MEMBERSHIP OF WELLCOME TRUST CASE CONTROL CONSORTIUM 2

Management Committee

Peter Donnelly (Chair)[1,2], Ines Barroso (Deputy Chair)[3], Jenefer M Blackwell[4,5], Elvira Bramon[6] , Matthew A Brown[7] , Juan P Casas[8] , Aiden Corvin[9], Panos Deloukas[3], Audrey Duncanson[10], Janusz Jankowski[11], Hugh S Markus[12], Christopher G Mathew[13], Colin NA Palmer[14], Robert Plomin[15], Anna Rautanen[1], Stephen J Sawcer[16], Richard C Trembath[13], Ananth C Viswanathan[17], Nicholas W Wood[18]

Data and Analysis Group

Chris C A Spencer[1], Gavin Band[1], Céline Bellenguez[1], Colin Freeman[1], Garrett Hellenthal[1], Eleni Giannoulatou[1], Matti Pirinen[1], Richard Pearson[1], Amy Strange[1], Zhan Su[1], Damjan Vukcevic[1], Peter Donnelly[1,2]

DNA, Genotyping, Data QC and Informatics Group

Cordelia Langford[3], Sarah E Hunt[3], Sarah Edkins[3], Rhian Gwilliam[3], Hannah Blackburn[3], Suzannah J Bumpstead[3], Serge Dronov[3], Matthew Gillman[3], Emma Gray[3], Naomi Hammond[3], Alagurevathi Jayakumar[3], Owen T McCann[3], Jennifer Liddle[3], Simon C Potter[3], Radhi Ravindrarajah[3], Michelle Ricketts[3], Matthew Waller[3], Paul Weston[3], Sara Widaa[3], Pamela Whittaker[3], Ines Barroso[3], Panos Deloukas[3].

Publications Committee

Christopher G Mathew (Chair)[13], Jenefer M Blackwell[4,5], Matthew A Brown[7], Aiden Corvin[9], Chris C A Spencer[1]

1 Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK; 2 Department of Statistics, University of Oxford, Oxford OX1 3TG, UK; 3 Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK; 4 Telethon Institute for Child Health Research, Centre for Child Health Research, University of Western Australia, 100 Roberts Road, Subiaco, Western Australia 6008; 5 Cambridge Institute for Medical Research, University of Cambridge School of Clinical Medicine, Cambridge CB2 0XY, UK; 6 Department of Psychosis Studies, NIHR Biomedical Research Centre for Mental Health at the Institute of Psychiatry, King's College London and The South London and Maudsley NHS Foundation Trust, Denmark Hill, London SE5 8AF, UK; 7 University of Queensland Diamantina Institute, Brisbane, Queensland, Australia; 8 Dept Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT and Dept Epidemiology and Public Health, University College London WC1E 6BT, UK; 9 Neuropsychiatric Genetics Research Group, Institute of Molecular Medicine, Trinity College Dublin, Dublin 2, Eire; 10 Molecular and Physiological Sciences, The Wellcome Trust, London NW1 2BE; 11 Centre for Digestive Diseases, Queen Mary University of London, London E1 2AD, UK and Digestive Diseases Centre, Leicester Royal Infirmary, Leicester LE7 7HH, UK and Department of Clinical Pharmacology, Old Road Campus, University of Oxford, Oxford OX3 7DQ, UK; 12 Clinical Neurosciences, St George's University of London, London SW17 0RE; 13 King's College London Dept Medical and Molecular Genetics, King's Health Partners, Guy's Hospital, London SE1 9RT, UK; 14 Biomedical Research Centre, Ninewells Hospital and Medical School, Dundee DD1 9SY, UK; 15 King's College London Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Denmark Hill, London SE5 8AF, UK; 16 University of Cambridge Dept Clinical Neurosciences, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK; 17 NIHR Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London EC1V 2PD, UK; 18 Dept Molecular Neuroscience, Institute of Neurology, Queen Square, London WC1N 3BG, UK.

Funding: This work was supported by Wellcome Trust awards 090532/Z/09/Z, 075491/Z/04/B and 084575/Z/08/Z. PD was supported in part by a Royal Society Wolfson Merit Award and CCAS by a Nuffield Department of Medicine Scientific Leadership Fellowship.

METHOD DETAILS

We considered m summary statistics $S_1(X_i), \ldots, S_m(X_i)$ ($i = 1, \ldots, n$ with n the number of individuals) and assumed that each individual is either normal or an outlier, which we indexed by $Z_i \in \{0, 1\}$. The joint distribution of the m summary statistics is supposed to be described by independent Gaussian distributions in both the normal and outlier class:

$$S_j(X_i)|Z_i, \mu_{Z_i,j}, \sigma^2_{Z_i,j} \sim N(\mu_{Z_i,j}, \sigma^2_{Z_i,j})$$

Gibbs sampling with uninformative priors

It is possible to specify uninformative priors for most of the nuisance parameters of the Gibbs sampling by assuming

$$P(Z, \mu, \sigma^2_0) = P(Z)P(\mu)P(\sigma^2_0) \propto \sigma_0^{-2}$$

As described in the main paper, we also assumed (with λ a parameter fixed a priori)

$$P(\sigma^2_{1,j}|Z, \mu_j, \sigma^2_{0,j}) = P(\sigma^2_{1,j}|\sigma^2_{0,j}) = \begin{cases} 1 \text{ if } \sigma^2_{1,j} = \lambda^2 \sigma^2_{0,j} \\ 0 \text{ otherwise} \end{cases}$$

and

$$P(\sigma^2_{0,j}|S_j(X), Z, \mu_{0,j}, \mu_{1,j}) \approx P(\sigma^2_{0,j}|S_j(X_{Z=0}), Z, \mu_{0,j})$$

Under the assumptions above, the full conditional distribution of each of Z_i, $\mu_{Z,j}$ and $\sigma^2_{0,j}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$ is known and can be sampled from using standard numerical methods. A (correlated) sample from the posterior distribution can then be obtained using the following algorithm:

Step 1. For each summary statistic j, sample $\mu_{0,j}$ from

$$\mu_{0,j}|S_j(X), Z, \sigma^2_{0,j} \sim N\left(\frac{1}{n_0}\sum_{i/Z_i=0} S_j(X_i), \frac{\sigma^2_{0,j}}{n_0}\right)$$

and $\mu_{1,j}$ from

$$\mu_{1,j}|S_j(X), Z, \sigma^2_{0,j} \sim N\left(\frac{1}{n_1}\sum_{i/Z_i=1} S_j(X_i), \frac{\lambda^2 \sigma^2_{0,j}}{n_1}\right)$$

where $n_z = \sum_i I(Z_i = z)$, with I the indicator function, is the number of individuals in each class.

Step 2. For each summary statistic j, sample $\sigma^2_{0,j}$ from

$$\sigma^2_{0,j}|S_j(X), Z, \mu_{0,j} \sim \text{Scale-Inv-}\chi^2\left(n_0, \frac{\sum_{i/Z_i=0}(S_j(X_i)-\mu_{0,j})^2}{n_0}\right)$$

and set $\sigma^2_{1,j} = \lambda^2\sigma^2_{0,j}$

Step 3. For each individual i, sample Z_i from

$$Z_i|S_1(X), \ldots, S_m(X), \mu, \sigma^2 \sim \text{Bernouilli}(\theta)$$

where

$$\theta = \frac{p_1}{p_0+p_1} \quad \text{and} \quad p_l = \prod_j P(S_j(X_i)|Z_i = l, \mu_{l,j}, \sigma^2_{0,j})$$

A sample from the posterior distribution is obtained by repeating Steps 1 to 3 for T iterations. The posterior probability of the i^{th} individual being an outlier is then estimated as

$$\frac{1}{T}\sum_{t=1}^{T} I(Z_i^{(t)} = 1)$$

where $Z_i^{(t)}$ is the class membership of the i^{th} individual at iteration t.

Alternative prior distributions

The above model considers uninformative priors for the parameters. If we find ourselves with no individual in the outlier class when exploring the posterior distribution, the conditional distribution of the means of the Gaussian distributions describing the variation in the summary statistics in the outlier class is not well defined. To circumvent this problem, we added a hierarchical component by assuming that means of both the normal and outlier class come from another Gaussian distribution with mean $\underline{\mu}_j$ and variance $\underline{\sigma}^2_j$ assigned to each summary statistic. Updating those two new hyper-parameters requires priors on $\underline{\mu}_j$ and $\underline{\sigma}^2_j$. We chose the normal prior on the mean with parameters $\mu_{H,j}$ and $\sigma^2_{H,j}$, and the conjugate Scaled Inverse Chi-Square prior on the variance with parameters ν_j and Γ^2_j. We also assumed that the variance of the normal class $\sigma^2_{0,j}$ comes from a Scaled Inverse Chi-Square distribution with k_j degrees of freedom and scale parameter ψ_j. To incorporate the new prior structure on the summary statistic means and on the variance of the normal class, the algorithm is modified as follows:

Step 1. For each summary statistic j sample $\mu_{0,j}$ from

$$\mu_{0,j}|S_j(X), Z, \sigma^2_{0,j}, \underline{\mu}_j, \underline{\sigma}^2_j \sim$$

$$N\left(\left(\frac{\underline{\mu}_j}{\underline{\sigma}^2_j} + \frac{\sum_{i/Z_i=0} S_j(X_i)}{\sigma^2_{0,j}}\right) \Big/ \left(\frac{1}{\underline{\sigma}^2_j} + \frac{n_0}{\sigma^2_{0,j}}\right), \left(\frac{1}{\underline{\sigma}^2_j} + \frac{n_0}{\sigma^2_{0,j}}\right)^{-1}\right)$$

with a similar update for the mean of the class representing outlier individuals.

Step 2a. For each summary statistic j, sample $\sigma^2_{0,j}$ from

$$\sigma^2_{0,j}|S_j(X), Z, \mu_{0,j}, k_j, \psi_j \sim$$

$$\text{Scale-Inv-}\chi^2\left(n_0 + k_j, \frac{k_j\psi_j + \sum_{i/Z_i=0}(S_j(X_i)-\mu_{0,j})^2}{n_0+k_j}\right)$$

and set $\sigma^2_{1,j} = \lambda^2\sigma^2_{0,j}$

Hyper-parameters $\underline{\mu}_j$ and $\underline{\sigma}^2_j$ are then updated with an additional sampling step:

Step 2b. For each summary statistic j sample $\underline{\mu}_j$ from

$$\underline{\mu}_j|\mu_{0,j}, \mu_{1,j}, \underline{\sigma}^2_j, \mu_{H,j}, \sigma^2_{H,j} \sim$$

$$N\left(\left(\frac{\mu_{H,j}}{\sigma^2_{H,j}} + \frac{\mu_{0,j}+\mu_{1,j}}{\underline{\sigma}^2_j}\right) \Big/ \left(\frac{1}{\sigma^2_{H,j}} + \frac{2}{\underline{\sigma}^2_j}\right), \left(\frac{1}{\sigma^2_{H,j}} + \frac{2}{\underline{\sigma}^2_j}\right)^{-1}\right)$$

and $\underline{\sigma}^2_j$ from

$$\underline{\sigma}^2_j|\mu_{0,j}, \mu_{1,j}, \underline{\mu}_j, \nu_j, \Gamma^2_j \sim$$

$$\text{Scale-Inv-}\chi^2\left(\nu_j + 2, \frac{\nu_j\Gamma^2_j + (\mu_{0,j}-\underline{\mu}_j)^2 + (\mu_{1,j}-\underline{\mu}_j)^2}{\nu_j+2}\right)$$

The more complicated model which places a hierarchical prior on the means of the distributions describing the variability of the summary statistics solves the problem of the empty outlier class and can be used to enforce the constraint that we expect outliers at both end of summary statistic range. To achieve this, hyper priors on the variance of the within class means can be specified with $\nu_j >> 2$ and Γ^2_j small, which has the effect of juxtaposing the two distributions.

The inference framework described above also has the capacity to include prior information on the proportion of outliers. This is potentially useful to reflect the fact that outlying individuals are likely to be a small fraction of the sample, and that each individual has a low prior probability of being in the outlier class. To incorporate this information, we introduce a new parameter q which specifies the prior probability that an individual is an outlier. The natural distribution through which to specify this prior is a Beta distribution, which is conjugate to the Bernoulli distribution sampled from in Step 3 of the algorithm. Parameters of this prior Beta distribution are denoted α and β. In the case of each individual having equal prior probability q of being an outlier, Step 3 becomes:

Step 3. For each individual i, sample Z_i from

$$Z_i|S_1(X),\ldots,S_m(X),\mu,\sigma^2,q \sim \text{Bernoulli}(\theta)$$

where

$$\theta = \frac{q\,p_1}{(1-q)\,p_0 + q\,p_1}$$

and update q using

$$q|Z,\alpha,\beta \sim \text{ Beta } (n_1+\alpha, n_0+\beta)$$

Looking back at the original model, we see that without explicitly modeling q we were in fact implicitly assuming that each individual had a prior probability of one half of being an outlier, irrespective of the fraction of individuals currently in the outlier class.

Extension to correlated summary statistics

The approach easily generalizes to correlated summary statistics:

$$S(X_i)|Z_i,\mu_{Z_i},\Sigma_{Z_i} \sim N(\mu_{Z_i},\Sigma_{Z_i})$$

with Σ_{Z_i} the covariance matrix. We consider an Inverse-Wishart prior with parameters k and Ψ for Σ_0. We also assume that μ_0 and μ_1 come from a Gaussian distribution with mean $\underline{\mu} = (\underline{\mu}_1,\ldots,\underline{\mu}_m)$ and a diagonal covariance matrix $\underline{\Sigma} = \text{diag}(\underline{\sigma}_j^2)$. We use the same priors as before for the hyper-parameters $\underline{\mu}_j$ and $\underline{\sigma}_j^2$.

The algorithm is then modified as follows:

Step 1a. Sample μ_0 from

$$\mu_0|S(X),Z,\Sigma_0,\underline{\mu},\underline{\Sigma} \sim$$

$$N\left(\underline{\Sigma}^*\left(\underline{\Sigma}^{-1}\underline{\mu} + \Sigma_0^{-1}\sum_{i/Z_i=0} S(X_i)\right), \underline{\Sigma}^*\right)$$

where $\underline{\Sigma}^* = \left(\underline{\Sigma}^{-1} + n_0\Sigma_0^{-1}\right)^{-1}$,with a similar update for the mean of the class representing outlier individuals.

Step 2a. Sample Σ_0 from

$$\Sigma_0|S(X),Z,\mu_0,k,\Psi \sim$$

$$\text{I-W}\left(n_0+k, \Psi + \sum_{i/Z_i=0}(S(X_i)-\mu_0)(S(X_i)-\mu_0)^T\right)$$

and set $\Sigma_1 = \lambda^2\Sigma_0$

Other steps remain the same as in the case of uncorrelated summary statistics.

Note: In the case of uninformative priors, the algorithm is reduced to the following steps:

Step 1a. Sample μ_0 from

$$\mu_0|S(X),Z,\Sigma_0 \sim N\left(\frac{1}{n_0}\sum_{i/Z_i=0} S(X_i), \frac{\Sigma_0}{n_0}\right)$$

with a similar update for the mean of the class representing outlier individuals.

Step 2. Sample Σ_0 from

$$\Sigma_0|S(X),Z,\mu_0 \sim \text{I-W}\left(n_0, \sum_{i/Z_i=0}(S(X_i)-\mu_0)(S(X_i)-\mu_0)^T\right)$$

and set $\Sigma_1 = \lambda^2\Sigma_0$

Step 3. For each individual i, sample Z_i from

$$Z_i|S_1(X),\ldots,S_m(X),\mu,\sigma^2 \sim \text{Bernouilli}(\theta)$$

where

$$\theta = \frac{p_1}{p_0+p_1}$$

Using prior information

The prior distributions described above allow the program to incorporate information about the typical distribution of the inlier summary statistics. When the proportion of outliers is large this information can help ensure that the correct cluster is assigned to be the inliers. The algorithm is also aided by using this prior information for initialisation. The current implementation of the algorithm either allows only a prior on the fraction by specifying α and β (and uninformative priors on the rest), or a full specification of the prior by setting the mean $\mu_{H,j}$ and variance $\sigma^2_{H,j}$ of hyper mean, the conjugate Scaled Inverse Chi-Square prior on the hyper variance with parameters ν_j and Γ_j^2, and a prior on the normal class covariance as another Inverse-Wishart prior with k degrees of freedom and scale matrix ψ.

As an example we simulated 1000 observations from two bivariate normal distribution with zero covariance and means $(0,0)$ and $(0,5)$, with the former assumed to be the inlier samples. We specified prior information to reflect the fact that we think the mean of the inlier distribution is centred on $(0,0)$ with weak prior information on the variance of the hyper distribution and the covariance of the normal individuals. The results of the clustering are shown in figure 1. This toy example demonstrates that the approach can correctly identify the inlier samples even though the fraction of outliers is 50% and the distribution have the same variance. It illustrates the use of prior information in helping obtain sensible outlier identification.

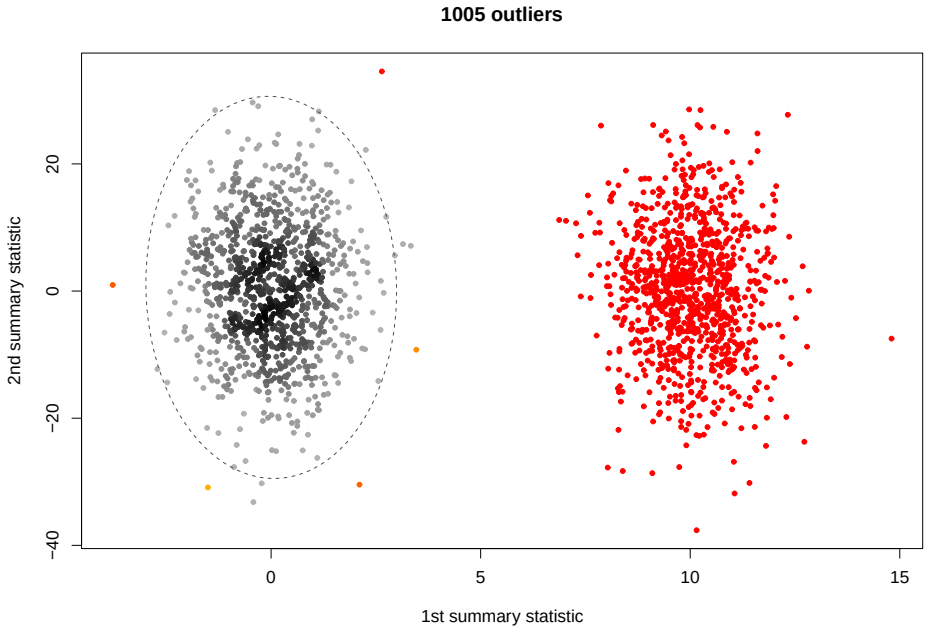

Fig. 1. Simulated data in which the proportion of inliers and outliers is the same and prior information is used to identify the correct cluster as ”normal” samples. 1000 individuals were simulated from the same bi-variate normal distribution with a different mean. The prior of hyper distribution mean was chosen to be centred on $(0, 0)$.

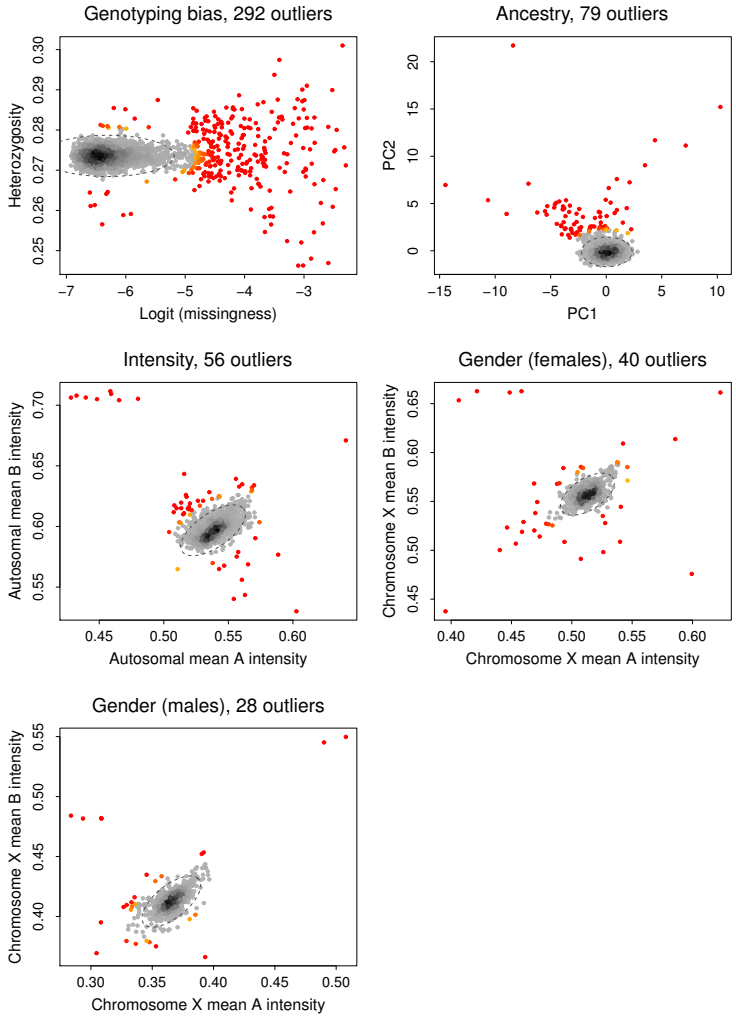

Fig. 2. Outlier identification for 58C samples genotyped on Illumina custom Human 1.2M-Duo according to 4 different criteria. ”Normal” individuals are coloured with a gradation from black to grey, with darker colours denoting higher density of individuals. Outliers are coloured with a gradation from orange to red, with darker colours denoting higher posterior probability of being an outlier. 99% confidence ellipse of the inferred inlier distribution is shown as a dashed grey line.

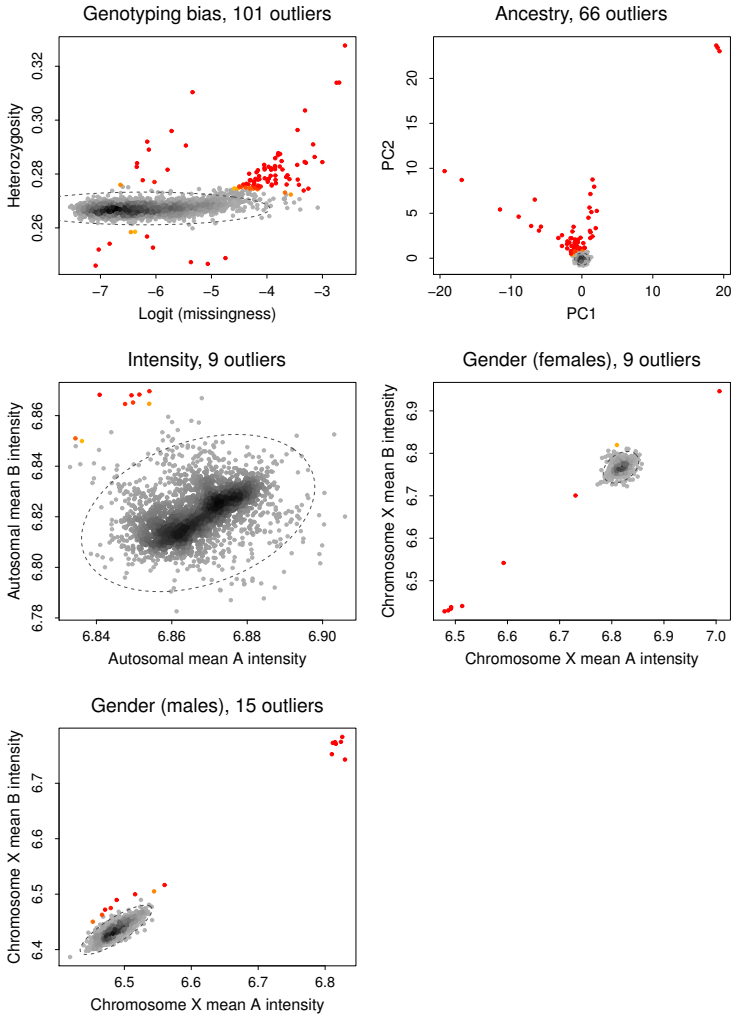

Fig. 3. Outlier identification for NBS samples genotyped on Affymetrix Genome-Wide Human SNP 6.0 according to 4 different criteria. ”Normal” individuals are coloured with a gradation from black to grey, with darker colours denoting higher density of individuals. Outliers are coloured with a gradation from orange to red, with darker colours denoting higher posterior probability of being an outlier. 99% confidence ellipse of the inferred inlier distribution is shown as a dashed grey line.

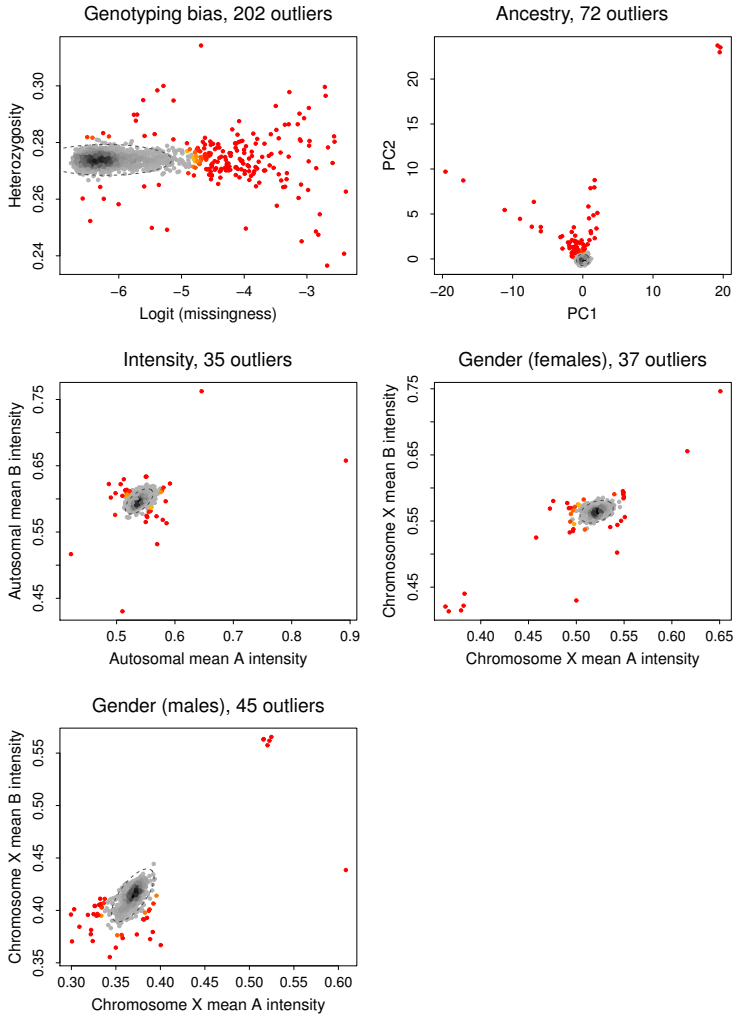

Fig. 4. Outlier identification for NBS samples genotyped on Illumina custom Human 1.2M-Duo according to 4 different criteria. ”Normal” individuals are coloured with a gradation from black to grey, with darker colours denoting higher density of individuals. Outliers are coloured with a gradation from orange to red, with darker colours denoting higher posterior probability of being an outlier. 99% confidence ellipse of the inferred inlier distribution is shown as a dashed grey line.