Supplementary Note: quality control of summary statistics

For summary statistics of binary traits derived from a logistic regression,

$$\mathrm{sd}(\boldsymbol{g_j}) \approx \frac{2}{\mathrm{se}(\hat{\gamma}_j)\,\sqrt{n_{\text{eff}}}}\ , \tag{S1}$$

where $n_{\text{eff}} = \dfrac{4}{1/n_{\text{case}} + 1/n_{\text{control}}}$. Anyway, we recommend to verify this assumption and to perform some quality control. Indeed, in simulations, the approximation of equation (S1) seems valid (Figure S1). However, in real data applications, where summary statistics comes from a meta-analysis of many external datasets, this approximation can be invalidated (Figure S2). Let us denote by SD_{ss} the standard deviations derived from the summary statistics and by SD_{val} the standard deviations of genotypes of individuals in the validation set. We removed variants with $\text{SD}_{\text{ss}} < 0.5 \cdot \text{SD}_{\text{val}}$ or $\text{SD}_{\text{ss}} > 0.1 + \text{SD}_{\text{val}}$ or $\text{SD}_{\text{ss}} < 0.1$ or $\text{SD}_{\text{val}} < 0.05$ (Figure S2).

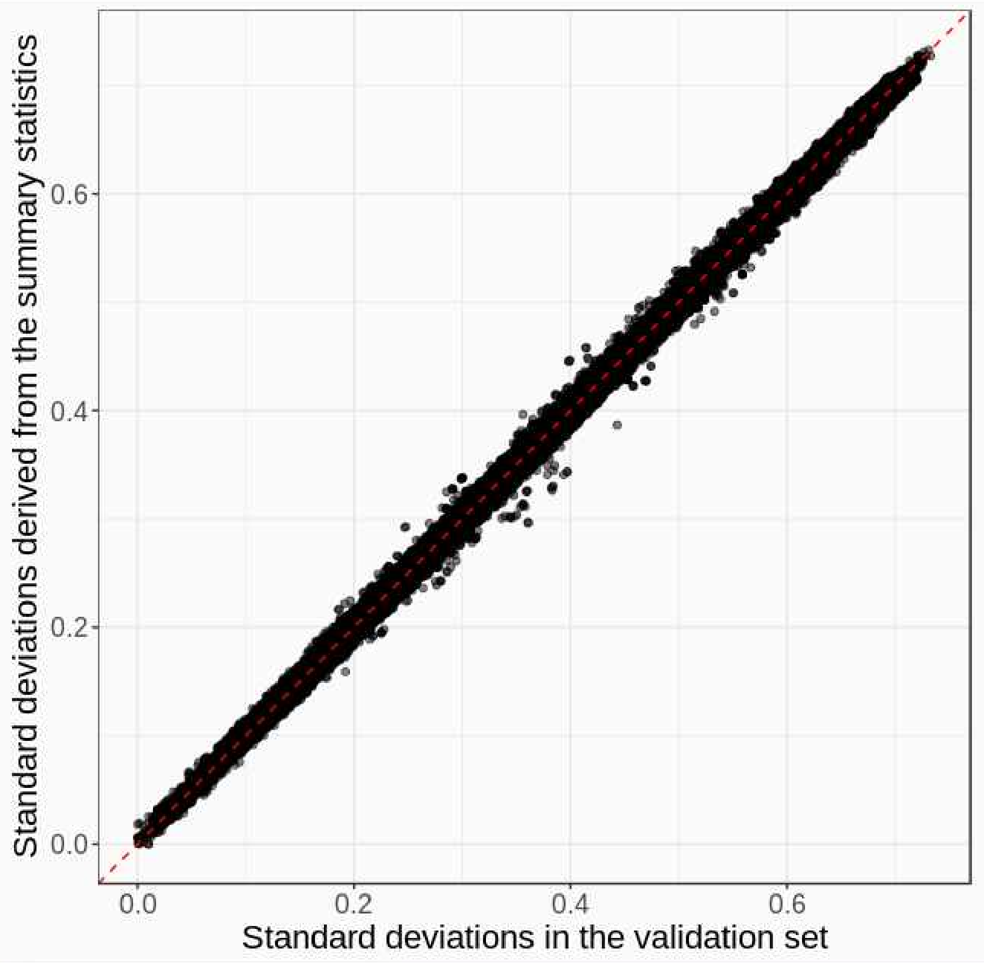

Figure S1: In simulations, standard deviations derived from summary statistics based on equation (S1) versus the standard deviations of genotypes of individuals in the validation set.

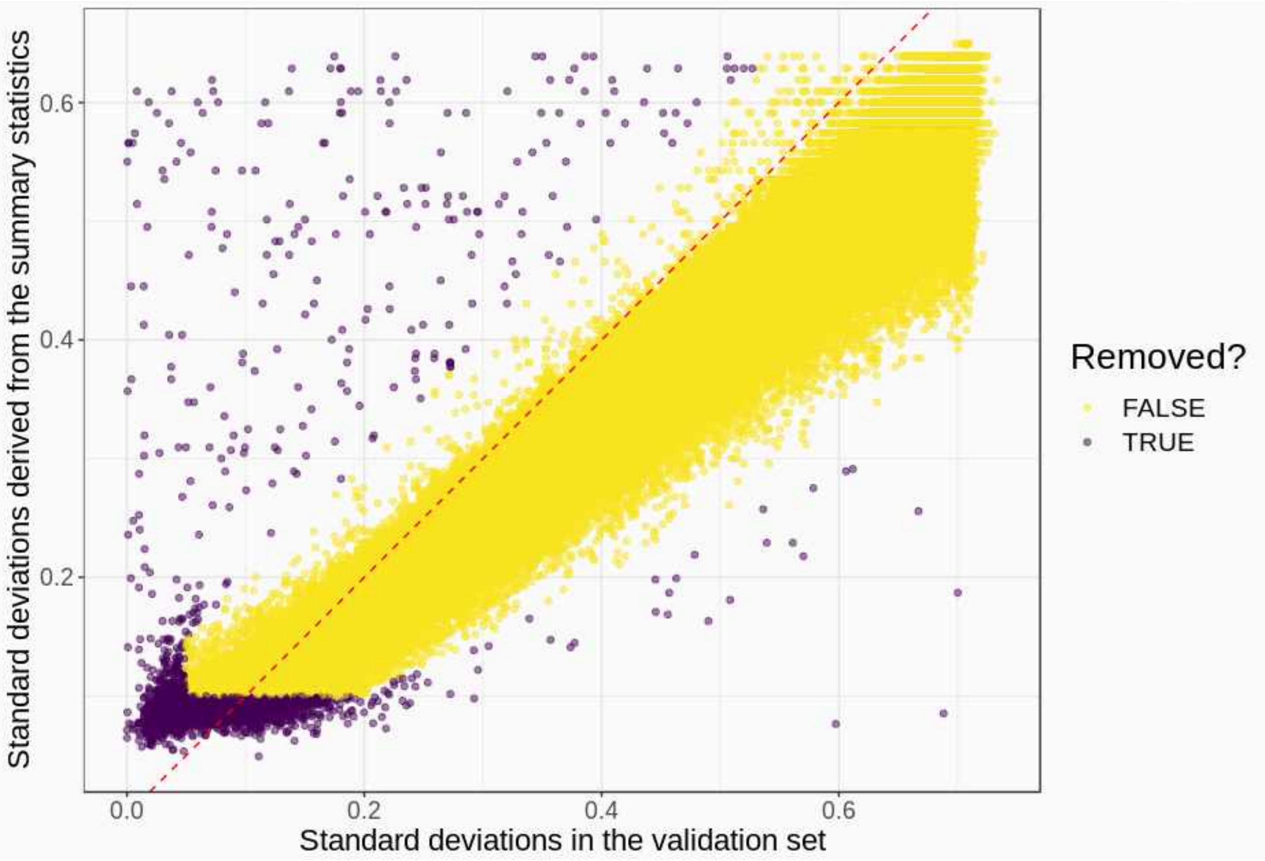

Figure S2: Standard deviations derived from summary statistics of breast cancer based on equation (S1) versus the standard deviations of genotypes of individuals in the validation set. Coloring shows the quality control applied in this paper.