Genotyping on the Affymetrix platform using Birdseed

Finny Kuruvilla, MD PhD
Broad Institute of Harvard and MIT
Massachusetts General Hospital

October 17, 2007

In 2006, Affymetrix genotyping microarrays utilized the same basic chip design.

24 or 40 probes were used to interrogate a single SNP

Half of these probes were “mismatch probes” that intentionally did not bind to the SNP in question, but were used for background correction

Of the perfect match probes (12 or 20), half would bind to the A allele and half to the B allele

Of the 6 or 10 probes that interrogate a given allele, each of them differ according to which strand they bind to and what offset the SNP base is on the probe (position 12, 13, 14, etc.)

End result: not a single SNP would be interrogated by the same probe more than once

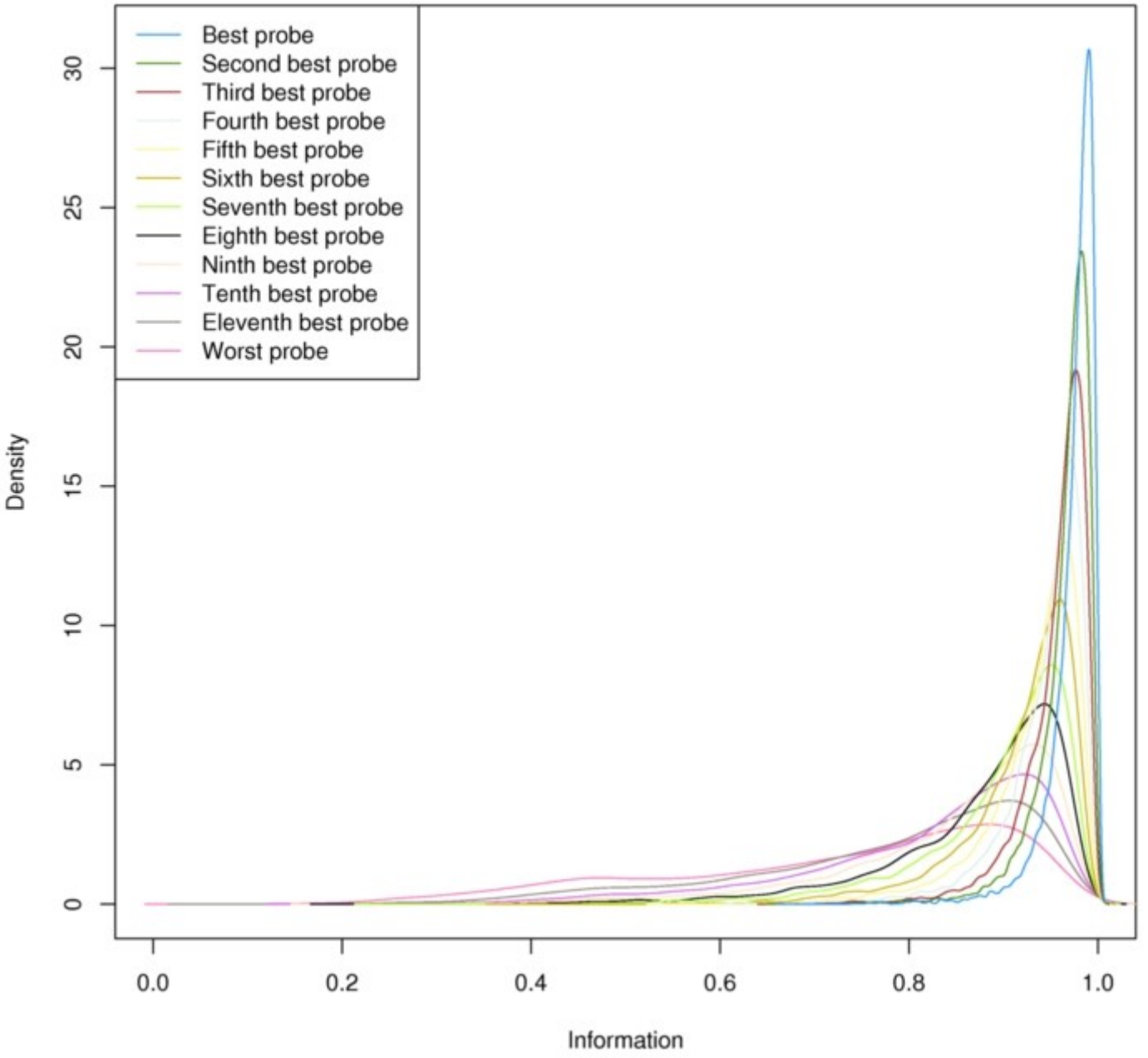
In early 2006, there were a number of flaws with several commercially available platforms

1. The 100K array had too few SNPs to have sufficient power for a GWAS

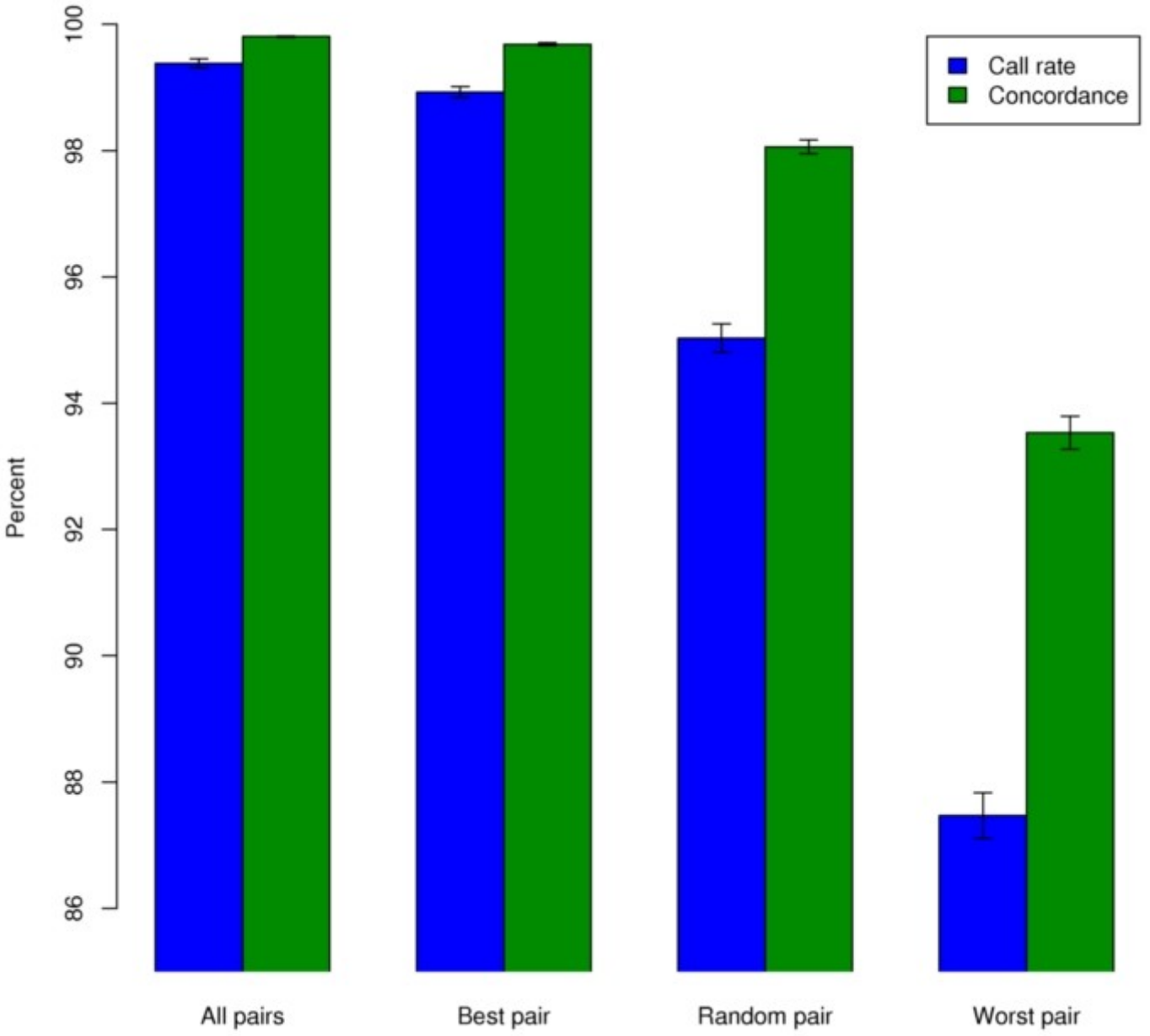
2. The 500K was not working properly for a number of reasons.

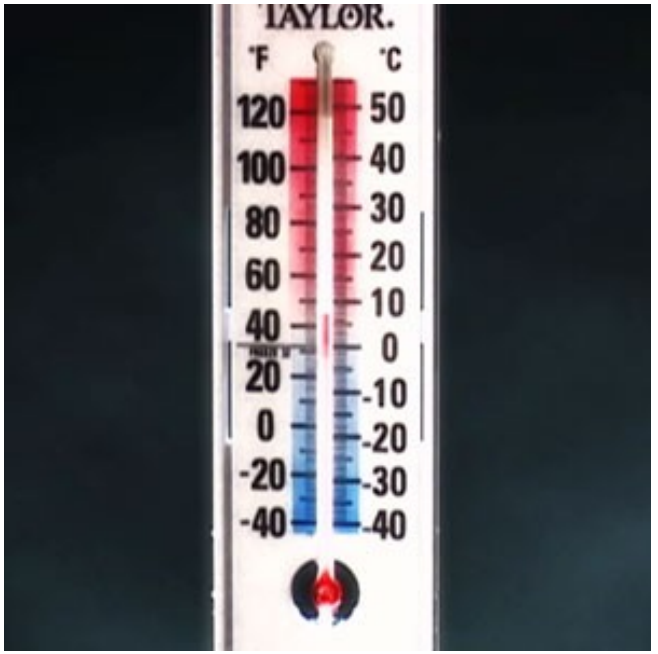
Also, copy-number analyses of the genome had to occur on separate platforms.

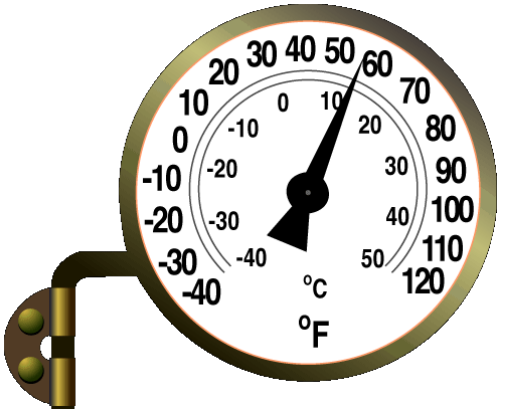
Logistic regression furnishes a means to rank information content in each probe

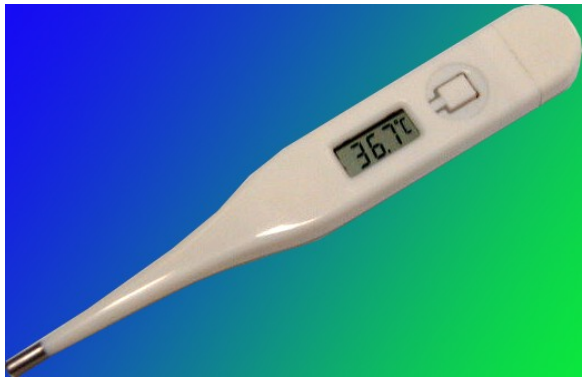


The information metric naturally selects which probe pairs perform best




TAYLOR.
°F
°C
120
100
80
60
40
20
0
-20
-40
50
40
30
20
10
0
-10
-20
-30
-40

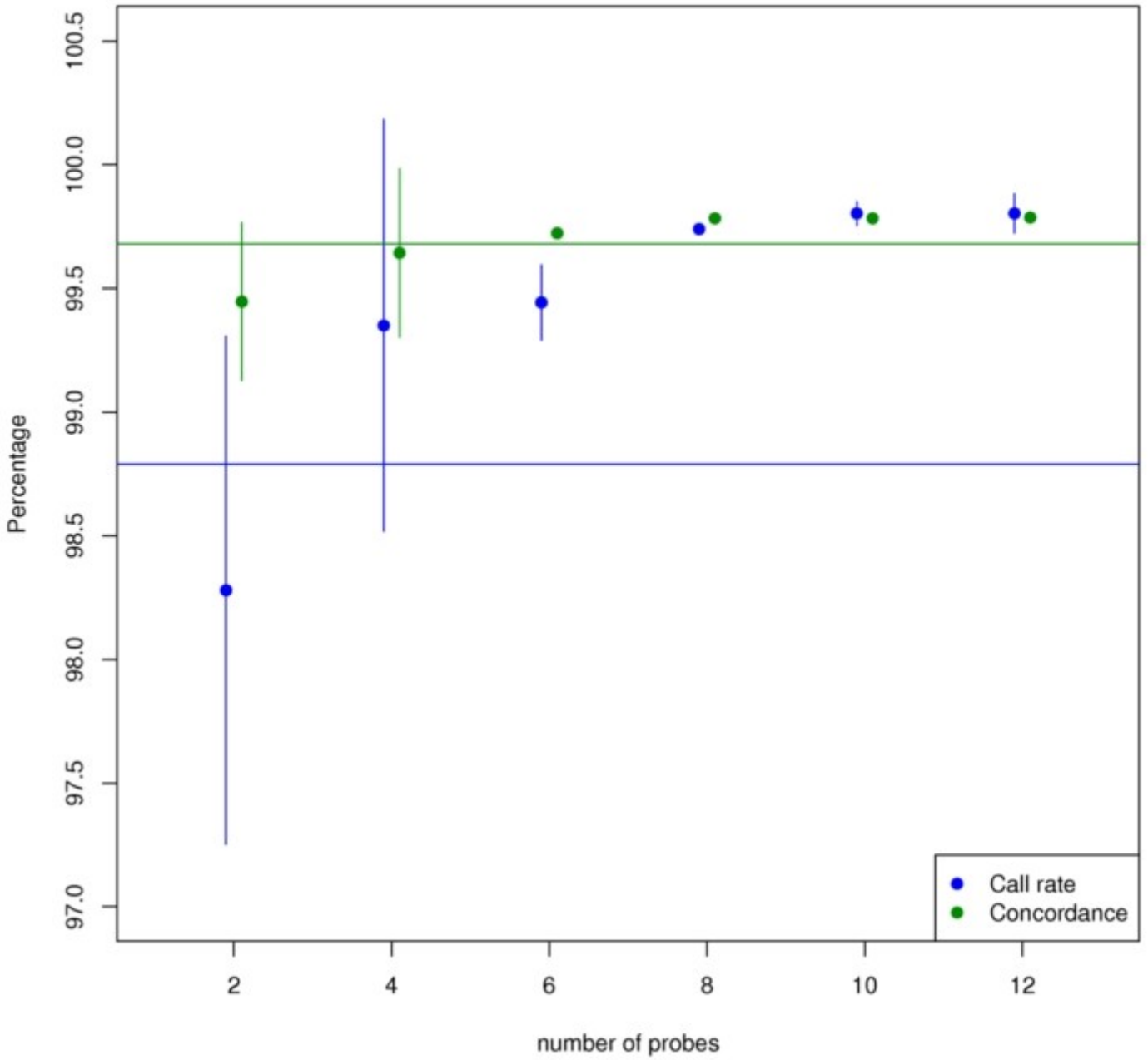

°C
°F


36.7°C


31.6°C
in
out
ice warner


37.0°C

Simulation experiments suggested that a new design would outperform the existing 500K design



Two chips were based on this new design

- Working closely with Affymetrix, with all the extra space, the two chip design (Nsp and Sty) could be reduced to one chip

- Copy-number probes were added with the still remaining space. (Steve McCarroll & Josh Korn)

- SNP6.0 has > 900,000 SNPs, with the new SNPs picked according to a multimarker tagging approach (Paul de Bakker).

% SNPs on Hapmap phase II captured with $r^2 > 0.75$ (multimarker tagging)

	500K	SNP6.0
YRI	59%	82%
CEU	80%	93%
EAS	78%	91%

- SNP5.0 was released in February 2007, SNP6.0 in June 2007. Tens of thousands of these arrays have now been used all over the world to study various diseases (at the Broad: autism, schizophrenia, lupus, heart disease, etc.)

Birdseed is a new tool to genotype SNPs on the Affymetrix SNP5.0 and SNP6.0 arrays

Existing algorithms (DM, BRLMM) could not work on the new chip design

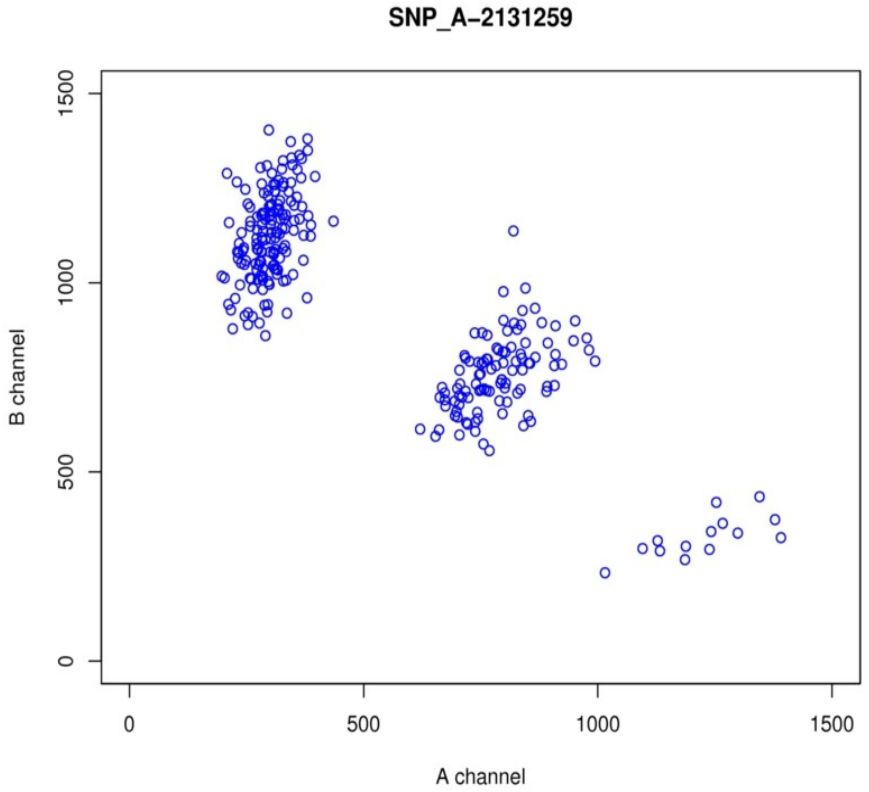
High-level overview of the Birdseed algorithm

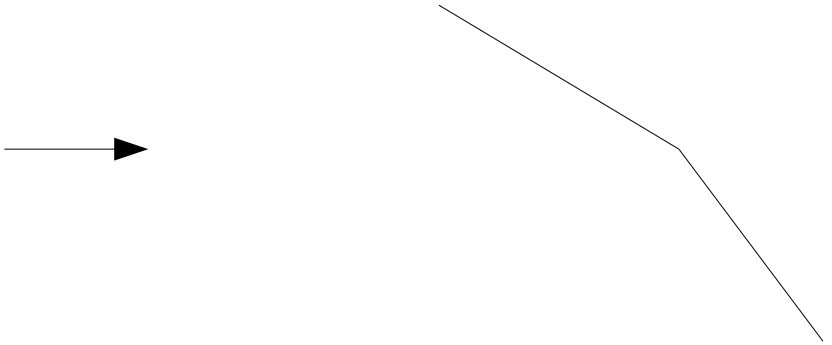
Phase 1: Building models from training data for every single SNP on the array

Phase 2: Genotyping SNPs on never-seen data using those models

(NOTE: be sure that you are running the latest version, which is not yet distributed by Affymetrix.)

In Phase I, Birdseed builds models of all SNPs by using a training data set (Hapmap)

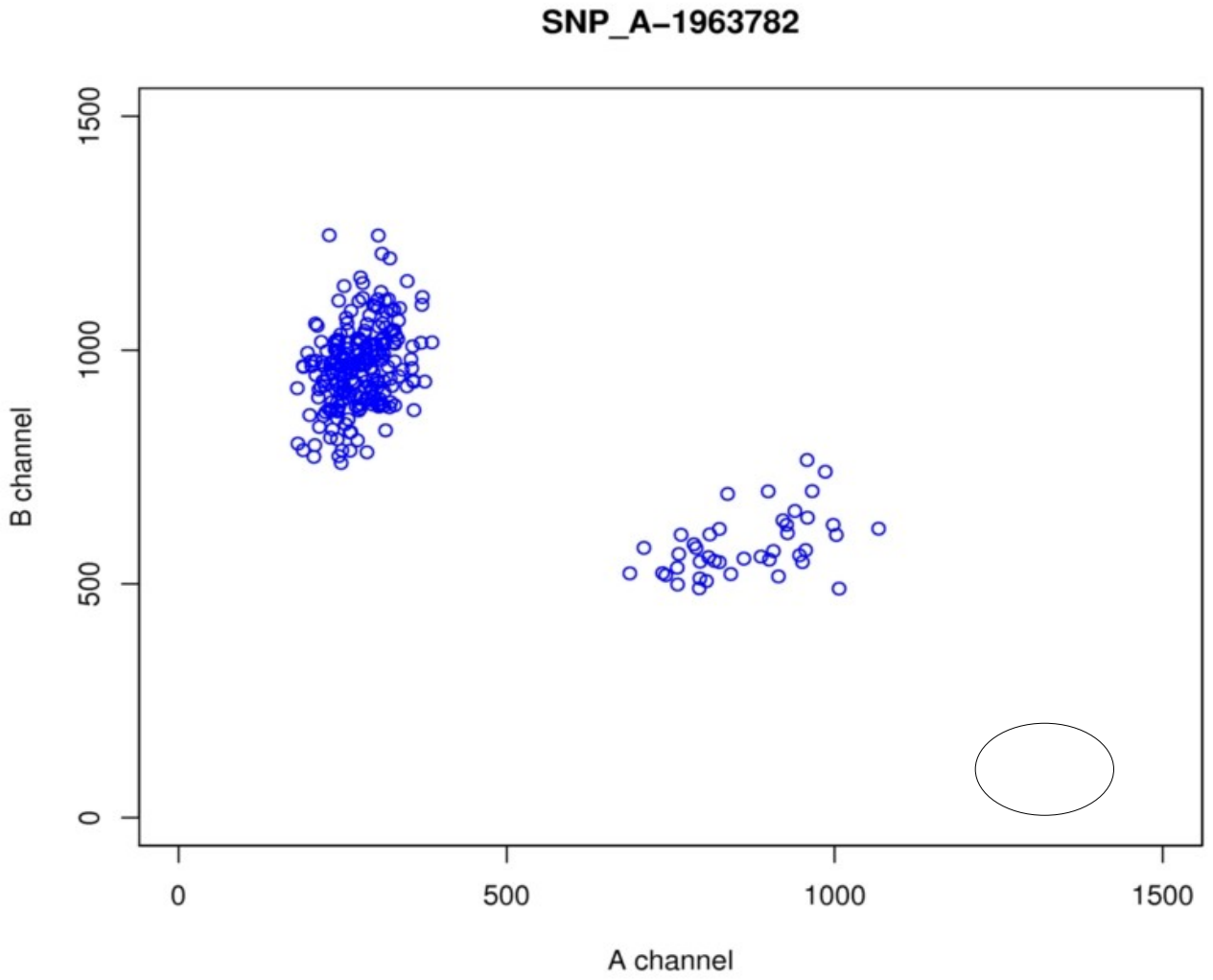




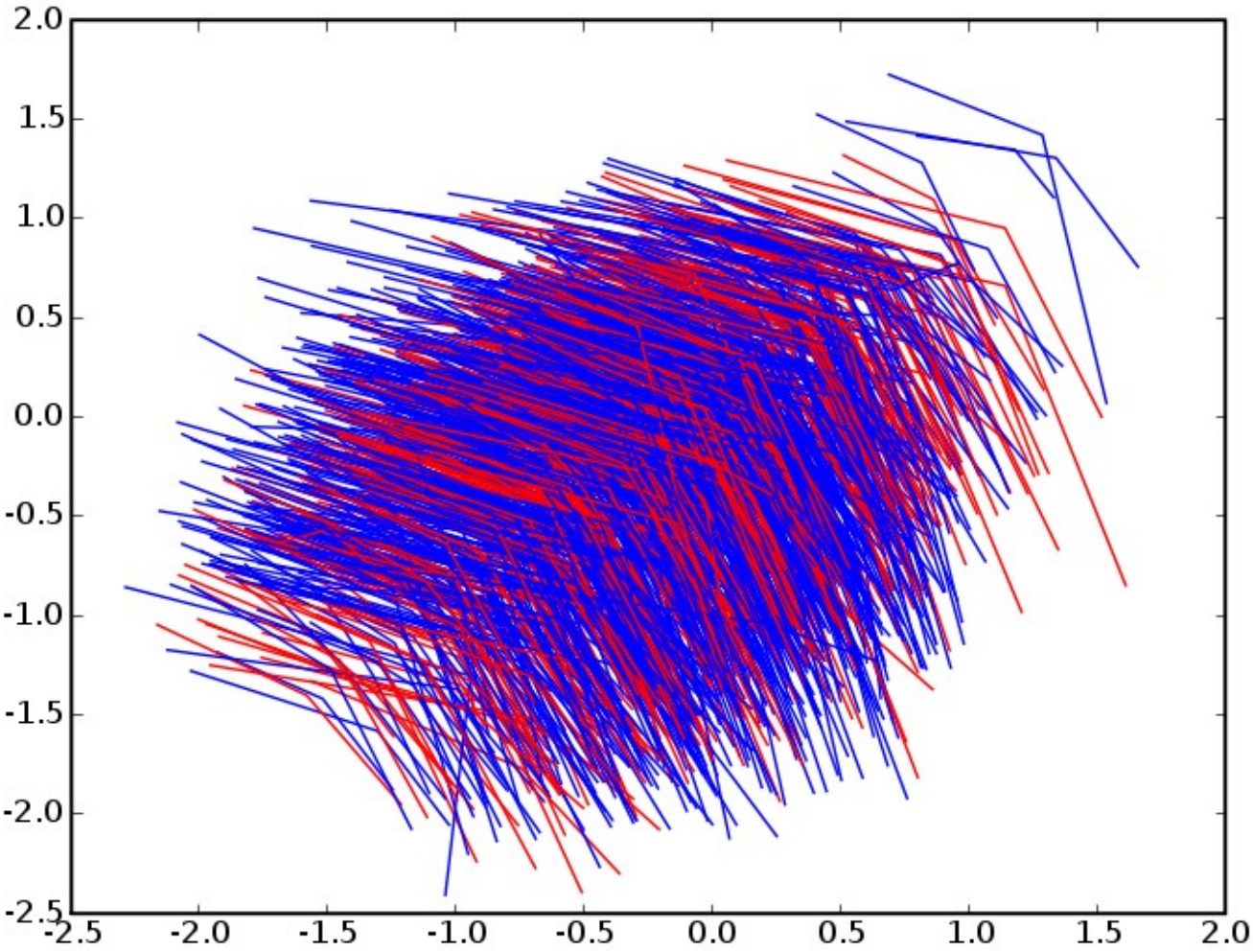
Each SNP can be thought of as a bird. The wingtips are AA and BB, the body is AB. Birds are computed for all SNPs.

AA: 1.1671 0.3133 0.0108 0.0039 0.0028 14
AB: 0.7499 0.7224 0.0056 0.0034 0.0089 102
BB: 0.2852 1.0713 0.0018 0.0019 0.0125 154

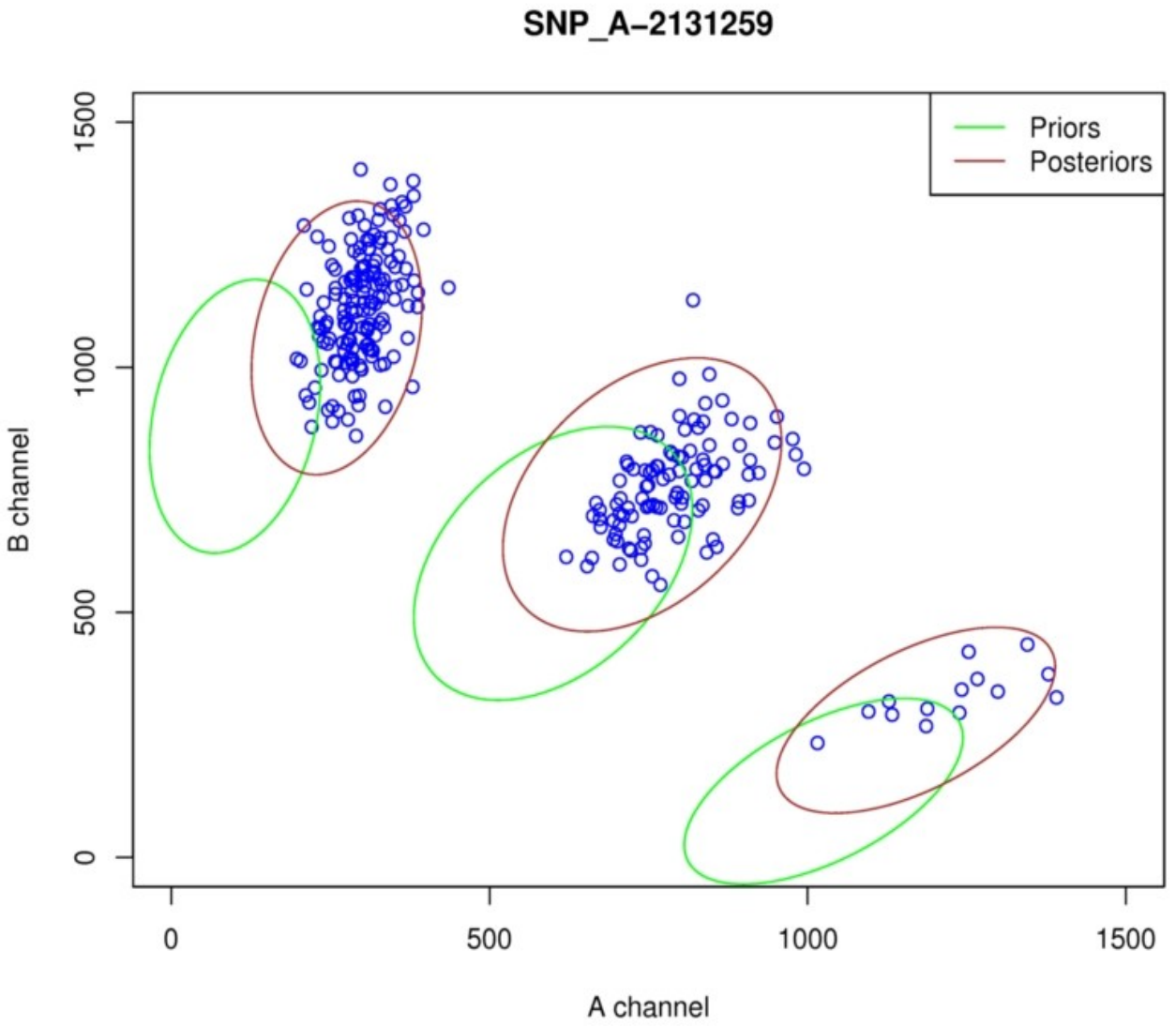
Since not all the clusters are present in training data, birdseed estimates cluster centers and covariance matrices



Birdseed can make highly accurate predictions because it has learned cluster morphology patterns by studying flocks of birds



In Phase II Birdseed uses a highly customized EM algorithm using the SNP-specific bird as the “seed” (hence the name) & as cluster anchors



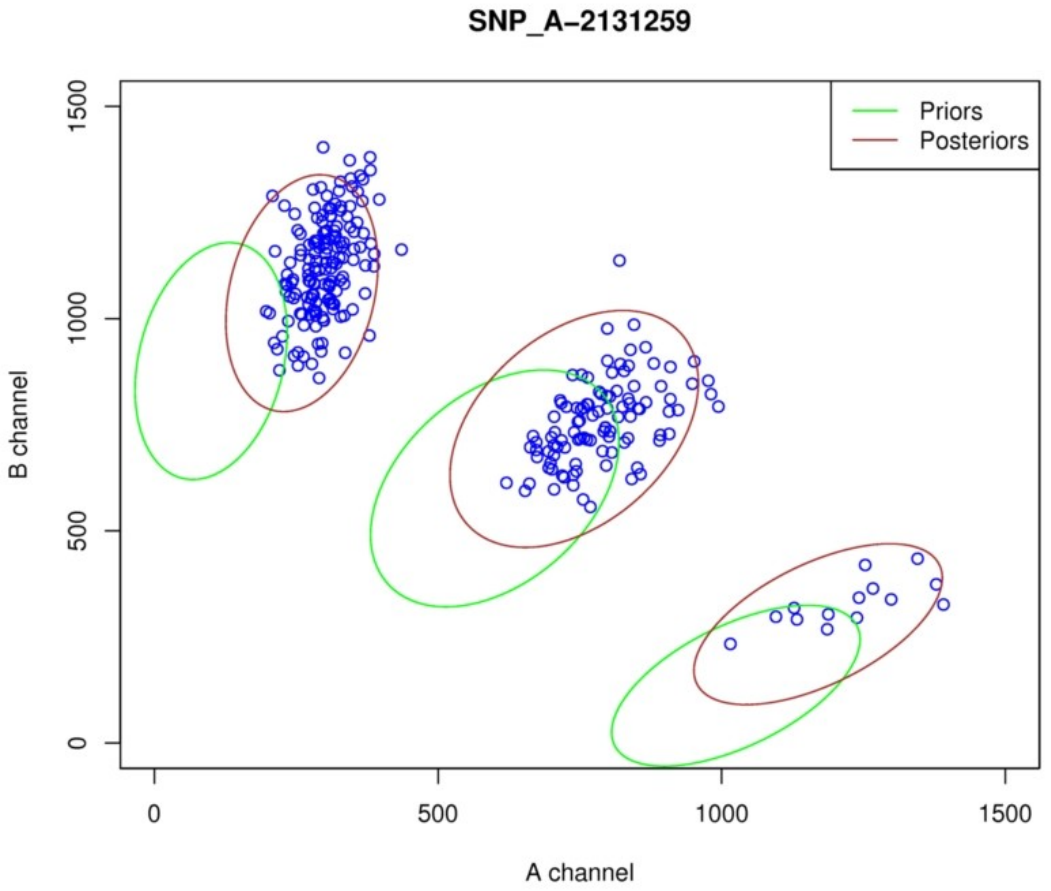
Josh Korn and Alec Wysoker

Birdseed provides a confidence score for every genotype it makes (0 is the best, and 1 is the worst)

Confidence = 80% * E1 + 20% * E2

E1 = posterior to 2nd closest peak / posterior to closest
E2 = deviation penalty from closest peak

Quality score = $-\log_{10}(\text{confidence} + 0.00001) * 2000$



Birdseed performance has been outstanding on diverse samples at many centers

Typical unfiltered call rates: ~99%
Typical unfiltered concordance with Hapmap: ~99.6%

Example data set (GoKinD)

Confidence:	0.1	0.056	0.032	0.018
Quality score:	2000	2500	3000	3500
Number of samples:	3051	3051	3051	3051
Number of SNPs:	500568	500568	500568	500568

Apply three filters for individual call rate, SNP call rate, and SNP allele freq.

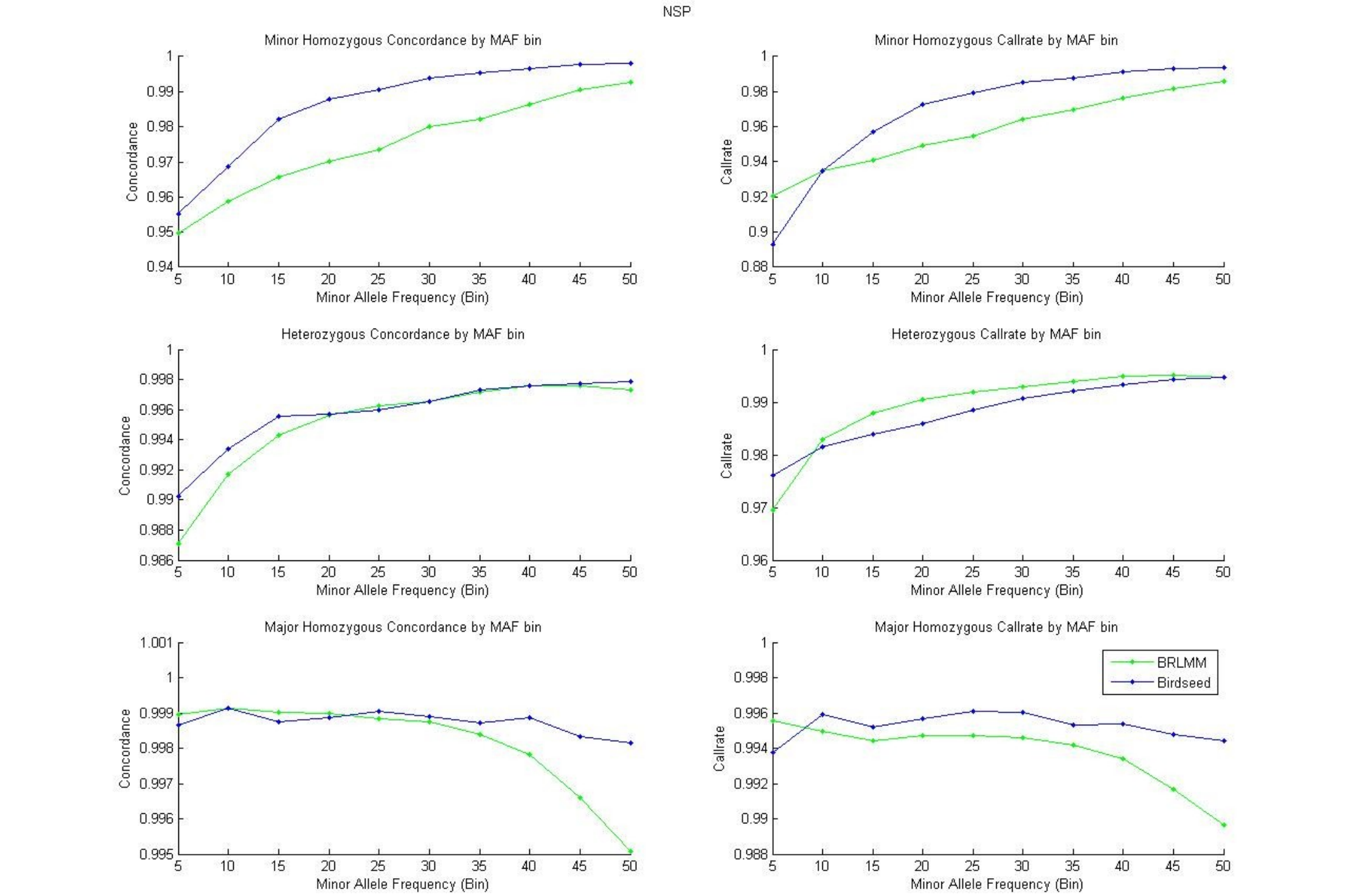
# indiv CR < 90%	1	16	57	287
# SNPs CR < 95%	31293	42101	53893	72079
# SNPs AF < 1%	60958	62805	64760	66164

Post-filter statistics:

# indiv remaining	3050	3035	2994	2764
# SNPs remaining	413060	403021	392346	377527
Ave Call Rate	99.6	99.5	99.2	98.0
# SNPs HWE <1e-3	25050	24891	25522	24707
# SNPs > 4 ME	18619	12175	5479	1737
Ave ME / indiv	529	373	192	63
#SNPs failing PA	2173	1465	675	253

Amanda Elliott

Birdseed even outperforms BRLMM on 500K



Josh Korn

Acknowledgements

David Altshuler
Mark Daly
Stacey Gabriel

Steve McCarroll
Josh Korn
Alec Wysoker
Paul de Bakker
Amanda Elliott
Julian Maller

Simon Cawley
Steve Lincoln