SNiPer-HD: improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays

Jianping Hua, David W. Craig, Marcel Brun, Jennifer Webster, Victoria Zismann, Waibhav Tembe, Keta Joshipura, Matthew J. Huentelman, Edward R. Dougherty, and Dietrich A. Stephan

Bioinformatics 2007 23: 57-63

Journal Club, 05/01/2007

Overview

- Developed for Affymetrix platform but should be usable for other genotyping plaforms as well
- Genome-Wide Association studies (GWA) are susceptible to low-quality genotyping
- Both per-SNP and per-call quality scores are presented to be used to screen results
- SniPer uses Expectation-maximization algorithm with training set to build model
- SNiPer-HD can be used together with other calling algorithms, such as BRLMM
- http://www.tgen.org/neurogenomics/data

Affymetrix is such a wonderful platform...

- Because the default calling algorithm (Dynamic Modeling - DM) used is such a low quality one...
- So bionformaticicists can earn their bread by improving it
- Another high-quality algorithm RLMM (Robust Linear Model with Mahalanobis distance classifier was presented previous year. The modified variant BRLMM is used for comparison.

GWA

- About 250 000 SNP-s are sufficient to cover caucasian or Asian populations
- GWA studies easily exceed billions of genotype calls
- Only handful of these billons are relevant for association study Others are neutral (good case) or noise (ugh!)
- For example, GWA was performed using Affymetrix 100K chip with 96 cases and 50 controls. Only 2 SNP-s survived bonferroni correction. One of these was the result of a genotyping error.

Klein,R.J. *et al*. (2005) Complement factor H polymorphism in age-related macular degeneration. *Science*, 308, 385-389

GWA II

- Affymetrix 500K chip with default dynamic modeling (DM) genotype calling
- Individuals were placed randomly into case and control groups (100 in each group)
- On average 6% of SNP-s were out of Hardy-Weinberg Equilibrium (HWE) ($p < 0,05$)
- If SNP-s were ranked by Fischr exact test P-value, 45% of top 100 SNP-s failed HWE.
- Permuting genotypes between classes retained these characteristics
- One probable cause may be systematic miscalling of heterozygotes at certain SNP-s

SNiPer-HD

- Genotyping as classification procedure
- Uses expectation-maximization (EM) clustering to estimate distribution parameters
- Provides 2 quality scores: quality index (per-SNP score) and confidence score (per-call score, similar to p-value)
- Uses vector of relative allele signals (RAS) as input (Affymetrix specific feature)

$$x_d = A_d/(A_d + B_d)$$

  - x_d - RAS value for given probe quartet d
  - A_d - Perfect match signal for allele A
  - B_d - perfect match signal for allele B

SNiPer-HD

- Mismatch signals are not used (so it can be easily applied to other genotyping platforms)
- SNP is represented by RAS vector

$$X_i = (x_1, x_2, \ldots x_D)$$

- One can expect these vectors to form 3 mass-concentrations in D-dimensional space (AA, AB, BB genotypes)
- It is assumed, that these RAS vectors are generated from a mixture of 3 Gaussian distributions
- If distribution is given, the genotype having the highest posterior probability is assigned according to Bayesian rule

Algorithm I

- 3 genotypes have prior probabilities of τ_0, τ_1, τ_2
- A SNP with genotype k has its RAS vector X generated by Gaussian distribution

$$f_k(X \mid \mu_k, \Sigma_k) = \frac{\exp\left[-\frac{1}{2}(X-\mu_k)^T \Sigma_k^{-1}(X-\mu_k)\right]}{(2\pi)^{D/2} \mid \Sigma_k \mid^{1/2}}$$

- μ_k - the mean vector
- Σ_k - covariance matrix
- τ_k, μ_k and Σ_k, $k = 1, 2, 3$ are to be estimated
- Covariance matrixes are assumed to be equal and spherical

$$\Sigma_0 = \Sigma_1 = \Sigma_2 = \lambda \boldsymbol{I}.$$

Algorithm II

- We start with z_{ik} = 1 if DM estimated genotype is k and 0 otherwise

$$\tau_k = n_k / N$$

$$\mu_k = \sum_{i=1}^{N} z_{ik} X_i / n_k$$

$$\lambda = tr(W)/ND,$$

$$n_k = \sum_{i=1}^{N} z_{ik}, \ W = \sum_{k=0}^{2} \sum_{i=1}^{N} z_{ik}(X_i - \mu_k)(X_i - \mu_k)^T$$

- new z_{ik} is calculated with the following formula

$$z_{ik} = \tau_k f_k(X_i \mid \mu_k, \Sigma_k) / \sum_{j=0}^{2} \tau_j f_j(X_i \mid \mu_j, \Sigma_j)$$

Implementation details

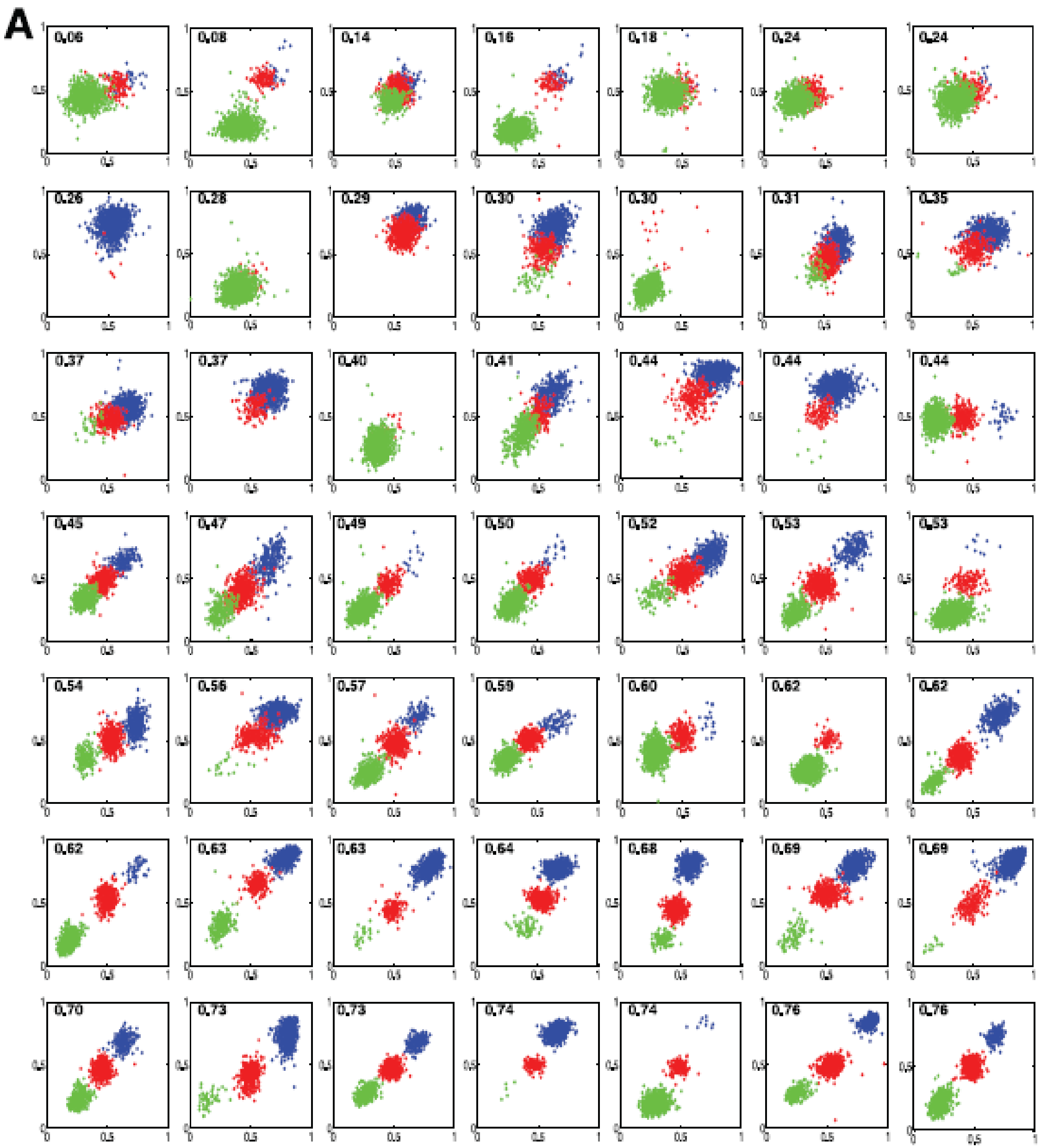
- The calculation is iterated, until the relative change in overall likelihood (not shown) is below treshold (0,001), or the maximum number of iterations (30) is reached.
- If tie occurs while evaluating genotypes, call is made randomly from the classes having maximum Z_{ik} - the resulting call has always confidennce <0,5
- The accuracy of estimated parameters is dependent on the quality of training samples. Call rate > 85% should be used, but resulting SNP quality index should indicate bad training samples anyways.
- Covariation matrixes Σ_k are assumed to be spherical and having equal volumes. This helps with low MAF SNPs.

Scores

- Confidence of a call is directly derived from the posterior probability of given signal vector
  $C = 1 - Z_{ik}$
- The quality index of a SNP is derived from the median silhouette width of given SNP (minimum silhouette width is used as parameter)

$$S(X_i) = \frac{\mathrm{b}(X_i) - \mathrm{a}(X_i)}{\max\left[\mathrm{b}(X_i), \mathrm{a}(X_i)\right]}$$

- $S(X_i)$ – silhouette width
- $a(X_i)$ – the average euclidean distance between Xi and all other sample points of gien genotype
- $b(X_i)$ – the minimum of the two average distances between Xi and the points of another genotype



Example of SNP signal plots, orderd by quality score (0,06 – 0,76)
X and Y are RAS values of diferent alleles

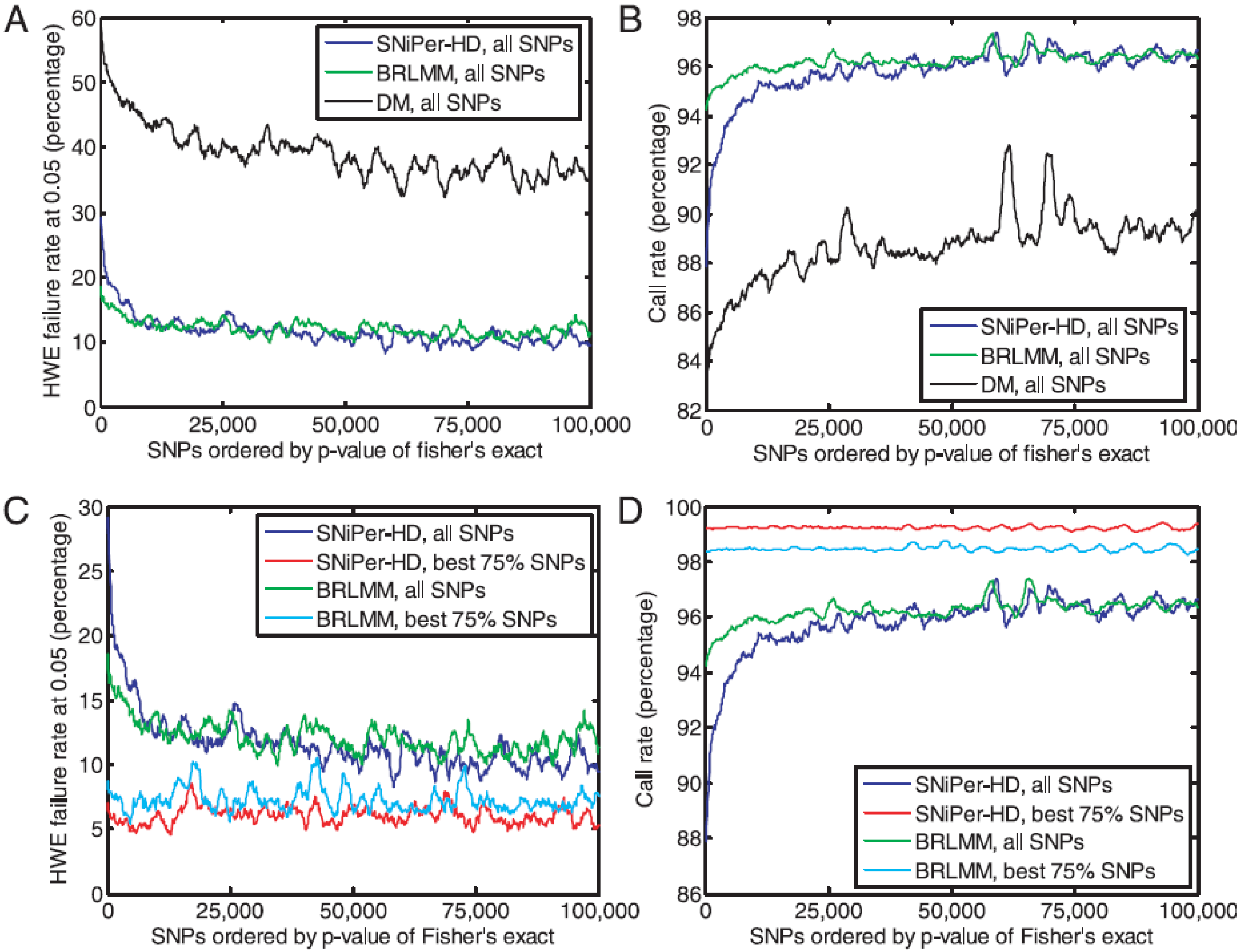
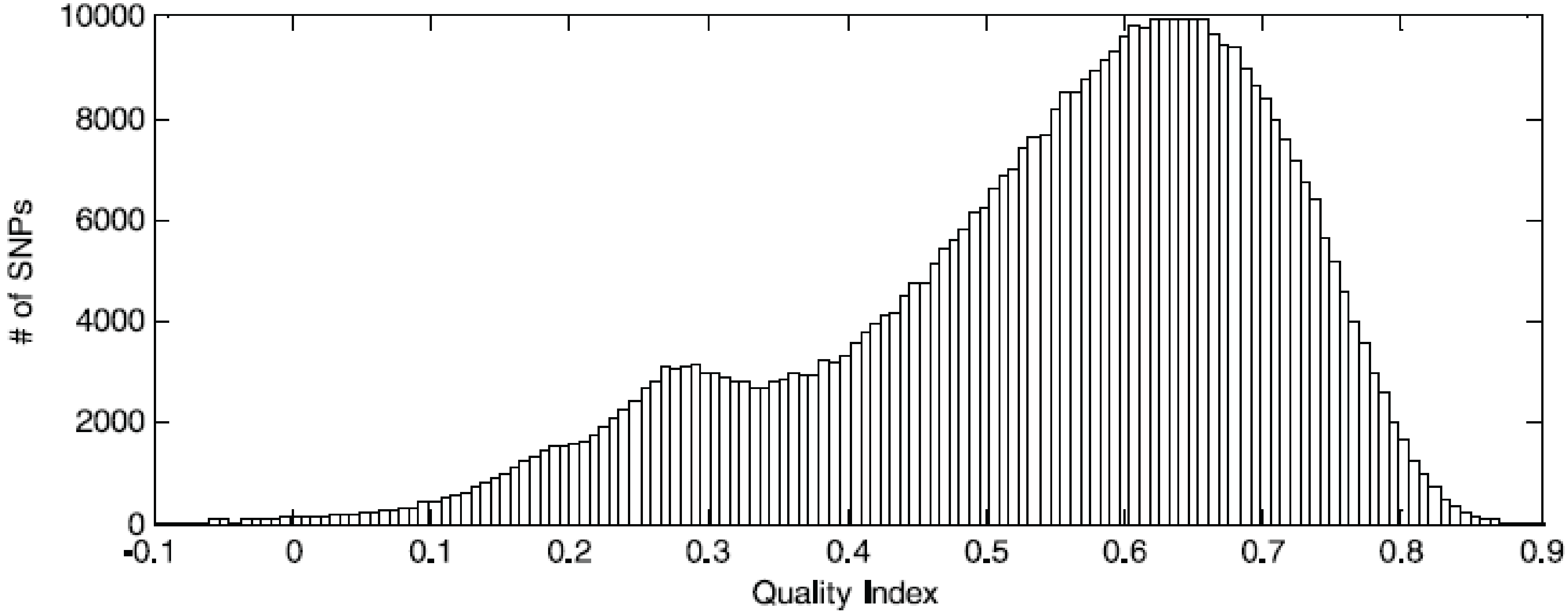


Fig. 2. HWE failure rate/call rate versus top SNPs ordered by P-value of Fisher's exact on DM, SNiPer-HD and BRLMM calls of all samples. Exact test ofHWEis used. The x-axis is the SNPs ordered by the P-value of Fisher's exact on a case-control study, and the y-axis is the percentage of SNPs that fail HWE at 0.05 for control samples in (A and C), and call rate on all individuals in (B and D). Default settings are used for DM and BRLMM to set 'NoCall'. For SNiPer-HD, 'NoCall' is set to any call with confidence score >0.05.



The histogram of SNP quality index (clipped from both ends)
Notice, that there is clear peak of low-quality SNP-s at about 0.28

Results

- About 900 individuals were tested with approximately 2/3 being cases and 1/3 being controls (some chips were left out because of low call rate).
- SNP-s with low p-value have clear tendency of HWE failure and low call rate
- Quality index of 0.45 was used to filter out badly behaving SNPs, preserving about 76% of total SNPs
- Good SNPs do not show any correlation of HWE failure and bad call rate with p-value
- Both SNiPer and BRLMM are vastly superior to DM. If filtering is performed, SniPer is slightly better than BRLMM

Comparison to BRLMM

- Both rely on training set with assigned genotype calls
- SNiPer utilizes EM to correct wrong calls and no-calls
- If initial seeds are low quality, EM may amplify the errors and give low-quality model
- BRLMM uses the database of other SNPs to polish model
- BRLMM cannot effectively correct training data and thus relies heavily on correct initial data
- Ideally they complement each other