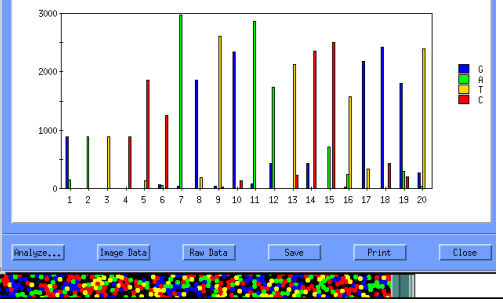
Repeat Cycle





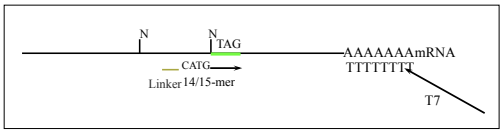
Swiss Institute of Bioinformatics - LF 11.2010

Sequencing a Specific Bead Address



From signatures to genes

- Database searching
- RT-PCR
- cDNA library screening.



- Functional hypotheses and analyses

Swiss Institute of Bioinformatics - LF 11.2010

Problems with SAGE/MPSS data

- Sequencing errors in the libraries
- Sequencing errors in the ESTs used to derive the signatures
- Incomplete digestion by Rx enzymes
- Ambiguities in signature to gene maps

Unreliable sequences

- 1% error rate in sequence means that there is 10% chance a signature is wrong in either library or EST
- Correction in libraries by elimination of low-frequency signatures (singlets) and by merging of neighbours of abundant tags
- Correction in ESTs and detection of SNPs by aligning to genome data

Getting signatures from the transcriptome (NCBI)

- Separate out individual species (e.g., human) sequences from GenBank submission records.
- Assign a SAGE tag to each sequence, by:
  - assigning sequence orientation through a combination of identification poly-adenylation signal (ATTAAA or AATAAA), poly-adenylation tail, and sequence label, and
  - extracting a 10 base tag 3'-adjacent to the 3'-most NlaIII site (CATG).
- Use information from NCBI's UniGene project, assigning an UniGene identifier to each species sequence with a SAGE tag.

The problem of poly(A)

- There is one chance in 256 that the first four nt upstream of the poly(A) are CATG
- There is one chance in 25 that CATG is found within 10 nt of the poly(A)
- Therefore, tags containing multiple A's at their 3' end may not be mapped correctly
- In fact, tags consisting of only A's are found very commonly in SAGE/MPSS libraries

Multiple tags per gene and genes per signature

- Many (probably most) genes have more than one polyadenylation site, and may be associated with multiple UniGene clusters
- About 1% of the tags originate from partially cleaved cDNA (i.e. from the 2nd restriction site)
- The same tag can appear in more than one mRNA, and have a different probability of being generated from each

Getting signatures from the genome

- Extract poly(A) proximal 3' tags from EST trace files and map to the genome
- Map exons on the genome from the transcriptome
- Follow the exons from the 3' tag to find NlaIII site and SAGE tag
- This identifies 120'000 reliable tags on the human genome

Efficiency of tag annotation

	HB4a	HCT-116	Combined
Total tags	17354	24065	27965
Contaminants [a]	160	264	276
Match virtual transcripts [b]	12109 (70%)	14699 (62%)	17992 (65%)
Match NCBI models [c]	9476 (55%)	10883 (46%)	12326 (44%)
Match Ensembl transcripts	8561 (50%)	9842 (41%)	11105 (40%)

a) Contaminants include mitochondrial and ribosomal RNAs and repetitive elements.
b) Percentages are calculated relative to total tags minus contaminants.
c) Combination of experimental and predicted transcripts (NM and XM identifiers) in RefSeq (November 8, 2002).

Distribution of tag abundance

Abundance (tpm) [a]	# of tags	# identified [b]	# of genes
>10000	7	7	3
>5000	25	24	14
>1000	154	149	120
>500	298	280	229
>100	1719	1600	1397
>50	3261	3060	2631
>10	10519	9608	8018
>5	15145	13517	10876
>1	27965	25779	17992

Most of the tags that cannot be identified are derived from lowly expressed genes

a) Mean between the HB4a and HCT-116 libraries.
b) Identification includes a match to a gene, to a known contaminant, or to a potential sequencing error or polymorphism; it does not include a match to the genome.

Jongeneel CV et al. Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing.Proc Natl Acad Sci U S A. 2003 Apr 15;100(8):4702-5.

Conclusion

- What is your biological question?
- Choose the right technique
- Compare the prices
- Do the right experiment
- Analysis the results, does it answer your question?
- Always ask for help - before - doing a mistake

- Questions ?