Bayesian Identity Metric for Genotyping

Iain Bancarz, ib5@sanger.ac.uk

May 2016

1 Introduction

This document describes the theoretical background to the Bayesian identity check, which is a QC metric introduced in pipeline release 1.11.5.

An identity check compares calls on a set of QC plex SNPs, between Infinium and another genotyping platform, such as Sequenom and/or Fluidigm. We observe *concordance* between Infinium and the alternate platform on the QC plex, and use it to evaluate *identity*. If concordance is too low, we are not confident that the production and QC calls derive from the same sample; this may indicate a sample swap or poor data quality.

Key questions include:

- How do we handle no-calls? How many non-null calls are required to evaluate identity?
- What is an appropriate concordance threshold for sample identity?
- How do we handle multiple QC plex runs, potentially on different platforms?

Previous versions of the identity check set an *ad hoc* threshold on concordance, which was used to mark samples as passing or failing QC. The Bayesian approach seeks to place the identity check on a sound probabilistic footing, in which the underlying assumptions are made explicit.

2 Bayesian Inference

Bayesian inference is a widely used statistical method in which we specify *prior* assumptions, and update the probability of an event based on observed data, finding what is known as the *posterior* probability.

This is done using *Bayes' rule*:

$$\Pr(H|E) = \frac{\Pr(E|H)\Pr(H)}{\Pr(E)}$$

where:

- H is a *hypothesis*; "samples are identical" or "samples are not identical"
- E is *evidence* which has been observed; in this case, the production and QC calls
- $\Pr(H|E)$ is probability of hypothesis given the evidence; this is the *posterior probability* which interests us.
- $\Pr(H)$ is probability of the hypothesis when evidence is not known; this is the *prior probability* of the hypothesis.
- $\Pr(E|H)$ is known as the *likelihood*. This is the probability of evidence given the hypothesis; if the samples are identical, what is the probability of observing these calls?
- $\Pr(E)$ is prior probability of the evidence, independent of the hypothesis. Serves as a normalising constant.

3 Framework for sample identity

We define the terms as follows:

Hypotheses

- H_1: Samples are identical
- H_2: Samples are not identical

Evidence

- Production calls: $p_1, p_2, \ldots, p_m$ where p_i is the production call on the ith of m SNPs in the QC plex, on the Infinium platform.
- QC calls: $q_{11}, q_{12}, \ldots, q_{mn}$, where q_{ij} is the call on the ith SNP in the jth of n QC runs, using Fluidigm, Sequenom, or any other platform.

Likelihood

Defining events

Suppose the evidence E is made up of multiple, independent events E_i. In that case:

$$\Pr(E|H) = \prod_i \Pr(E_i|H)$$

Let E_i consist of a production call and n QC calls on the ith SNP:

$$E_i = \{p_i, q_{i1}, q_{i2}, \ldots, q_{in}\}$$

where as before, p_i is the production call on Infinium, and q_{ij} are the QC calls on Fluidigm, Sequenom, or both. Each QC call q_{ij} is either equivalent to the production call p_i, or not equivalent. Our model will use the probability of equivalent calls.

Binomial distribution

Let us consider whether each of the QC calls q_{ij} is equivalent to the production call p_i. We have n_i calls on the ith SNP, excluding no-calls so that $n_i \leq n$. Suppose that out of these n_i calls, we have k_i matches. Let u_i be the probability that production and QC calls are identical on the ith SNP, given that the samples are equivalent; in other words, given H_1. Let v_i be the probability of identical calls, given that the samples are not equivalent, ie. H_2 holds.

So, we are interested in the likelihood of observing k_i equivalent calls out of n_i trials, with probability of equivalence u_i or v_i. This is simply the binomial distribution. Omitting the subscripts for i to simplify notation, the binomial distribution is:

$$\Pr(k) = \binom{n}{k} u^k (1-u)^{(n-k)}$$

and similarly for v, where:

$$\binom{n}{k} = \frac{m!}{k!(n-k)!}$$

Probability of equivalence

Now, for our two hypotheses H_1 and H_2, we have probabilities u and v as follows:

- u_i is the probability of equivalent calls, given that the samples are identical. If the samples are identical, any mismatches must be the result of a genotyping error. We then have $u_i = 1 - r$, where r is the expected error rate.
- v_i is the probability of equivalent calls, given that the samples are *not* identical. So we have $v_i = \hat{v}_i + \theta$, where $\hat{v}_i$ is the "true" probability of equivalent calls on non-identical samples for the ith SNP, and θ is a correction factor to account for experimental error.

In general, v_i will depend on the expected degree of relatedness between samples, as well as the heterozygosity and minor allele frequency (MAF) of the ith SNP.

Normalising constant

We can now find the normalising constant $\Pr(E)$ as follows:

$$\Pr(E) = \Pr(E|H_1)\Pr(H_1) + \Pr(E|H_2)\Pr(H_2)$$

Computing the identity metric

Our desired metric is the probability that the production and QC samples are identical. This happens exactly when hypothesis H_1 holds. We can compute the probability of H_1 given the evidence E as follows:

$$\Pr(H_1|E) = \frac{\Pr(E|H_1)\Pr(H_1)}{\Pr(E)} = \frac{\Pr(E|H_1)\Pr(H_1)}{\Pr(E|H_1)\Pr(H_1) + \Pr(E|H_2)\Pr(H_2)}$$

where:

$$\Pr(E|H_1) = \prod_i \Pr(E_i|H_1) = \prod_i b(k_i; n_i, u_i)$$

where $b(k_i; n_i, u_i)$ is the binomial distribution for k_i successes in n_i trials with probability u_i, and similarly for H_2 and v_i.

4 Required parameters

Model inputs

In order to compute the identity metric as above, the following parameters must be supplied or estimated from data:

- Prior probability of sample mismatch, $\Pr(H_2)$. We also have $\Pr(H_1) = 1 - \Pr(H_2)$.
- Error rate of genotype calls, r. Probability of equivalent calls on identical samples, $u = 1 - r$.
- Probability of identical calls on non-identical samples for the ith SNP, v_i.

The attached graphs simulate the effects of varying the above parameters while keeping the others at constant default values. The defaults used were:

- $\Pr(H_2) = 0.01$
- $r = 0.01$
- $v_i = 0.40625$ for all SNPs

The default v_i assumes heterozygosity 0.5 and minor allele frequency 0.25. Unless otherwise stated, the number of SNPs in the QC plex was 24.

Pass/fail threshold

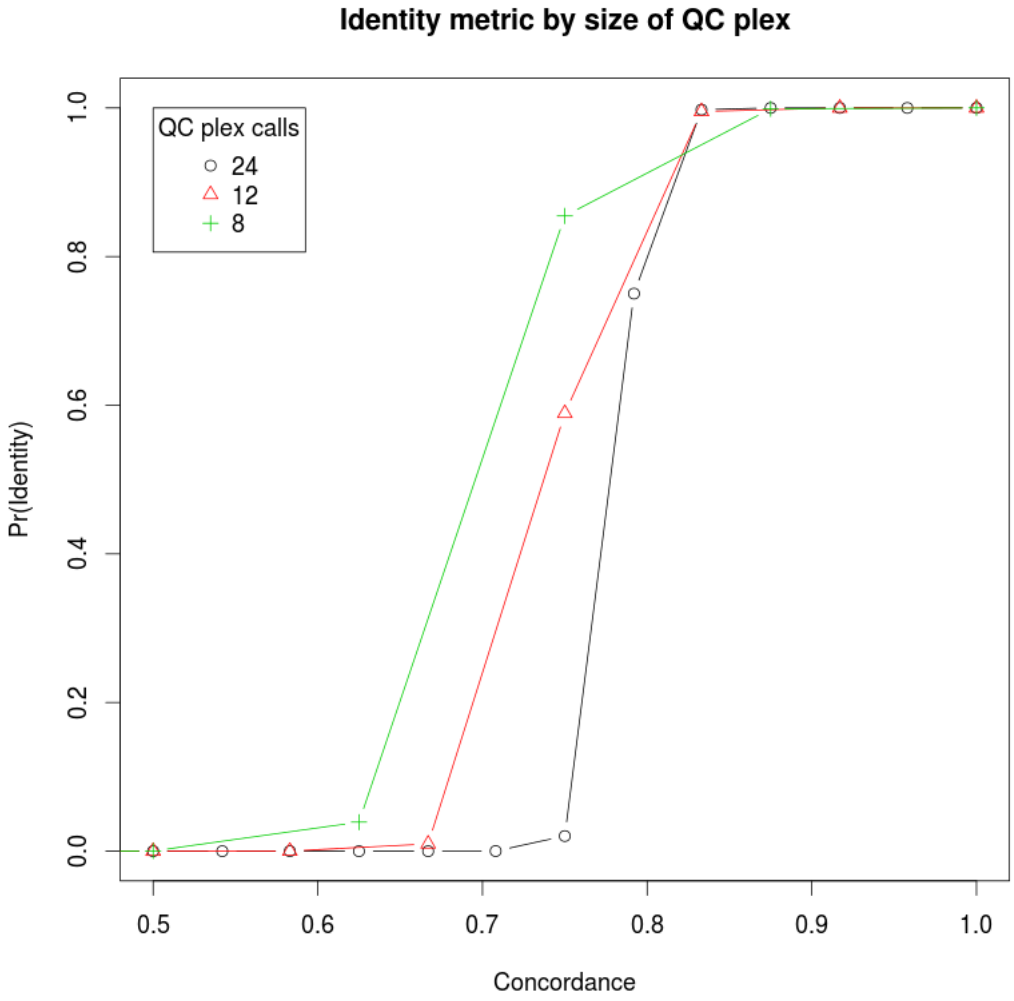
To apply the metric in practice, we need a threshold on the posterior probability $\Pr(H_1|E)$. Samples pass if they are above the threshold, and fail otherwise.

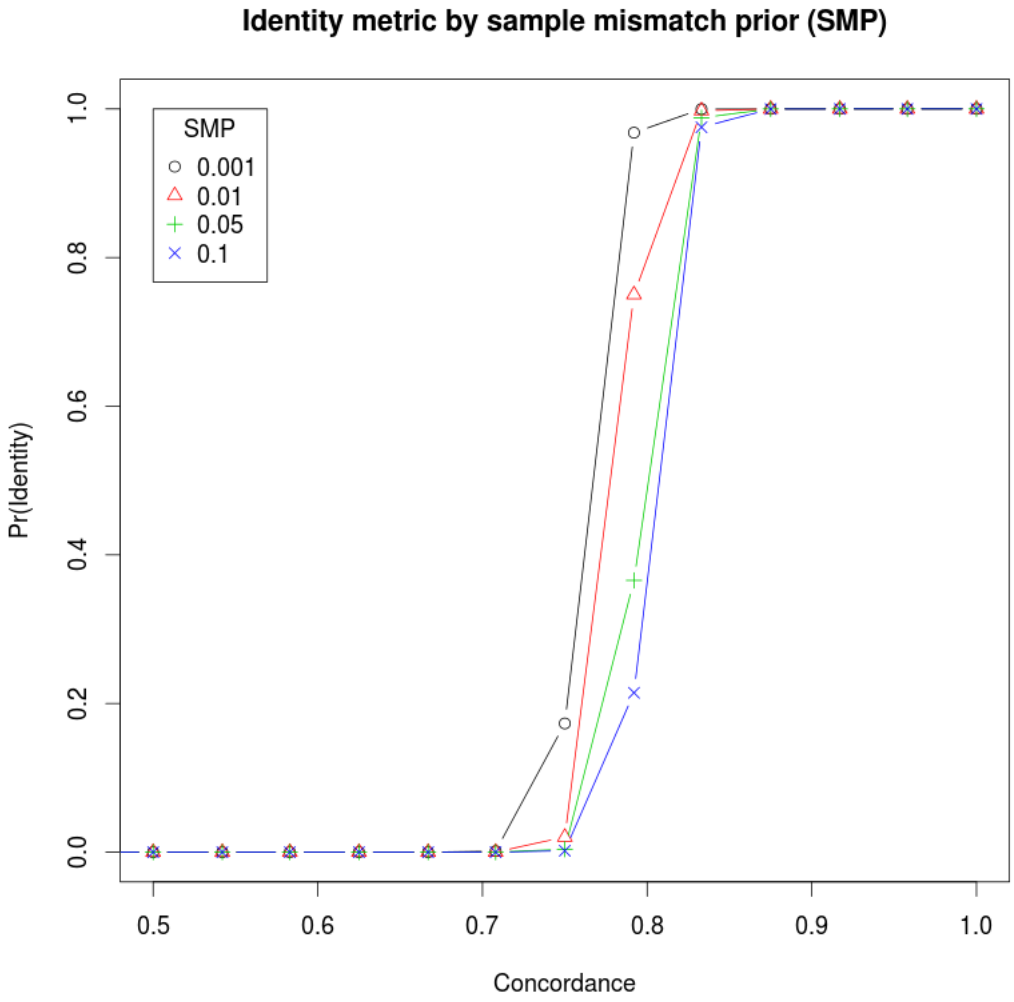
There are two plausible motivations for choosing a threshold:

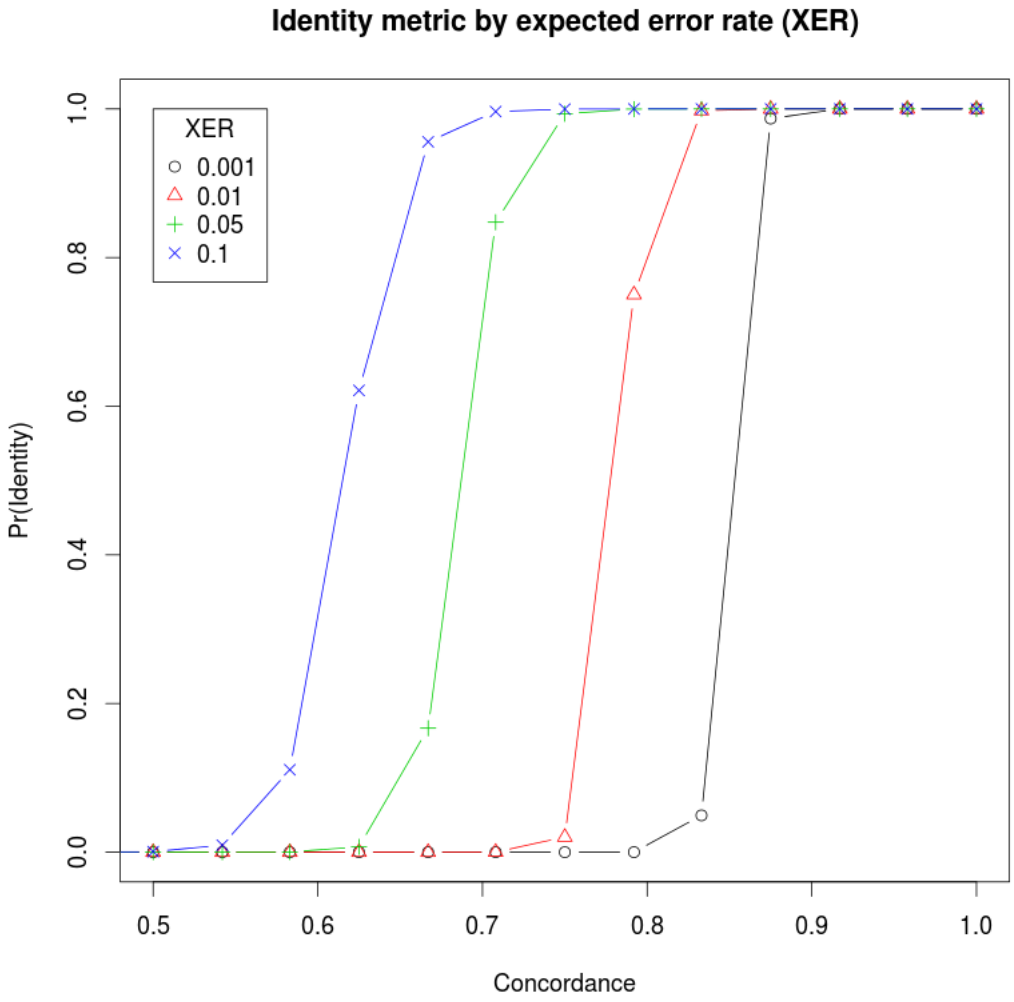
- **Optimistic:** Require convincing evidence that a sample swap *has* occurred. Set threshold to a "significant" probability of non-identical samples, for example 0.95.
- **Pessimistic:** Require convincing evidence that a swap *has not* occurred. Specifically, insist that posterior probability of identity is higher than prior probability; that is, the QC plex calls have made us *more certain* that the samples are not identical. Set threshold to $1 - \Pr(H_2)$, where as above H_2 is the sample mismatch hypothesis. With default parameters, this gives us a threshold of 0.99.

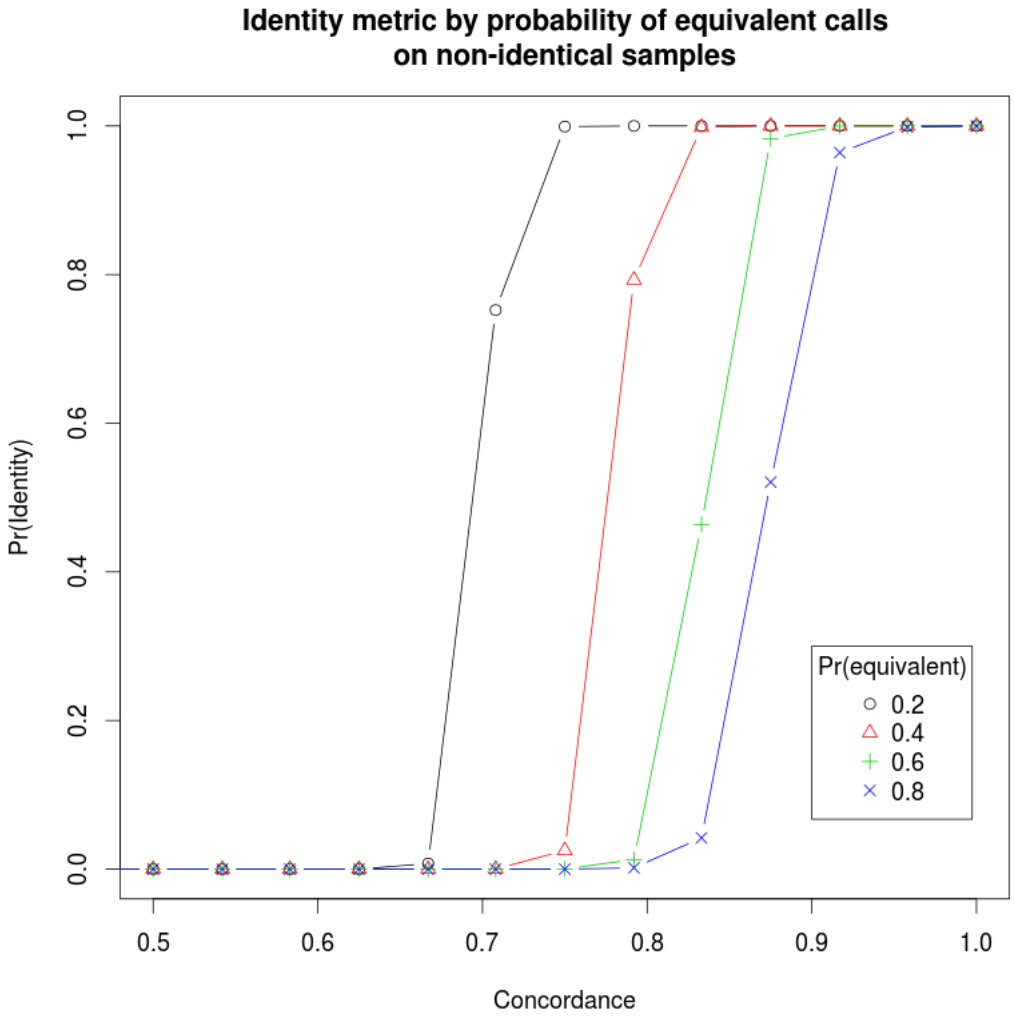
In the pipeline implementation of the Bayesian identity check, the second, "pessimistic" scenario has been adopted.

5 Graphs for simulated data




Identity metric by sample mismatch prior (SMP)
SMP
0.001
0.01
0.05
0.1
Pr(Identity)
0.0
0.2
0.4
0.6
0.8
1.0
0.5
0.6
0.7
0.8
0.9
1.0
Concordance


Identity metric by expected error rate (XER)
XER
0.001
0.01
0.05
0.1
Pr(Identity)
0.0
0.2
0.4
0.6
0.8
1.0
0.5
0.6
0.7
0.8
0.9
1.0
Concordance


Identity metric by probability of equivalent calls
on non-identical samples
Pr(Identity)
0.0
0.2
0.4
0.6
0.8
1.0
0.5
0.6
0.7
0.8
0.9
1.0
Concordance
Pr(equivalent)
0.2
0.4
0.6
0.8