SNP Genotyping at Low DNA Concentrations

Nielsen, M.B.[1] (maltebn@math.aau.dk), **Andersen, M.M.**[1,2] (mikl@math.aau.dk), **Eriksen, P.S.**[1] (svante@math.aau.dk), **Mogensen, H.S.**[2] (helle.smidt@sund.ku.dk), **Morling, N.**[1,2] (niels.morling@sund.ku.dk)

[1] Department of Mathematical Sciences, Faculty of Engineering, Aalborg University, Denmark
[2] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

Motivation

The **HID_SNP Genotyper** v.4.3.1 plug-in (Thermo Fisher Scientific) is efficient at calling SNP genotypes when the amount of DNA is sufficiently high. However, for biological traces, the DNA examined may be sparse or degraded, in which case the number of no-calls increases to reduce the risk of wrong genotype classifications.

At the Section of Forensic Genetics at the University of Copenhagen (UCPH), we have observed wrong genotype classifications from the variant caller of the HID_SNP Genotyper plug-in with DNA concentrations below 62.5 pg.

To overcome this problem for biological traces, UCPH has introduced two additional criteria for SNP calling with the HID_SNP Genotyper plug-in: the coverage must be above 100 reads, and the heterozygous balance must lie between $^2/_3$ and $^3/_2$. If these criteria aren't met, UCPH declares a no-call, leading to quite a lot of no-calls.

Purpose and Applications

A probabilistic approach to SNP genotyping (or biallelic genotyping) can reduce the number of wrong genotype classifications and no-calls.

Genotype probabilities from a probabilistic model are direct measures of the reliability of the called SNPs. By introducing a **probability threshold**, q, we will declare a no-call if the probability for the genotype with the highest estimated probability is lower than q.

Genotype probabilities can be incorporated into analysis software, such as Genogeographer, to account for the probability of incorrect calls.

Our model can be used for biallelic data if the possible allele signals are interchangeable.

Material and Methods

We made four dilution series of DNA from five individuals in concentrations of 500 pg, 250 pg, 125 pg, 62.5 pg, 31.8 pg, and 16 pg DNA. The data included 165 autosomal biallelic markers (SNPs) and was arranged in two columns of reads - one for each of the possible alleles. If both signals from an observation (SNP genotype) were zero (i.e., zero reads), the observation was removed before analysis.

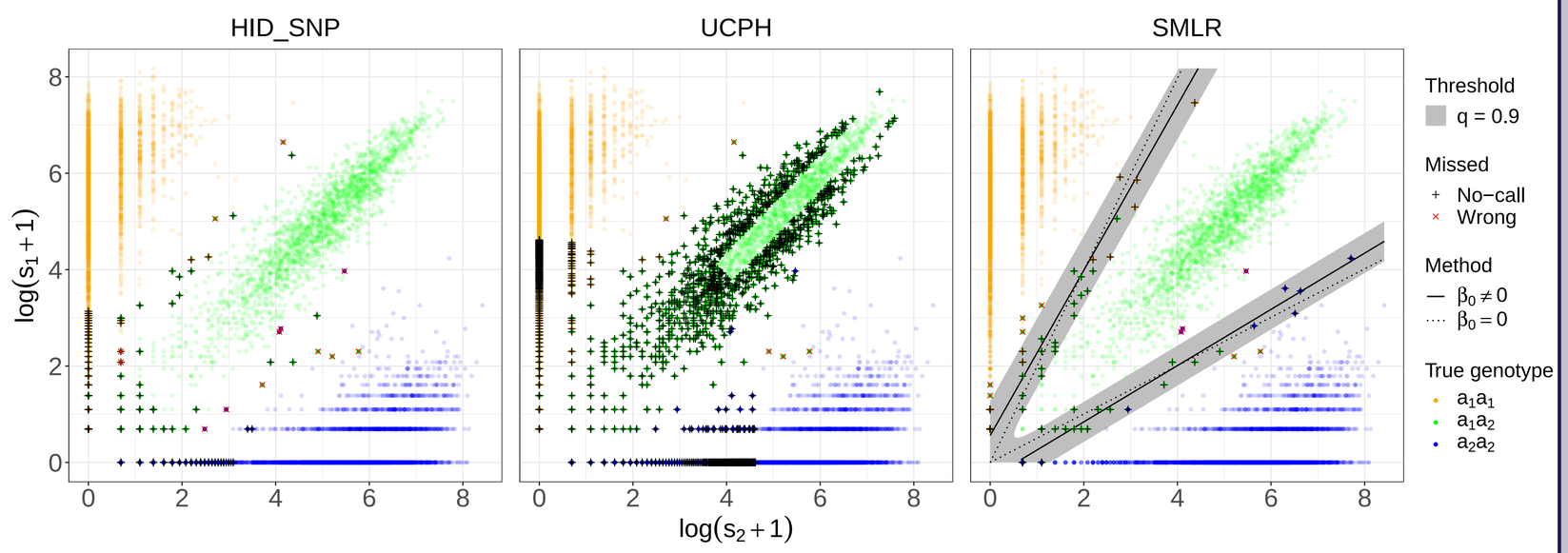
If s_1 and s_2 are the two possible signals, we applied a variance stabilizing transformation, $f(x) = log(x + 1)$, and then used a **symmetric multinomial logistic regression** (SMLR), where the genotype, G, was regressed on the transformed signals, and the heterozygous genotype was chosen as the baseline:

$$\log \frac{P(G = a_1a_1 \mid s_1, s_2)}{P(G = a_1a_2 \mid s_1, s_2)} = \beta_0 + \beta_1 f(s_1) + \beta_2 f(s_2)$$

$$\log \frac{P(G = a_2a_2 \mid s_1, s_2)}{P(G = a_1a_2 \mid s_1, s_2)} = \beta_0 + \beta_2 f(s_1) + \beta_1 f(s_2)$$

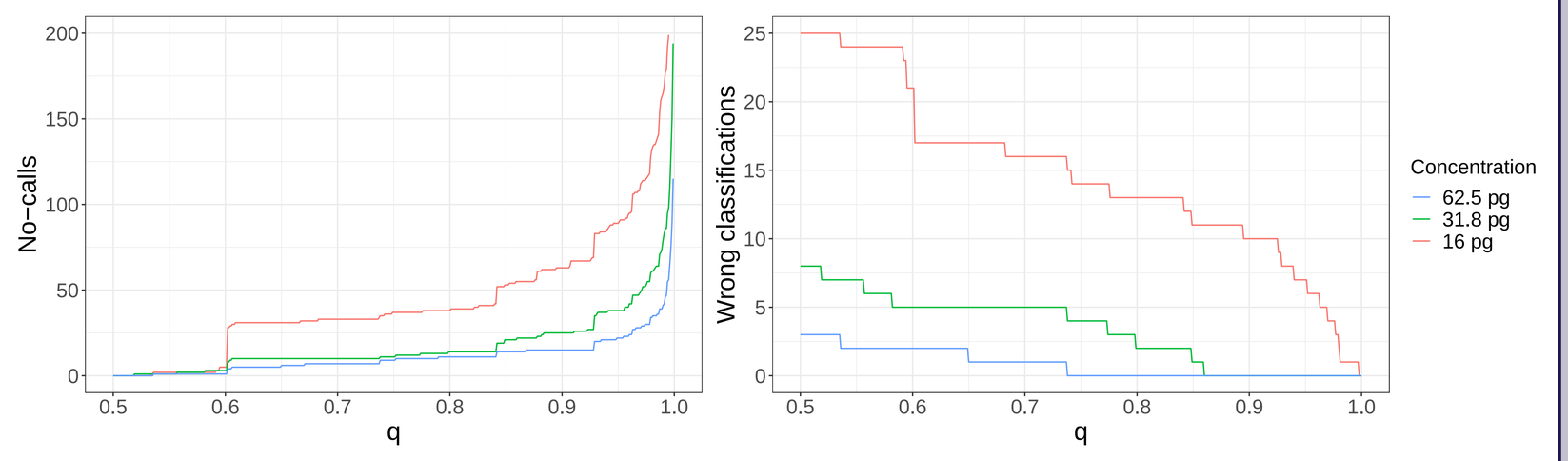
The HID_SNP Genotyper plug-in gave identical genotype classifications for the optimal concentrations from all four dilution series, so we assumed these to be the true genotypes and used them as the response in our regression.

Methods' classifications



No-calls and wrong classifications for the HID_SNP Genotyper plug-in, the UCPH criteria, and the SMLR model. The grey band around the separation lines of the SMLR model corresponds to a probability threshold of $q = 0.9$. Therefore, signal-pairs $(\log(s_1 + 1), \log(s_2 + 1))$ inside this band were declared as no-calls. The model was fitted to the full dilution series. For graphical reasons, we plotted only the low concentrations (62.5 pg, 31.8 pg, and 16 pg DNA), as the wrong classifications were observed with these DNA amounts.

No-calls and wrong calls vs. *q*



No-calls and wrong classifications for the SMLR model as a function of the probability threshold, q.

No-calls

Method	62.5 pg	31.8 pg	16 pg
HID_SNP	48	83	157
UCPH	536	705	874
SMLR, q=0.900	15	25	63
SMLR, q=0.965	27	47	107
SMLR, q=0.998	89	150	257
Total calls	3,298	2,633	1,967

Wrong calls

Method	62.5 pg	31.8 pg	16 pg
HID_SNP	1	2	10
UCPH	0	0	5
SMLR, q=0.900	0	0	10
SMLR, q=0.965	0	0	5
SMLR, q=0.998	0	0	0
Total calls	3,298	2,633	1,967

Conclusion

We have demonstrated that analysis with symmetric multinomial logistic regression can reduce the number of no-calls and wrong classifications of SNP genotypes at low DNA amounts.

For 62.5 pg and 31.8 pg DNA, we were able to reduce the number of no-calls by more than 95% compared to the UCPH criteria, while still maintaining zero wrong calls.

For 16 pg DNA, we were able to reduce the number of no-calls by 70% compared to the UCPH criteria, while simultaneously reducing the number of wrong calls by 100%.

Future work

We intend to test the SMLR model on data obtained with other kits and on data from crime scene samples with degraded and compromised DNA.

Online poster and figures

https://github.com/maltebn/ISFG2022