Supplementary materials for CRMA v2

Henrik Bengtsson et al.

November 3, 2008

Contents

1 Annotation data files

Although the features on the arrays never change, their annotations might get updated as the Human Genome databases get updated. In this study, we have used the latest annotation data available on the Affymetrix website as of August 2008. We will list all annotation files needed by CRMA v2 for analyzing data for the GenomeWideSNP_5 (GWS5) and the GenomeWideSNP_6 (GWS6) chip types. First, we specify all relevant files available from Affymetrix (Tables 1 & 2). Then we specify the files compiled from the Affymetrix files for use in aroma.affymetrix (Tables 3 & 4).

1.1 Affymetrix annotation data

File name	GenomeWideSNP_5,r2.cdf
File size	239,670,727 bytes
MD5 checksum	e062cb42a392dd74721d5b7cf9a286f1
Notes	Affymetrix default chip definition file (CDF).
File name	GenomeWideSNP_5,Full,r2.cdf
File size	261,578,689 bytes
MD5 checksum	79f7a8353b4978dedbeff05a7897ff6e
Notes	Affymetrix full chip definition file (CDF).
File name	GenomeWideSNP_5.CN_probe_tab
File size	43,252,969 bytes
MD5 checksum	e661544a27163d1242ca690a486cb6b5
Notes	Affymetrix file probe sequence file for CN probes.
File name	GenomeWideSNP_5.probe_tab
File size	233,706,497 bytes
MD5 checksum	69b66720591fde9333acf5eb4d1b3e68
Notes	Affymetrix file probe sequence file for SNPs.
File name	GenomeWideSNP_5.cn.na26.annot.csv
File size	180,849,309 bytes
MD5 checksum	ec4f6cb4b482923d73d07f1b07faefe4
Notes	Affymetrix NetAffx v26 annotation file for CN probes.
File name	GenomeWideSNP_5.na26.annot.csv
File size	755,337,946 bytes
MD5 checksum	af59235b6fccada7f871257149a89215
Notes	Affymetrix NetAffx v26 annotation file for SNPs.

Table 1: Details of all Affymetrix specific GenomeWideSNP_5 annotation files used in the study.

File name	GenomeWideSNP_6.cdf
File size	484,489,553 bytes
MD5 checksum	223f3cd9141404b2a926a40cf47d6f1a
Notes	Affymetrix default chip definition file (CDF).
File name	GenomeWideSNP_6.Full.cdf
File size	493,291,745 bytes
MD5 checksum	3fbe0f6e7c8a346105238a3f3d10d4ec
Notes	Affymetrix full chip definition file (CDF).
File name	GenomeWideSNP_6.CN_probe_tab
File size	96,968,290 bytes
MD5 checksum	3dc2d3178f5eafdbea9c8b6eca88a89c
Notes	Affymetrix file probe sequence file for CN probes.
File name	GenomeWideSNP_6.probe_tab
File size	341,479,928 bytes
MD5 checksum	2037c033c09fd8f7c06bd042a77aef15
Notes	Affymetrix file probe sequence file for SNPs.
File name	GenomeWideSNP_6.cn.na26.annot.csv
File size	482,222,873 bytes
MD5 checksum	948eb406774aa5097590debd0d667a22
Notes	Affymetrix NetAffx v26 annotation file for CN probes.
File name	GenomeWideSNP_6.na26.annot.csv
File size	1,628,608,540 bytes
MD5 checksum	323f9afa0c180c146260b5eb689d0bd2
Notes	Affymetrix NetAffx v26 annotation file for SNPs.

Table 2: Details of all Affymetrix specific GenomeWideSNP_6 annotation files used in the study.

1.2 Imported annotation data

Several of Affymetrix annotation files use file formats that are not intended to be used directly in a computational system, but rather be imported once. The *aroma.affymetrix* framework utilizes its own binary files that are more compact and faster to access. The contents of these are imported from the above Affymetrix annotation data files. Further details on the source files used for each compiled file is given in the file footer of each file. To download these and for further details, see the aroma.affymetrix webpage.

File name	GenomeWideSNP_5,HB20080710.acs
File size	121,981,027 bytes
MD5 checksum	bd0da64b09ea164082e066798774a3c5
Notes	Aroma.affymetrix probe sequence data from above Affymetrix sequence files.
File name	GenomeWideSNP_5,Full,r2,na26,HB20080822.ufl
File size	3,684,511 bytes
MD5 checksum	a533ac3ba64f36902d13bed8fe2a9a5b
Notes	Aroma.affymetrix fragment length data mapping to the full CDF. Compiled from above Affymetrix NetAffx files.
File name	GenomeWideSNP_5,Full,r2,na26,HB20080822.ugp
File size	4,605,439 bytes
MD5 checksum	9827c9fad08144d6e590b87984350a26
Notes	Aroma.affymetrix genome location data mapping to the full CDF. Compiled from above Affymetrix NetAffx files.
File name	GenomeWideSNP_5,r2,na26,HB20080822.ufl
File size	3,445,230 bytes
MD5 checksum	566812ef309486baf0a8496bb8a35ef5
Notes	Aroma.affymetrix fragment length data mapping to the default CDF. Compiled from above Affymetrix NetAffx files.
File name	GenomeWideSNP_5,r2,na26,HB20080822.ugp
File size	4,306,339 bytes
MD5 checksum	9330fdcb30b3a3b4c13cbdeb8de4ffb5
Notes	Aroma.affymetrix genome location data mapping to the default CDF. Compiled from above Affymetrix NetAffx files.

Table 3: Details of all aroma.affymetrix specific GenomeWideSNP_5 annotation files used in the study.

File name	GenomeWideSNP_6,HB20080710.acs
File size	179,217,531 bytes
MD5 checksum	f04f081e0a1900653d957a8f320744c0
Notes	Aroma.affymetrix probe sequence data from above Affymetrix sequence files.
File name	GenomeWideSNP_6,Full,na26,HB20080722.ufl
File size	7,526,454 bytes
MD5 checksum	6f11e9bd3a7a0cb060d5fcf671b0776a
Notes	Aroma.affymetrix fragment length data mapping to the full CDF. Compiled from above Affymetrix NetAffx files.
File name	GenomeWideSNP_6,Full,na26,HB20080821.ugp
File size	9,407,937 bytes
MD5 checksum	5a7bef30a458cb238ae2167aa41f5bd6
Notes	Aroma.affymetrix genome location data mapping to the full CDF. Compiled from above Affymetrix NetAffx files.
File name	GenomeWideSNP_6,na26,HB20080821.ufl
File size	7,425,058 bytes
MD5 checksum	522b89d875f39832f5423e78cffba8c8
Notes	Aroma.affymetrix fragment length data mapping to the full CDF. Compiled from above Affymetrix NetAffx files.
File name	GenomeWideSNP_6,na26,HB20080821.ugp
File size	9,281,127 bytes
MD5 checksum	ad63ef009b44f1274f6e2a35cb951dbc
Notes	Aroma.affymetrix genome location data mapping to the default CDF. Compiled from above Affymetrix NetAffx files.

Table 4: Details of all aroma.affymetrix specific GenomeWideSNP_6 annotation files used in the study.

2 Summary of annotation data

This section provides a summary of the relevant annotation data available for the GWS5 and the GWS6 chip types. We distinguish between the annotation data available in the Chip Definition Files (CDFs) and the NetAffx files, because the former are less likely to change over time whereas the latter gets updates when the genome annotations get updated. Affymetrix' NetAffx annotation data are updated several times a year.

2.1 Affymetrix CDF data

Affymetrix provides one "default" and one "full" CDF for each of the GWS5 and GWS6 chip types. See Table 2 for these files. The default CDF contains a subset of the full CDF where certain SNP units have been filtered out due to homology to other regions and poor performance (private communication with Affymetrix). Tables 5 & 6 provides a summary of these CDFs. All summaries presented here and in the main paper refer to the full CDFs, unless stated otherwise.

2.1.1 Unit annotations

	GWS5 default	GWS5 full	GWS6 default	GWS6 full
UNIT TYPES				
CN units	417,269	417,269	945,826	945,826
SNPs	440,794	500,568	906,600	931,946
Subtotal	858,063	917,847	1,852,426	1,877,772
AFFX-SNPs	3,022	3,022	3,022	3,022
Subtotal	861,085	920,859	1,855,448	1,880,794
Others	24	69	621	621
Total	861,109	920,928	1,856,069	1,881,415
EXCLUDED FROM FULL				
CN units	0	-	0	-
SNPs	59,744	-	25,346	-
AFFX-SNPs	0	-	0	-
Others	45	-	0	-
CN UNIT STRANDNESS				
Sense only	0	0	0	0
Antisense only	417,269	417,269	945,826	945,826
Opposite strands	0	0	0	0
Both strands	0	0	0	0
SNP STRANDNESS				
Sense only	234,449	260,266	477,538	491,830
Antisense only	174,549	194,126	429,062	440,116
Opposite strands	31,796	46,176	0	0
Both strands	0	0	0	0
AFFX-SNP STRANDNESS				
Sense only	145	145	145	145
Antisense only	129	129	129	129
Opposite strands	0	0	0	0
Both strands	2,748	2,748	2,748	2,748
SNP ALIGNMENT				
Aligned allele pairs	285,984	308,169	906,600	931,946
Non-aligned allele pairs	154,810	192,399	0	0
AFFX-SNP ALIGNMENT				
Aligned allele pairs	3,022	3,022	3,022	3,022
Non-aligned allele pairs	0	0	0	0

Table 5: Summary of the unit annotation available in the CDF files. All values are in counts.

2.1.2 Probe annotations

	GWS5 default	GWS5 full	GWS6 default	GWS6 full
PROBES				
CN probes	417,269	417,269	945,826	945,826
SNP probes	3,526,352	4,004,544	5,660,710	5,833,210
Subtotal	3,943,621	4,421,813	6,606,536	6,779,036
AFFX-SNP probes	81,504	81,504	81,504	81,504
Subtotal	4,025,125	4,503,317	6,688,040	6,860,540
Other probes	178,055	188,239	32,420	32,420
Excluded from full	488,376	0	172,500	0
Total	4,691,556	4,691,556	6,892,960	6,892,960
EXCLUDED FROM FULL				
CN probes	0	-	0	-
SNP probes	478,192	-	172,500	-
AFFX-SNP probes	0	-	0	-
Other probes	10,184	-	0	-
CN UNIT PROBES				
CN units with 1 probe	417,269	417,269	945,826	945,826
SNP PROBE PAIRS				
SNPs with 3 pairs	0	0	796,045	811,179
SNPs with 4 pairs	440,794	500,568	110,555	120,767
AFFX-SNP PROBE PAIRS				
SNPs with 12 pairs	2,461	2,461	2,461	2,461
SNPs with 20 pairs	561	561	561	561

Table 6: Summary on the probe annotation available in the CDF files. All values are in counts.

2.2 Affymetrix NetAffx data

Affymetrix makes so called NetAffx CSV files available for download. These files contain annotation data exported from their NetAffx data base. For each chip type there exists one or more NetAffx CSV files, e.g. for the GWS chip types there is one for the CN units and one for all other units on the chip. See Tables 1 & 2 for which these files are. The NetAffx data base is updated frequently and the following information is likely to get slightly outdated over time.

2.2.1 Genome positions

In Table 7 we summarize how many SNPs and CN loci have known annotations according to NetAffx. The genome location per chromosome is summarized in Table 8 for each of the default and the full CDF.

We do not know why the location is unknown for some loci, but from our investigation we believe it is mainly because such loci map to multiple positions in the genome. For instance, in NetAffx release 26, there is no genomic location reported for SNP_A-4228947 (on GWS5, GWS6 and Mapping250K_Nsp), and according to the NCBI SNP data base it (rs11261805) maps to the two locations 41,240,208 and 43,493,496 on Chr9.

Furthermore, for GWS5 the NetAffx annotation files currently available only contain data on the units in the default CDF. This is confirmed by comparing the names of the units with known locations in the default and the full GWS5.

	GWS5		GWS6	
	default	full*	default	full
	#loci	#loci	#loci	#loci
SNPs with known locations	440,094	440,094	905,386	929,967
SNPs with unknown locations	700	60,474	1,214	1,979
CN probes with known locations	312,384	312,384	945,806	945,806
CN probes with unknown locations	104,885	104,885	20	20
AFFX-SNPs with known locations	3,012	3,012	3,012	3,012
AFFX-SNPs with unknown locations	10	10	10	10
Total with known locations	**755,490**	**755,490**	**1,854,204**	**1,878,785**
Total with unknown locations	165,369	92,121	1,244	2,009
Total	861,085	920,859	1,855,448	1,880,794

Table 7: Summary of genomic location data that is available in the NetAffx files of contents in GWS5 and GWS6 with respect to unit and probe class and availability of annotation data. The effective sets of units available for CN analysis are emphasized in bold. *See text for why the default and the full GWS5 are identical. NetAffx v26 was used for this summary.

		GWS5		GWS6	
		default	full*	default	full
chromosome	seq. length (Mbs)	#loci	#loci	#loci	#loci
1	245.2	58,548	58,548	144,499	146,401
2	243.3	63,380	63,380	151,902	153,663
3	199.4	53,120	53,120	126,337	127,766
4	191.6	50,594	50,594	118,933	120,296
5	181.0	48,661	48,661	114,333	115,672
6	170.7	47,329	47,329	111,440	112,825
7	158.4	39,325	39,325	99,818	100,996
8	145.9	41,134	41,134	97,040	98,277
9	134.5	31,563	31,563	81,036	82,168
10	135.5	39,140	39,140	92,331	93,592
11	135.0	37,948	37,948	88,295	89,525
12	133.5	36,422	36,422	86,209	87,321
13	114.2	28,119	28,119	65,310	66,067
14	105.3	23,621	23,621	56,339	57,103
15	100.1	20,968	20,968	52,810	53,556
16	90.0	20,718	20,718	53,329	54,182
17	81.7	17,411	17,411	46,024	46,632
18	77.8	21,870	21,870	51,510	52,093
19	63.8	10,631	10,631	29,855	30,299
20	63.6	17,911	17,911	43,052	43,628
21	47.0	10,412	10,412	24,787	25,111
22	49.5	9,346	9,346	24,000	24,484
X	152.6	26,373	26,373	86,064	87,198
Y	51.0	946	946	8,841	9,485
Mitochondrial	16.6kb	-	-	110	445
total		755,490	755,490	1,854,204	1,878,785

Table 8: Distribution of loci by chromosome for the different CDFs of GWS5 and GWS6. *See text for why the default and the full GWS5 are identical. NetAffx v26 was used for this summary.

2.2.2 Restriction enzymes and PCR fragment lengths

	GWS5		GWS6	
	default #loci	full* #loci	default #loci	full #loci
SNPs on NspI only	116,979	116,979	240,001	246,080
SNPs on StyI only	74,135	74,135	154,884	160,899
SNPs on both	248,980	248,980	510,330	522,472
SNPs with known lengths	440,094	440,094	905,215	929,451
SNPs with unknown lengths	700	60,474	1,385	2,495
CN probes on NspI only	140,099	140,099	451,191	451,191
CN probes on StyI only	1,208	1,208	0	0
CN probes on both	171,077	171,077	494,615	494,615
CN probes with known lengths	312,384	312,384	945,806	945,806
CN probes with unknown lengths	104,885	104,885	20	20
AFFX-SNPs with known lengths	0	0	0	0
AFFX-SNPs with unknown lengths	3,022	3,022	3,022	3,022
Total with known lengths	752,478	752,478	1,851,021	1,875,257
Total with unknown lengths	108,607	168,381	4,427	5,537
Total	861,085	920,859	1,855,448	1,880,794

Table 9: Summary of fragment-length data that is available in the NetAffx files of contents in GWS5 and GWS6 with respect to unit and probe class and availability of annotation data. *See text for why the default and the full GWS5 are identical. NetAffx v26 was used for this summary.

3 Multi-enzyme digestion

For the 100K as well as the 500K SNP-only assays, DNA is prepared in two parallel processes, each digesting the DNA using a unique restriction enzyme, amplifying the fragments by PCR, and hybridizing the products to separate arrays. In the GWS assays, which like 500K use enzymes *Nsp*I and *Sty*I, the two mixes of PCR products are no longer hybridized to separate arrays but instead hybridized in aliquot to the same array (Affymetrix Inc., 2007a,b). Consequently, SNP target DNA of PCR products originating from different digestions may hybridize to the same probe, which is something that has to be taken into account when, for instance, fitting the fragment-length normalization. For CN probes the situation is somewhat different. Affymetrix selected the CN probes from a large pool of CN probes based on their performance on copy numbers (private communication). This pilot study was conducted on a specially designed in-house chip set containing probes that are known to be on an *Nsp*I fragment. For this reason, some of the selected probes are exclusively on *Nsp*I fragments, some are by chance both on *Nsp*I and *Sty*I fragments, but none are exclusively on *Sty*I fragments. Note, when annotation for the human genome is updated, some of the probes might by chance be re-annotated to become *Sty*I-only probes. We have found that it is important that the preprocessing models these differences, otherwise there is a substantial risk for getting systematic biases between SNPs and CN probes due to enzymatic mixing imbalances. See Table 9 for details on fragment-length information for the two chip types and the two enzymes.

4 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).

4.1 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 plier algorithm. The total CNs are obtained by summing $\theta_{ij} = \theta_{Aij} + \theta_{Bij}$. Log ratios are calculated as in Eqn (15), 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.

4.2 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)'. For GWS6, array 'NA12750' was suggested by dChip to be used as the baseline array for normalization, because it had the median median (sic!) probe signal. As suggested, we verified that the spatial intensity plot of this array was not abnormal. 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).

4.3 dChip*

Due to 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 flavor of the dChip method by adding an asterisk to the label.

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

5 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 for how many loci there are on each chromosome. Loci in pseudo-autosomal regions (PARs) are excluded. Each of the two sex-chromosomes have to PARs (Blaschke and Rappold, 2006). See Table 11 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 11: 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. (2007a). *Genome-Wide Human SNP Nsp/Sty 6.0 User Guide*. Affymetrix Inc. Rev 1.

Affymetrix Inc. (2007b). *Genome-Wide Human SNP Nsp/Sty Assay 5.0*. Affymetrix Inc. Rev 2.

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.