Dissecting the details of probe response

Cynthia Gibas

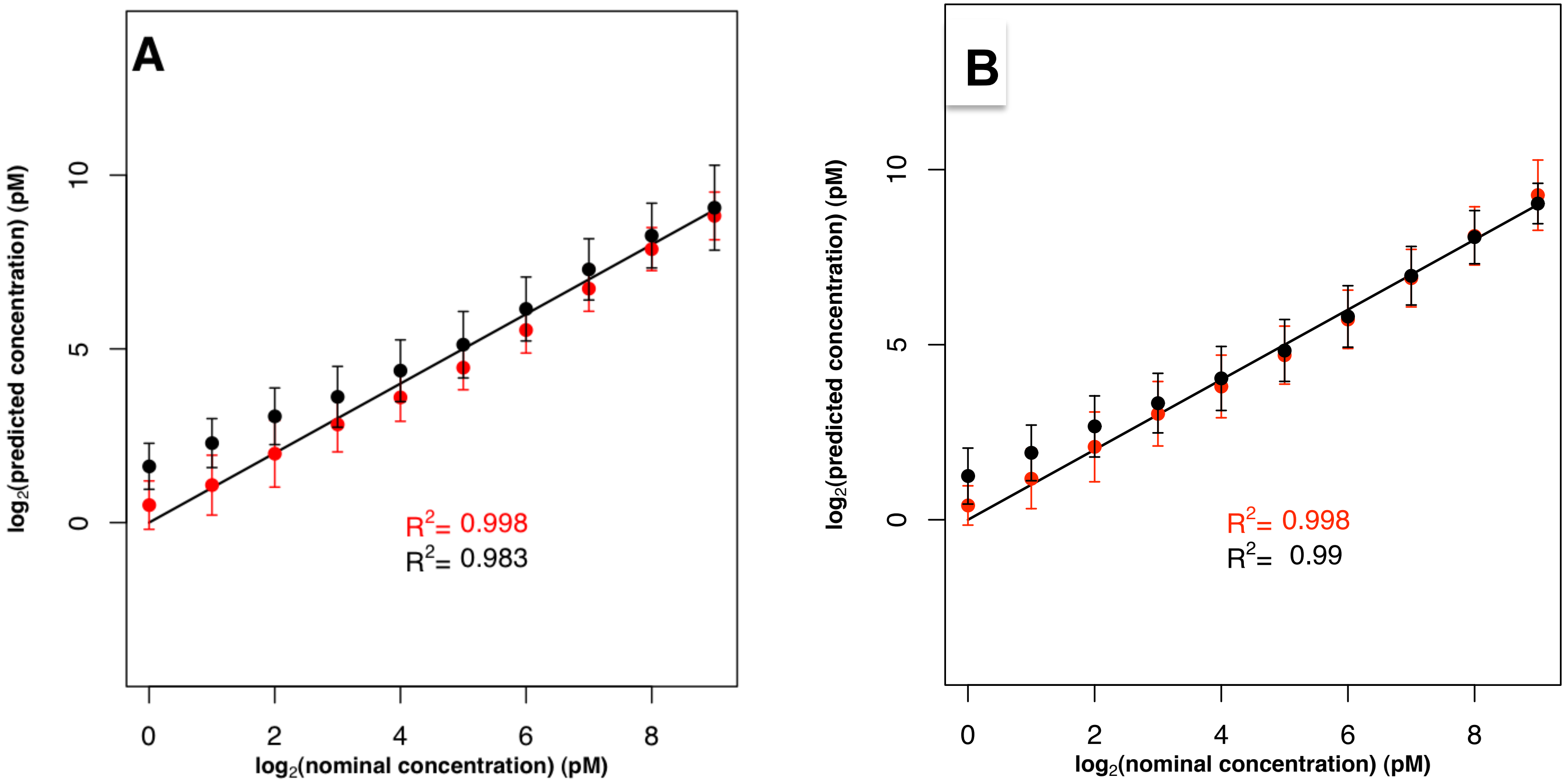
UNC Charlotte

GLAM

- In a model array system, a generalized Langmuir model can be used to estimate concentration from intensity
- Can train on a subset of the data to generalize Langmuir parameters from the whole array – opens up possibilities for internal controls
- Performs well in the linear range of the experiment
- Performance degrades only somewhat on real data (e.g. MAQC data sets)

Gharaibeh et al, 2010a

GLAM

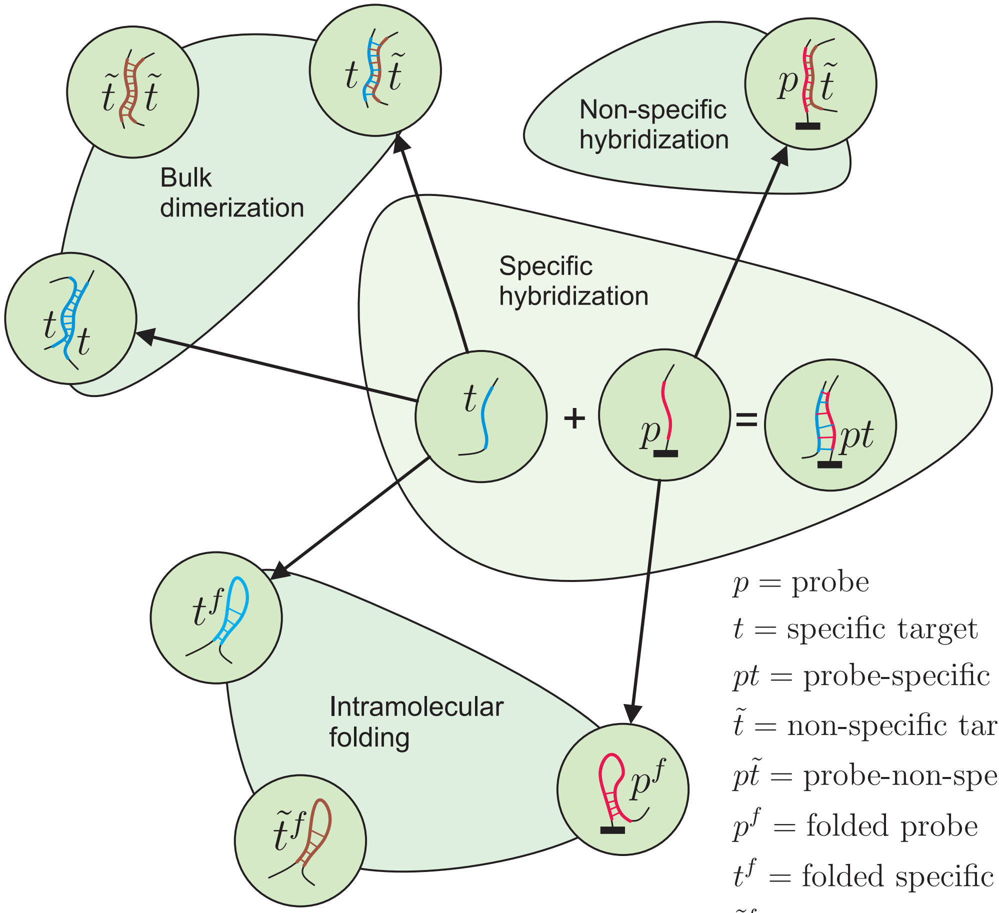


Compares favorably to hybridization-based and probe property dependent approaches.

Hypotheses (2005)

- We can use modified solution hybridization models to accurately predict surface behavior
- If we can model the situations that lead to low-performing probes or problem interactions on the microarray, we can “fix” them in the design or interpretation phase.

Hybridization modeling



p = probe

t = specific target

pt = probe-specific target duplex

$\tilde{t}$ = non-specific target

$p\tilde{t}$ = probe-non-specific target duplex

p^f = folded probe

t^f = folded specific target

$\tilde{t}^f$ = folded non-specific target

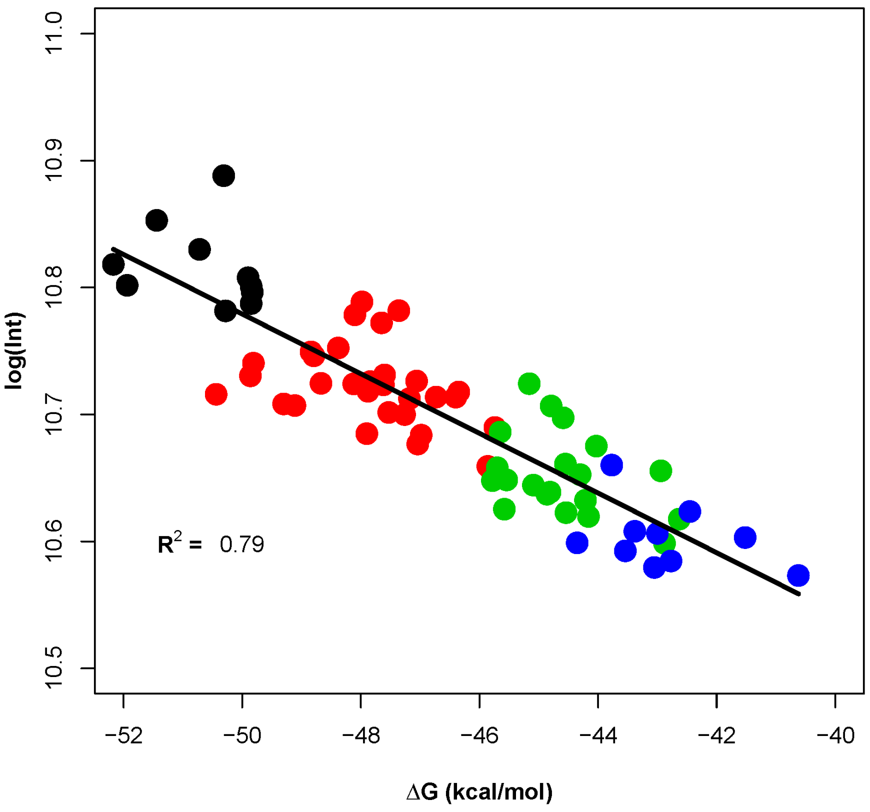
Four components

- Polymorphism
- Cross-hybridization
- Kinetic interference (target)
- Kinetic interference (probe)

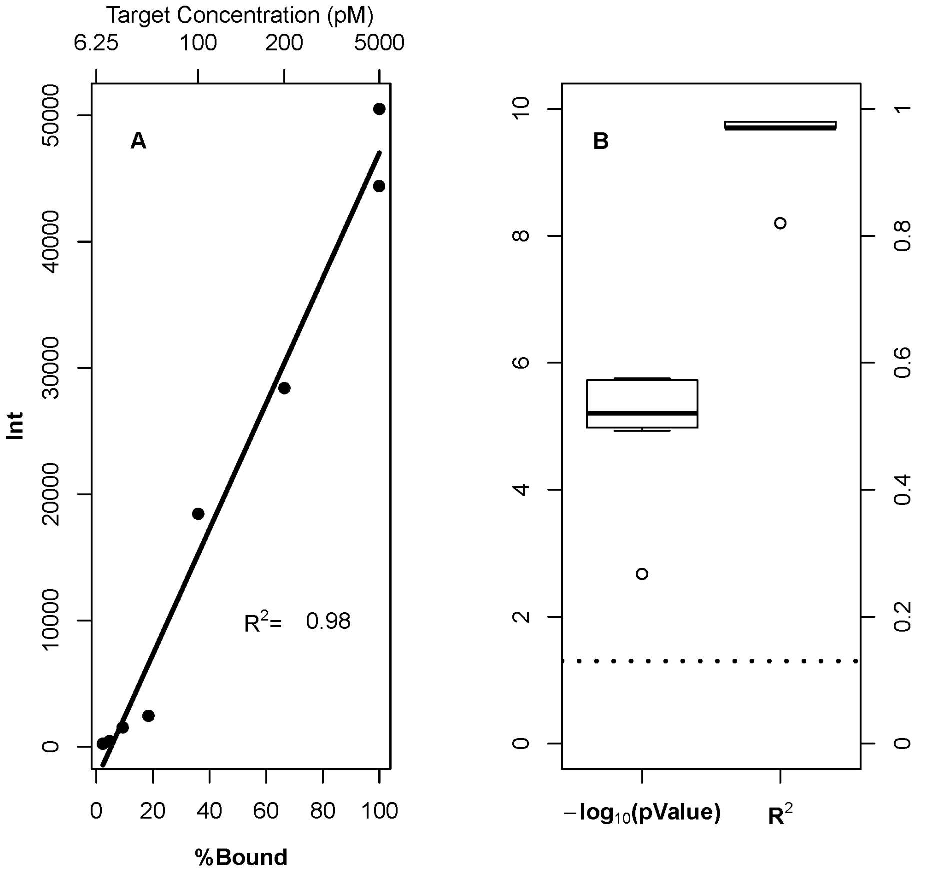
Effect of small-number mismatches

- We modeled 1- 2- and 3-mer central mismatches in 50mers
- Solution n-state model is predictive; mismatches are detectable

Intensity vs. hybridization free energy



Intensity vs. predicted fraction bound

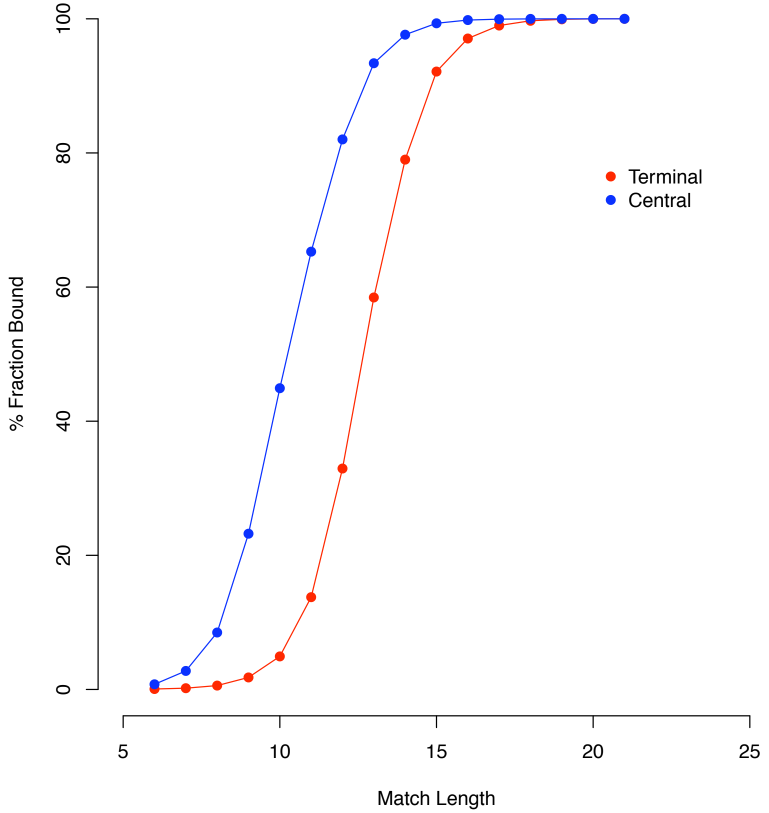


Gharaibeh et al, 2010b

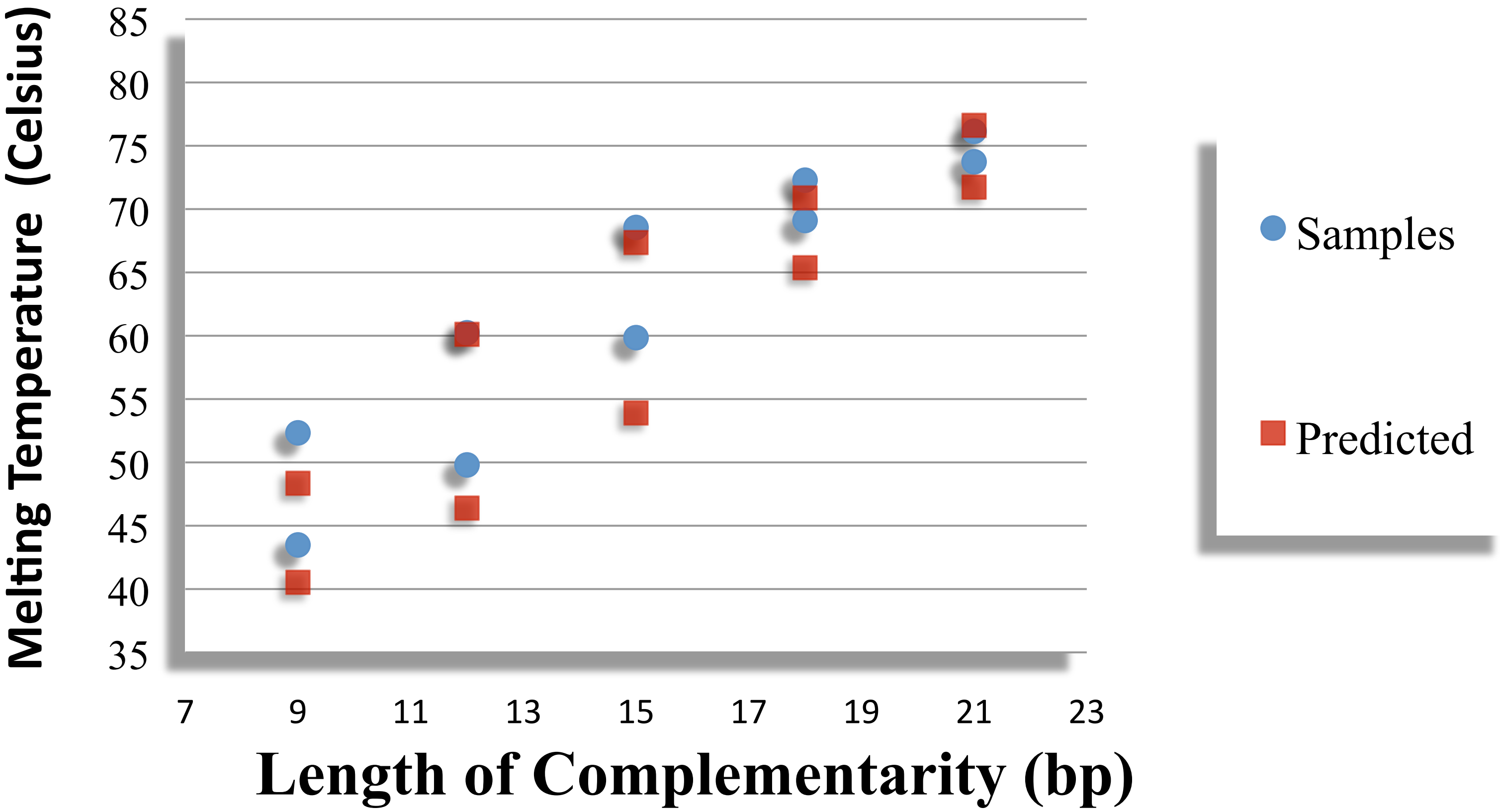
Where does cross-hybridization start?

- Classically referred to as the “minimum nucleation length”
- We simulated hybridization of a few thousand 50mer probe-target pairs, permuting the sequence of one binding partner so that only a central, or terminal stretch of complementarity remained
- In simulations at 60C, we began to see >50% of the pair form duplexes even with a central match of 9nt. At 12 nt, 100% were in duplex form

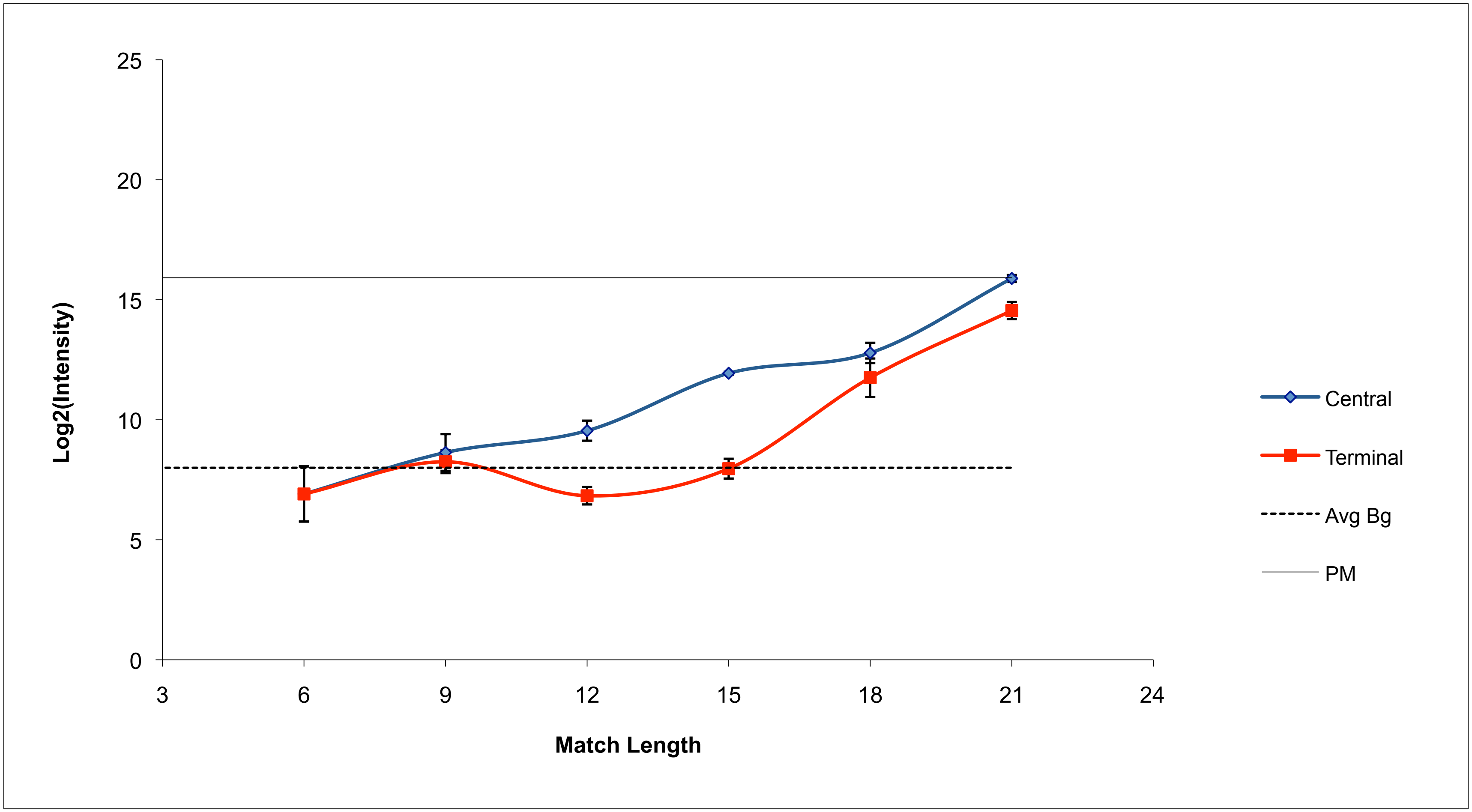
Minimum nucleation, N-state model



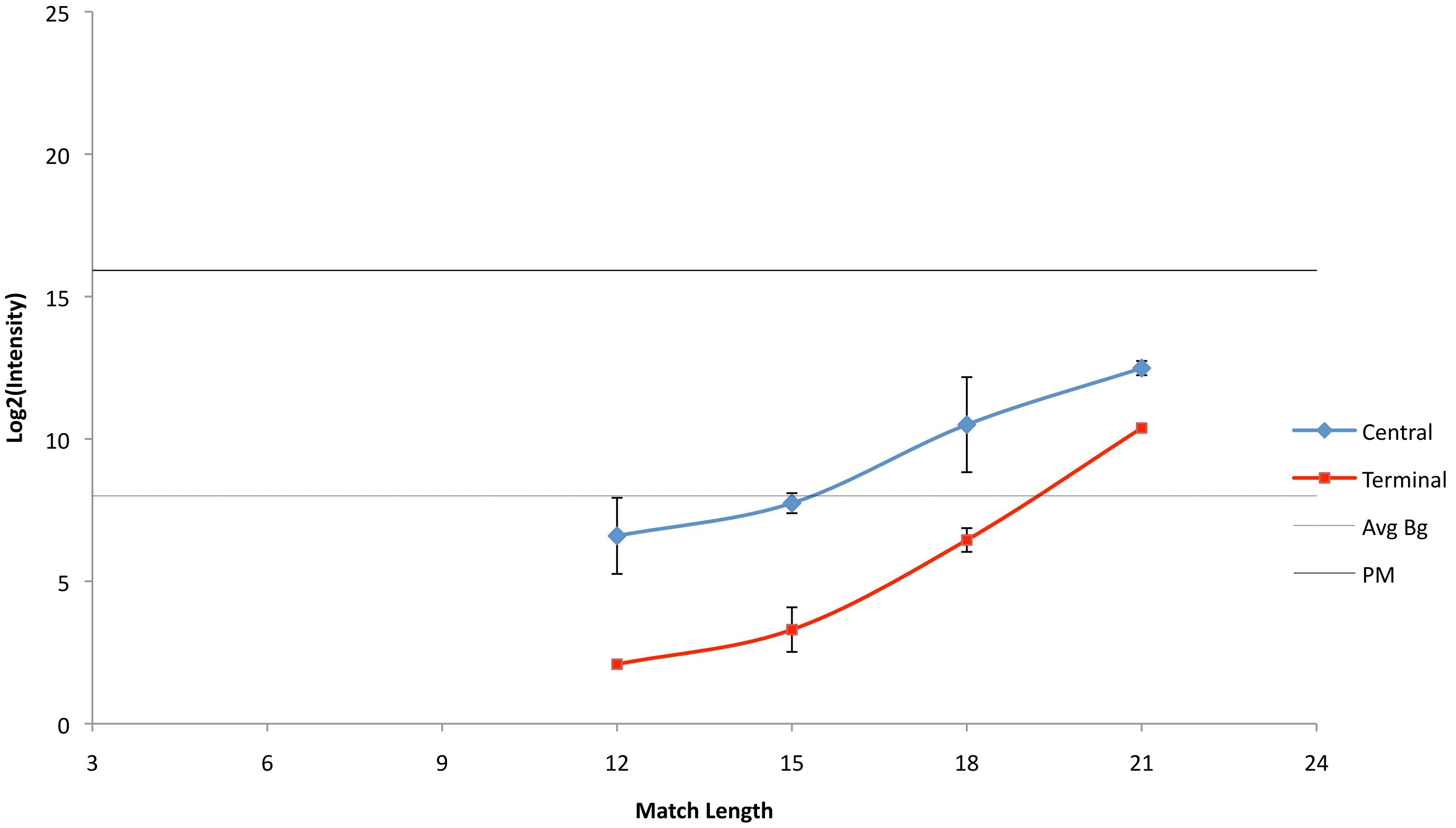
Minimum nucleation in solution



On the array: labeled target, alone…

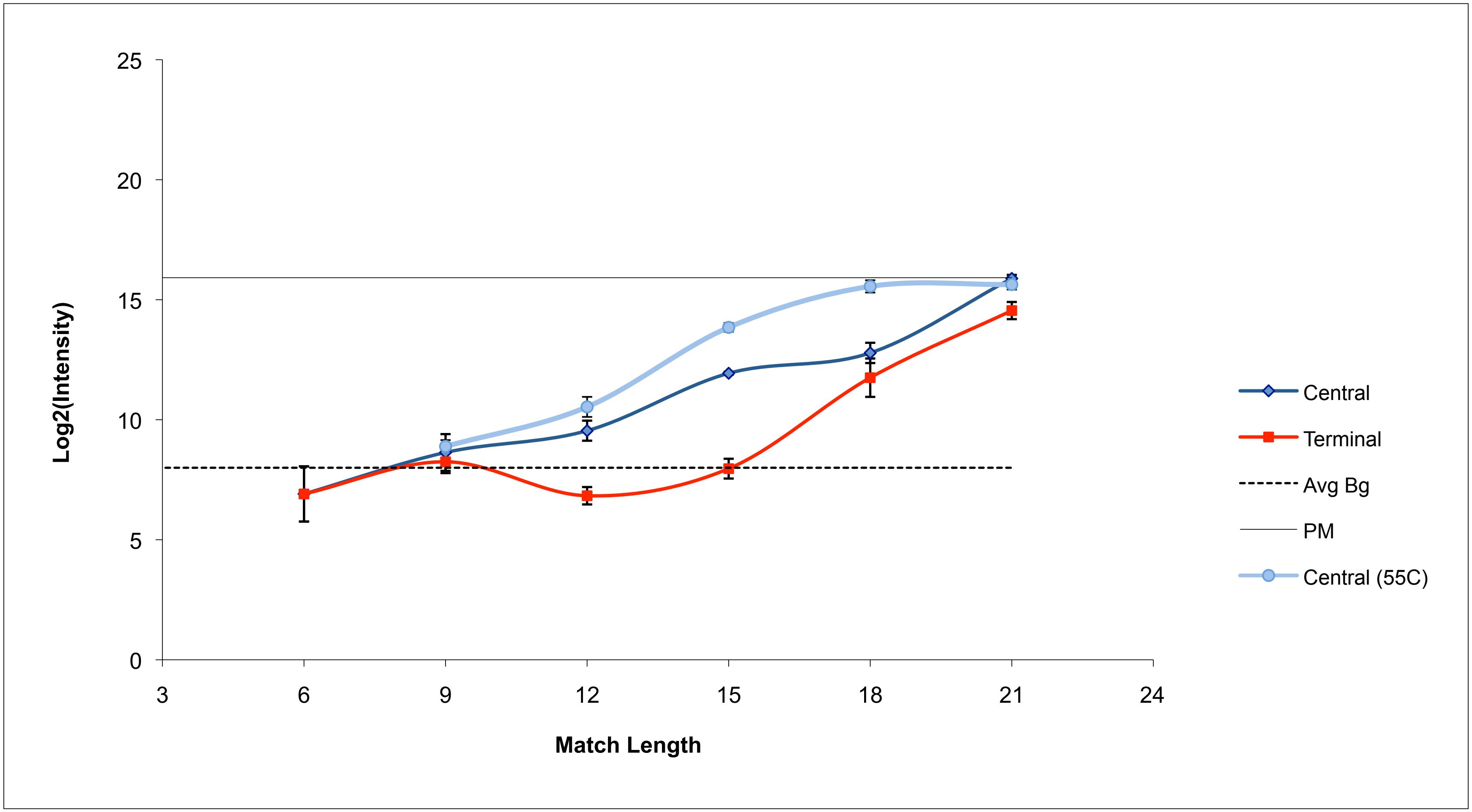


…and with perfect match competitor



Not as troubling for the technology as we thought 10 days ago –
but still points to the need for more accurate models

And at a lower temperature

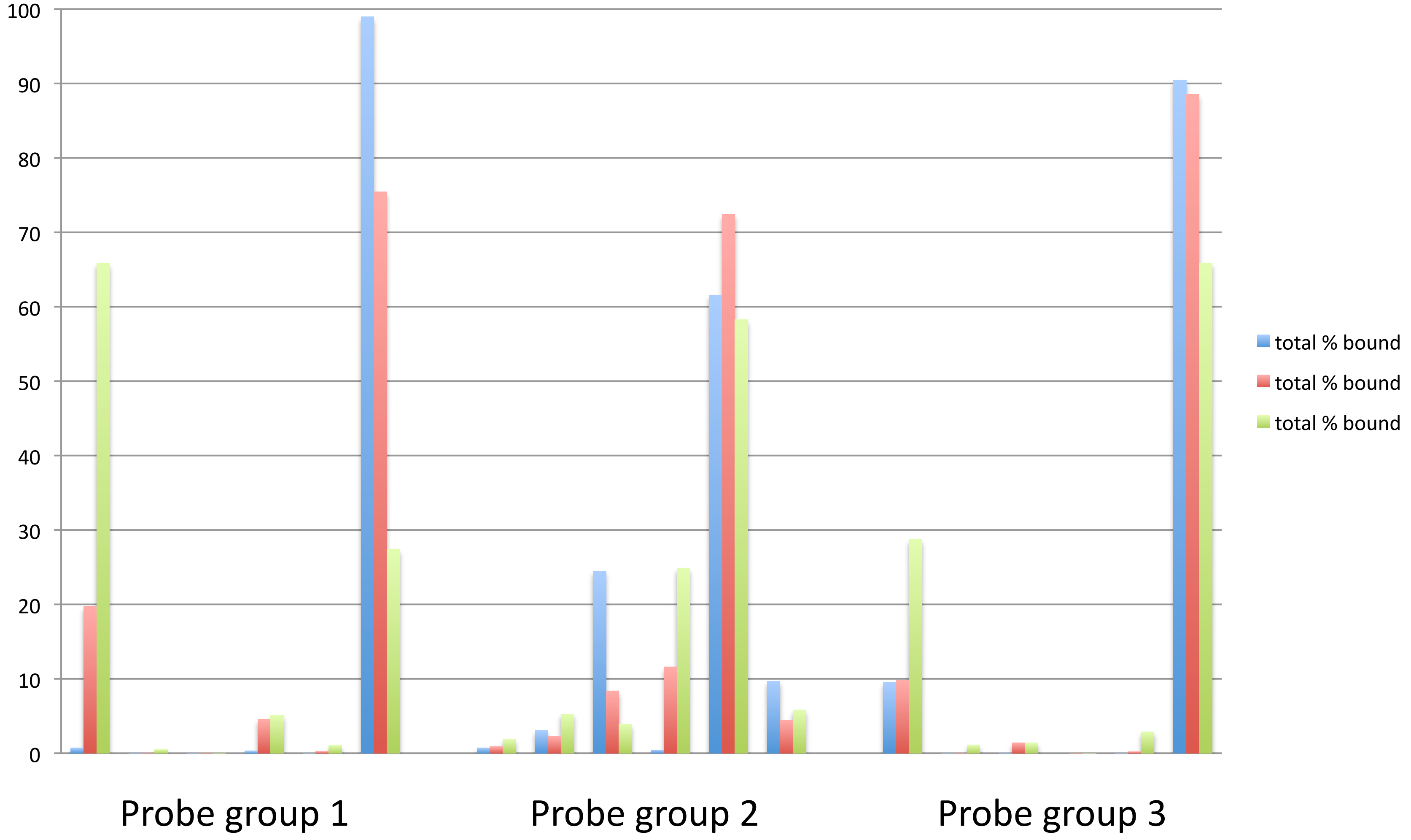


The implication is that melting temperatures calculated using modified solution-state models may be as much as 10 degrees off

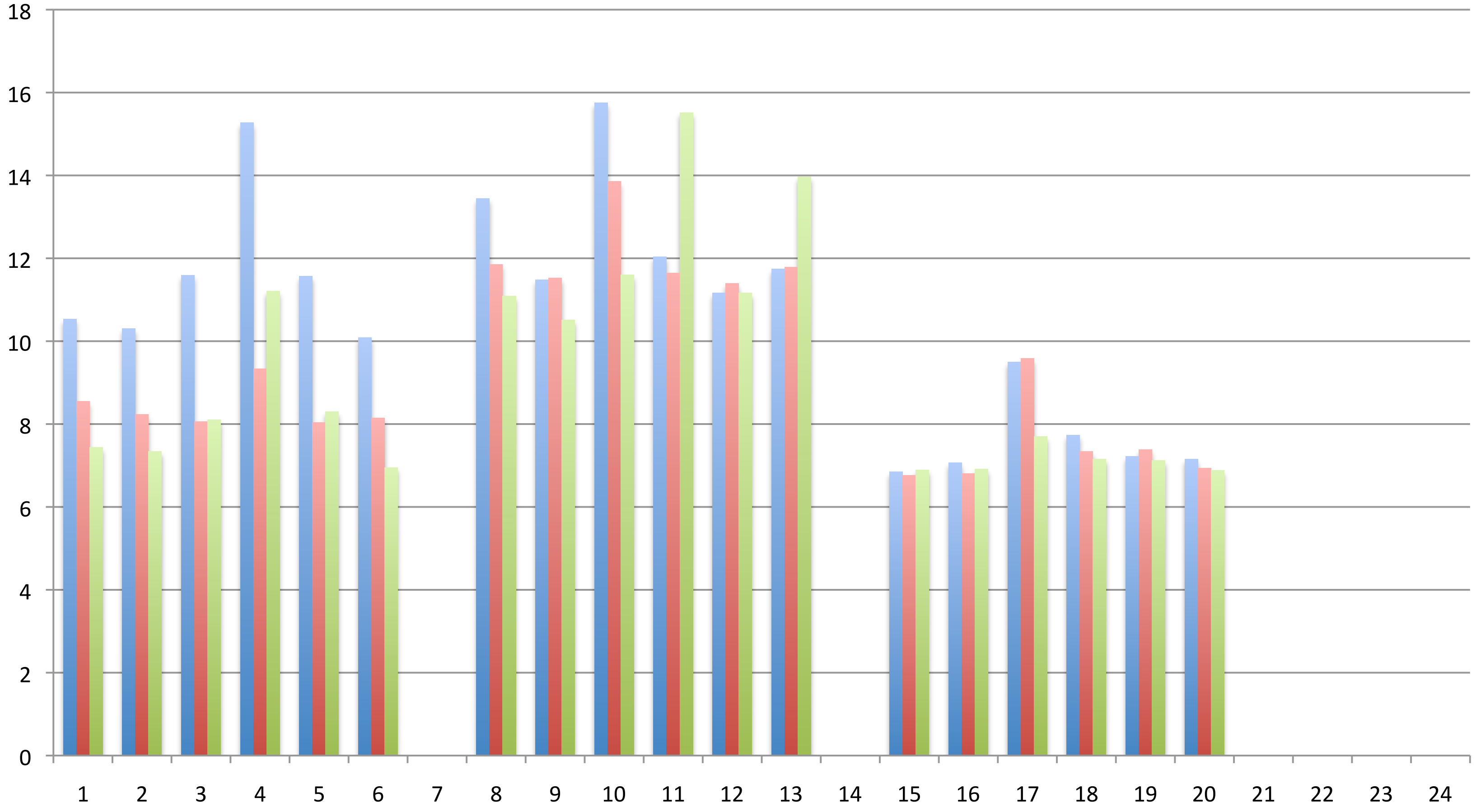
Target Secondary Structure

- Hypothesis: structure formation in the target (ss cDNA or cRNA) has the potential to compete with probe hybridization
- Predict dramatic binding differences over sequence using OMP endpoint model
  - Experiment 1: design multiple probes to modeled folded and structure-free target regions; hybridize to full-length target
  - Experiment 2: construct simulated sheared target fragments; hybridize (Weller)

Predictions:



VERY preliminary results:



Probe group 1 Probe group 2 Probe group 3

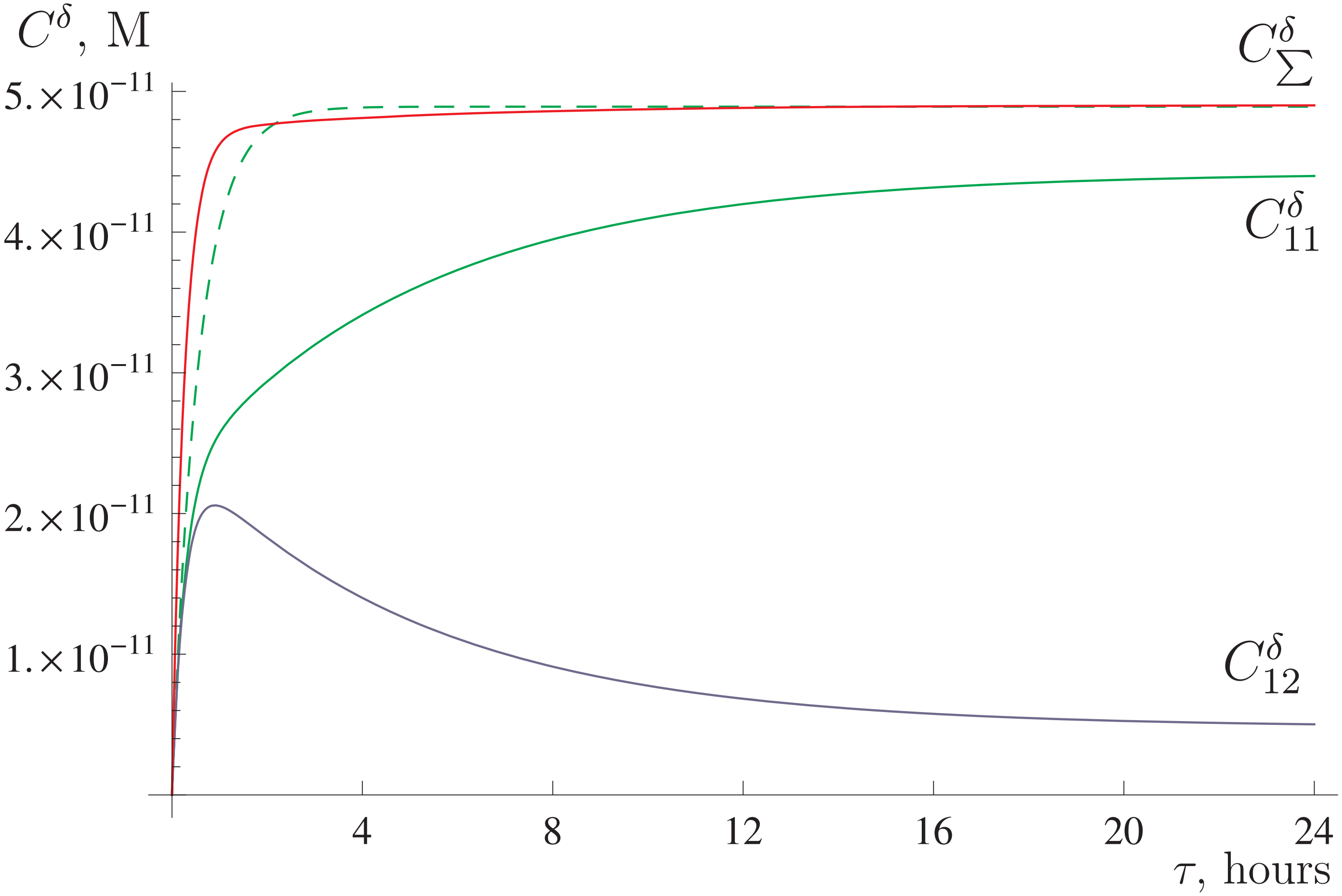
Preliminary conclusions:

- Hybridization efficiency does vary significantly across the target
- It’s temperature dependent, as expected
- But the details of the prediction are totally wrong

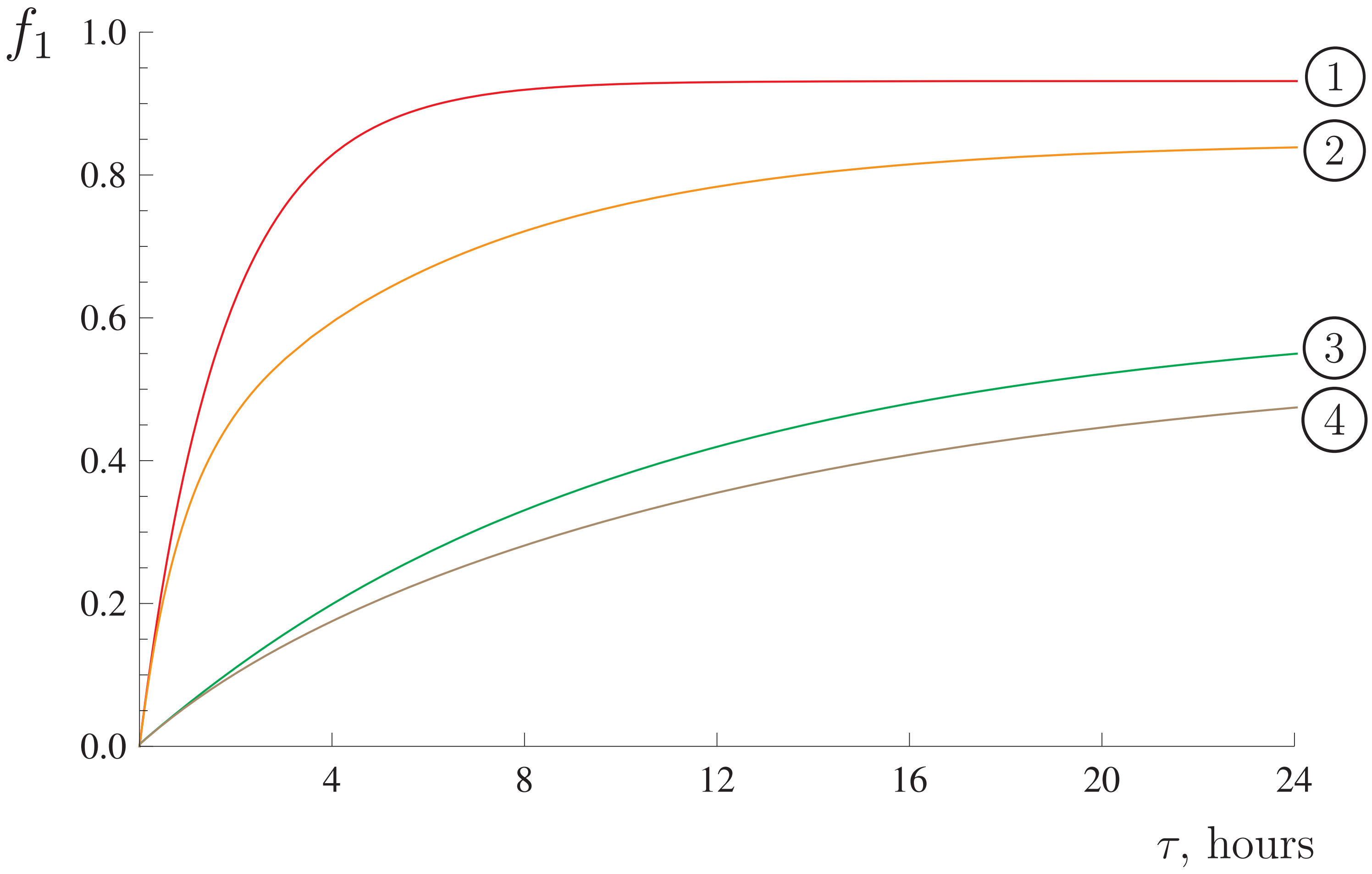
Kinetic model

- A generalized kinetic model that incorporates both hybridization and unimolecular structure formation is possible
- Physical models generally applied to microarrays are special cases of this model

Multiplex competition



Adding unimolecular competition



Findings (2011)

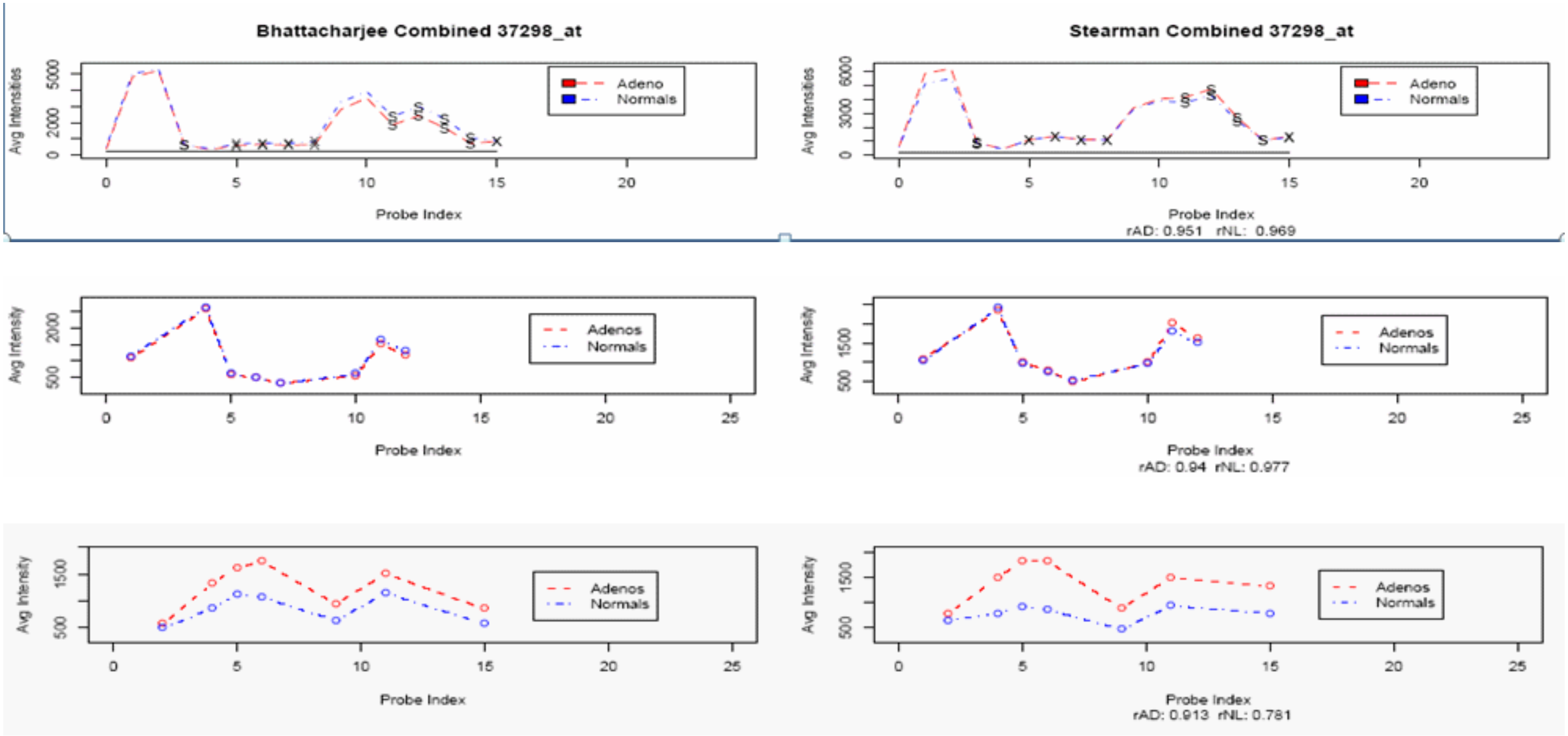
- Solution hybridization models predict idealized surface hybridization behavior rather well, but fail (so far) unpredictably for non-ideal duplexes
- BUT -- If we can model the situations that are most likely to lead to problem probes or problem interactions on the microarray, we can discard data selectively and improve results.

Biophysics-based data cleansing

- Measurement filtering:
  - probes within linear range(200-20,000 f.u.) for Affymetrix scanners and labeling
- Sequence and structure filtering:
  - Matching sequence has known SNPs
  - Probes cross-hybridize in the genome
  - Stable probe or target internal structure lowers availability for duplex
  - Runs of Gs in the probe (4 or more - tetraloops?)
  - T7 primer motifs

Thompson et al 2009 (BaFL pipeline)

Biophysics-based probeset cleansing



Acknowledgements

- Dr. Raad Gharaibeh
- Dr. Vlada Ratushna
- Vladimir Gantovnik
- Jaishree Garhyan
- Austin Craven
- Stephen McGee
- Josh Newton
- …and a fleet of interns

Collaborators

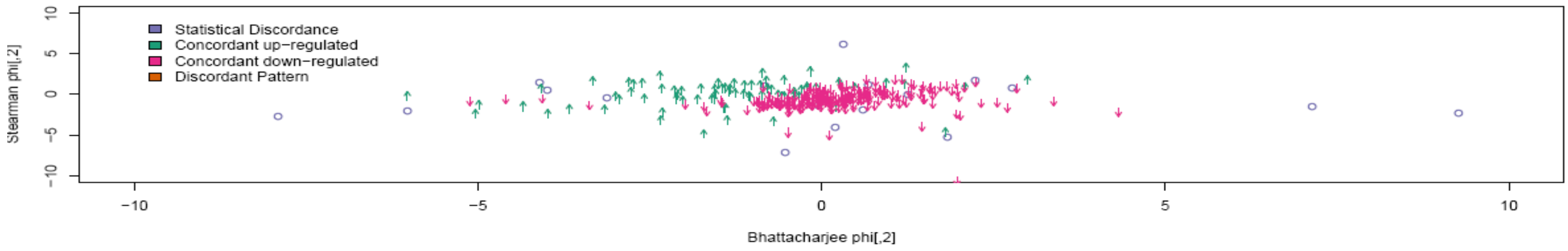
- Jennifer Weller
- Anthony Fodor

Funding

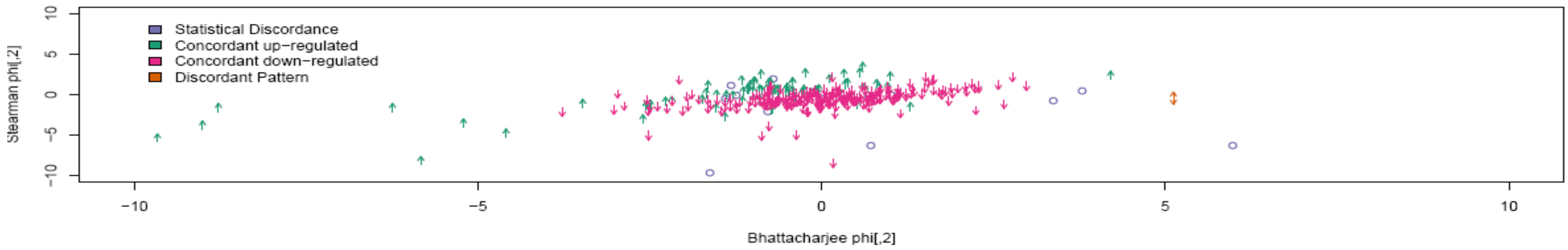
- NIH R01-GM072619
- NC Biotechnology Center

This may not be as useful for the talk, it really has to do with what happens when you leave in the bad probes and let statistics ‘take care of it’. We compare the same 325 ProbeSets. For RMA and dCHIP we have to let the algorithm use all of the probes, and we used the values delivered by the algorithms, while with BaFL we just used the probes we approved and did a simple mean of the constituent intensities. Then we used a non-linear methods to separate by sample class – it is far better with BaFL. This leads into the need for ArrayInitiative – just published – RMA and dCHIP require the array description file (cdf) and there was no nice graphical tool for constructing them. Now there is.

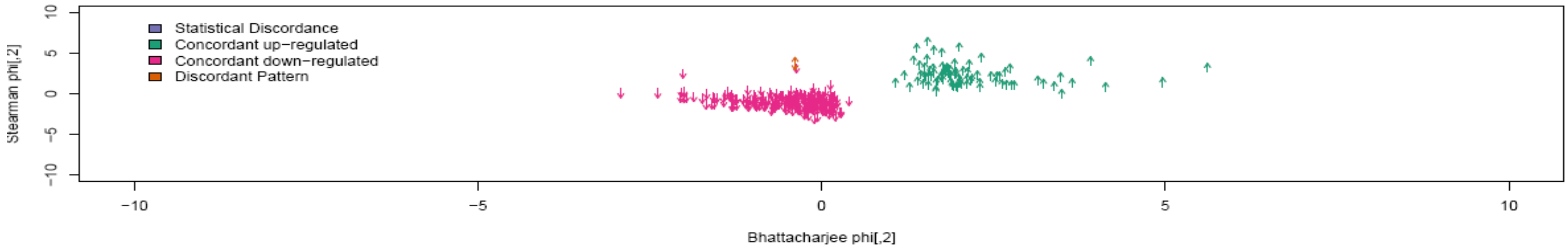
Non Linear Reduction of 325 RMA ProbeSets



Non Linear Reduction of 325 dCHIP ProbeSets



Non Linear Reduction of 325 BaFL ProbeSets


🟪 Statistical Discordance
🟩 Concordant up−regulated
🟥 Concordant down−regulated
🟧 Discordant Pattern
Stearman phi[,2]
Bhattacharjee phi[,2]