



Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Federal Department of Economic Affairs DEA

Agroscope Changins-Wädenswil Research Station ACW

Diagnostic Microarrays: How to get there

J. E. Frey

COST853 Zürich, 19. September 2006



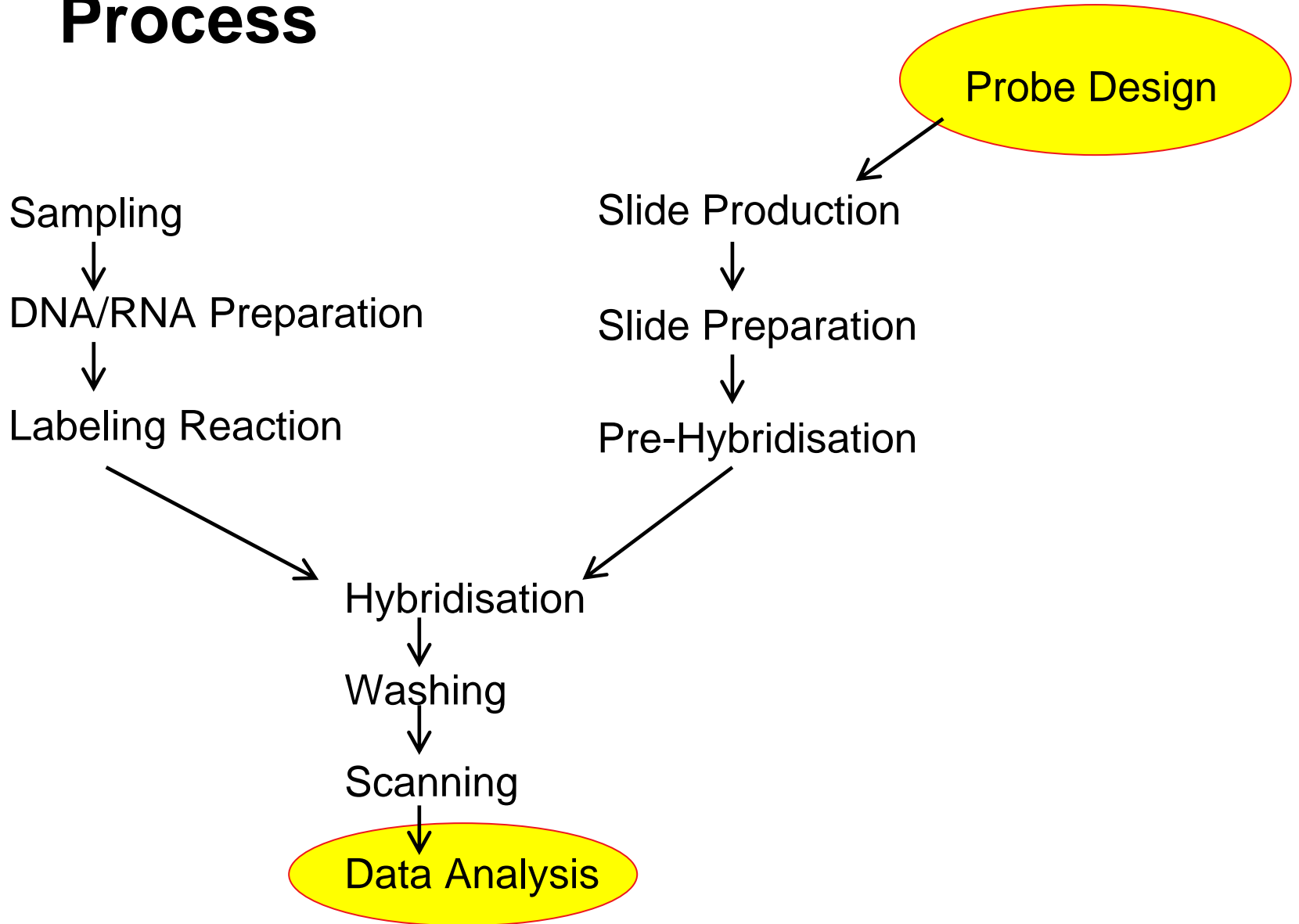
Overview

- **Process**
- **Array types, methods and applications**
- **Strategies for probe design**
- **ACWW – What do we do and which Q related to bioinformatics do we have**

Naïve?



Process





Types, Methods & Applications

... Infinity ...

... how can we use it?





Array-Types

• DNA Arrays

- DNA – Detection
 - SNP (single nucleotide polymorphism)
 - Genotyping / Diagnostics
 - Re-Sequencing
 - Short Oligonucleotides for detection of mismatches
- RNA – Detection
 - Establish transcript profile
 - Long Oligonucleotides (60 mers)
 - Short Oligonucleotides (20 mers)

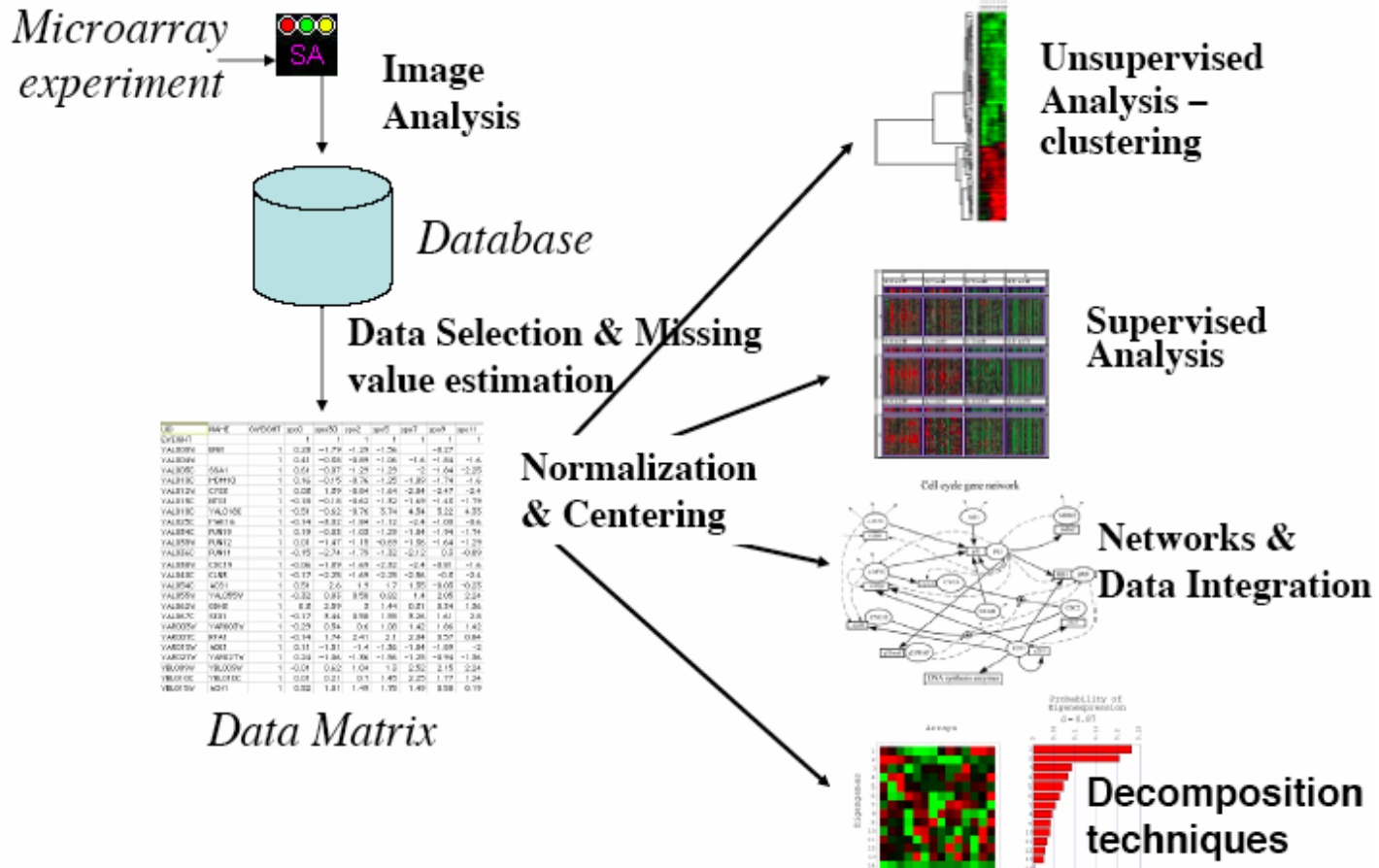
• Protein Arrays

- Epitopes / other chemical structures



Analysis & Quantification

Microarray Data Flow



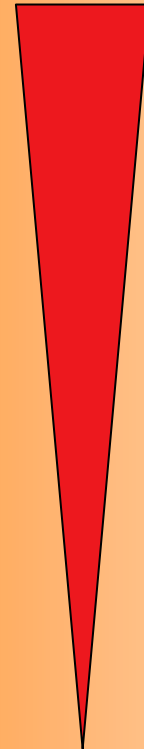
http://stein.cshl.org/genome_informatics/guest_lectures_2003/OlgaLecture.pdf



Applications – DNA Arrays

- **Change in gene expression**
 - Drug development
 - Reaction to drugs
 - Development of therapies
- **Change in genome content**
 - Tumor-classification
 - Risk assessment
 - Prognosis
- **Mutations / Polymorphisms (Diagnostics)**
 - Drug development
 - Development of therapies
 - Tracking disease progression
- **Diagnostics**
 - Detection and identification of pathogens

Screening: Gene



High vs. Low Density

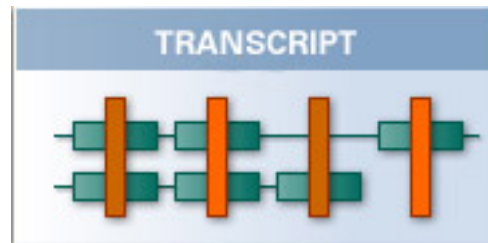
Screening:
Organisms



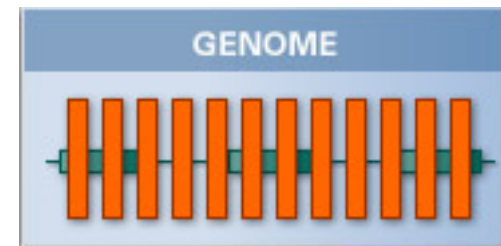
Probe design



- Gene expression



- Gene expression
- Alternative splicing



- Transcription mapping
- SNPs



Oligonucleotides



Strategy 1: Precision by design
(long oligos)



Applied Biosystems
Human Genome Survey Microarray V2.0

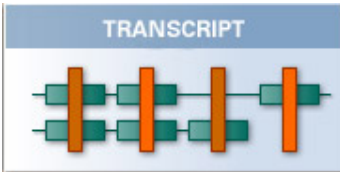
29098 Gens – 60mer Probes, <1500 bp from 3'end (robust labeling)

Probes/Gene	Genes in the V2.0 Array
1	25,299 (~87%)
2	3,096
3	535
4	105

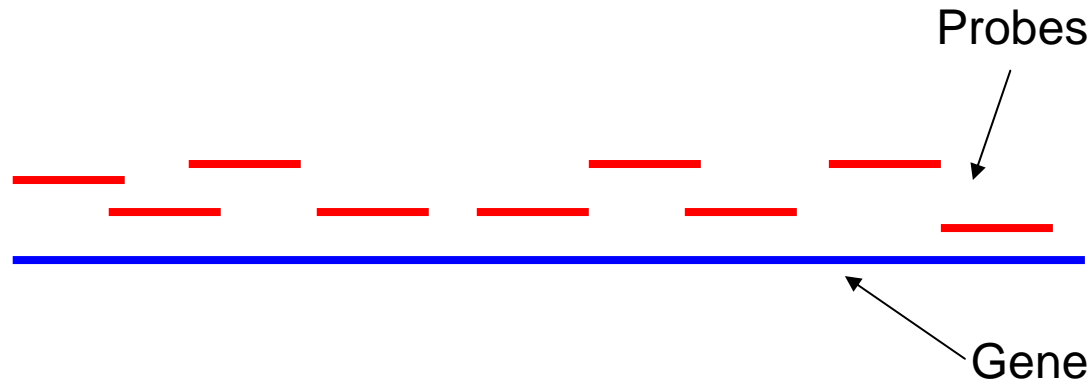
- **Biological processes**
(zB 0.4% Apoptosis genes)
- **Molecular funktion**
(zB 0.2% Ion channels)



Oligonucleotides

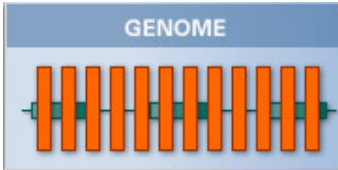


Strategy 2: Precision by redundancy (Transcript)
(short oligos)

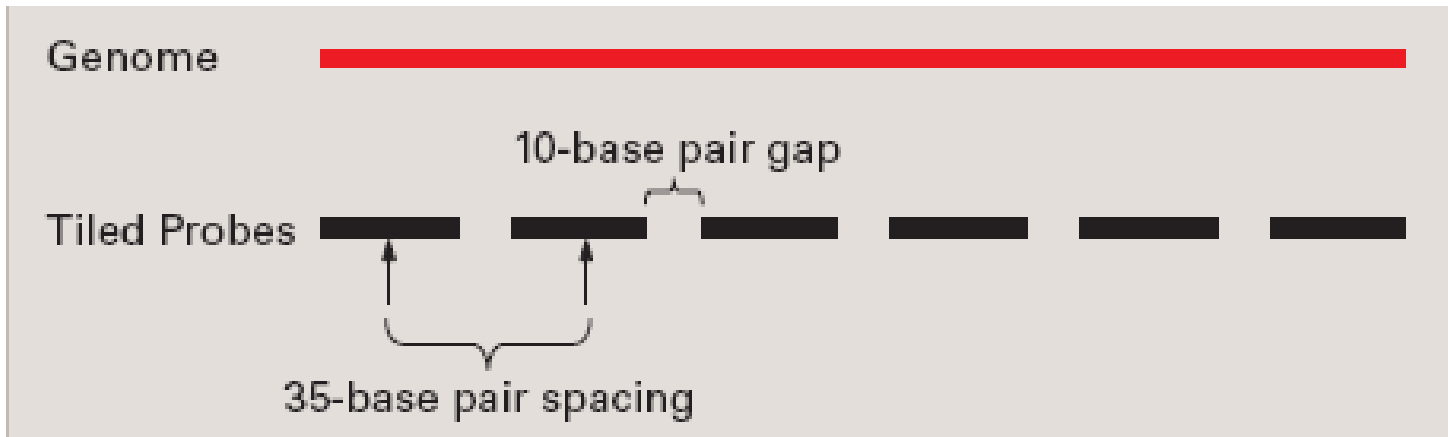




Oligonucleotides



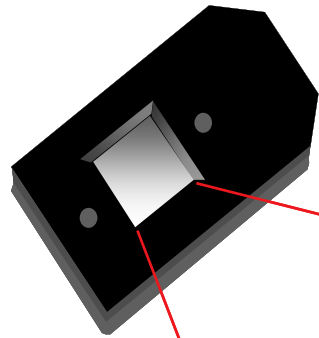
Strategy 3: Whole Genome Tiling Arrays (Genome)
(short oligos)





Examples – Affymetrix

GeneChip Probe Array



1.28cm

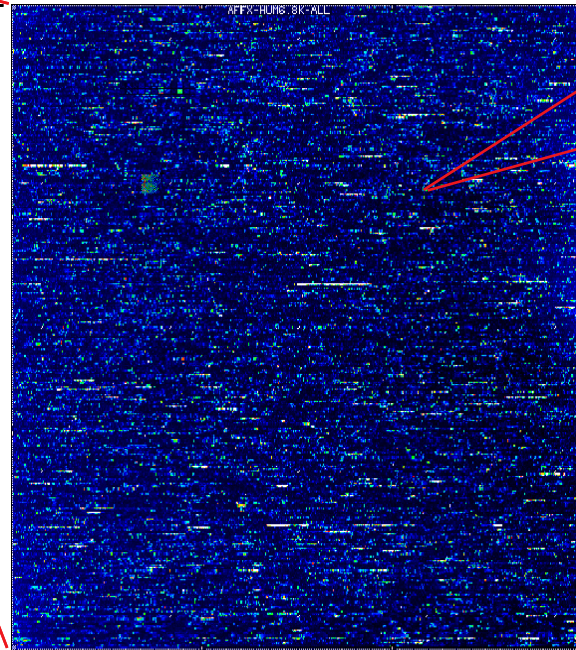
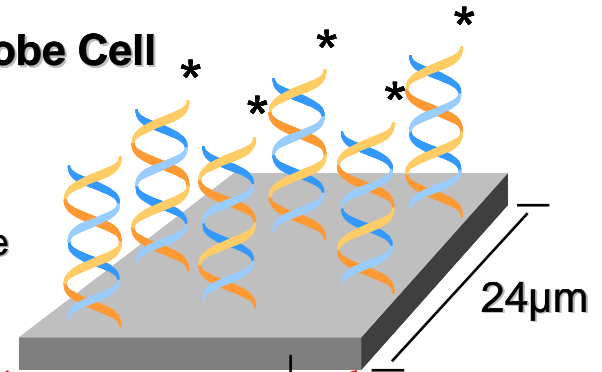


Image of Hybridized Probe Array

Hybridized Probe Cell

Single stranded, fluorescently labeled DNA target
Oligonucleotide probe



Each probe cell or feature contains millions of copies of a specific oligonucleotide probe

Over 200,000 different probes complementary to genetic information of interest

Courtesy: Affymetrix

Deepti Malhotra
Biological Sequence Analysis



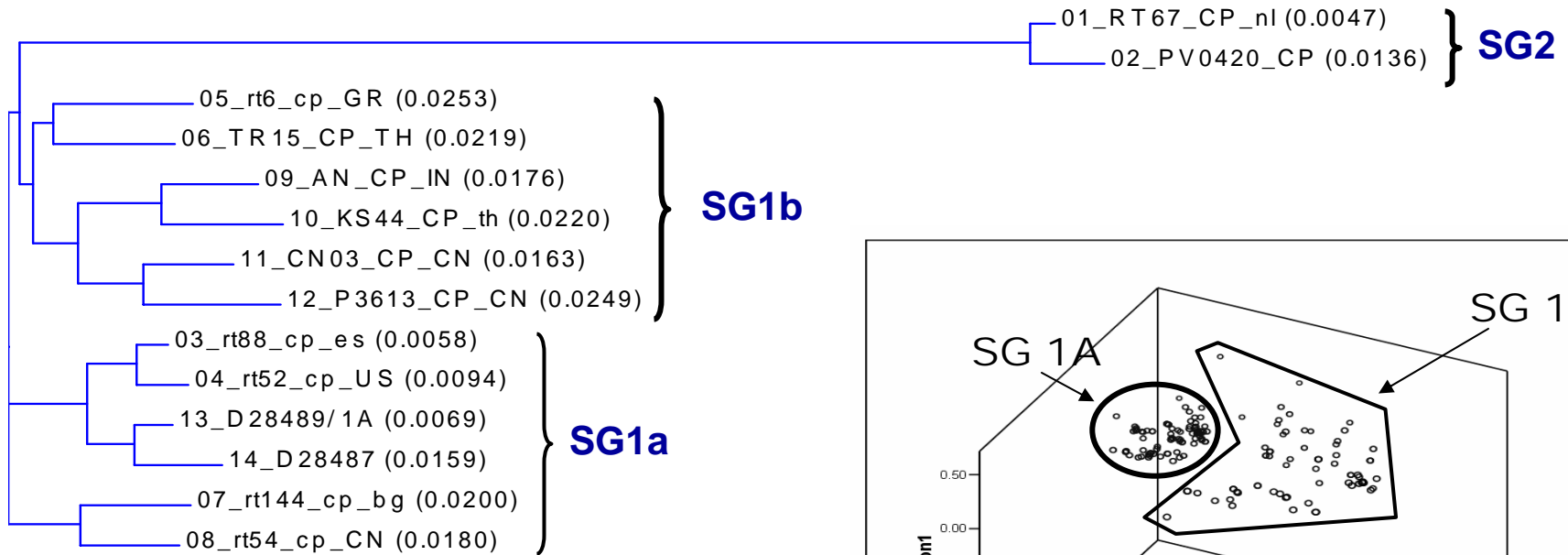
Diagnostic Arrays

Development of diagnostic Microarrays for agriculture in ACW Wädenswil

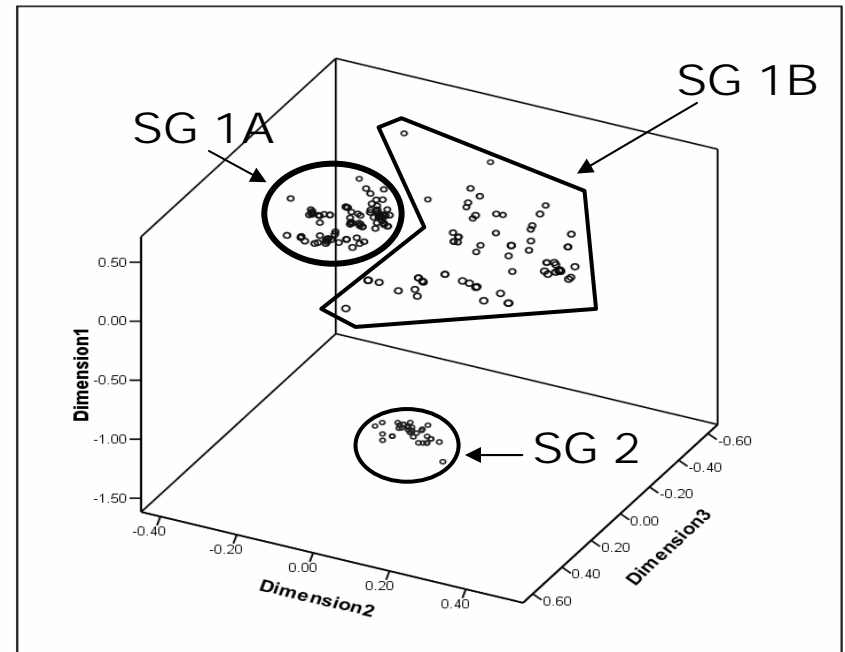
- **2 Approaches**
- **Gene Chip**
 - **Probe design (oligo-sequence)**
 - **Probe design (strategy hierarchial)**
- **Genome Chip**
 - **Probe design (strategy?)**



Probe Design: Which groups?



Unrooted Neighbour-Joining Tree



Three-dimensional scaling



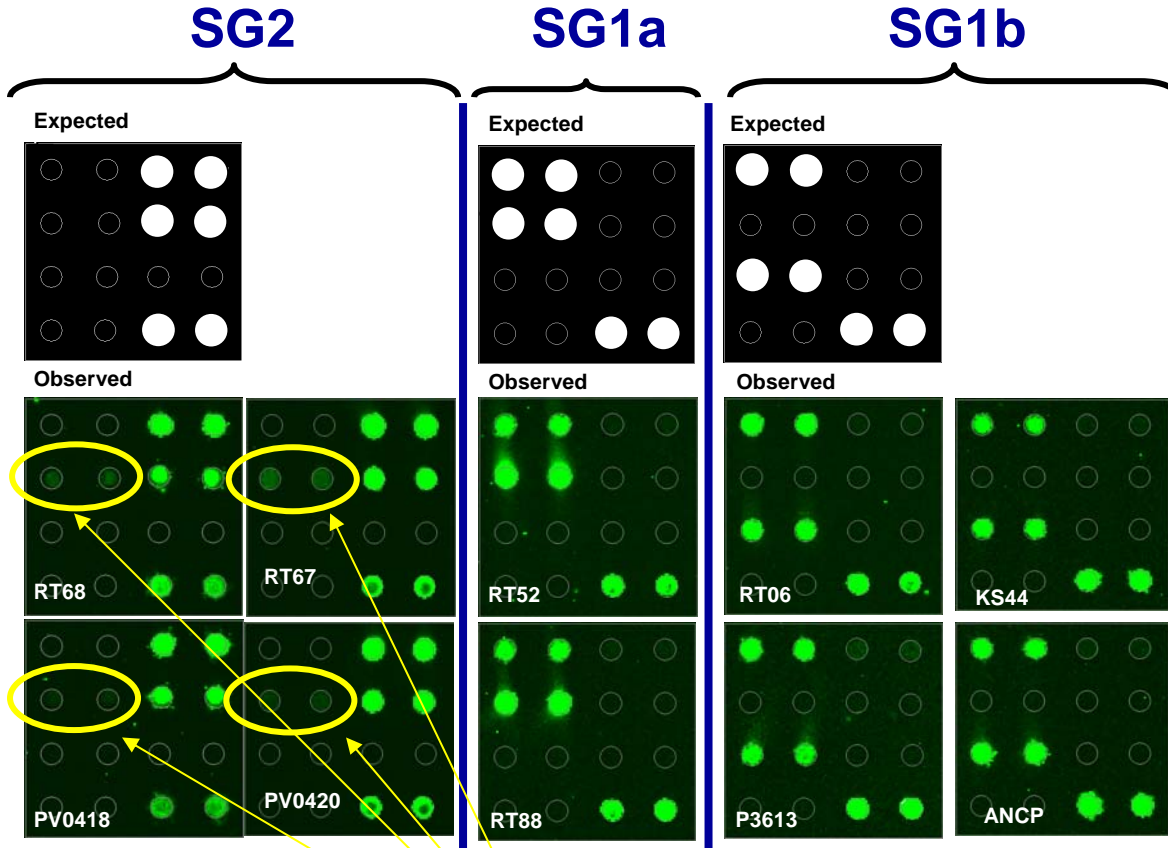
Probe Design: Which targets?

Serogroup 2 Probes

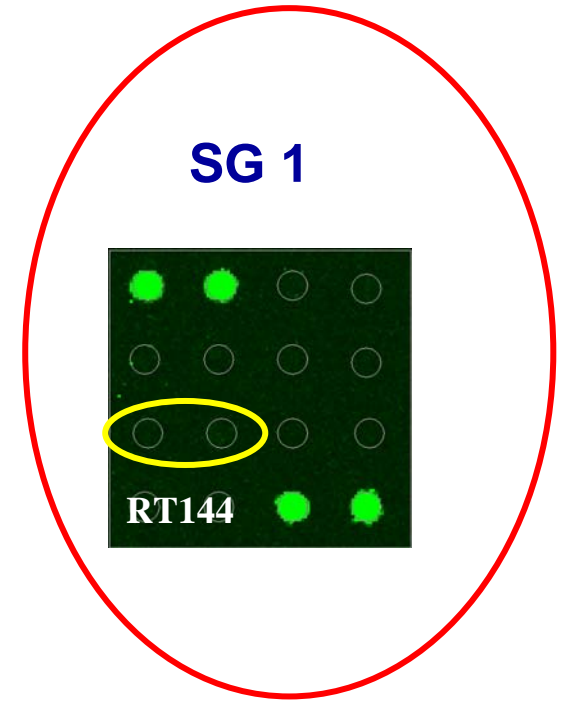
	88	100	113	134	140	159	
RT67	GATCCAGGGTTGCGTGCTTTGACTCA			AAACCCTCGC	CATTGGTCGTCC	CACT	SG2
PV0420	GATCCAGGGTTGCGTGCTTTGACTCA			AAACCCTCGC	CATTGGTCGTCC	CACC	
RT144	GATGCTAAGTTTAGAGTCTTGTCGCA			AGACGTTAGCAGCTGGTCGTCC	TACC	SG1a	
RT54	GATGCTAAGTTTAGAGTCTTGTCGCA			AGACGTTGGCAGCTGGTCGTCC	TACC		
ANCP	GATGCTAAGTTTAGAGTCTTGTCGCA			AGACGTTAGCAGCTGGTCGTCC	TACT		
KS44	GATGCTAAGTTTAGAGTCTTGTCGCA			AGACGTTAGCAGCTGGTCGTCC	TACT		
CN03	GATGCTAAGTTTAGAGTCTTGTCGCA			AGACGTTAGCTACTGGTCGTCC	AACT		
P3613	GATGCTAAGTTTAGAGTCTTGTCGCA			AGACGTTAGCAACTGGTCGTCC	TACT		
TR15	GATGCTAAGTTTAGAGTCTTGTCGCA			AGACATTAGCAGCTGGTCGTCC	AACT		
RT06	GATGCTAAGTTTAGAGTCTTGTCGCA			AGACGTTAGCAGCTGGTCGTCC	AACT		
RT88	GATGCTAAGTTTAGAGTCTTGTCGCA			AGACGTTAGCAGCTGGTCGTCC	AACT	SG1b	
RT52	GATGCTAAGTTTAGAGTCTTGTCGCA			AGACGTTAGCAGCTGGTCGTCC	AACT		
D28489	GATGCTAAGTTTAGAGTCTTGTCGCA			AGACGTTAGCAGCTGGTCGTCC	AACT		
D28487	GATGCTAAGTTTAGAGTCTTGTCGCA			AGACGTTGTAGCCGGTCGTCC	AACT		
SG1	-----			-----			
SG1a	-----			-----			
SG1b	-----			-----			
SG2-1	-----			-----			
SG2-2	-ATCCAGGGTTGCGTGCTTTGACTC-			-AACCCTCGC	CATTGGTCGTCC	CAC-	



Hybridization



weak cross-hybridization



False-negative



Reason for cross-hybridization

Serogroup 1 Probes

	287	300	311	432	440	456	
RT67	CAGTCACGGAC	CTATGATAAGAAGCT		TGGCGATGGTAATTCACCGGTTT		TC	SG2
PV0420	CAGTCACGGAC	CTATGATAAGAAGCT		TGGCGATGGTAATTCACCGGTTT		TC	
RT144	CAGTTACAGAATTCGAC	CAAGAACT		TGCGGACGGAGCCTCACCGGTACTG			SG1a
RT54	CAGTCACGGAATTCGATAAGAAGCT			TGCGGACGGAGCCTCACCGGTACTG			
ANCP	CAGTCAC TGAGTTCGATAAGAAGCT			C GCGGACGGAGCCTCACCAGTACTG			
KS44	CAGTCAC TGAGTTCGATAAGAAGCT			TGCGGACGGG GCCTCACCGGTACTG			
CN03	CAGTCACAGAGTTCGATAAGAAGCT			C GCGGACGGAGCCTCACCGGTACTG			
P3613	CAGTCACAGAGTTCGAC	CAAGAAGCT		C GCGGACGGAGCCTCACCGGTACTG			
TR15	CAGTCACGGAGTTCGATAAGAAGCT			TGCGGACGGAGCCTCACCGGTACTG			
RT06	CAGTCACGGAGTTCGATAAGAAGCT			TGCGGACGGAGCCTCACCGGTACTG			
RT88	CAGTCACGGAATATGATAAGAAGCT			C GCGGACGGAGCCTCACCGGTACTG			SG1b
RT52	CAGTCACGGAATATGATAAGAAGCT			C GCGGACGGAGCCTCACCGGTACTG			
D28489	CAGTCACGGAATATGATAAGAACT			TGCGGACGGAGCCTCACCGGTACTG			
D28487	CAGTCACGGAATATGATAAGAACT			TGCGGACGGAGCCTCACCGGTACTG			
SG1	-----			- GCGGACGGAGCCTCACCGGTACTG			
SG1a	CAGTCACNGAATATGATAAGAAGC			-----			
SG1b	CAGTCACNGAGTTCGATAAGAAGC			-----			
SG2-1	-----			-----			
SG2-2	-----			-----			



Hierarchical Design



Bacteria

16S rDNA Probes

Kl.

Betaproteobacteria

Gamma proteobacteria

Ord.

Burkholderiales

Enterobacteriales

Pseudomonadales

Xanthomonadales

Actinomycetales

Gram negative

Gram positive

Fam.

Burkholderiaceae

Enterobacteriaceae

Pseudomonadaceae

Xanthomonadaceae

Microbacteriaceae

Gat.

Ralstonia

antonia

Erwinia

Pseudomonas

Xanthomonas

Clavibacter

Curtobacterium

Sp.

solanacearum

stewartii

amylovora
chrysanthemi

syringae

vesicatoria

axonopodis

arboricola

oryzae

fragariae

flaccumfaciens

subsp. sepedonicus
subsp. insidiosus
subsp. michiganensis

Subsp. Pathov.

pv. stewartii

pv. persicae

pv. citri
pv. phaseoli
pv. vesicatoria
pv. dieffenbachiae

pv. pruni
pv. fragariae

pv. oryzae
pv. oryzicola

pv. translucens

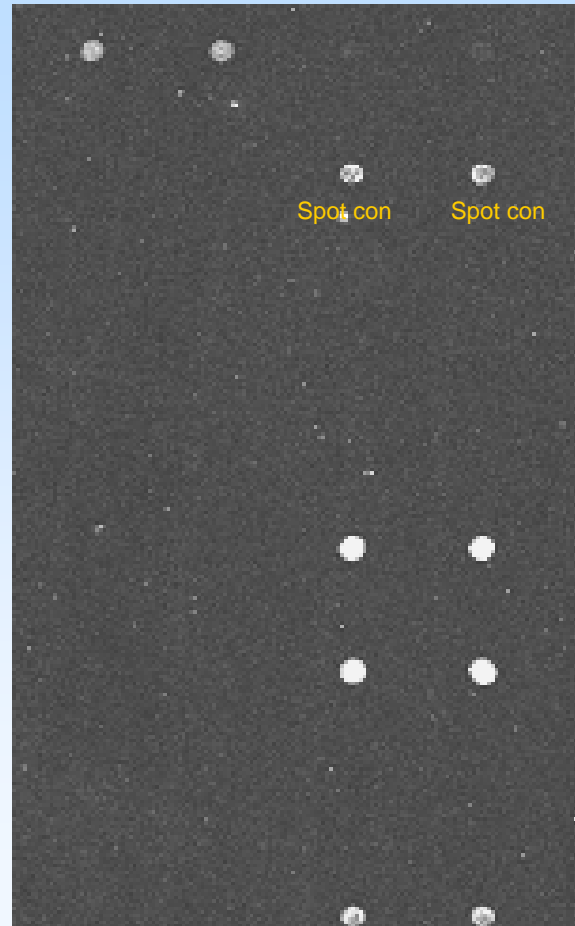
pv. flacc.

DNA Probe



16S-rDNA Probes (20bp)

Bacteria	Bacteria	Archaeob.	Archaeob.
H2O	H2O	spot. con.	spot. con.
Pseudo.	Pseudo.	Xanth.	Xant
Xanthom.	Xanthom.	Xylella	Xylella
Burk.	Burk.	Ral.	Ral.
B. cary.	B. cary.	Ral. sol.	Ral. sol.
Xyl. amp.	Xyl. amp.	H2O	H2O
H2O	H2O	spot. con.	spot. con.



Ralstonia solanacearum



Problem

- **Q: In the absence of an ARB background – how do we assess the quality of our probes?**
- **Example:**
 - **Lepidoptera >150'000 spp (>3000 M-Europe)**
 - **Typical “Key”:**
30 spp, 10-50 ind./sp., 3 Haplotypes/sp.

>99% unknown (sequence, but also variance patterns)

• A: we can't!



Potential Solution

- **Increase of redundancy**
 - Increase the number of probes per amplicon
 - Increase the number of queried genes
- **Q: How many probes are required to be 95% sure to have a correct ID?**

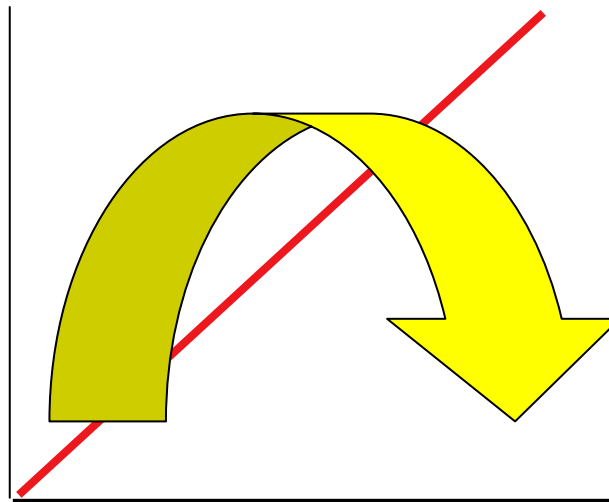


Bioinformatics project idea?



Solutions may cause problems

- **False positives**
 - Increased number of probes will increase the potential for false positives
- ***Q: With increasing probe number, how does***
- ***This potential affect data quality?***



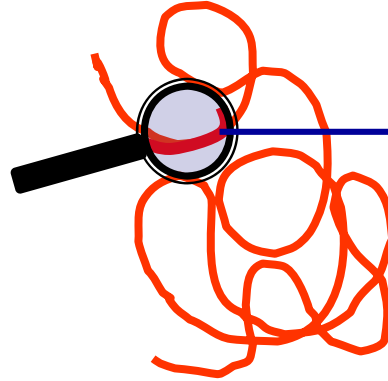
Bioinformatics project idea?



Gene vs. Genome Principle

Gene (intelligent design)

- sequence information on a small subsample

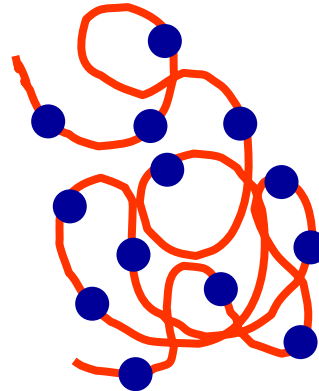


	186	190	200	210	22
AY332664	186	CCATTCTAGGGTCAATCAACTTCATTACCACAATTA			
AY332669	186	CCATTCTAGGGTCAATCAATTTTCATTACCACAATTA			
AY332671	186	CAATTCTCGGCTCAATCAATTTTCATTACCACAATTA			
AY332682	186	CAATCCTTGGGGCAATTAATTTTCATTACCACATATCA			
AY332685	186	CAATCCTTGGGGCAATTAATTTTCATTACTACTATCA			
AY332684	186	CAATCTTGGGGCAATCAACTTTATTACCACAATTA			
AY332686	186	CAATCCTTGGAGCAATTAATTTTATTACCACAATTA			

Co-dominant markers

Genome (random design)

- presence/absence information on the entire genome / target DNA (in mixed samples)
- potential for SNP detection



Org1: 1000101101001
 Org2: 1001001101100
 Org3: 1000010101101
 Org4: 0100110101111

Dominant markers

SNPs: ddA / ddG / ddT / ddC

Co-dominant markers



Genome Chip Design

- **Use of many primers with random sequence**
 - **No need for sequence information > can be used for any species**
 - **Many potential targets per genome > High resolution**
- **Analysis on microarray**
 - **De-multiplexing**
- **Identification by comparison to database**
 - **E.g., Clusteranalysis**



Genome Chip Design

In situ selection of non-selfcomplementary random oligos

using Visual Basic

- GC content 50-55%
- 3'-base a G or C
- Tm 55 - 58°C (Sugimoto 1996)
- min 5 bp difference and
max 7 consecutive base complements among any probe pairs
- no poly-N > 4 bp
- no hairpins > 4 bp => 4386 oligonucleotides
- go up to Tm 56 – 58°C => 824 oligonucleotides
- Take every second => 412 oligonucleotides



Genome Chip Design

Virtal Genom-Chip - low density

- 768 Markers (Random Sequence)
- 2 fully sequenced Strains of 4 Microorganisms:
 - *Escherichia coli* K12 MG1655
 - *Escherichia coli* O157H7
 - *Staphylococcus aureus* Mu50
 - *Staphylococcus aureus* N315
 - *Mycobacterium tuberculosis*
 - *Mycobacterium tuberculosis* CDC1551
 - *Neisseria meningitidis* serogroup A strain Z2491
 - *Neisseria meningitidis* serogroup B strain MC58
- All by 12.9.2001 publicly available MO sequences > 2 Mio bp

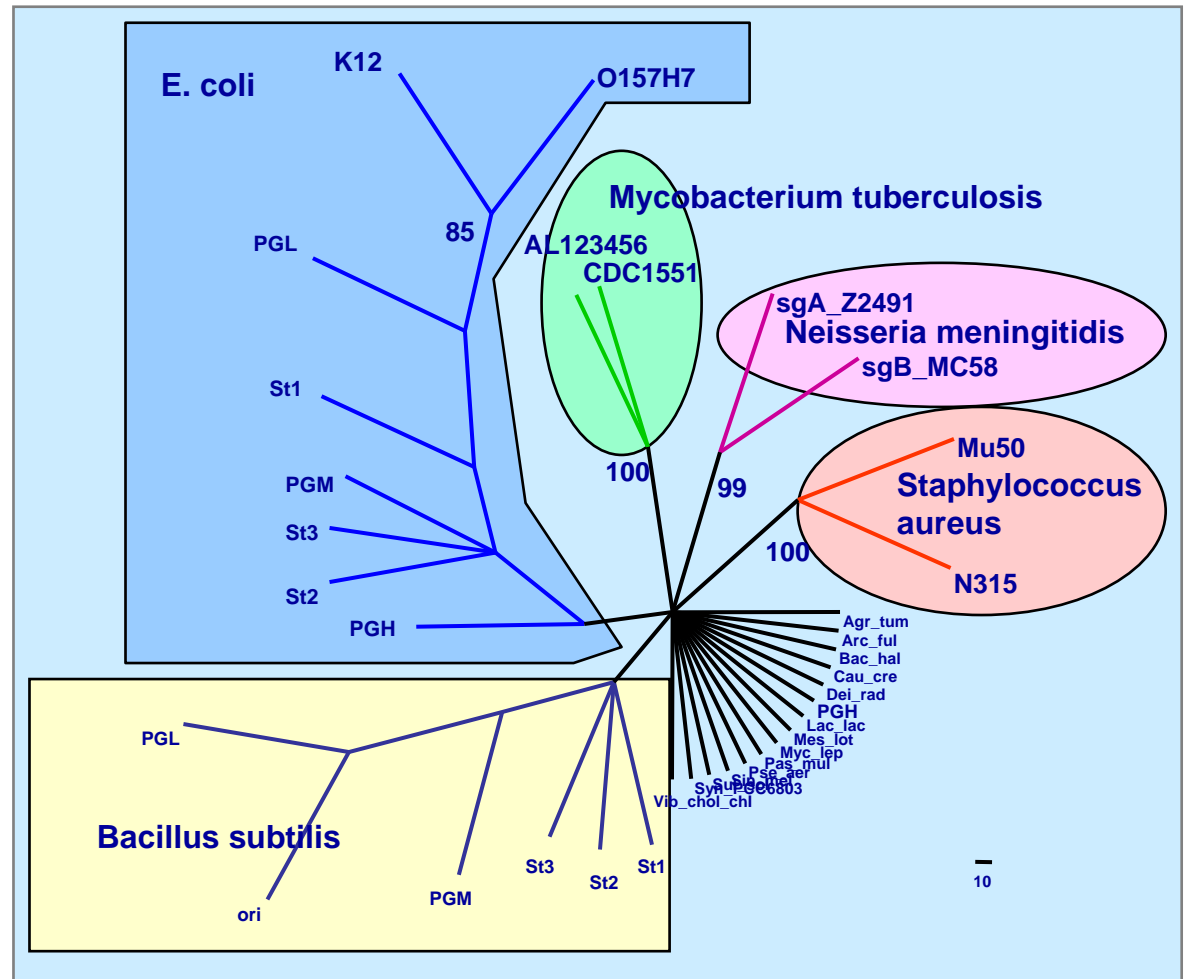


Genome Chip Design

Virtual Random Chip - Robustness

Computer Simulation

Program input:
Two different fully sequenced strains of each of four species, complemented with ca. 40 other full sequences of other species of microorganisms





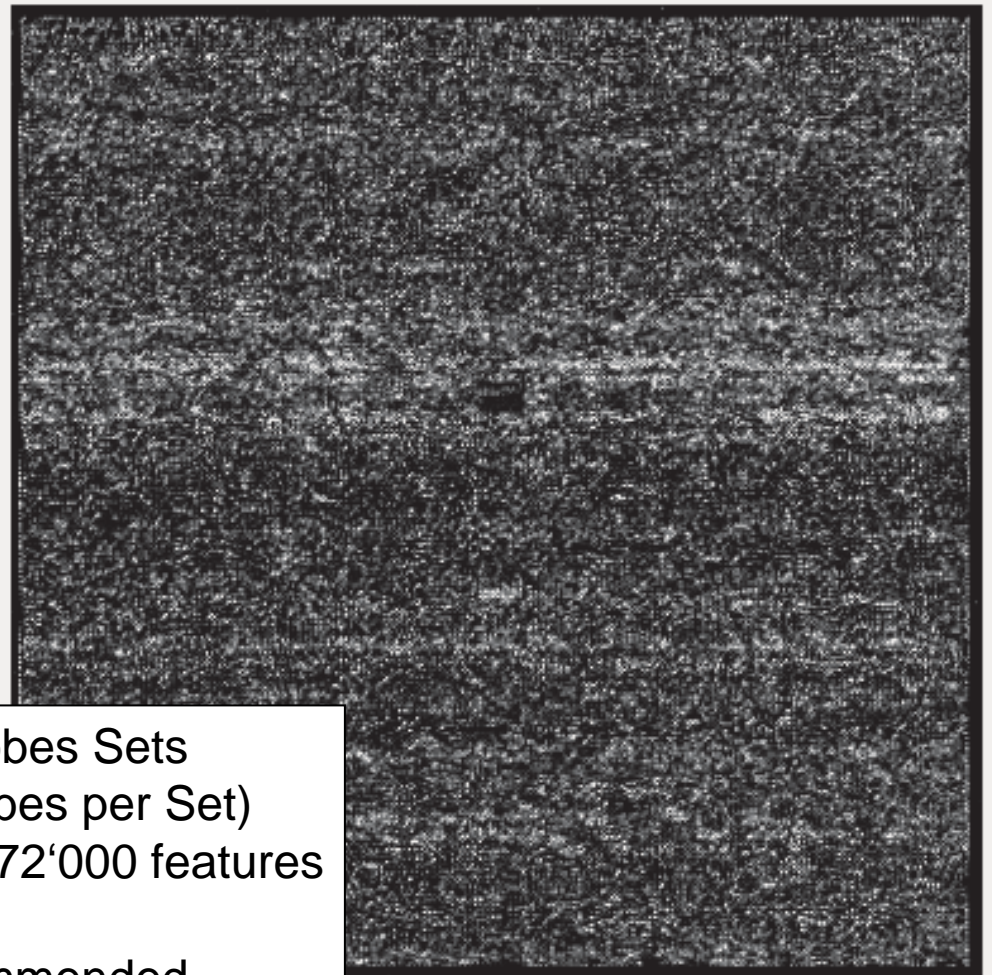
How to realize?

Nimblegen?

**10'000 probes
@50/probe**

= 500'000

NimbleExpress™ Array Image



Array Format: 12'000 Probes Sets
(11 Probes per Set)
Total 272'000 features
Spot Size: 17 um
Oligo length: 25mer recommended



Problem I

Good news: No need for >10'000 probes

- **Short oligos:**

- **Specificity** 
- **Reproducibility** 

- **Q: Which probe design strategy makes best use of the available resource?
(>500K probes = 50x redundancy)**



Bioinformatics project idea?



Problem II

Good news: With $>10'000$ probes there may be a potential for individual identification in mixed templates

- **Q: Which analysis strategy would be suitable?**
 - **Eg iterating through data subsamples**



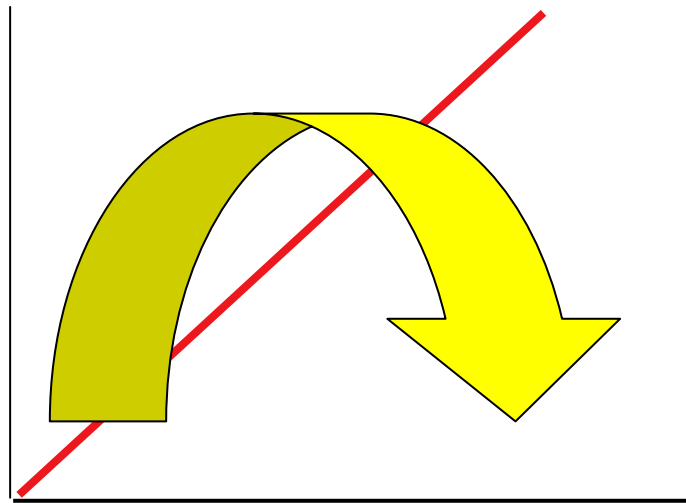
Bioinformatics project idea?



Problem III

@>2Mio probes – can good news turn into bad news?

- Q: How does increased number of probes affect the resolution power?



Bioinformatics project idea?



Team

Markus Oggenfuss
Beatrice Frey
Cosima Pelludat
Frédérique Pasquer
Visiting scientists
Apprentices
Jürg E. Frey



COST Action 853

