Supplementary Notes #2:
Details on the methods used for the CRMA v2 study

Henrik Bengtsson et al.

April 12, 2009

Contents

1 How raw copy numbers were estimated by other models

In addition to CRMA v2, two external methods were evaluated in this paper. The first is Affymetrix' *CN5* method (Affymetrix Inc., 2008), and the second is implemented in the dChip software (Li and Wong, 2001).

1.1 CN5

The CN5 method is implemented in the 'apt-copynumber-workflow' software part of the Affymetrix Power Tools (APT) v1.10.0. The Affymetrix Genotyping Console (GTC) v3.0 (build 3.0.3083.25494) software (Affymetrix Inc., 2008) utilizes APT for CN5 estimates. We choose to run GTC, because it is not fully documented what settings should be used for APT. According to Affymetrix both approaches produce identical results (Affymetrix Scientific Community Forums, Thread: 'copy number: Genotyping Console 3.0 vs. apt 1.10.0?' on August 15, 2008). In CN5, probe signals are normalized ('adapter-type background correction') for systematic variation due to so called *enzyme recognition-sequence class.* Next, all probe signals (excluding control probes) are quantile normalized using the Affymetrix 'sketch' algorithm. For SNPs, chip effects $\{(\theta_{Aij}, \theta_{Bij})\}$ (as in the log-additive model of RMA) are estimated separately for the two alleles using the Probe Logarithmic Intensity Error (PLIER) algorithm. The total CNs are obtained by summing $\theta_{ij} = \theta_{Aij} + \theta_{Bij}$. Log ratios are calculated as in Eqn (15) [in the CRMA v2 manuscript], where the reference is $\theta_{Rj} = \text{median}_i\{\theta_{ij}\}$ with the important difference that for ChrX (ChrY) it is only samples that empirically are found to females (males) that are included. Finally, the raw CNs (log-ratios) are shifted such that the median of all median autosomal signals is zero. (Affymetrix Inc., 2008) There are some *limitations/restrictions* in CN5 worth knowing about:

1. The CN5 method is available only for GWS6. Affymetrix explicitly says that neither GTC nor APT implements CN5 for GWS5.

2. The CN5 method is limited to the default GWS6 CDF, that is, it cannot be used with the full GWS6 CDF.

3. The CN5 method use only females (males) when calculating reference on ChrX (ChrY). In the current implementation of GTC is not possible to force CN5 to estimate raw CN ratios on ChrX (ChrY) using all samples.

4. The GTC software does not export $\{\theta_{ij}\}$ but only log-ratio CNs.

It is because of the latter two restrictions we choose to calculate the CRMA v2 and dChip estimates on ChrX and ChrY the same way as in CN5. This is the only way a comparison of methods can be done.

1.2 dChip

For the dChip model, we used the *dChip 2008* (Build: July 10, 2008, http://www.dchip.org/). Probe-level data was normalized using the *invariant-set method* (Li and Wong, 2001), and PM signals were background corrected by '5th percentile of region (PM-only)'. By default, dChip suggests to use the array which has the median median (sic!) probe signal as the baseline array for normalization. We chose to follow this suggest after verifying that the spatial intensity plot of this array was not abnormal. For the HapMap data set, the baseline array was 'NA12750'. For probe summarization, the dChip multiplicative model was used, with $\text{PM} = \text{PM}_\text{A} + \text{PM}_\text{B}$ for SNPs ("Compute signals separately for A and B allele" unchecked), returning MBEI scores (corresponding to $\{\theta_{ij}\}$). For maximal comparison, the MBEI scores were imported to *aroma.affymetrix* and raw CNs where calculated as in Eqn (15) [of the CRMA v2 manuscript].

1.3 dChip*

Due to the odd performance of dChip for SNPs, we also ran the analysis where the MBEI probe summarization was replaced by averaging the signals while keeping everything else the same. We denote this variant of the dChip method by adding an asterisk to the label.

2 Summary of CN methods and their supported chip types

	CRMA (v1)	CRMA v2	dChip	CNAG	CN4	CN5
Mapping10K_Xba131	yes	yes	yes	-	-	-
Mapping10K_Xba142	yes	yes	yes	-	-	-
Mapping50K_Hind240	yes	yes	yes	yes	yes	-
Mapping50K_Xba240	yes	yes	yes	yes	yes	-
Mapping250K_Nsp	yes	yes	yes	yes	yes	-
Mapping250K_Sty	yes	yes	yes	yes	yes	-
GenomeWideSNP_5 (default)	-	yes	yes	-	-	-
GenomeWideSNP_5 (full)	-	yes	yes	-	-	-
GenomeWideSNP_6 (default)	-	yes	yes	-	-	yes
GenomeWideSNP_6 (full)	-	yes	yes	-	-	-
Custom SNP & CN chip types	yes	yes	?	-	?	?

Table 1: Summary of methods that estimate raw CNs for the different Affymetrix SNP & CN chip types.

3 Methods for the evaluation

We base all the performance assessments using relative copy numbers (chip effects) on the non-logarithmic scale, that is, $C_{ij} = 2 \cdot \theta_{ij}/\theta_{Rj}$. This is contrary to Bengtsson *et al.* (2008), where we used log-ratios $M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$. We use ChrX and ChrY loci for the evaluation. See Table 8 in Supplementary Notes #1 for how many loci there are on each chromosome. Loci in pseudo-autosomal regions (PARs) are excluded. Each of the two sex-chromosomes have two PARs (Blaschke and Rappold, 2006). See Table 2 for details. In addition to excluding PARs, regions known to be CN polymorphic (Redon *et al.*, 2006) are excluded. There are 48 such regions on ChrX and and 7 on ChrY. We use a safety margin of 100kb on each side. For further details on the evaluation methods are available in Bengtsson *et al.* (2008).

chromosome	PAR 1	PAR 2
X	1-2,692,881	154,494,747-154,824,264
Y	1-2,692,881	57,372,174-57,701,691

Table 2: Pseudo-autosomal regions on ChrX and ChrY according to Blaschke and Rappold (2006). The regions are specified as base positions where the first position of the chromosome is index one.

References

Affymetrix Inc. (2008). *Affymetrix Genotyping Console 3.0 - User Manual*. Affymetrix Inc.

Bengtsson, H., Irizarry, R. A., Carvalho, B., and Speed, T. P. (2008). Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics*, **24**(6), 759–767.

Blaschke, R. J. and Rappold, G. (2006). The pseudoautosomal regions, shox and disease. *Curr Opin Genet Dev*, **16**(3), 233–239.

Li, C. and Wong, W. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA*, **98**(1), 31–6.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., and ... (2006). Global variation in copy number in the human genome. *Nature*, **444**, 444–454.