NimbleGen® SYSTEMS, INC.

High Definition Genomics

Todd Richmond, PhD

NimbleGen Systems, Inc.

NimbleGen® SYSTEMS, INC.

Topics

- NimbleGen Array Synthesis and Design
- Upcoming products and platforms
- Empirical Probe Optimization

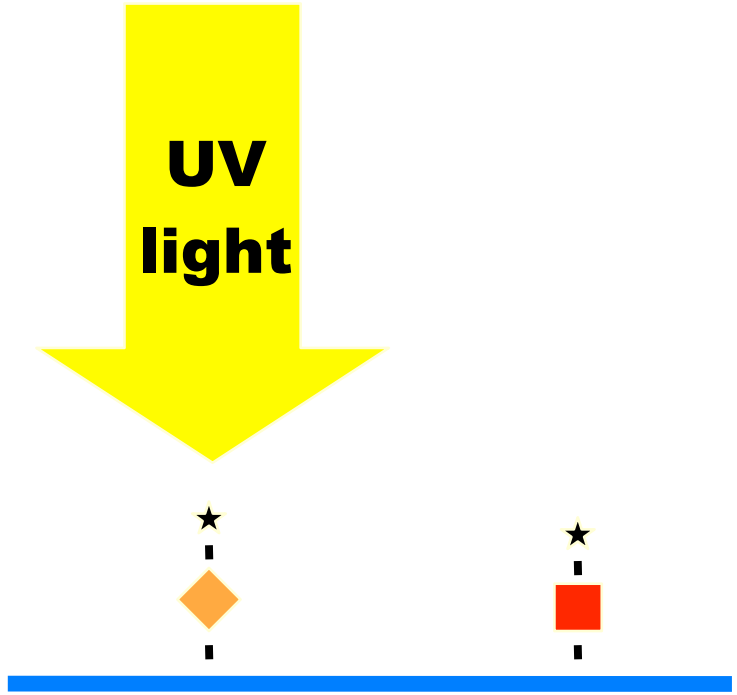
NimbleGen®
SYSTEMS, INC.

NimbleGen Array Synthesis and Design

Oligo Synthesis Controlled By Light

★ Photo labile protecting group

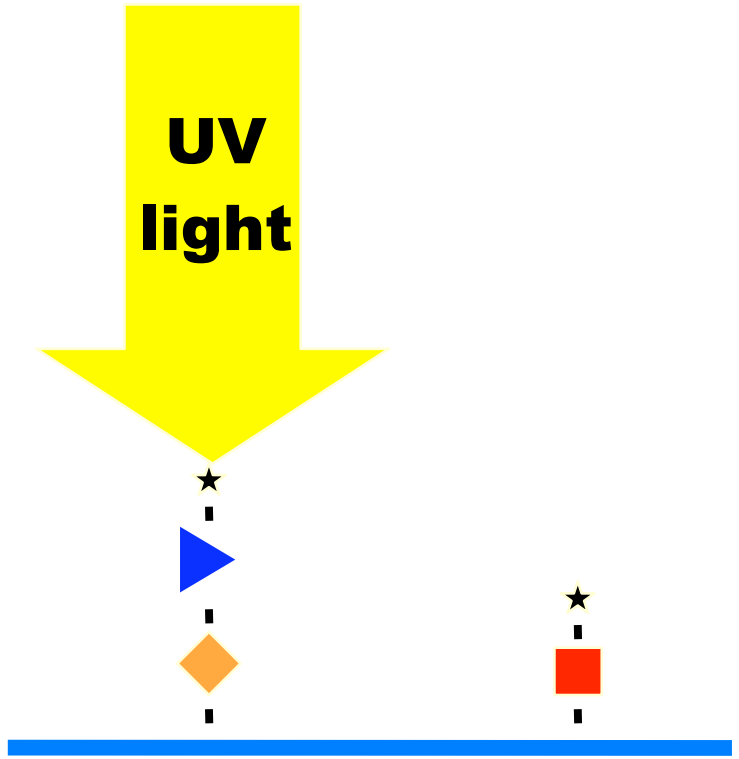
◆ DNA monomer



Oligo Synthesis Controlled By Light

★ Photo labile protecting group

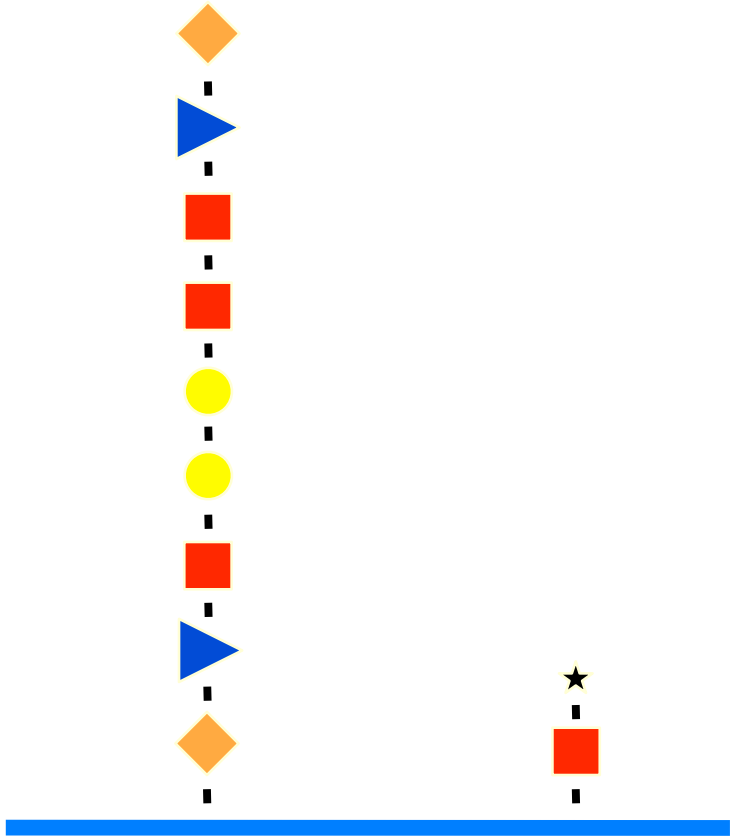
◆ DNA monomer



Oligo Synthesis Controlled By Light

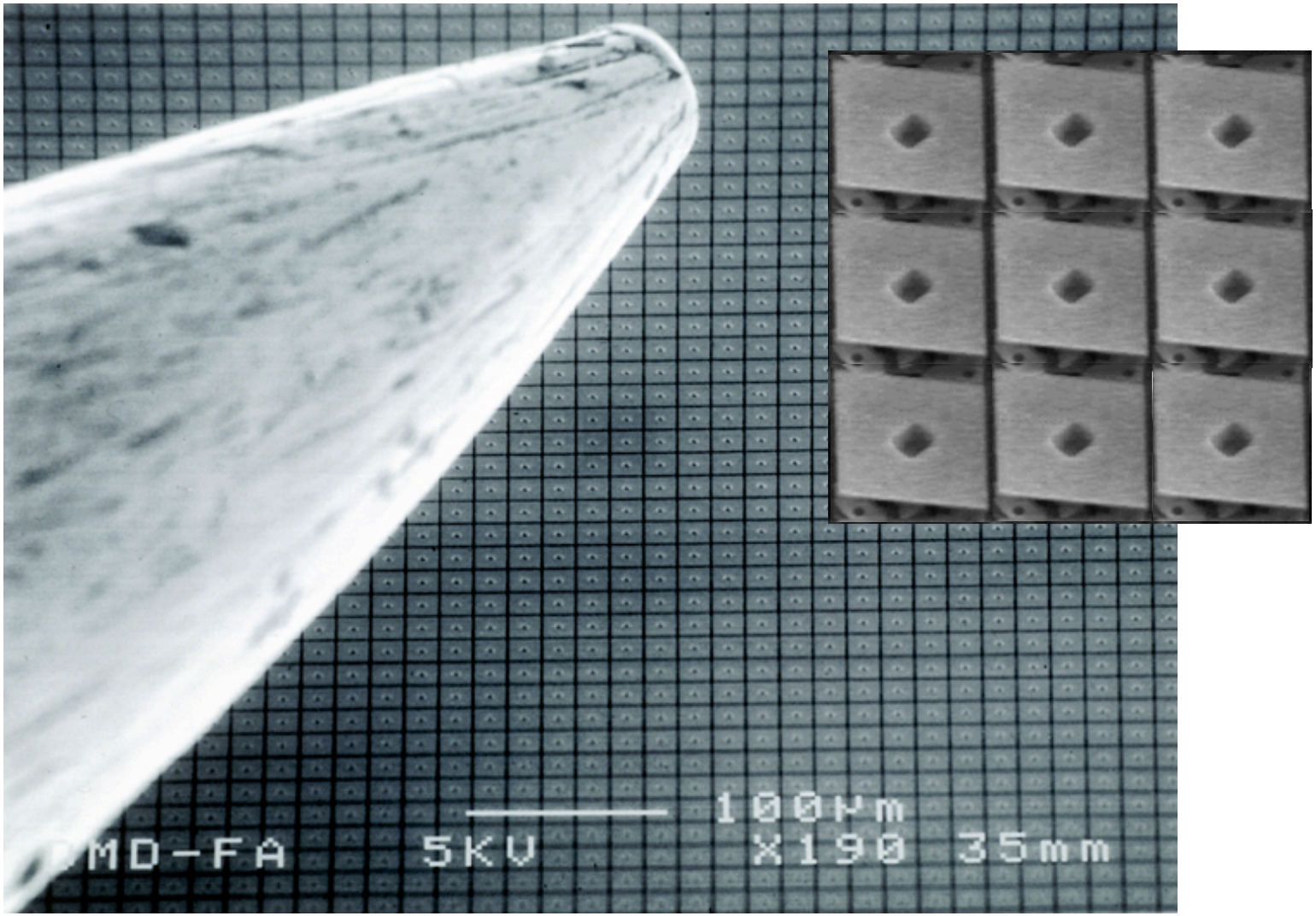
★ Photo labile protecting group

◆ DNA monomer



NimbleGen® SYSTEMS, INC.

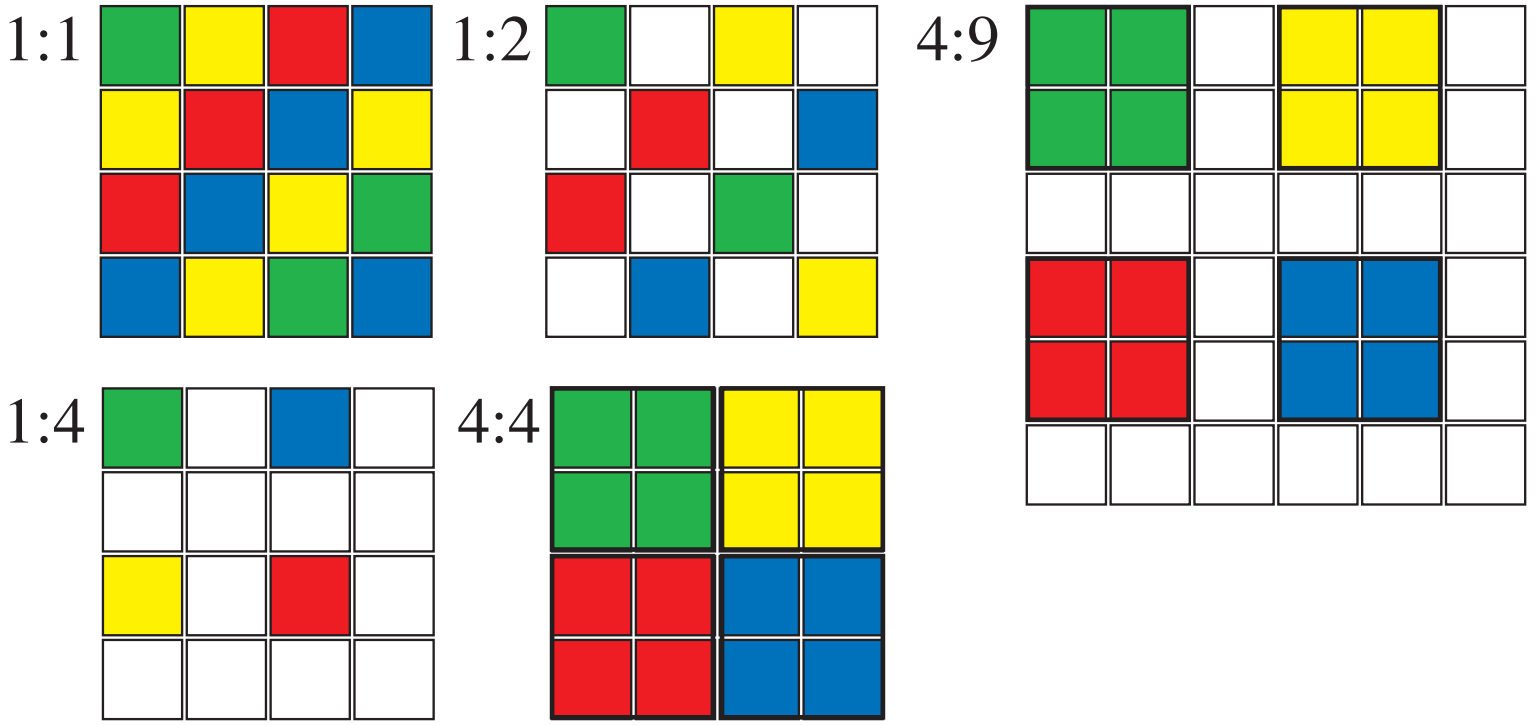
DMD: Digital Mirror Device



High – Definition Genomics SM

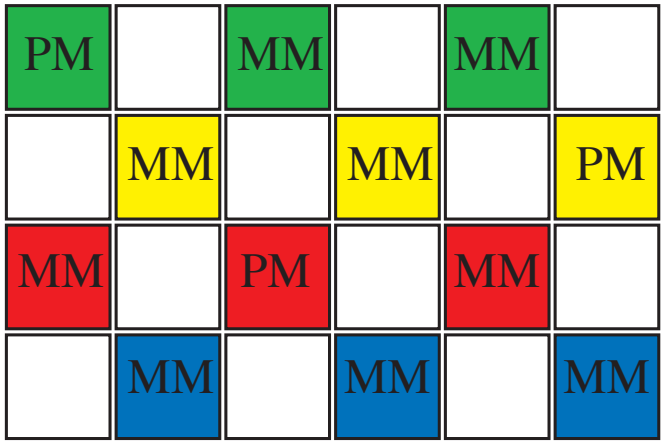
Feature Formats

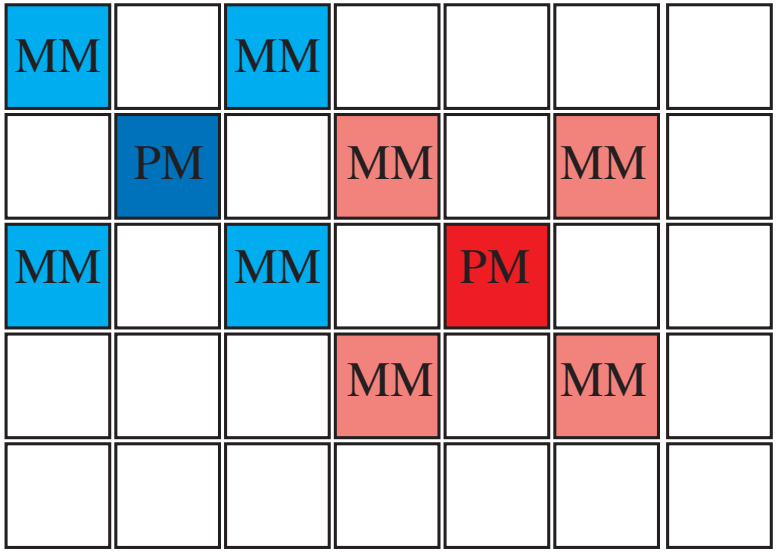
- Flexibility allows different feature sizes and spacing



Feature Formats

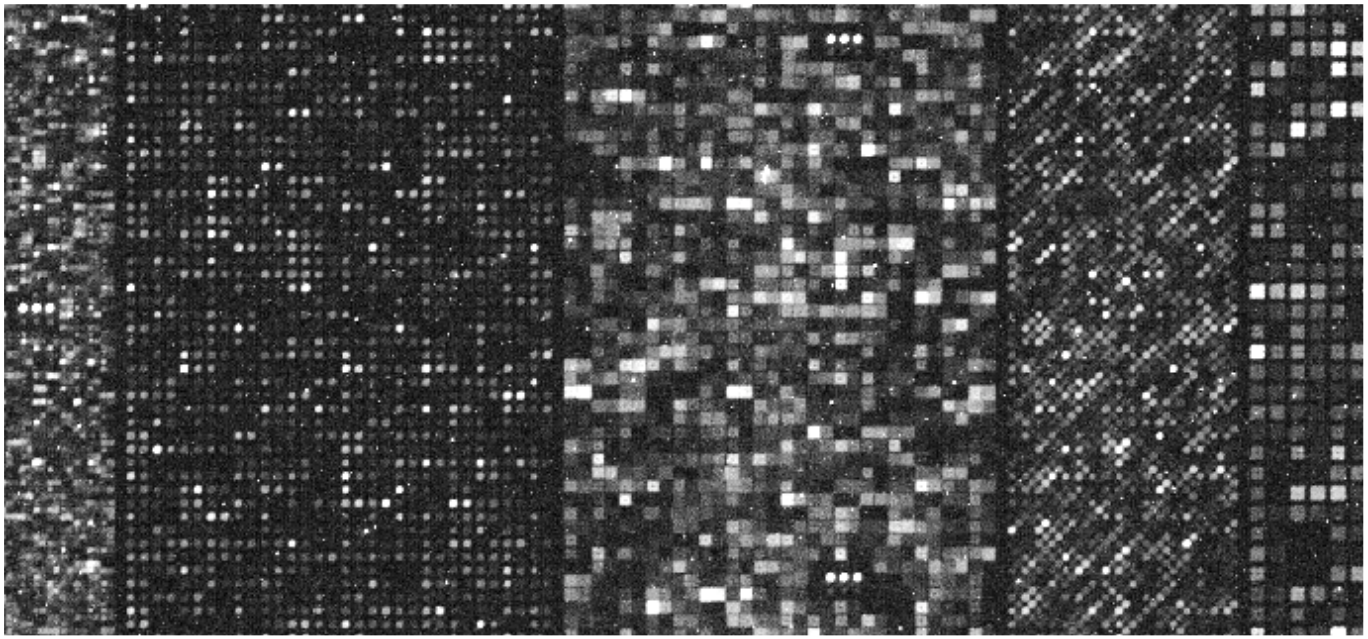
- Can produce custom patterns as well





Multiple Feature Formats

- Can have multiple feature sizes and densities on a single array



Analysis Considerations/Challenges

- Expression
  - feature sizes/density
  - replicate probe sets
  - variable number of probes
  - variable length probes
  - mismatches
  - one or two color

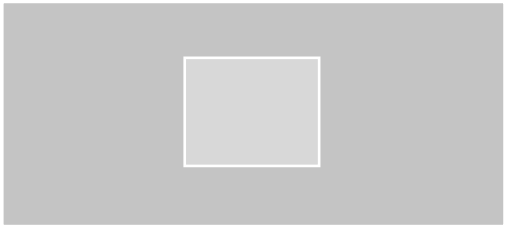
Analysis Considerations/Challenges

- Genomic tiling
  - replicate probe sets
  - variable probe spacing - 1 bp to 6 kb
  - gaps due to repeat masking or uniqueness
  - mismatches
  - one or two color

NimbleGen® SYSTEMS, INC.

Upcoming products and platforms

Path to Higher Probe Density



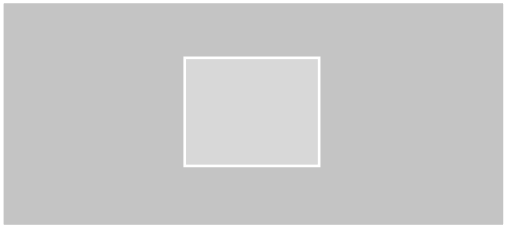
TODAY

786,000 mirrors

1:2 layout

390,000 features/array

Path to Higher Probe Density

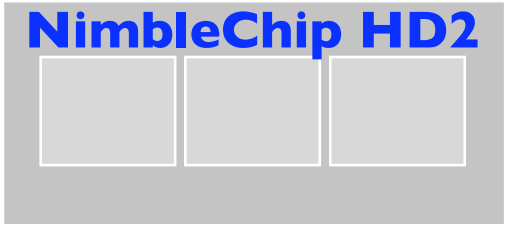


TODAY

786,000 mirrors

1:2 layout

390,000 features/array



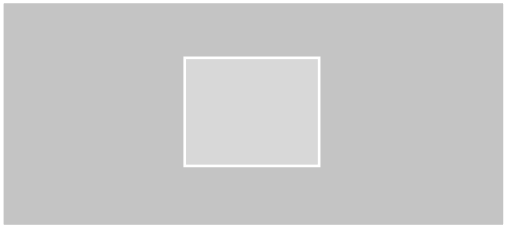
Q1 '07

4,200,000 mirrors

1:2 layout

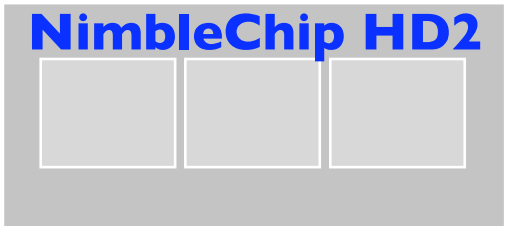
2,100,000 features/array

Path to Higher Probe Density



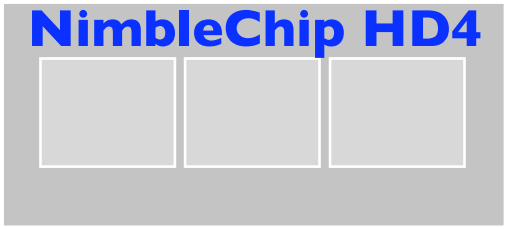
TODAY

786,000 mirrors
1:2 layout
390,000 features/array



Q1 '07

4,200,000 mirrors
1:2 layout
2,100,000 features/array



Future

4,200,000 mirrors
1:1 layout
4,200,000 features/array

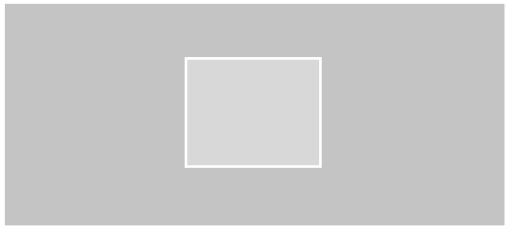
HD2

- 2.1 million feature arrays will target genomic tiling applications first
- 100 bp step for entire human genome in 7 arrays
  - ChIP
  - CGH
  - Methylation
  - Expression tiling (future)

HD2 Data Analysis

- Need to evaluate the impact of processing 6X as much data
- Still un-resolved issues
  - Coordinates for 3 DMDs - 1 or 3 sets
  - Continuous or discontinuous coordinates - do we account for the gap between arrays.
  - Randomization issues - across entire feature space or within each DMD.

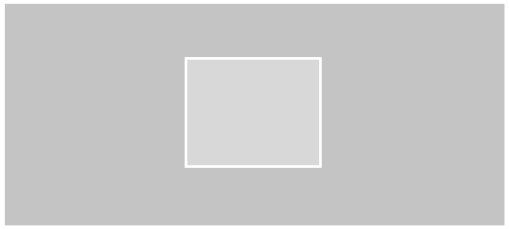
Path to Higher Throughput



TODAY

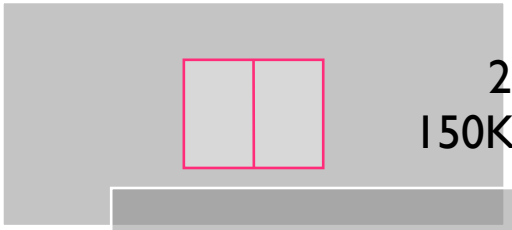
One hybridization chamber

Path to Higher Throughput

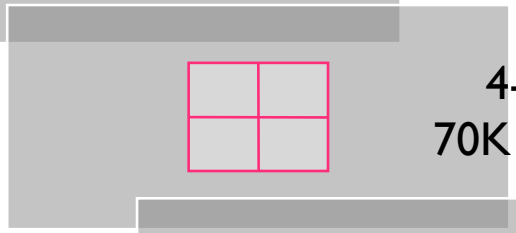


TODAY

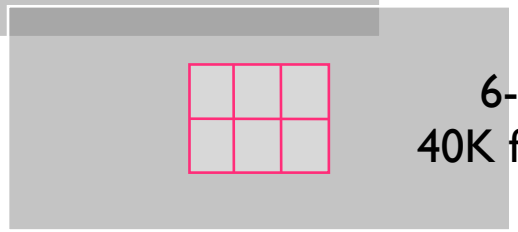
One hybridization chamber



2-plex
150K features



4-plex
70K feature

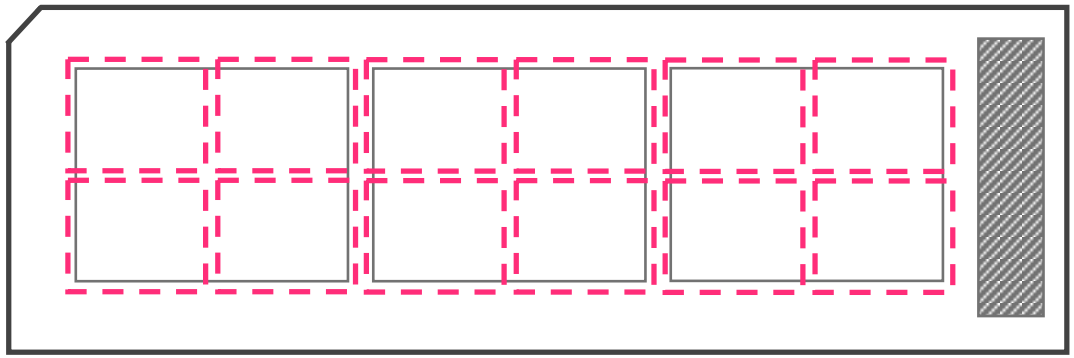


6-plex
40K features

Q4 '06

Multiplex hybridization chambers

Path to Higher Throughput



Combination of 4,200,000 features and multiplex will create a 12plex with ~250,000 features/subarray

NimbleGen SYSTEMS, INC.

Multiplex Arrays

- Current plan is to "explode" multiplex arrays into separate arrays with their own identifiers at scan time.
- Existing software will not have to be changed to support multiplex format
- All current applications will be supported in multiplex format

Other Products Coming

- 2-color expression arrays
  - new protocols for synthesizing full-length cDNA
  - random prime labeling
    - loss of strand information

NimbleGen® SYSTEMS, INC.

Empirical Probe Optimization

Goal

- For each gene in a bacterial genome, find *N* optimal probes so that the expression profile of the entire genome may be assayed in a single array of NimbleGen's multiplex format
  - *N* is generally 2-7 probes
  - previous studies have shown that 2-3 probes is sufficient (as long as they are the "right" ones)

Approach

- From previous studies, we know the following:
  - Bright probes are not always the best
    - While bright, they are often unresponsive
  - Reproducible probes are not always the best
    - While consistent, they sometimes do not respond to changing concentrations
- The only way to find probes that measure changes in DNA concentration is to subject them to different concentrations and watch how they respond

Approach

- Hybridize probes to different genomic DNA concentrations and select those that are bright, reproducible, **and** **follow the concentration curve**.
  - Ensures that all probes see equal copy number
- Need normalization controls to ensure that we can get the appropriate concentration series
  - Genomic DNA from another species used as a standard
  - Random probes used as low end standard

Initial Probe Selection

- Generate all possible 24-mer probes from CDS features
- Gather normal information for standard probe selection
  - Uniqueness check against both strands of whole genome
- E. coli probes are checked against target genome
- Organism probes not checked against E. coli. Hyb will eliminate those that, by chance, share similarity to E. coli.

Initial Probe Selection

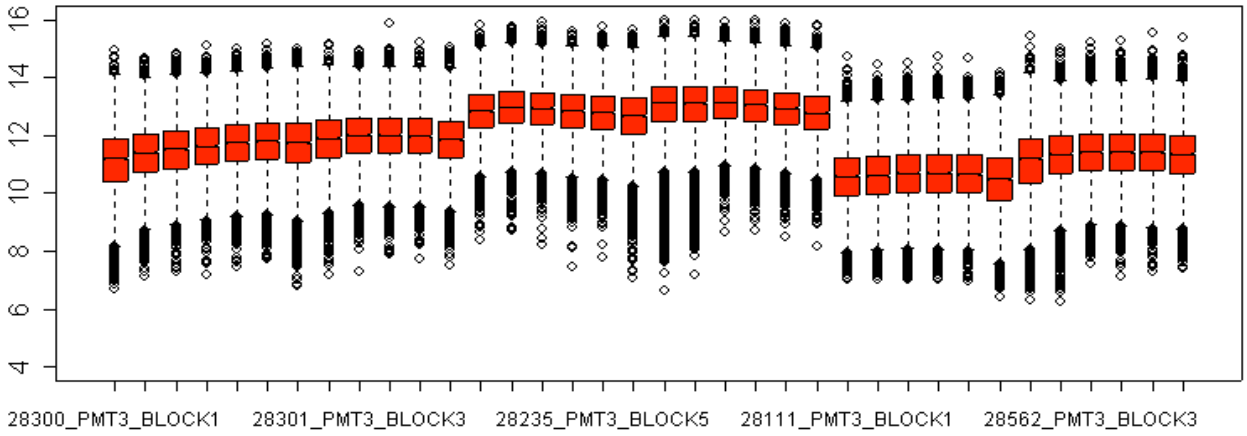
- Do standard probe selection
- 4-5X oversampling so we have enough probes to evaluate
- E. coli normalization probes = 15% of total
- Random probes = 5% of total

Normalization

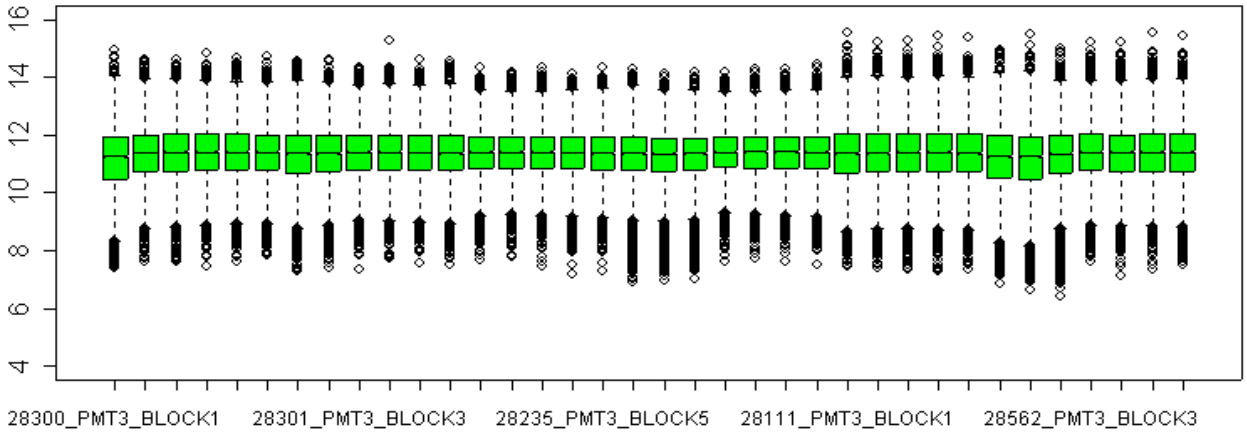
- Data collected from 0.25X, 1X, and 4X concentration series
- Arrays scanned at 3 PMTs – 50V apart
- Signal intensities rearranged by replicate, one column per replicate per array
  - 2700 genes; 21 probes per gene
  - 6 replicate blocks; 2 arrays per concentration
- vsn normalization used to correct array data.
  - E. coli and random probes used to generate parameters which are then assigned to the organism probes

NimbleGen® SYSTEMS, INC.

Ecoli probes before variance stabilizing normalization (PMT3)

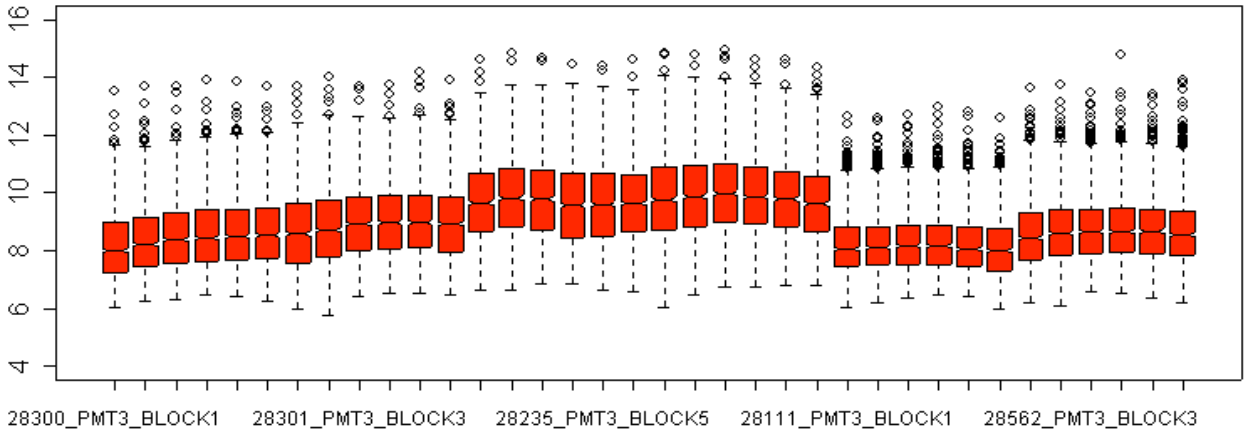


Ecoli probes after variance stabilizing normalization (PMT3)

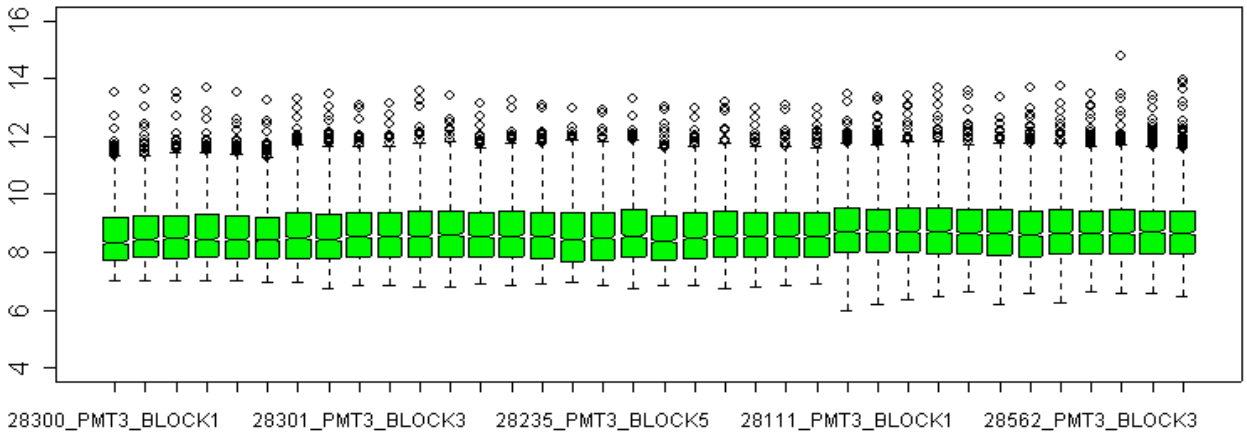


NimbleGen® SYSTEMS, INC.

Random probes before variance stabilizing normalization (PMT3)

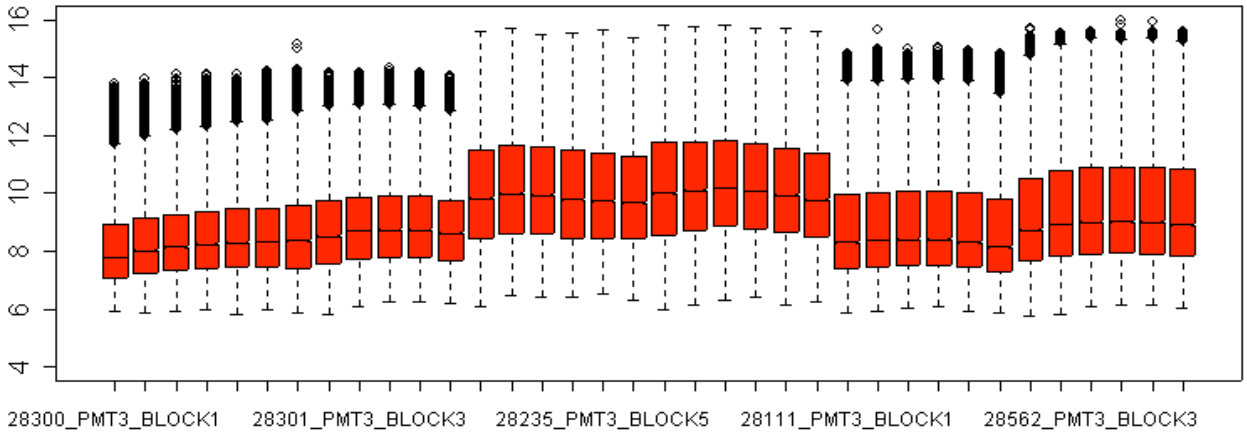


Random probes after variance stabilizing normalization (PMT3)

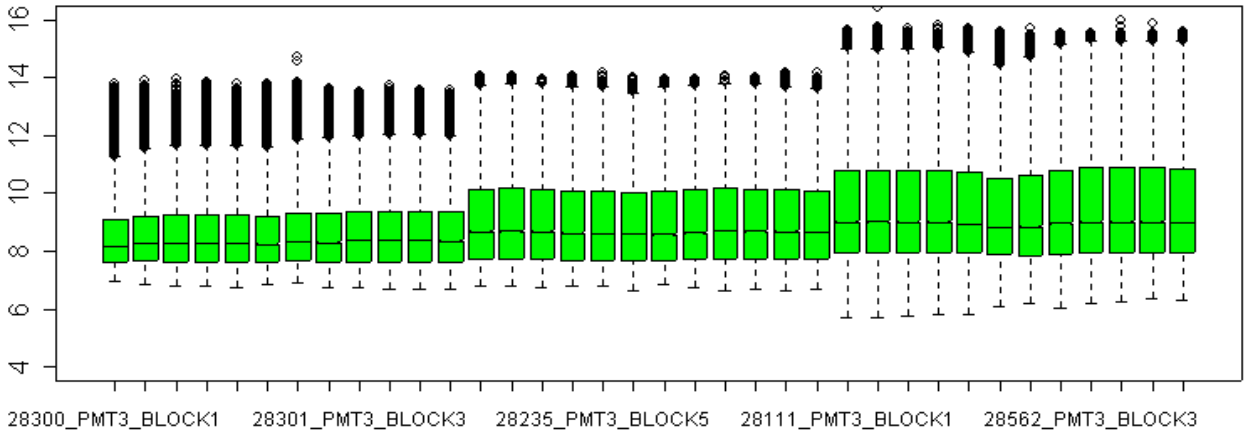


NimbleGen® SYSTEMS, INC.

Organism probes before variance stabilizing normalization (PMT3)



Organism probes after variance stabilizing normalization (PMT3)



Optimized Probe Selection

- Want probes that best follow the dilution series, and have maximal brightness and consistency
- Weighted linear regression fitted to the dilution series for each probe
  - log2 transformation and weighted linear regression were used to minimize the effect of outliers on the data.
  - weights calculated by fitting the line, calculating the residuals, and then using the weights from a tukey biweight mean calculation on the residuals to fit a weighted regression.
- Tukey biweight mean of the signal intensities of the probe at the 4X dilution represents the overall signal intensity of the probe.

Optimized Probe Selection

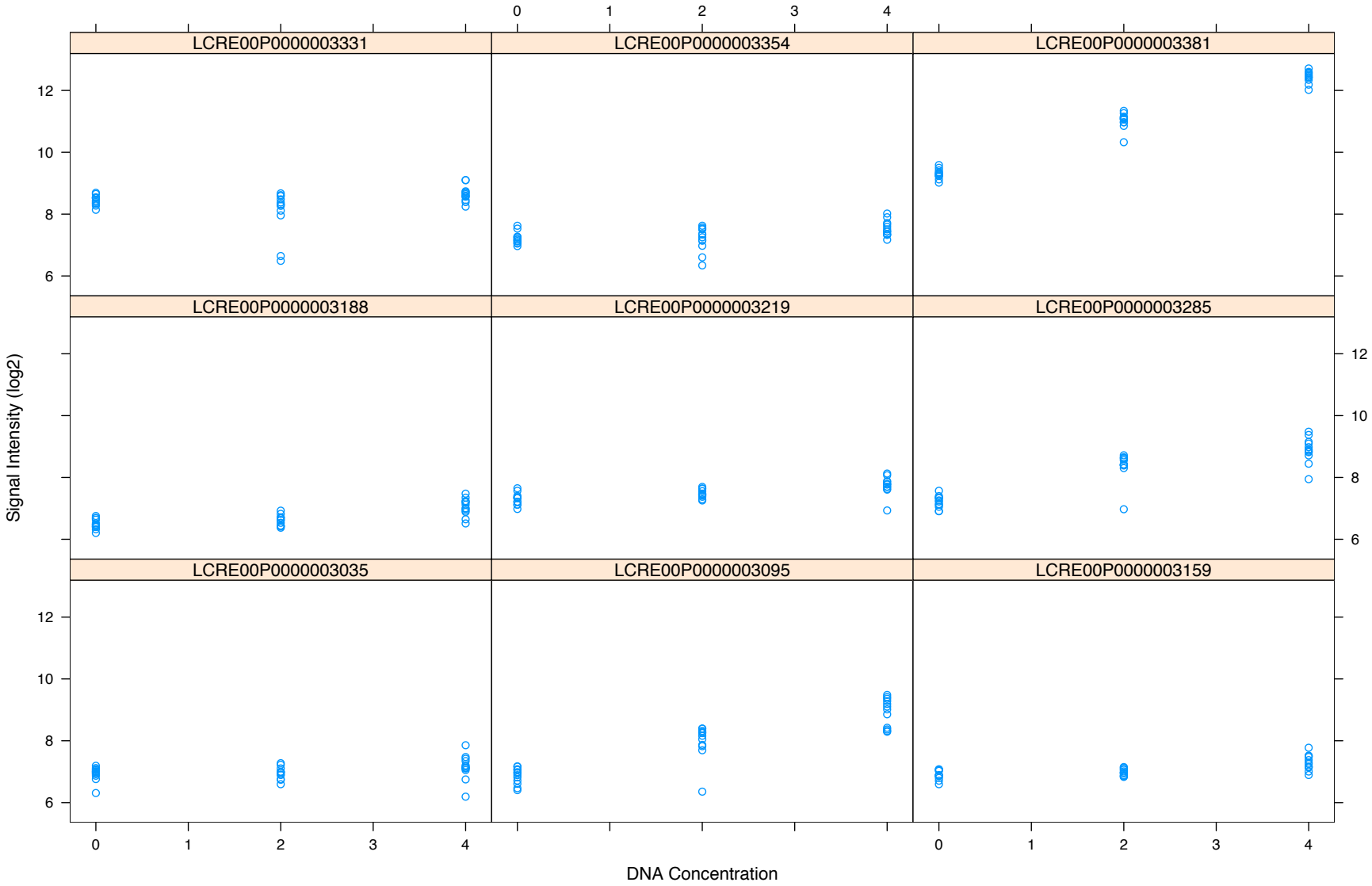
- Have four main parameters for second round of rank selection:
  - Slope of the regression line (range: 0 to 1)
    - Never reaches theoretical maximum of 2, and sometimes is negative
  - r^2 of the regression line (range: 0 to 1)
  - $\log_2$ signal intensity at 4X concentration (range: 0 to 16)
  - $\log_2$ position from 3' end (range: 0 to 13)
- Goal is to have the first three make the major contributions, contributing ~1/3 to the score.
  - Position is secondary, primary selection provides spacing

Optimized Probe Selection: Score Contributions

- slope * 100 = 0-100 contribution
- r^2 * 100 = 0-100 contribution
- $\log_2$ signal * 6 = 0-96 contribution
- $\log_2$ position * -3 = -39-0 contribution
  - after first probe, penalty becomes a bonus as distance is measured from nearest selected probe

NimbleGen® SYSTEMS, INC.

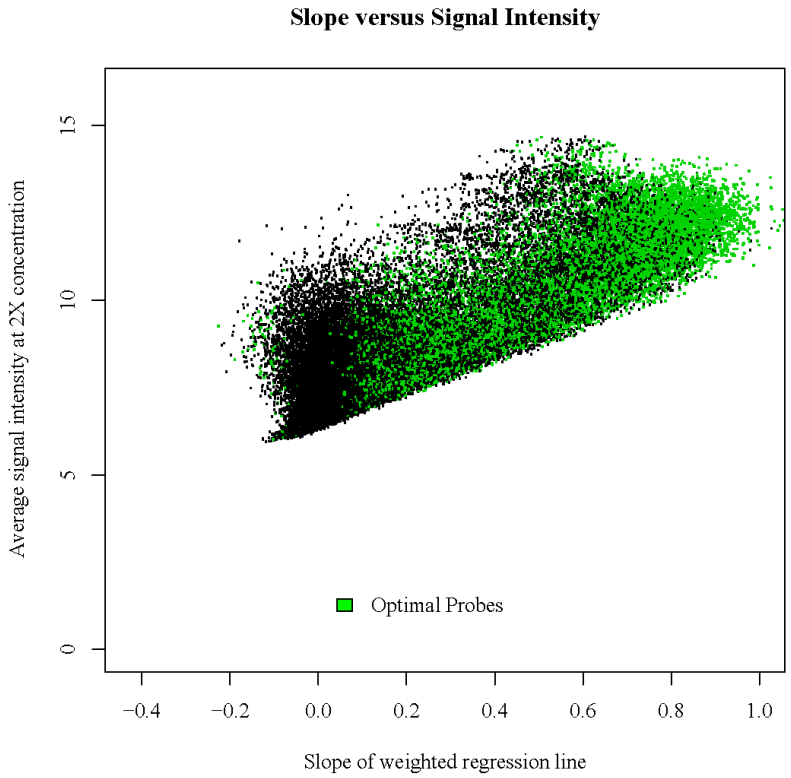
Example Data



High – Definition Genomics SM

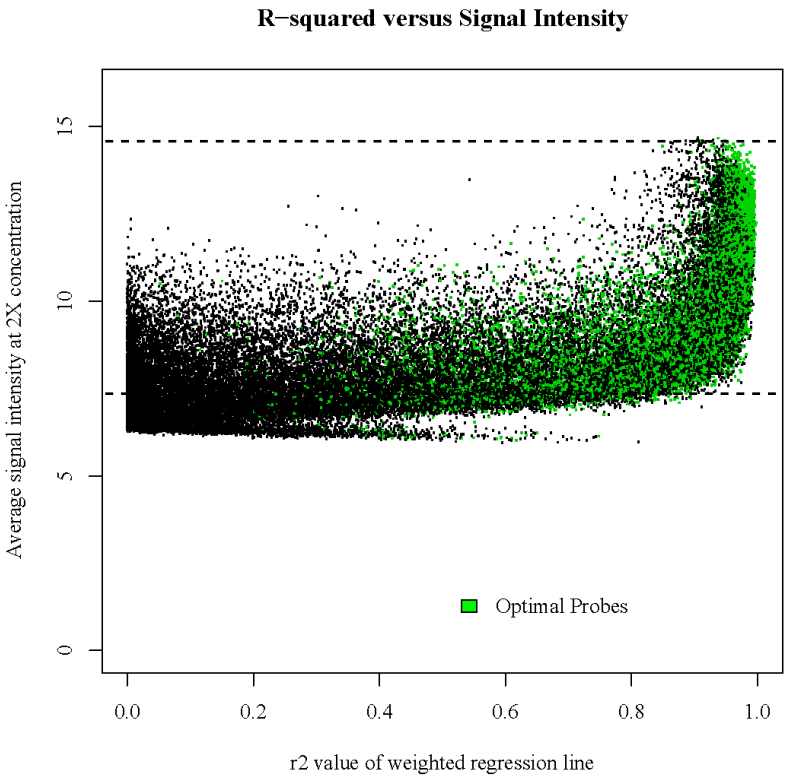
Slope versus Signal Intensity

- The probes that are best at following the concentration series span a 16-fold range in intensity



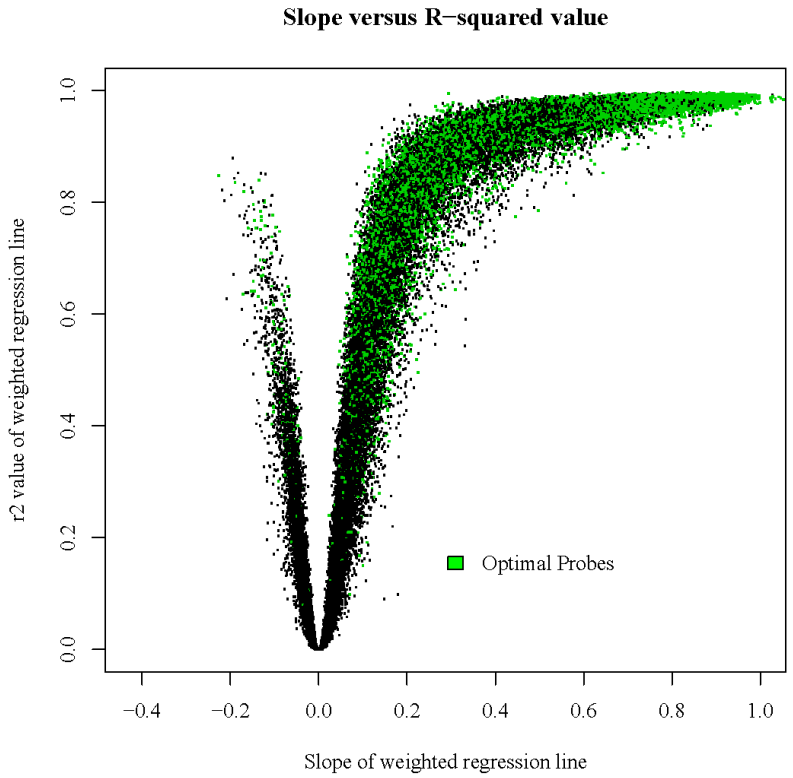
r^2 versus Signal Intensity

- The most reproducible probes span a 250-fold range of intensities



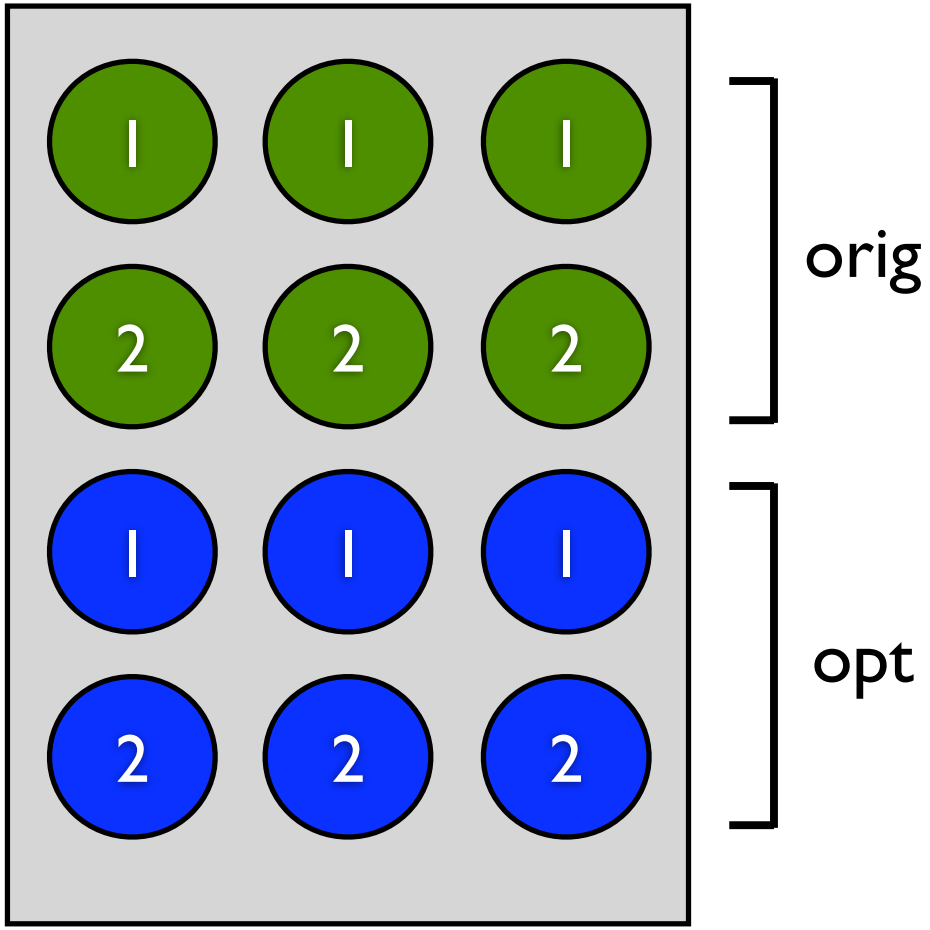
Slope versus r^2

- Probes that best follow the concentration series are also some of the most reproducible



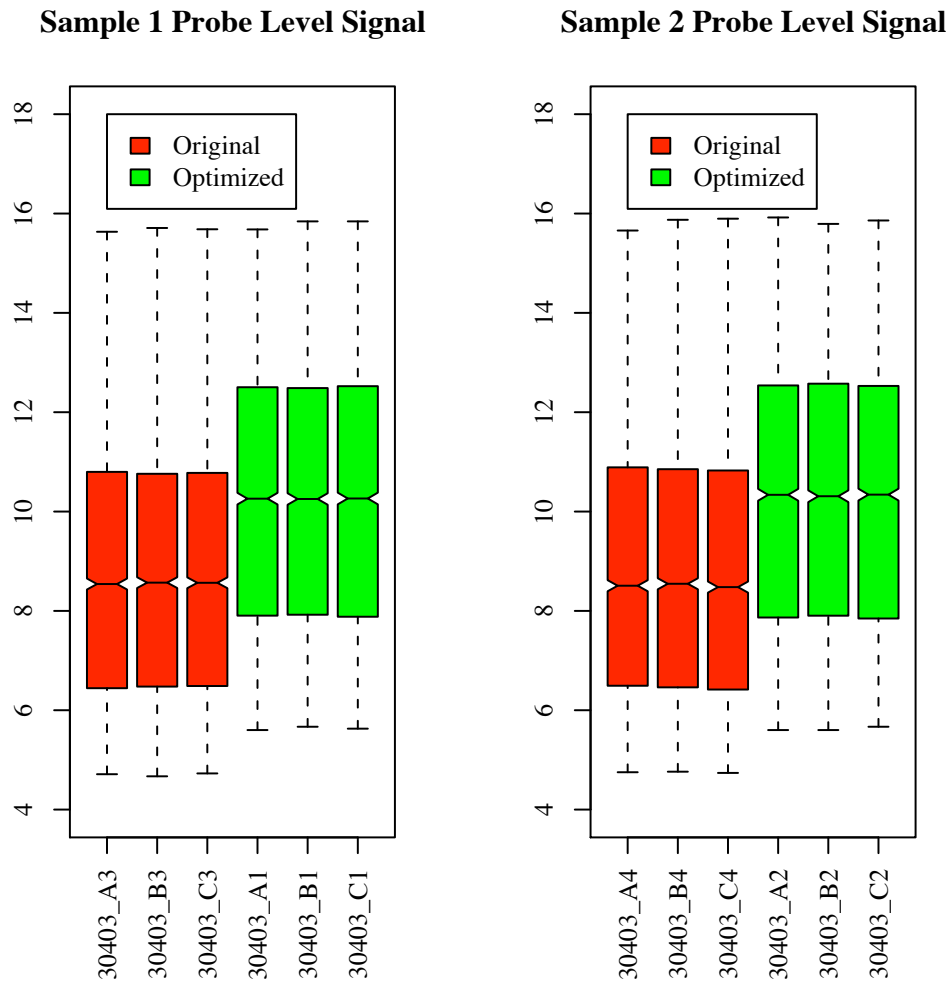
Side-by-side Comparison

- Optimized probe set vs original probe set
- 2700 genes; 5 probes per gene
- Two RNA samples
- 12plex design



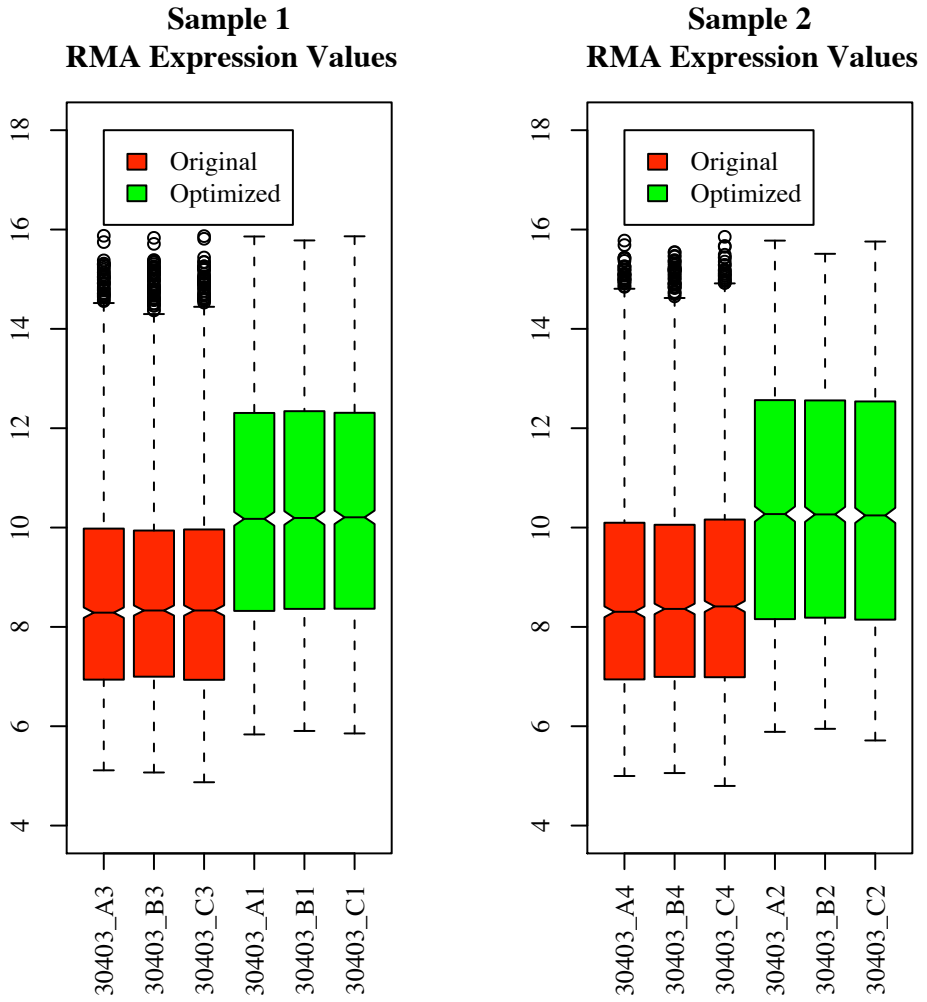
Probe Level Comparison

- Data separated by well
- Normalized by sample/probe set combination
- Median signal intensity of optimized probes ~4X greater than non-optimized

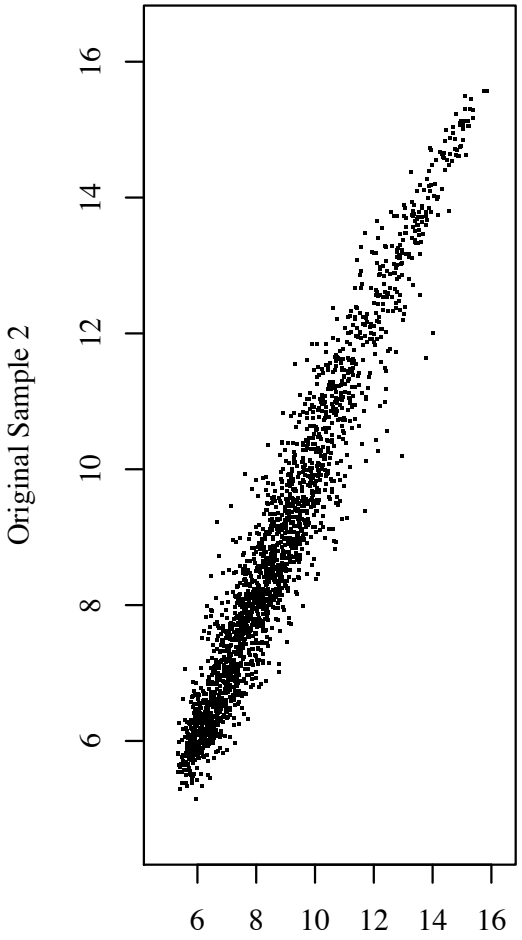


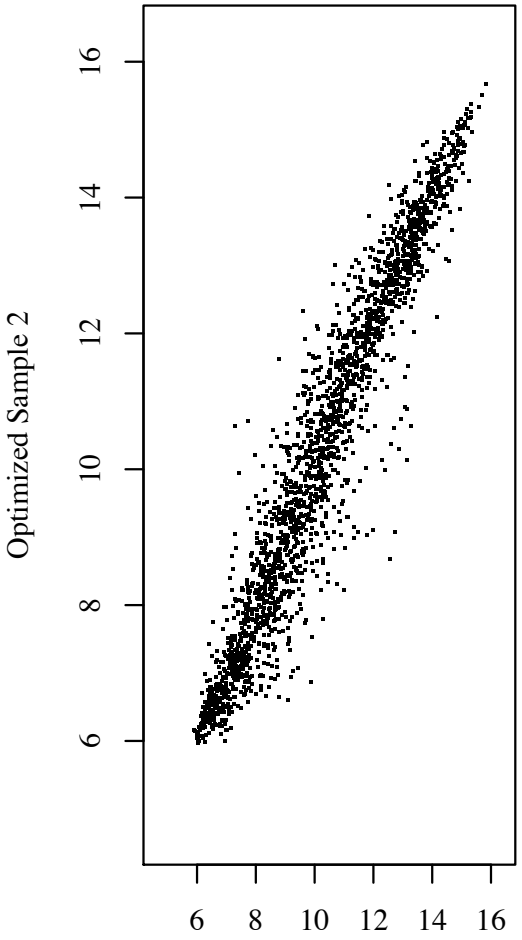
Expression Summary

- Expression summary using RMA
- Median expression level also 4X



Differential Expression





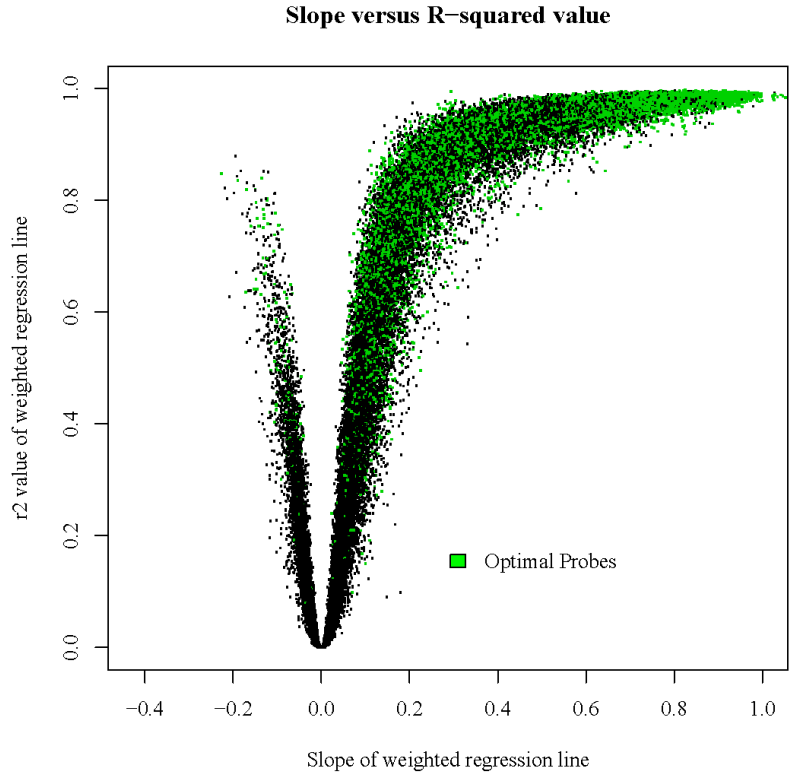
Orig	Rep1	Rep2	Rep3
Rep1	-	0.975	0.974
Rep2		-	0.980
Rep3			-

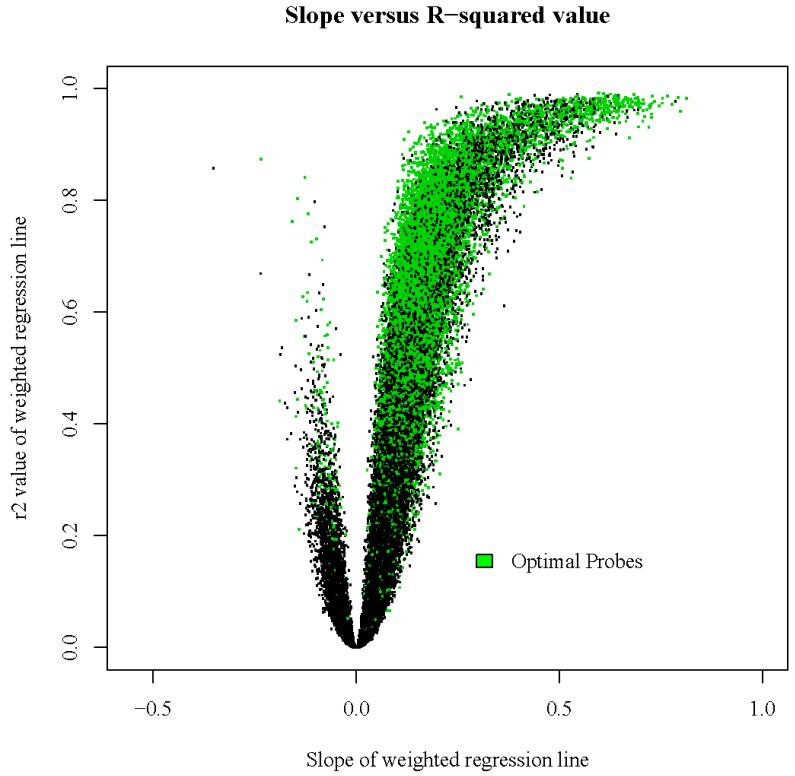
Opt	Rep1	Rep2	Rep3
Rep1	-	0.989	0.987
Rep2		-	0.989
Rep3			-

	>2-fold	>4-fold	>8-fold
Original	194	8	0
Optimized	267	39	4

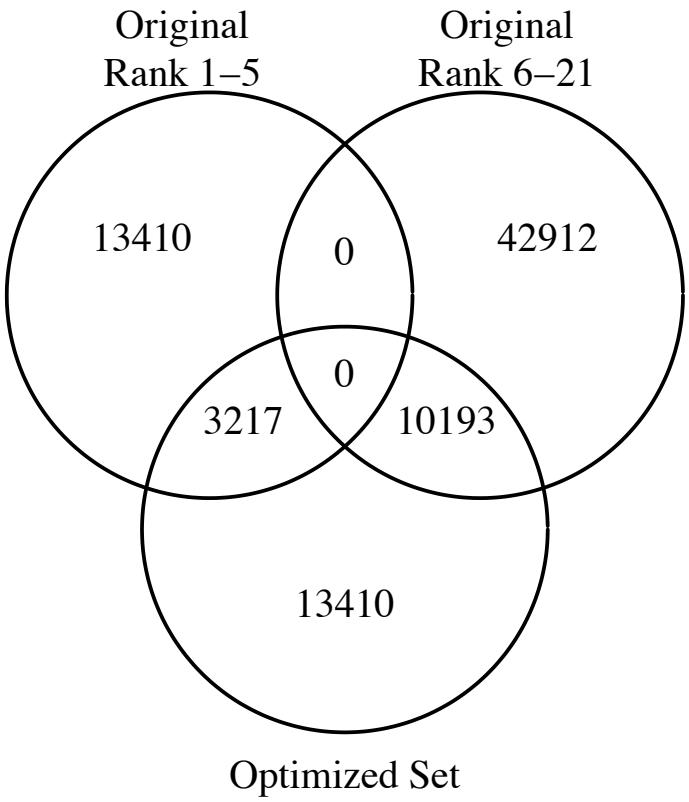
High Definition Genomics℠

Suboptimal Example





in silico vs empirical



Conclusions

- Empirical performance preferred over *in silico* prediction
- To find probes that measure concentration changes, must submit them to changing concentrations
- Useful for bacterial expression and genomic tiling applications (CGH, ChIP, methylation, exon arrays)
  - greater utility when probe choice is not constrained
- Less useful for eukaryotic expression
  - probe sequences that span splice junctions are not present in genomic sample

Acknowledgments

- Bioinformatics
  - Steve Smith
  - Michael Hogan
  - Kyle Munn
  - Melodee Patterson
  - Nan Jiang
  - Michael Molla
  - Jacob Kitzman
- R&D
  - Roland Green
  - Rebecca Selzer
  - Tom Albert
  - Jason Norton
- Business Development
  - Emile Nuwaysir