CN.FARMS: a probabilistic model to detect DNA copy numbers

Djork-Arné Clevert[1,3], Andreas Mitterecker[1], Andreas Mayr[1], Robert Burger[1], An De Bondt[2], Marianne Tuefferd[2], Willem Talloen[2], Hinrich Göhlmann[2], and Sepp Hochreiter[1]

[1]Institute of Bioinformatics, Johannes Kepler University Linz 4040 Linz, Austria; [2]Johnson & Johnson Pharmaceutical Research & Development, Janssen Pharmaceutica n.v., Beerse, Belgium; [3]Department of Nephrology and Internal Intensive Care, Charité University Medicine, Berlin, Germany

MOTIVATION: Existing pre-processing methods for DNA microarrays designed to detect copy number variations (CNVs) lead to high false discovery rates (FDRs). High FDRs misguide researchers especially in the medical context where CNVs are wrongly associated with diseases. We propose a probabilistic latent variable model, cn.FARMS, for array-based CNV analysis which controls the FDR without loss of sensitivity. At a DNA region, cn.FARMS constructs a model by a Bayesian maximum a posteriori estimation where the unobserved, latent variable represents the copy number that is measured by observed genetic markers (probes). The latent variable's prior prefers parameters which represent the null hypothesis, (same copy number for all samples), from which the posterior can only deviate by a high information content in the data. The more probes agree on the region's copy number, the less is the uncertainty about the latent variable's value, the higher is the information content.

RESULTS: We compared CN.FARMS on a HapMap Mapping250K_Nsp and SNP6.0 benchmark data set to CRMAv2 and dChip. The comparison is based on the sex determination based on the data from the X chromosome, where males possess one copy and females two. The ROC curve serves to compare the FDR for different true positive rates. In both experiments cn.FARMS yielded the best classification results.

AVAILABILITY: This approach is publicly available in **R** at http://www.bioinf.jku.at/software.

CN-ANALYSIS AS A THREE-STEP PIPELINE

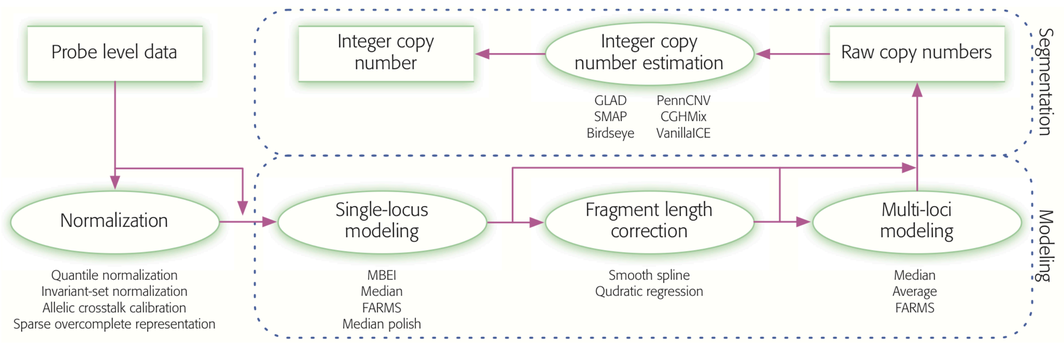

FIGURE 1: Copy number analysis for (Affymetrix) DNA genotyping arrays as a three-step pipeline: (1) Normalization, (2) Modeling, and (3) Segmentation. Modeling is divided into "single-locus modeling" and "multi-loci modeling" with "fragment length correction" as an optional intermediate step.

THE MODEL

Our approach is based on a linear model with Gaussian noise. Denote the actually observed sum of allele A and B to zero mean normalized and log-transformed PMs by $\boldsymbol{x}$ and the copy number variation in the hybridization mixture by z. Then we assume that the log-observations $\boldsymbol{x}$ depend on the true copy number variation z via

$$\boldsymbol{x} \;=\; \boldsymbol{\lambda} z \;+\; \boldsymbol{\epsilon}\;, \tag{1}$$

where

$$\boldsymbol{x}, \boldsymbol{\lambda} \in \mathbb{R}^n \;,\; z \sim \mathcal{N}(0,1) \;,\; \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})\,. \tag{2}$$

In equation (1), z models the latent factor in the data $\boldsymbol{x}$, while $\boldsymbol{\epsilon}$ models the independent noise in each probe of each array. According to the model, the observation vector $\boldsymbol{x}$ is Gaussian distributed:

$$\boldsymbol{x} \sim \mathcal{N}\left(\mathbf{0} \;,\; \boldsymbol{\lambda}\boldsymbol{\lambda}^T \;+\; \boldsymbol{\Psi}\right) \tag{3}$$

Model parameters were estimated with the expectation-maximization (EM) algorithm, modified to maximize the Bayesian posterior

$$p(\boldsymbol{\lambda}, \boldsymbol{\Psi} \mid \{\boldsymbol{x}\}) \;\propto\; p(\{\boldsymbol{x}\} \mid \boldsymbol{\lambda}, \boldsymbol{\Psi})\; p(\boldsymbol{\lambda}, \boldsymbol{\Psi}). \tag{4}$$

SPARSE OVERCOMPLETE REPRESENTATION

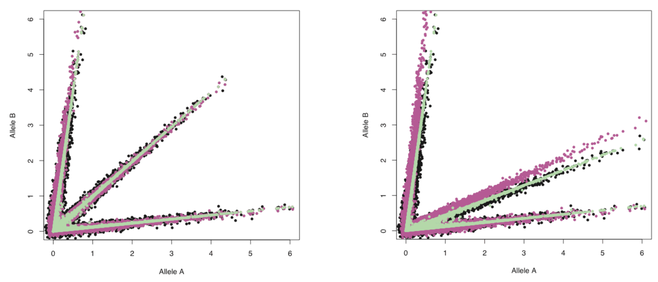

FIGURE 2: The ACC model (Bengtsson et al, Bioinformatics, 2008) assumes symmetric crosstalk (black dots) and therefore expects the AB genotypes on the diagonal between the AA (horizontal) and BB (vertical) clusters. Therefore ACC (magenta dots) systematically overestimates the BB and AB clusters if the AB cloud is not diagonal between AA and BB clusters, whilst the SPARSE OVERCOMPLETE REPRESENTATION (green dots) fits the discrete clusters much better.

Allelic crosstalk correction is done by SPARSE OVERCOMPLETE REPRESENTATION. The observed data can be described using an overcomplete basis by (1). Sparseness on the latent coefficients is forced by a Laplacian distributed prior. The likelihood of an observation x is given by:

$$p(\boldsymbol{x}|\boldsymbol{\lambda}, \boldsymbol{\Psi}) = \int p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\lambda}, \boldsymbol{\Psi}) p(\boldsymbol{z}) d\boldsymbol{z}$$

With the factorable Laplacian:

$$p(\boldsymbol{z}) = (2)^{-\frac{L}{2}} \prod_{l=1}^{L} \exp(-\sqrt{2}|z_l|)$$

META-PROBE SETS

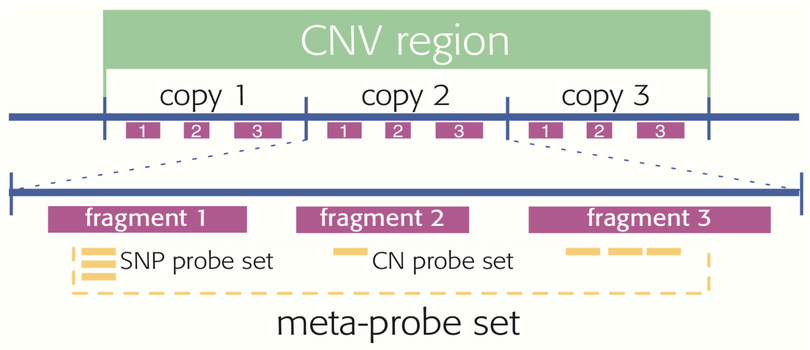

FIGURE 3: The concept of "multi-loci modeling" and meta-probes. Lower panel: The probes which target a fragment (often at a SNP position) are summarized to a raw copy number of this fragment. Upper panel: The raw fragment copy numbers are in turn the meta-probes for a DNA region. Meta-probes are summarized to a raw region copy number

DATASETS

We compared cn.FARMS on a HapMap Mapping250K_Nsp and SNP6.0 benchmark data set to CRMA and dChip. The aim described in this poster is to distinguish males from females based on the X chromosome copy numbers, where males possess one copy and females two. The benchmark data is publicly available at http://ftp.hapmap.org

RESULTS

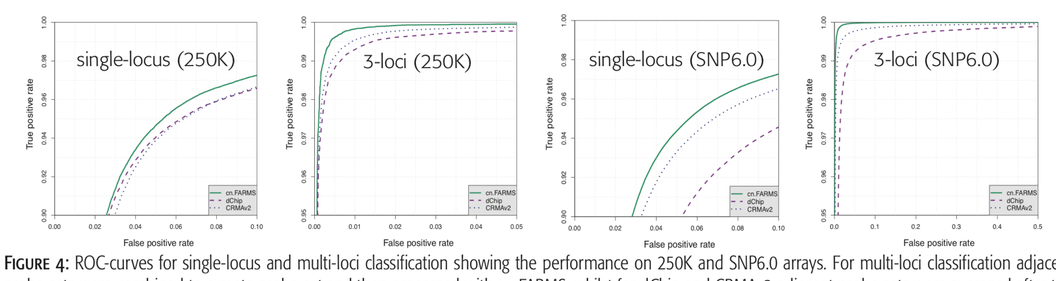

FIGURE 4: ROC-curves for single-locus and multi-loci classification showing the performance on 250K and SNP6.0 arrays. For multi-loci classification adjacent probe sets were combined to a meta-probe set and then processed with CN.FARMS, whilst for dChip and CRMAv2 adjacent probe sets were averaged after the summarization.

TABLE 1 60 Affymetrix 250K_NSP and SNP6.0 HapMap CEU founders arrays were used for the assessment as proposed by (Bengtsson et al, Bioinformatics, 2008). We compared CN.FARMS to CRMAv2 and dChip. The performance is assessed by how well males can be distinguished from females based on the X chromosome signal intensities, where males possess one copy and females two. To estimate the ROCs on ChrX, we excluded the odd-performing female NA12145 and all loci, which refer to copy number - or pseudoautosomal regions on the X chromosome.

ARRAY TYPE	AFFYMETRIX 250K			AFFYMETRIX SNP6.0		
METHOD	CN.FARMS	DCHIP	CRMAV2	CN.FARMS	DCHIP	CRMAV2
ROC-AUC single-locus	0.9852	0.9818	0.9820	0.9838	0.9721	0.9807
ROC-AUC 2-loci	0.9983	0.9969	0.9974	0.9983	0.9894	0.9963
ROC-AUC 3-loci	0.9998	0.9995	0.9992	0.9998	0.9990	0.9953
ROC-AUC 4-loci	1.0000	0.9998	0.9999	0.9999	0.9976	0.9995
COMPUTATIONAL COSTS [S]	1,055	2,493	3,363	3,657	11,850	6,210

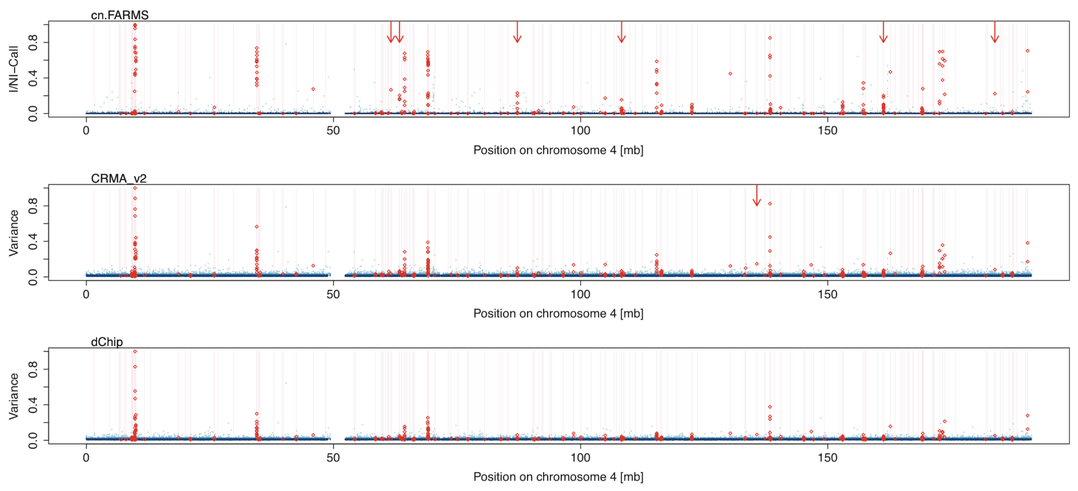

FIGURE 5: CNV calling plots across chromosome 4 for 3-loci regions. The y-axis gives the I/NI-call ($\text{INI} = (1+\lambda^T\Psi^{-1}\lambda)^{-1}$) estimated by cn.FARMS and for both CRMA v2 and dChip the variance. Calling values are scaled such that the maximum is one. Local calling densities are encode by blue color shades. True CNVs (reported in Conrad et al. (2010)) are marked as light-rose bars and calls at these loci by red circles. A perfect calling method maximally calls all true CNVs (red circles at 1) and does not call others (dark-blue background at 0). cn.FARMS better separates called true CNVs (true positives) from true negatives (less variance indicated by dark-blue density the bottom). The red arrows, e.g. at positions 65 or 85mb in the upper cn.FARMS panel, indicate verified CNVs which were detected by one method, in this case cn.FARMS, but not by both others. cn.FARMS identifies true CNVs with a lower FDR than CRMA v2 and dChip.

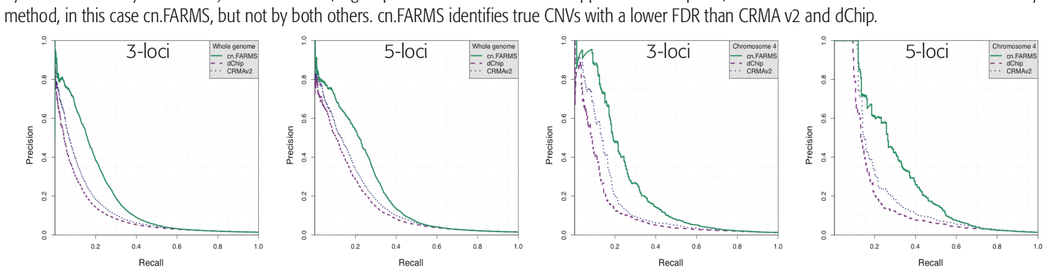

FIGURE 6: Precision-recall curves (PRCs) on HapMap SNP 6.0 arrays for cn.FARMS, CRMA v2, and dChip at detecting previously multiple confirmed CNVs reported in Conrad et al. (2010). cn.FARMS detection criteria is the I/NI call whereas CRMA v2, and dChip use the variance of raw copy numbers. A PRC more in the upper-right-hand corner performs better. Note, that precision is (1-FDR) thus the FDR is the distance of the curve to the upper limit

TABLE 2: Area under the precision-recall curves on HapMap SNP 6.0 arrays for cn.FARMS, CRMAv2, and dChip at detecting previously multiple confirmed CNVs reported in Conrad et al. (2010). A larger value means that the method has lower FDR averaged over different given recall values.

	CHROMOSOME 1			CHROMOSOME 3			CHROMOSOME 4			WHOLE GENOME		
# of combined loci	3	5	7	3	5	7	3	5	7	3	5	7
CN.FARMS [PR-AUC]	0.20	0.23	0.25	0.27	0.34	0.39	0.25	0.31	0.34	0.20	0.24	0.26
CRMAV2 [PR-AUC]	0.16	0.19	0.23	0.16	0.23	0.29	0.16	0.22	0.26	0.13	0.18	0.21
DCHIP [PR-AUC]	0.14	0.19	0.22	0.13	0.20	0.25	0.12	0.19	0.21	0.11	0.16	0.19

REFERENCES

W. TALLOEN, D.-A. CLEVERT, S. HOCHREITER, D. AMARATUNGA, L. BIJNENS, S. KASS, H. GÖHLMANN; Calling probe sets informative or non-informative for the experiment: A highly effective gene filtering tool for microarray data. Bioinformatics 2007 23(21):2897-2902

S. HOCHREITER, D.-A. CLEVERT, K. OBERMAYER; A new summarization method for Affymetrix probe level data. Bioinformatics 2006, 22: 943-949.

H. BENGTSSON, R. IRIZARRY, B. CARVALHO, T.P. SPEED; Estimation and assessment of raw copy numbers at the single locus level, Bioinformatics, 2008. [pmid: 18204055] [doi: 10.1093/bioinformatics/btn016]

H. BENGTSSON, P. WIRAPATI, T.P. SPEED; A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6, Bioinformatics 2009, 10.1093/btp371

ACKNOWLEDGMENT

I acknowledge the support of the ISCB, Department of Energy, and National Science Foundation in the form of an International Travel Grant, which enabled me to attend this conference.