Improved gridding algorithms in AGCC produce equivalent biological results

In the transition from GCOS to AGCC, a minor change with a negligible effect on the biological data produced has occurred in the method that transforms DAT files to CEL files. Less than one percent of the time GCOS finds a slightly suboptimal grid location (off by one pixel at one corner). In AGCC the grid finding methodology has been improved to find the optimal grid location for all DAT files. AGCC produces results that are more consistent than those provided by GCOS. The algorithmic changes are minimal in effect, at most one pixel difference in the placement of grid locations, and have small effects compared to experimental noise. In the few rare circumstances where the new CEL file results are slightly different they produce equivalent biological results. The list of affected arrays is given in Appendix A.

To verify that the downstream changes are minimal in scope, two representative arrays from the array designs that would be most dramatically affected by the sub-optimal grid placement were investigated: a genotyping array Xba50K, and an expression array U133A. Amongst genotyping arrays, the Xba50K array is most likely to be subject to possible impact by the gridding difference as it has the smallest feature among arrays that do not employ a gridding optimization known as "feature extraction". Feature extraction optimizes the location of every feature and provides an extra layer of robustness to small variances in the position of the grid. Similarly for expression, the U133A array has the smallest feature size among arrays that do not employ the feature extraction optimization and hence is the expression array that will be most affected by the potential gridding differences.

For each chosen array type the use of the suboptimal corner finding routine was forced to maximize the number of times this rare event occurred. This increased the frequency of affected CEL files to approximately one in two, much higher than the typical rate of less than 1 in 100. Even in this worst-case scenario, while many of the individual intensities in a CEL file change, only 2% of individual features typically change by more than 10%, and the quality of the biological results is equivalent (in part due to the robustness of the analytic methods used).

For the representative genotyping Xba50K array, 63 experiments (arrays) were analyzed. Forcing the use of the suboptimal grid optimization resulted in 18 arrays with a sub-optimal grid position. For these 18 arrays, approximately 9% (median across the affected arrays) of the features intensity changed by at least 0.1 (the smallest reportable change). These small changes to a minority of features will have a minimal impact on the genotyping results. In this generation of genotyping arrays, the DM algorithm (Di et al, Bioinformatics 2005) is used for genotyping. The DM algorithm uses a robust combination of probe quartets to evaluate the genotype, and thus is resistant to small changes in a minority of probes. For these 18 arrays, the overall genotyping performance was found to be essentially identical (Figure 1). This figure plots the proportion of correct calls (y-axis) against the proportion of SNPs for which no call is made (x-axis) for all possible settings for the no-call threshold. This enables the comparison of results even when the scores for any particular SNP or collection of SNPs may change with respect to the threshold. For example, at a 5% no-call rate, the arrays show 99.8% concordance (red lines) with HapMap over all SNPs in all experiments, whether sub-optimally gridded or not. Some individual SNPs may change from calls to no-calls (or vice versa) in a given experiment, or change genotype, but the overall accuracy and call-rate are very minimally different. Therefore the impact of using the updated optimal routine is negligible on genotyping performance.

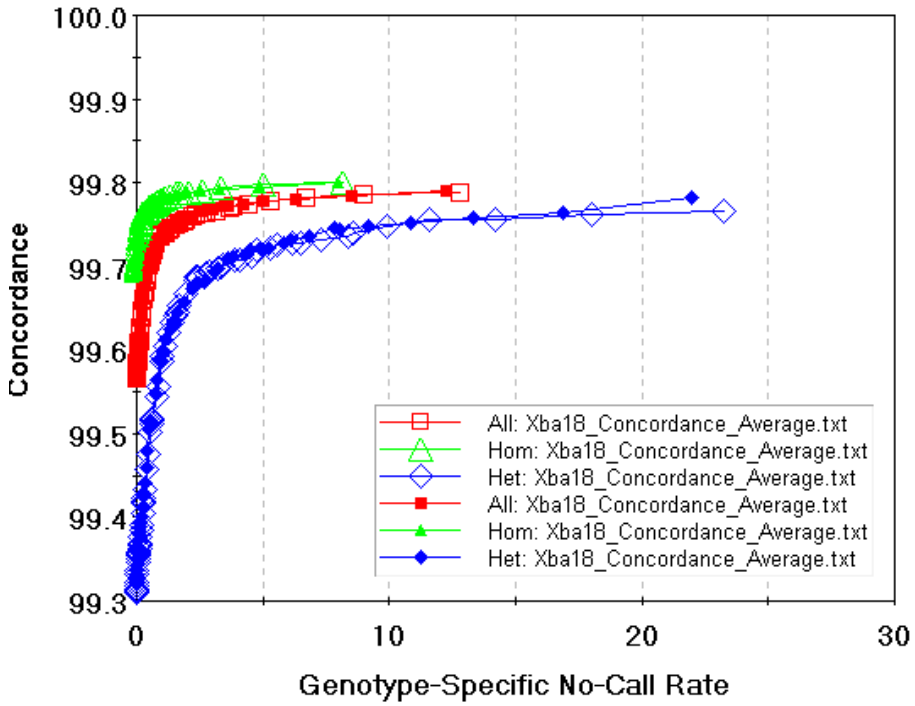

Figure 1: Comparison of the concordance of genotype calls to HapMap reference genotypes for the optimal and sub-optimal grid placements.
The open symbols reflect optimal gridding and the closed symbols represent the forced suboptimal gridding. Each symbol on the line represents a given score threshold for no-calls. For a range of possible confidence thresholds on genotype calls the no-call rate and the concordance with HapMap reference genotypes is plotted. The red curves summarize the overall performance, the green and blue curves decompose performance into homozygous specific and heterozygous-specific performance, respectively. The two gridding modes yield almost identical performance.

For the representative expression array, U133A, 42 experiments (arrays) from the public Latin square data set were analyzed. Forcing the use of the suboptimal grid optimization resulted in 27 arrays with a sub-optimal grid position (again, this is very much a worst-case scenario given that the frequency of the suboptimal grid location is less than one percent). While approximately 30% of the features change intensity by at least 0.1 (the smallest reportable change) only 2% of individual features typically change by more than 10% in intensity, which is small compared to the experimental noise (Figure 2). Figure 2 (A-D) also demonstrates that the larger intensity changes are predominantly limited to the lowest-intensity features. Furthermore, most summarization routines (*e.g.*, MAS 5, RMA, and PLIER) are robust to individual feature changes, and therefore even less change is observed at the signal level. The typical correlation of the PLIER signal intensities is 0.999 between the sub-optimal and optimal gridding algorithms (Figure 2a). This compares very favorably to the correlation between technical replicates in this experiment which ranged from 0.988-0.9996 (data not shown). While the Percent Present (% P) call rate produced by the MAS5 algorithm is not always an easily-interpretable metric, people generally monitor its behavior. Figure 3 displays the % P for the optimal grid placement in red and for the sub-optimal grid placement in blue for the 27 arrays with the sub-

optimal grid placement in GCOS. The largest % P difference observed for the sub-optimal grid placement was 0.5% compared to 4.6% for the largest % P difference between any of the technical replicates (data not shown). After comparing the signal values and %P calls from the optimal and sub-optimal grid placements, we conclude that the effect of always using the optimal routine is negligible on expression performance.

In summary, AGCC fixes an issue that causes GCOS to use a sub-optimal grid location (off by at most 1-pixel in the corner) for less than one percent of DAT files for the susceptible arrays listed Appendix 1. This difference in gridding results in small changes in the intensities calculated for approximately 9% of the features. Both expression and mapping analysis use summarization methods that robustly combine multiple features into a single probeset minimizing the observed effect at the signal level. Examination of genotyping experiments analyzed with both grid alignments determined that the overall accuracy and call-rate are equivalent but not identical. Similarly, examination of the signal values and the %P calls for expression arrays revealed very minimal differences. In fact, the technical variation between replicates seen in the experiments was greater than the variation caused by the sub-optimal grid locations. Therefore, we conclude that the impact of using the updated optimal gridding routine is negligible on genotyping and expression performance.

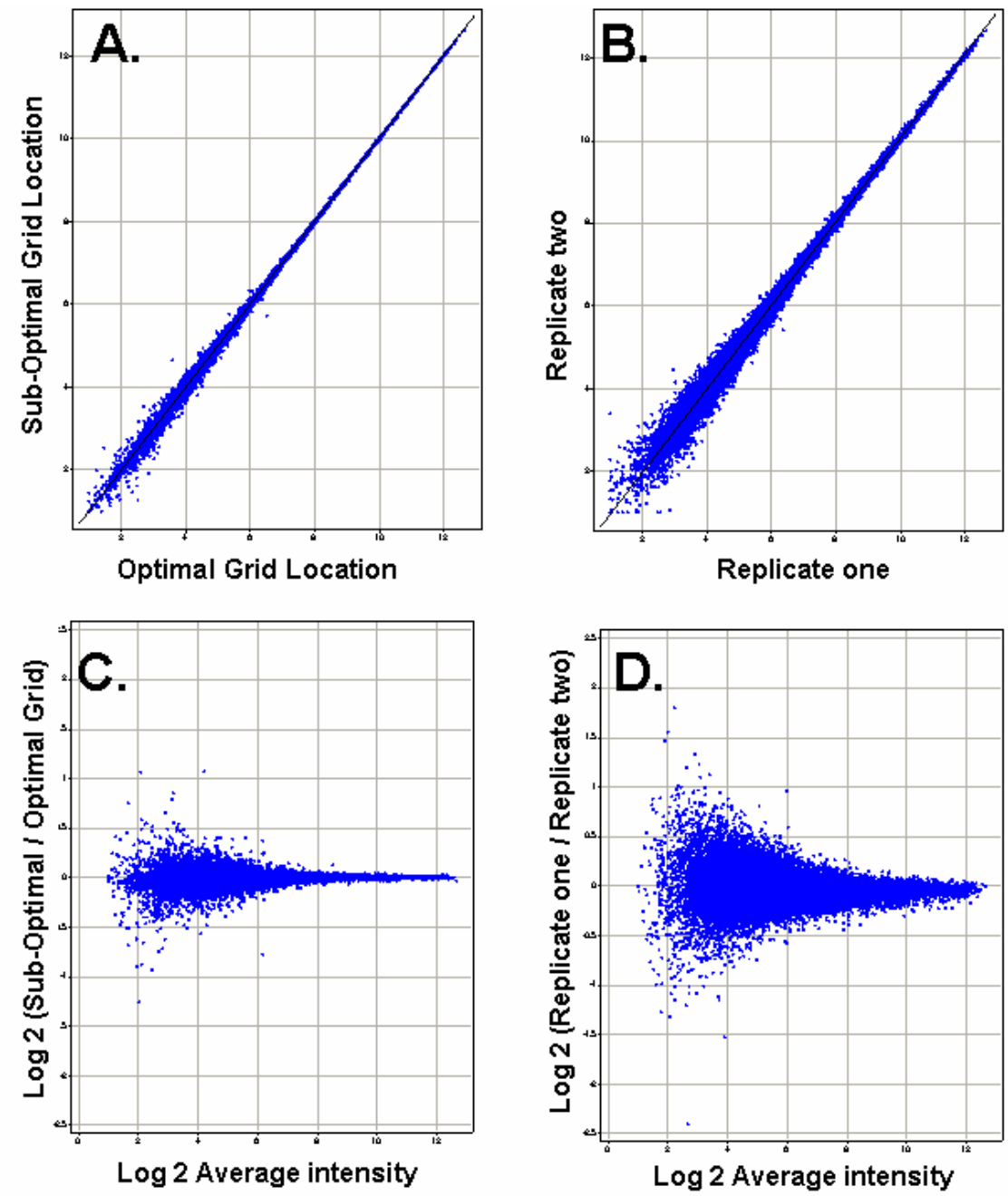

Figure 2: Comparison of the technical variation that due to sub-optimal grid placement.

The PLIER signals from a sub-optimal grid location are plotted against the PLIER signals from an optimal grid location for a representative expression array in **A**. The PLIER signal values had 2 added to them to avoid taking the log of near-zero signal results. The R^2 value is 0.999 and the linear fit is y=-0.021+1.003*x indicating that the two arrays are very highly correlated and there is no overall bias introduced by one of the grid placements (*i.e.*, no significant impact on derived fold changes). In **B** a similar plot for two technical replicates (including the one used in A) with the optimal grid place. The R^2 value is 0.995 and the linear fit is y=-0.067 + 1.013*x. **C** and **D** show MvA plots for arrays in **A** and **B** respectively. Note that the variation introduced by the sub-optimal grid placement is less than that observed for the technical replicates.

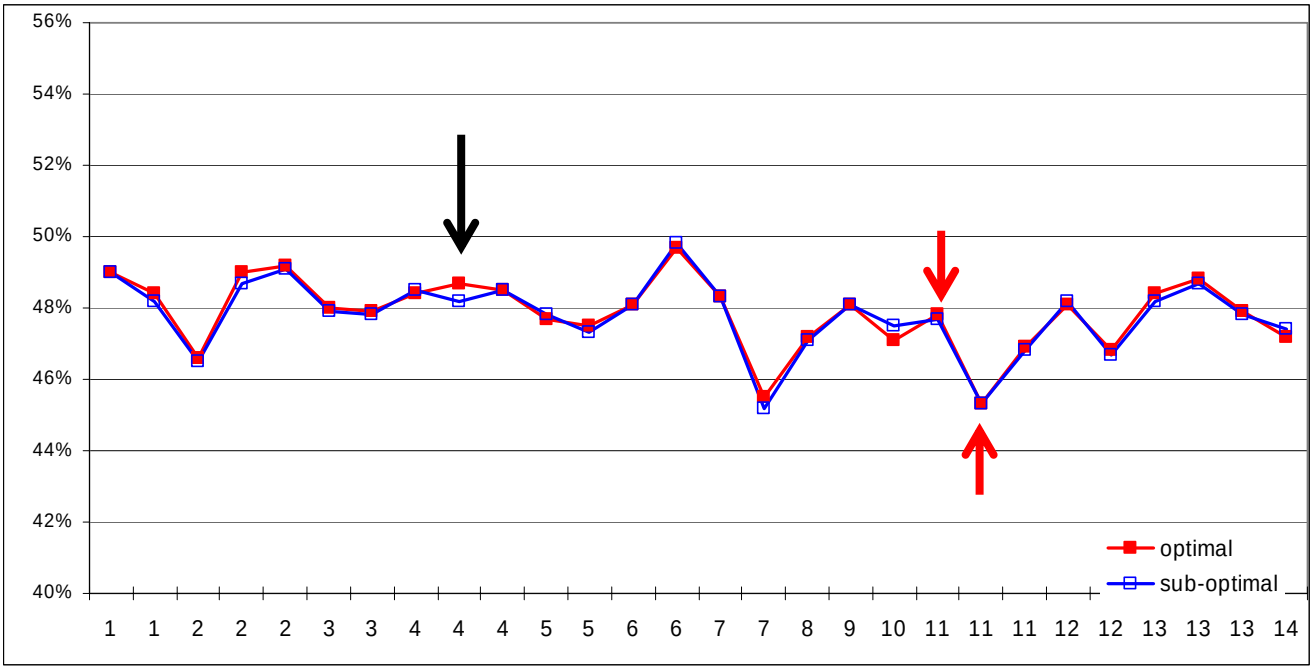

Figure 3: Comparison of the sub-optimal grid placement on the MAS5.0 %P calls.

The MAS5.0 algorithm was used to generate the %P calls for the 27 arrays affected by the sub-optimal grid placement with both grid placements. The y-axis contains the %P call and the x-axis is the individual experiments, with the technical replicates represented by the repetition of the experiment number. The %P for the optimal grid alignment is shown in red, and the sub-optimal are shown in blue. The largest difference observed difference between the optimal and sub-optimal grid placement was 0.05% (black arrow). In general the difference in %P between technical replicates for the same experiment was larger than the difference between the gridding algorithms (compare the difference between points represented by the two red arrows).

References:

[1] Xiaojun Di et al. *Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays*, Bioinformatics 2005 21(9):1958-1963

Appendix A

The following is a list of commercial array designs susceptible to GCOS selecting a sub-optimal grid location, resulting in a 1-pixel difference in grid location in less than one percent of the cases. For information about the affected Custom Arrays please contact Affymetrix Support.

Commercial Expression Arrays

1286_s01
A_tumefaci530019
Ag
ATH1-121501
B_anthraci530023
B_anthraci530025
B_longum530035
B_melitens530041
B_suis530042
Barley1
Bovine
Bsubtilis
C_diphther530061
C_trachoma530050
Canine
Canine_2
Celegans
Chicken
Chrom21-22A
Chrom21-22A_new
Chrom21-22B
Chrom21-22C
Citrus
CornChip0
Cotton
Drome_AntiSense
DrosGenome1
Drosophila_2
DrosophilaTiling-Forward
DrosophilaTiling-Reverse
E_coli_2
Ecoli
Ecoli_ASv2
encode01
ENCODE01-Forward
ENCODE01-Reverse
H_influenz530078
HC_G110
HG_U95A
HG_U95Av2
HG_U95B
HG_U95C
HG_U95D
HG_U95E
HG-Focus
HG-U133_Plus_2
HG-U133A
HG-U133A_2
HG-U133A_tag
HG-U133B
HG-U95_new
HT_HG-U133_Plus_A
HT_HG-U133_Plus_B
HT_HG-U133A
HT_HG-U133A_old
Hu6800
Human-CMM_1
L_monocyto530090
M_avium530100
M_bovis530101
M_jannasch530094
M_tubercul530104
Maize
Medicago
MendelQC_510nm34545ns
MendelQC_700nm34545ns
MG_U74Av2
MG_U74Bv2
MG_U74Cv2
MOE430A
MOE430B
Mouse430_2
Mouse430A_2
Mu11KsubA
Mu11KsubB
N_europaea530115
N_meningit530113
N_meningit530114
NuvoSelect_v1
P_gingival530124
P_putida530130
P_syringae530131
Pae_G1a
Pae_G1a_noIgnoreShiftRowOutliers
Par1
ParAllele_tags
ParAllele_tags_2
ParAllele_tags_3
Plasmodium_Anopheles
Poplar
Porcine
qcchip_2214_mod1
R_prowazek530139
RAE230A
RAE230B
Rat230_2
RatToxFX
RG_U34A
RG_U34B
RG_U34C
Rhesus
Rice
RN_U34
RT_U34
S_aureus
S_epidermi530152
S_flexneri530144
S_mutans530155
S_pneumoni530156
S_pneumoni530157
S_pyogenes530158
S_sp530168
S_typhimur530142
Soybean
Sugar_Cane
T_denticol530175

TB-All-Antisense
TB-All-Sense
Test3
Test3_Format_1
Test3_Format_2
Test3_Format_3
Test3_noIgnoreShiftRowOutliers
Tomato
TrueTag25KB_15
TrueTag25KB_25
U133_X3P
U133AAofAv2
V_cholerae530180
V_parahaem530181
Vitis_Vinifera
wheat
X_laevis_2
X_tropicalis
Xenopus_laevis
Yeast_2
YG_S98
Zebrafish

Commercial Mapping Arrays

AD082_169_511060
ax13339
Citrus_SNP
HT_Mapping50K_Xba
HU131_10K_Xba_Mapping
HuGeneFocused50K
HuGenomeWide50K
HuSNP
Mapping10K_Xba131
Mapping10K_Xba142
Mapping50K_Hind240
Mapping50K_Xba240
Xba142_EA

Commercial Universal Arrays

arabidopsis_tlgF
arabidopsis_tlgF_4x
arabidopsis_tlgR
arabidopsis_tlgR_4x
Chrom21_22A_F_4x
Chrom21_22A_F_v04
Chrom21_22A_R_4x
Chrom21_22A_R_v04
Chrom21_22B_F_4x
Chrom21_22B_F_v04
Chrom21_22B_R_4x
Chrom21_22B_R_v04
Chrom21_22C_F_4x
Chrom21_22C_F_v04
Chrom21_22C_R_4x
Chrom21_22C_R_v04
DrosophilaTlg-Fwd_4x
DrosophilaTlg-Rev_4x
ENCODE01_F_4C
ENCODE01_F_v03
ENCODE01_F_v04
ENCODE01_R_4C
ENCODE01_R_v03
ENCODE01_R_v04
ENCODE01-Forward_4x
ENCODE01-Reverse_4x
GenFlex
GenFlex_Tag_16K_dev
GenFlex_Tag_16K_v2
Human35bp_1_01F
Human35bp_1_01R
Human35bp_1_02F
Human35bp_1_02R
Human35bp_1_03F
Human35bp_1_03R
Human35bp_1_04F
Human35bp_1_04R
Human35bp_1_05F
Human35bp_1_05R
Human35bp_1_06F
Human35bp_1_06R
Human35bp_1_07F
Human35bp_1_07R
Human35bp_1_08F
Human35bp_1_08R
Human35bp_1_09F
Human35bp_1_09R
Human35bp_1_10F
Human35bp_1_10R
Human35bp_1_11F
Human35bp_1_11R
Human35bp_1_12F
Human35bp_1_12R
Human35bp_1_13F
Human35bp_1_13R
Human35bp_1_14F
Human35bp_1_14R
MIP_Tag_70K
Mouse35bp_1_01F
Mouse35bp_1_01R
Mouse35bp_1_02F
Mouse35bp_1_02R
Mouse35bp_1_03F
Mouse35bp_1_03R
Mouse35bp_1_04F
Mouse35bp_1_04R
Mouse35bp_1_05F
Mouse35bp_1_05R
Mouse35bp_1_06F
Mouse35bp_1_06R
Mouse35bp_1_07F
Mouse35bp_1_07R
Mouse35bp_1_08F
Mouse35bp_1_08R
Mouse35bp_1_09F
Mouse35bp_1_09R
Mouse35bp_1_10F
Mouse35bp_1_10R
Mouse35bp_1_11F
Mouse35bp_1_11R
Mouse35bp_1_12F
Mouse35bp_1_12R
Mouse35bp_1_13F
Mouse35bp_1_13R

Mouse35bp_1_14F
Mouse35bp_1_14R
Mouse35bp_1_15F
Mouse35bp_1_15R
Mouse35bp_1_16F
Mouse35bp_1_16R
ParAllele_Tags_Haw
PombeAlla520099
S_cerevisiaeF_4x
S_cerevisiaeR_4x
S_pombe
S_pombe_4x
TAG_3
Tag3_BCM
TrueTag_10K_A
TrueTag_10K_B
TrueTag_25K_A
TrueTag_25K_A-PI
TrueTag_25K_B
TrueTag_30K_A
TrueTag_3K_A
TrueTag_3K_A_570nm
TrueTag_5K_A
TrueTag_5K_B
TrueTag_5K_B1
Univ_70K_Tag

Custom Arrays

For information about the affected Custom Arrays please contact Affymetrix Support.