# Supplementary Material: Copy-number-aware differential analysis of quantitative DNA sequencing data

### Mark D. Robinson1,2\*, Dario Strbenac2, Clare Stirzaker2,5, Aaron L. Statham2, Jenny Song2, Terence P. Speed3,4, Susan J. Clark2,5

1Institute of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190 CH-8057 Zurich, Switzerland

2Epigenetics Laboratory, Cancer Research Program, Garvan Institute of Medical Research, Sydney 2010, New South Wales, Australia

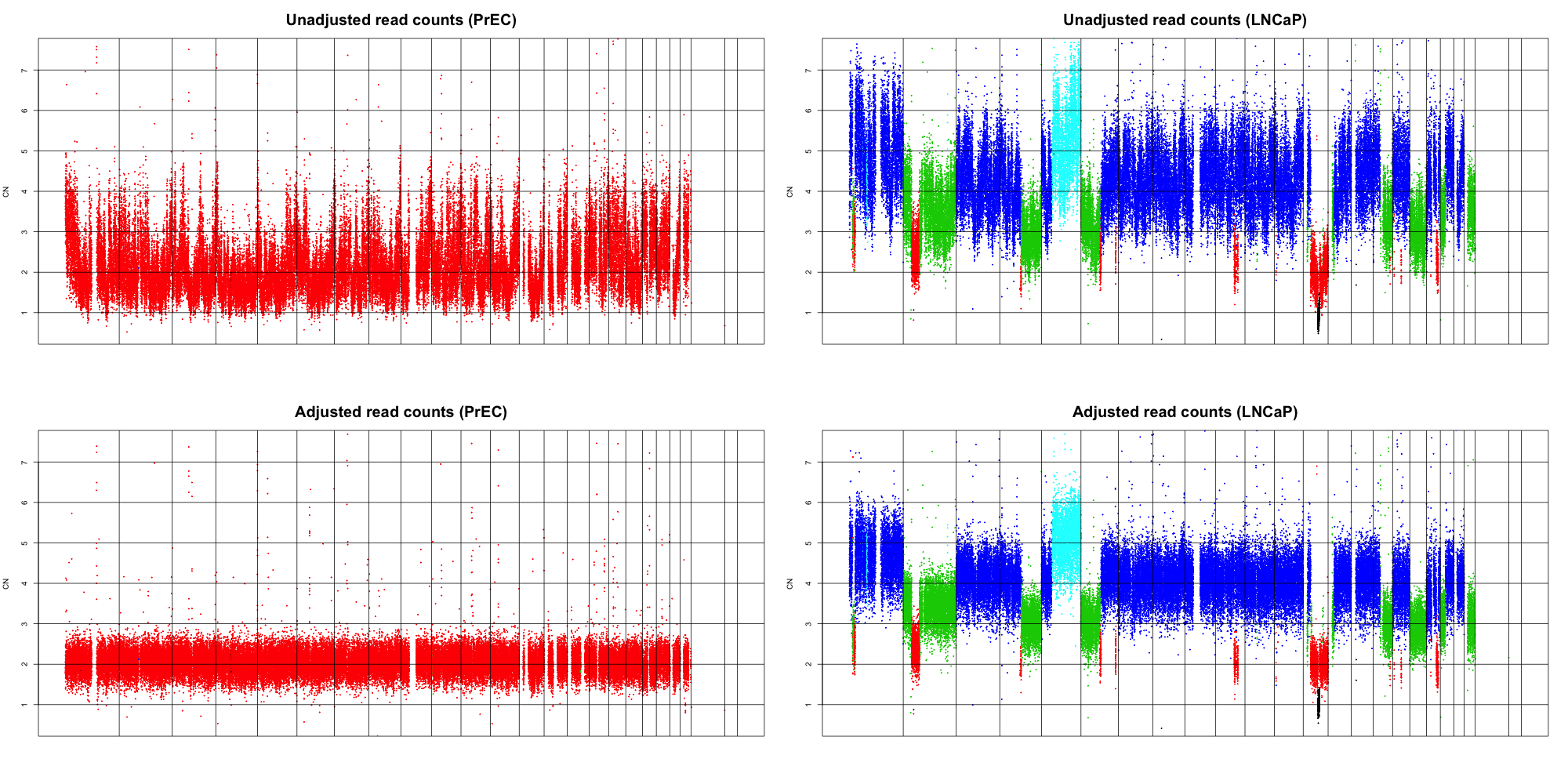
3Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Parkville, Melbourne 3052, Victoria, Australia

4Department of Medical Biology, University of Melbourne, 3050, Victoria, Australia

5St Vincent's Clinical School, University of NSW, Sydney, NSW, Australia

\*Corresponding author

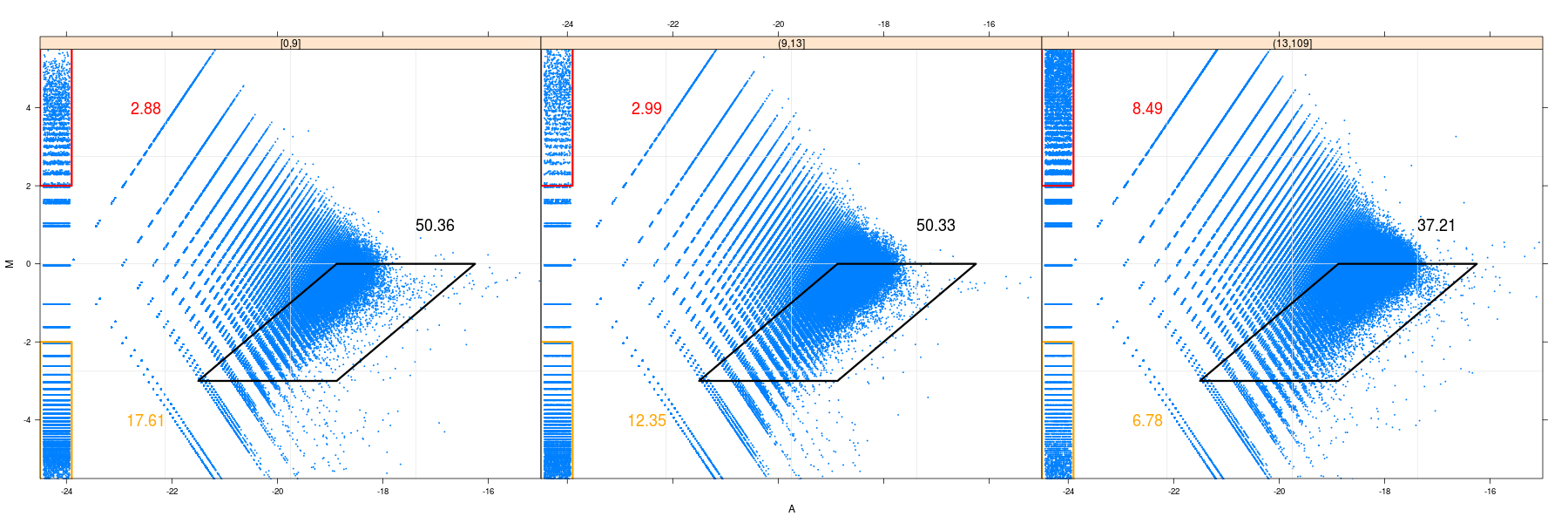
iMac-MarkR:Users:mark:projects:cn_paper:ABCD-DNA_Supplement_RCode_Data:Figures:Supplementary_Figure_1_cndistro.pdf**Supplementary Figure 1** – Genome-wide summary of the PICNIC estimates from LNCaP Affymetrix SNP 6.0 array data. Clearly, 4 copies is the most prominent state (~65% of the genome) and is therefore used for the normalization factor estimation.



### Supplementary Figure 2 – High concordance in CNV estimates between Affymetrix SNP 6.0 arrays and gDNA-seq read densities. The top panels show raw read densities (normalized to median and multiplied by 2 and 4, respectively for PrEC and LNCaP), coloured according to the inferred copy estimate from Affymetrix SNP 6.0 arrays (PICNIC algorithm; black-1 copy, red-2 copies, green-3 copies, blue-4 copies, cyan-5 copies). The bottom panel shows read densities after GC content and mappability correction (See Methods); these values are segmented and used as CNV offsets in ABCD-DNA.

### mark:projects:cn_paper:Additional_file_3_smear_by_CNV.pdf

### Supplementary Figure 3 – M versus A “smear plots” for various DNA-seq datasets, stratified by CNV state between LNCaP and PrEC cells. M (depth-normalized log-fold-change) versus A (depth-normalized average-log-count) plots for MBDCap-seq (a), H3K27me3-seq (b) and H3K4me3-seq (c). Blue lines are a suitably-chosen percentile of A values and red lines represent median M values for points to the right of the blue line. One can notice a consistent increase in the center of the M values, according to CNV state. The scaling factors (red horizontal lines), which are summarized in Figure 2, are used here for illustration purposes only and not in the ABCD-DNA method itself.



### Supplementary Figure 4 – Smear plots of normalized MBDCap-seq (after estimation of scaling factors), stratified by CpG density. CpG density is calculated for all copy-number-neutral regions. Regions are split into 3 equally-sized categories (low, medium, high). Red, black and orange polygons (red, black and orange numbers representing percentage of points, respectively) highlight the asymmetry in DNA methylation changes; yet, the center of the M values asymptotes at 0.

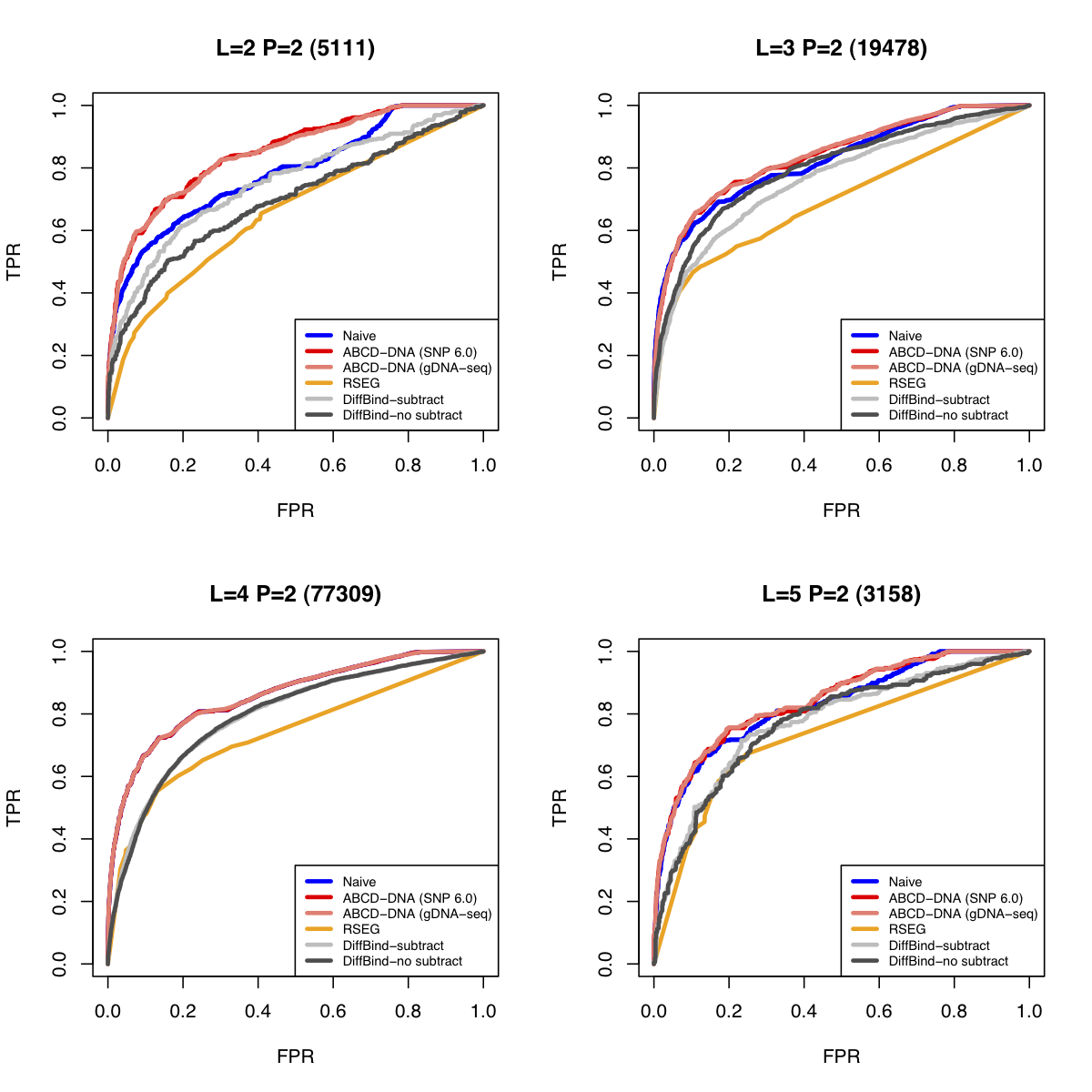
**mark:projects:cn_paper:Additional_file_1_inputs_at_ruikes_hyper_hypo.pdfmark:projects:cn_paper:Additional_file_1_fig1.pdf**

### Supplementary Figure 5 – Association between CNV and differential peak detection rates for Kim et al. (Jung H Kim et al. 2011) and Ruike et al. (Ruike et al. 2010). A) RRPD (see Figure 4) is plotted using the peak detection calls reported in Kim et al. (HPeak algorithm) and the relative CNV according to PICNIC estimates on the SNP arrays used for *our* study. B) For the list of differentially methylated regions between MCF7 and HMEC cells (“hyper” are regions with higher methylation in MCF7, “hypo” are regions with lower methylation), distributions of the ratio of read densities in corresponding regions between MCF7 inputs and HMEC inputs are shown as boxplots. The “null” category represents all remaining regions (i.e. not differentially methylated).

**mark:projects:cn_paper:Additional_file_5_6up.pdfSupplementary Figure 6 – Effect of CNV compensation.** Venn diagrams show the genome-wide overlap of regions deemed significantly differentially methylated by “Naïve” and “CNV-aware” GLM analyses. An almost perfect overlap for copy-number-neutral (L=4 P=2) regions is observed, while the overlap degrades as CNV increases.

mark:projects:cn_paper:Additional_file_6_Zscores_by_CNV_truth.pdf

**Supplementary Figure 7 – Boxplots of MBDCap-seq Z-scores (P-values mapped onto the standard normal distribution and signed according to the log-fold-change) by CNV state and “truth”.**  Each group of 3 boxes gives a CNV state, with label “a.b”, where “a” denotes the truth according to 450k array (-1: hypomethylated, 1: hypermethylated, 0: not differentially methylated) and “b”, the copy number state in LNCaP cells (PrEC copy number state is 2 for all); for example, the label “1.L=3” denotes hypermethylated genomic regions where LNCaP has 3 copies. Top panel shows Naïve Z-scores (GLM without CNV offsets), where all boxes increase in magnitu; bottom panel shows CNV-aware Z-scores, where the magnitude of all boxes remains constant, regardless of CNV state. Notably, the Naïve scores are positively correlated with differential CNV, whereas distributions of ABCD-DNA Z-scores are flat with respect to CNV.



**Supplementary Figure 8 – Extra ROC curves to highlight sensitivity to asymmetry in the “truth”**. See Description in Figure 5. This plot highlights that results are sensitive to the bias towards hypermethylated regions.