PHENOTYPE-ASSOCIATED GENE DISCOVERY WITH ALPHAGENOME

Anonymous authors
Paper under double-blind review

ABSTRACT

Identifying which genes are relevant to a phenotype is a prerequisite for editing the genome to treat disease, and the established tool, genome-wide association, has a known failure mode: where population structure correlates with the trait, a variant-level test cannot separate a causal locus from one that is merely ancestry-informative. We study this failure on a testbed built to make it unavoidable: a cohort from the 1000 Genomes Project relabelled by a phenotype, skin pigmentation, that coincides exactly with continental ancestry. For each candidate gene individually, we predict melanocyte RNA-seq with a frozen sequence-to-function model and train one classifier per gene on that signal alone; the balanced accuracy it reaches induces a ranking over genes. Over nine literature-curated pigmentation genes against thirty-three independently screened controls, the classifier's ranking recovers the panel at $\mathrm{AUC} = 0.880$ ($p < 0.001$, permutation null), where a variant-level association test on the same cohort and label reaches 0.724. The gap is largest where a candidate list is drawn: four of the classifier's five highest-ranked genes are panel genes, against two for the association test, and by the threshold recovering its 5th panel gene it admits 1 control of 33 where the two association estimators admit 6 and 20—an advantage significant against both estimators over the top 15% of the ranking, a cap fixed in advance ($p = 0.007$, $p = 0.006$). Over the whole ranking it persists but, at nine positives, is resolved only against the aggregating estimator. Since follow-up is spent only on the head of a candidate list, this means fewer ancestry-driven false leads per experiment, from a filter built into the representation rather than a model of the confound.

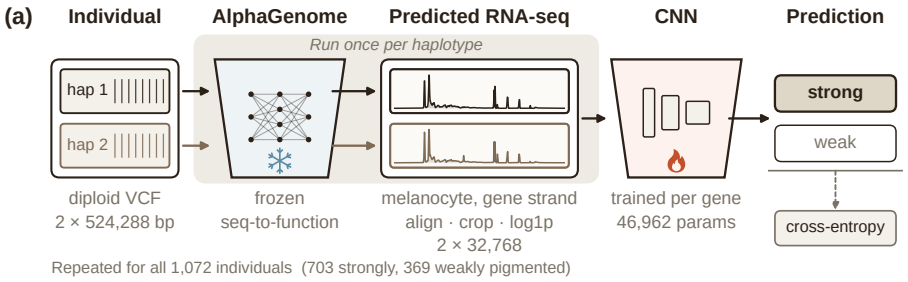


Figure 1: **One classifier per gene.** Each of the individual's two haplotypes contributes a $524{,}288$ bp window around the candidate gene, and frozen AlphaGenome predicts melanocyte (CL:1000458) RNA-seq on the gene's strand. The two tracks are aligned, cropped and normalised (Section 3.4), then stacked into one $2 \times 32{,}768$ tensor read by a CNN trained for that gene alone. This is repeated independently for each of the 42 genes, and the balanced accuracy each reaches orders them (Fig. 3).

1 INTRODUCTION

Editing the genome to treat disease presupposes knowing which genes are relevant to the phenotype being treated, and the established tool for finding them carries a failure mode that statistical care does not remove. Genome-Wide Association Studies (GWAS) test each variant against the trait and have mapped loci for a large number of human traits (Visscher et al., 2017); restricted to what a single

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

cohort can provide, however, the limitation is evident: where population structure is correlated with the trait, a variant-level test cannot separate a causal locus from one that is merely ancestry-informative. The usual defences model that correlation, through principal-component correction (Price et al., 2006), mixed models (Yang et al., 2011; Loh et al., 2015; Zhou et al., 2018; Mbatchou et al., 2021) or within-family designs, yet remain prone to false positives for *gene* assignment in small, highly correlated samples, because the statistic corrected is an association and the quantity wanted is a cause. This work proposes a route around that failure, and reports what it returns on a testbed built to make the failure unavoidable.

Our testbed is a cohort in which that correlation is at its most extreme. The 1000 Genomes Project distributes high-coverage sequencing for 3202 individuals across five superpopulations (Byrska-Bishop et al., 2022) with no phenotype label natively, so we construct a noisy binary one, weak or strong pigmentation, from population membership, which certain populations carry relatively uniformly (Candille et al., 2012; Crawford et al., 2017). Individuals in the same group share ancestry, so any statistical method is fated to find many valid explanations for the label, only some of which are pigmentation: an uncorrected test flags most of the chromosome (Fig. 5, Appendix G). The difficulty belongs to the data rather than to any estimator, and it fixes what a result can mean: every *strong* individual is African and every *weak* one European, so a gene scoring well here carries ancestry-informative signal at its locus, and its pigmentation reading is inherited from the literature the panel was curated from (Section 4.4).

A recent pipeline offers a different starting point. De Souza et al. (2026) classified ancestry from AlphaGenome-predicted RNA-seq over a panel of pigmentation-related genes, and a single-gene negative control—OCA2 against a randomly chosen non-pigmentation gene, FBXO5—favoured the pigmentation gene (0.64 against 0.37), read there as a first sign of biological plausibility. One gene against one leaves open whether the gap is gene biology or the two genes drawn, so we replace it with a ranking over the whole panel against randomly sampled controls.

Two properties of this pipeline should filter out false positives a variant-level method cannot. The classifier reads predicted RNA-seq, so a variant that correlates with ancestry but does not influence transcript abundance reaches it only through its predicted transcriptional consequence; and AlphaGenome predicts per ontology term, so restricting attention to the melanocyte requires a variant to lie in a region actively transcribed there in order to matter. The method is summarised in Fig. 1: for each candidate gene *individually*, we generate AlphaGenome tracks and train a CNN to classify the individual by phenotype, so that a gene cannot appear uninformative merely because another gene in the same input carried the label. The balanced accuracy each classifier reaches ranks the genes, and we read that ranking against a variant-level association test on the same individuals and label, on the metric that matters for prioritisation: the partial AUC over the top of the ranking, where a candidate list is drawn.

The contributions of this work are: *(i)* A gene-scoring methodology from AlphaGenome melanocyte RNA-seq predictions—one classifier per gene, on that gene's window alone—extending the single-gene comparison of De Souza et al. (2026) to a systematic ranking over nine panel and thirty-three randomly sampled control genes against a permutation null. *(ii)* A head-to-head protocol for comparing that ranking against a variant-level association test on the same cohort and label, read where a candidate list is drawn and calibrated by permutation. *(iii)* The finding that the gene-level ranking recovers the panel at $\mathrm{AUC} = 0.880$ ($p < 0.001$) and beats both variant-level estimators where a candidate list is drawn, the ablations locating that advantage in the representation's exclusions rather than in added capacity.

2 Related Work

Frozen sequence-to-function models as representations. Models predicting functional assays from DNA sequence have progressed from DeepSEA (Zhou & Troyanskaya, 2015) through Basenji (Kelley et al., 2018) and Enformer (Avsec et al., 2021) to Borzoi (Linder et al., 2025) and AlphaGenome (Avsec et al., 2026), which we use strictly frozen.[1] De Souza et al. (2026) arranges its predictions over a pigmentation gene panel into a track-by-position matrix per individual and trains

[1] Published in *Nature* in January 2026 (Avsec et al., 2026) and queried here at the all-folds checkpoint of its public inference API, between 26 November 2025 and 13 September 2026, never fine-tuned.

a CNN on it, establishing the panel, track selection and window geometry we reuse. Two caveats bear on our results: these models predict cross-locus variation considerably better than cross-individual variation at a fixed locus (Huang et al., 2023; Sasse et al., 2023; Karollus et al., 2023), and capture promoter-proximal determinants more reliably than distal enhancers (Karollus et al., 2023). Our method depends only on their annotating arbitrary sequence, which makes scoring an individual's own reconstructed genome possible.

Gene prioritisation within a single cohort. A first group of methods needs only the genotypes and label of the cohort at hand. GWAS test each variant against the trait and remove population structure through principal-component correction (Price et al., 2006), mixed models (Yang et al., 2011; Loh et al., 2015; Zhou et al., 2018; Mbatchou et al., 2021) or within-family designs; in our cohort the structure is the label itself (Section 3.1), so the baseline of Section 3.3 is run uncorrected. Ranking genes then requires reducing a window's p-values to one number, in two families (Neale & Sham, 2004; Kang et al., 2013): best-SNP methods take the extremum (VEGAS's top-SNP form (Liu et al., 2010), MAGMA's `snp-wise=top` (de Leeuw et al., 2015), PASCAL's max-chi-squared (Lamparter et al., 2016)), aggregating methods the mean or sum of the chi-squared statistics (MAGMA's default, PASCAL's and VEGAS's summing forms), and GATES (Li et al., 2011; Simes, 1986) interpolates between the two. Transcriptome-wide association instead scores a gene through expression imputed from genotype (Gamazon et al., 2015; Gusev et al., 2016): its imputation model is fitted once on a reference transcriptome, after which it too runs on a single cohort. Our method belongs to this group, scoring a gene through predicted melanocyte expression, and Section 4.3 runs a TWAS-style reduction of our own track as a baseline.

Gene prioritisation against gold standards. A second group is fitted or validated beyond any single cohort, on loci whose causal gene was established by functional experiments or on summary statistics meta-analysed across studies: SNP-to-gene linking (Mountjoy et al., 2021; Gazal et al., 2022; Schipper et al., 2025), PoPS (Weeks et al., 2023), colocalisation (Giambartolomei et al., 2014; Wang et al., 2020) and heritability-based scoring (Fine et al., 2019). None of them runs on a single cohort, and the first two draw on existing annotation of the very genes in our panel, so running them here would in part measure the literature the panel was curated from (Appendix C); colocalisation additionally needs expression measured in the relevant cell type and assumes one shared causal variant per locus, untenable in a continental contrast. Our comparison is therefore against the first group.

Pigmentation genetics. Pigmentation is among the best-characterised complex traits. SLC24A5 carries the largest-effect known light-skin allele in Europeans (Lamason et al., 2005); rs12913832 modulates OCA2 expression through a long-range HERC2 enhancer loop (Visser et al., 2012; Sulem et al., 2007); and GWAS in European, African and Latin American cohorts have implicated TYR, TYRP1, SLC45A2, MFSD12, MC1R and DDB1 (Candille et al., 2012; Crawford et al., 2017; Adhikari et al., 2019).

3 METHOD

3.1 COHORT AND LABELS

The 1000 Genomes Project high-coverage collection contains 3202 individuals, including 602 trios, across 26 populations and five superpopulations (Byrska-Bishop et al., 2022). We retain the eight populations whose typical skin pigmentation is established in the literature (Candille et al., 2012; Crawford et al., 2017): YRI, ESN, LWK, MSL and GWD are labelled *strong*, and FIN, CEU and GBR *weak*. This leaves 1 072 individuals, 703 strong and 369 weak (Fig. 4, Appendix A).

The split is family-aware, with individuals sharing a family identifier kept within a single split, giving 761 training, 149 validation and 162 test individuals; the test split's majority-class proportion, 0.7099, is the baseline every accuracy below is read against.

3.2 GENES

Nine pigmentation genes are taken from the eleven-gene panel of De Souza et al. (2026), picked there from the literature on pigmentation and hair (Section 2): TYR, TYRP1, SLC45A2, OCA2 and HERC2 in the melanin-synthesis and melanosome machinery, SLC24A5, MFSD12 and DDB1 with skin-pigmentation phenotypes, and MC1R, whose loss-of-function alleles give red hair and fair skin. The two excluded, EDAR and TCHH, carry hair rather than pigmentation phenotypes and belong to that panel only because it retained hair-follicle cell types, whereas this study retains the melanocyte alone.

The 33 controls are randomly sampled genes, each used whole: none was removed or substituted, since dropping a control that scored inconveniently high would turn the sample into a selection. They satisfy three conditions. No control carries a Gene Ontology term under pigmentation, melanin or melanocyte (Ashburner et al., 2000). None causes a Mendelian disorder with a pigmentary phenotype in ClinVar (Landrum et al., 2018). None appears in the KEGG melanogenesis pathway (Kanehisa & Goto, 2000). Appendix B reports each screen separately with its date and counts, together with the three genes flagged as borderline on grounds outside the screens and kept, and the one control that clears the annotation screen vacuously.

3.3 GWAS BASELINES

Our method (Section 3.4) ranks genes by how well a classifier trained on that gene's window alone predicts the phenotype label; the established alternative ranks them by how strongly the variants they contain associate with the trait. We run that alternative on the same genes, individuals and label, at the strength the field would run it at, to compare the two directly.

The per-variant test. Every variant inside the same $524{,}288$ bp window the classifier reads, with minor allele frequency at least 0.01, is tested against the binary pigmentation label with plink2 (Chang et al., 2015). The test is run both on all $1\,072$ retained individuals and on the 840 founders obtained by keeping one individual per family identifier, since the collection contains trios and an association test assumes independent samples. Significance is read against the conventional genome-wide threshold of 5×10^{-8}, with inflation quantified by the genomic control factor (Devlin & Roeder, 1999). The test is not corrected for population structure, because no correction is possible in this cohort: the structure is the label itself.

From variants to a gene score. Neither the window minimum nor a Bonferroni correction over the variant count is usable on its own: the first ranks variant density in part, the second treats variants in linkage disequilibrium as independent tests (Appendix C). We therefore report three gene scores over the same filtered variants, one from each of the two reduction families of Section 2—best-SNP and aggregating—and one interpolating between them, each as $S_g = -\log_{10} p_{\mathrm{gene}}$. The first is the window minimum corrected by Šidák (1967) over the effective number of independent tests, estimated from the eigenvalues of the variant correlation matrix by the method of Li & Ji (2005). The second is the window mean chi-squared that MAGMA applies by default (de Leeuw et al., 2015), the mean taken over the window's variants rather than over a genome-wide scan. The third is the extended Simes procedure GATES (Li et al., 2011), which aggregates over every variant yet reduces to the corrected minimum when a single variant carries the signal. Appendix C defines each in full.

3.4 CLASSIFIER

Representation. For each gene we take a window of $524{,}288$ bp centred on the gene midpoint, in the individual's own sequence, and pass it through AlphaGenome, once per haplotype. The predicted track retained is RNA-seq of CL:1000458 (melanocyte of skin) (Diehl et al., 2016) on the strand where the gene is located, and it is this track that reaches the classifier.

Alignment. AlphaGenome returns predictions on the coordinates of the individual's own sequence, so an indel displaces every later position relative to the reference and a convolution reading a fixed offset would read different biology in different individuals. We remove this with a Fit/Apply procedure: a coordinate system shared by every individual is fit once from the training split alone,

and every individual's prediction is written into that layout (Appendix D). Section 4.3 reports what it contributes by rebuilding the whole ranking without it.

Normalisation. The aligned signals are cropped to the central $32{,}768$ positions and compressed with $\log 1p$ against a maximum m fitted on the training split alone, each value x becoming

$$\tilde{x} \;=\; \frac{\log(1+x)}{\log(1+m)}, \tag{1}$$

which places the training split in $[0, 1]$. The maximum is reduced over position, individual *and* haplotype at once, so both haplotype channels share one divisor (Appendix D).

Model. The two haplotypes of the retained track are stacked, so the input is a $2 \times 32{,}768$ matrix per individual and the output is a binary label. Unlike De Souza et al. (2026), who scored a single haplotype per individual, we retain both phased haplotypes as input channels, so a heterozygous individual is represented by the two distinct signals its haplotypes actually predict rather than by one standing in for the whole genotype. The architecture follows De Souza et al. (2026) up to a first layer widened for the extra channel: three convolutional stages, global max pooling over position and a fully connected head with dropout, $46{,}962$ parameters in total (Appendix E). *One classifier is trained per gene, on that gene's window alone*, and for each we report the balanced accuracy along with the overall accuracy, against the majority-class baseline of the test split.

3.5 METRICS

Both methods return one score per gene, so both induce an ordering over the same 9 panel and 33 control genes. Throughout, cls denotes our classifier's ordering and gwas the variant-level baseline, the Šidák-corrected minimum of Section 3.3; other baselines are named where they appear. We compare the orderings with the Mann–Whitney statistic (Mann & Whitney, 1947) in its area-under-the-curve form, $\mathrm{AUC} = \Pr(\text{a panel gene outranks a control gene}) = U/(n_1 n_0)$, with $n_1 = 9$ and $n_0 = 33$.

The AUC weights every false-positive rate equally and so can be high for a ranking whose top is poor—the *early-retrieval* problem (Truchon & Bayly, 2007)—whereas a candidate list is drawn from the top alone. Our primary metric is therefore the partial AUC, the ROC curve integrated only up to a cap c on the false-positive rate (McClish, 1989), which is how a benchmark ranking a gene set against externally defined positives is read (Guala & Sonnhammer, 2017). Let $C_M(k)$ be the number of control genes method M ranks at or above its k-th best panel gene, a tie counting $1/2$, so that $C_M(k)/n_0$ is the false-positive rate at which M recovers its k-th panel gene. The partial AUC at cap c is the mean true-positive rate over the part of the curve with $\mathrm{FPR} \le c$,

$$\mathrm{pAUC}(c) \;=\; \frac{1}{c\,n_1} \sum_{k=1}^{n_1} \max\bigl(0,\; c - C_M(k)/n_0\bigr), \tag{2}$$

which lies in $[0, 1]$ and equals the AUC exactly at $c = 1$, so *AUC always means* $c = 1$. *pAUC without an argument always means* $c = 0.15$: five admitted false positives out of $n_0 = 33$ controls, a plausible size for a candidate list taken forward, fixed in advance and not the value that maximises the result. Because the cap is a free parameter supplied by the analyst (Hand, 2009), we report the whole interpolation over c (Table 5) rather than the declared point alone. Differences are written $\Delta\mathrm{AUC} = \mathrm{AUC}_{\mathrm{cls}} - \mathrm{AUC}_{\mathrm{gwas}}$ and $d = \mathrm{pAUC}_{\mathrm{cls}} - \mathrm{pAUC}_{\mathrm{gwas}}$. Two further readouts are computed from the identical scores and reported as auxiliary: P@5, the fraction of the five highest-ranked genes that are panel genes, the most directly interpretable quantity here but too coarse at $n_1 = 9$ to test a difference on, and AUPRC, average precision, which summarises the whole ranking as the locus-to-gene literature does (Mountjoy et al., 2021; Schipper et al., 2025). Appendix F gives the reasoning behind all three choices in full.

The nulls. Each ranking's AUC or pAUC is tested against a permutation of the panel/control assignment over the 42 genes, and p is the fraction of permutations at least as extreme as the observed value. The differences $\Delta\mathrm{AUC}$ and d are tested against two nulls: *label reassignment*, which draws a random set of 9 of the 42 genes as the panel and recomputes the difference, and *method exchange*, which keeps the panel fixed and swaps the two methods' ranks gene by gene, treating

Table 1: **The head-to-head against every ranking.** Each row is the identical statistic with only the second ranking's scores changed. *AUC* is that ranking's own panel-against-control separation with its permutation p; *pAUC* is the area at the pre-declared cap $c = 0.15$ and p *vs. classifier* the classifier's advantage over that row under the label-reassignment null, two-sided. *P@5* and *AUPRC* are the auxiliary precision–recall readouts of Section 3.5; the last row is the expectation under a random ranking; bold marks the classifier's figures and advantages significant at $\alpha = 0.05$; under the stricter method-exchange null no comparison clears 0.05 at this cap, though the mean-χ^2 advantage does at wider caps. †the only disagreement between the two readouts: the primary pAUC does not clear at this cap, and dropping HERC2 alone takes this p to 0.458 (Appendix I, which also sweeps c for every row).

	whole ranking		top of the ranking ($c = 0.15$)		precision–recall		
ranking	AUC	p	pAUC	p vs. classifier	P@5	AUPRC	p vs. classifier
classifier (this work)	**0.880**	<0.001	**0.499**	—	**0.800**	**0.729**	—
GWAS, Šidák minimum	0.724	0.021	0.155	**0.007**	0.400	0.446	**0.019**
GWAS, window mean χ^2	0.492	0.536	0.089	**0.006**	0.200	0.252	**0.001**
TWAS-style scalar	0.623	0.136	0.199	**0.030**	0.400	0.363	**0.006**
raw sequence	0.690	0.042	0.242	**0.028**	0.400	0.405	**0.007**
no alignment	0.763	0.008	0.388	0.082	0.600	0.610	**0.047**†
random ranking (expected)	0.500	—	0.088	—	0.214	0.278	—

every disagreement between them as chance. Label reassignment is primary, and a p reported without qualification is against it; method exchange is stricter and reported only as a sensitivity analysis. Appendix F defines both, along with the DeLong cross-check and the leave-one-gene-out repetition.

4 Results and Discussion

Every ranking reads the same 9 panel and 33 control genes on the same 1 072 individuals, so the question throughout is how well each separates the two classes, and where. Table 1 answers it in brief: the classifier recovers the panel at $\mathrm{AUC} = 0.880$ ($p < 0.001$) against 0.724, and at the top, where a candidate list is drawn, its margin over both is large and significant.

4.1 GWAS results

The uncorrected test is significant everywhere, and—what matters for a ranking—equally everywhere: about a third of the variants clear the genome-wide threshold whether their window holds a panel gene (32.3%) or a control (31.2%, pooled over the 33). At variant level the two classes are indistinguishable. The inflation behind this is extreme, $\lambda_{GC} = 27.9$, and it is real rather than artefactual: a random label of the same class balance returns $\lambda_{GC} = 0.92$, so the test is calibrated, and the 840 founders alone leave $\lambda_{GC} = 25.7$, so the trios do not carry it. Neither check separates population structure from a strongly polygenic trait, since real signal of either kind survives founder restriction and vanishes under permutation; what settles it here is the construction of the label, a deterministic function of superpopulation membership (Section 3.1). The question is what survives when such a window is reduced to one score.

The aggregating family ranks window architecture. Nothing survives: mean χ^2 reaches $\mathrm{AUC} = 0.492$ ($p = 0.536$), indistinguishable from chance. With a third of every window significant, the mean is dominated by a background common to all 42 windows, and what varies between them is how many variants they hold and how redundant those are: mean χ^2 correlates with the count m at $r = 0.48$ and the LD proxy m_{eff}/m at $r = 0.50$ (both $p = 0.001$), where Šidák and GATES correlate with neither ($|r| \leq 0.06$; Appendix C). Neither quantity differs between panel and control windows (Mann–Whitney $p = 0.976$ and 0.232), so the aggregating estimator orders the genes by an architectural property orthogonal to the biology, which is a ranking at chance by construction.

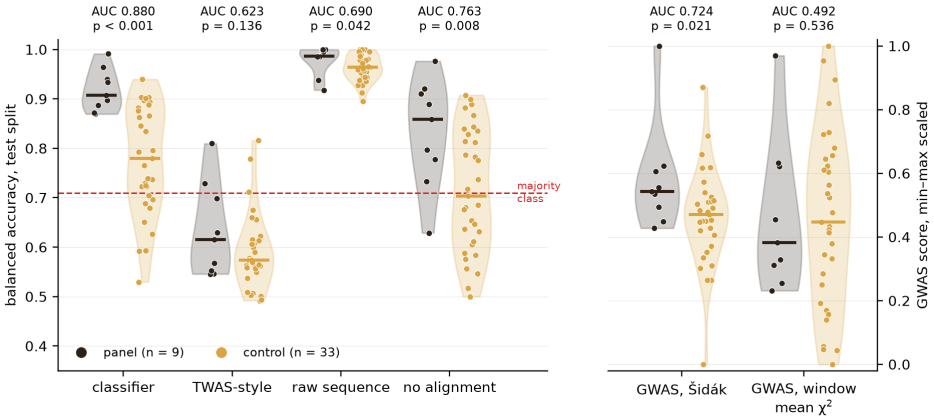


Figure 2: **Panel genes separate from controls on the test split.** Left: balanced accuracy of the 42 per-gene classifiers, for each classifier arm. Right: the two GWAS scores over the same 42 windows, min–max scaled to $[0, 1]$—a monotone rescaling, so ranking and annotation are unchanged. Panel (dark) and control (light) distributions, one point per gene; bars are group medians, the dashed line the split's majority-class baseline, and the annotation the probability that a panel gene outranks a control with its permutation p (per-gene numbers: Table 7).

The best-SNP family survives, for a reason specific to this trait. The extremum ignores that background, asking only whether one variant is more extreme than the rest, and the three best-SNP scores agree almost exactly on the resulting order ($\rho \geq 0.998$; raw minimum and GATES both reach $\mathrm{AUC} = 0.717$ against Šidák's 0.724), so we carry Šidák alone. That it separates panel from control at all ($\mathrm{AUC} = 0.724$, $p = 0.021$) on an ancestry-determined label is not evidence that it recovers pigmentation biology. The label is ancestry, so the test measures allele-frequency differentiation between the African and European samples, and pigmentation is precisely the trait whose largest-effect known alleles are also among the most strongly differentiated variants in the genome (Lamason et al., 2005; Crawford et al., 2017). Confound and truth point the same way here, so the extremum reads the first while appearing to find the second—a generous baseline, and one a less differentiated trait would not supply.

The separation is not where a candidate list is drawn. Its $\mathrm{AUC} = 0.724$ lives in the middle and tail of the order: 22 of the 33 controls outrank its worst panel arm, and the two largest-effect pigmentation genes are the clearest case, SLC24A5 and SLC45A2 placing 20th and 28th of 42 (Table 6), so the order among the strongest windows is not tracking effect size either. At the top, where a screen stops, it reaches $\mathrm{pAUC}(0.15) = 0.155$ against 0.088 at random—the early-retrieval problem of Section 3.5 in the data, and the comparison Section 4.2 draws.

4.2 Classifier results

Trained on nothing but its own gene's window, the 42 classifiers separate the two classes (Fig. 2, left-most pair): the median panel arm reaches 0.907 balanced accuracy against 0.780 for controls, giving $\mathrm{AUC} = 0.880$ at $p < 0.001$ against the permutation null. The separation shows in the whole distribution rather than one tail: all nine panel arms clear the majority-class baseline of 0.710, where nine of the thirty-three controls do not, and retraining under three further seeds leaves it stable (Appendix E).

The ranking the pipeline returns. Figure 3 draws the ordering this induces. The genes the method hands back first are SLC24A5 (0.991), SLC45A2 (0.964), HERC2 and SPRED2 (both 0.940) and TYRP1 (0.933): four panel genes and one control in the first five. That control is SPRED2, and what Section 3.2 establishes of it is an *absence*: no pigmentation GO term, no pigmentary Mendelian disorder, no place in the melanogenesis pathway. The doubt comes from one step away, its paralog SPRED1 causing Legius syndrome with hyperpigmented macules, not from SPRED2's own record. Extending it to ten adds MC1R and the controls FRA10AC1, KIAA0319, SEM1 and EGF, and

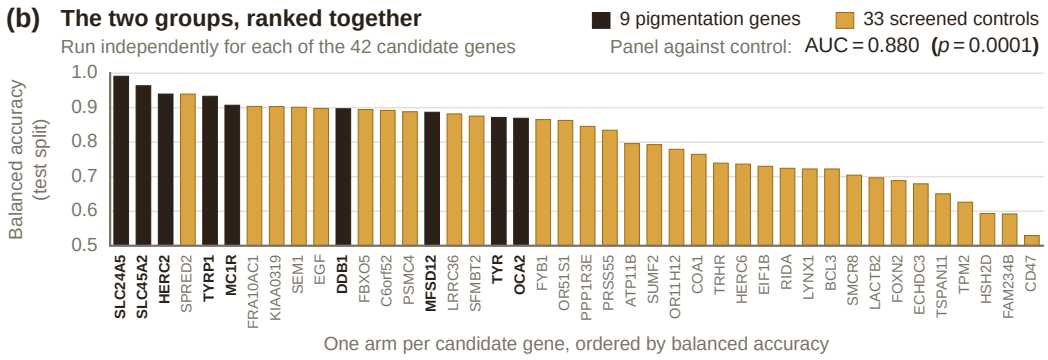


Figure 3: **The ranking the pipeline returns.** All 42 arms ordered by test-split balanced accuracy, one bar per candidate gene; panel genes dark and bold, screened controls gold. The axis floors at 0.5, which is chance for a balanced accuracy.

EGF is the same case: clean on all three screens, while the EGFR signalling it initiates modulates melanogenesis. Such a list carries five panel genes, three controls nothing touches, and two whose clean record is an absence of annotation rather than evidence of absence—the limitation that would also strike SLC24A5 (Appendix B). Appendix J gives all 42 arms in order.

Against the variant-level baseline, the gap is at the top. By the threshold recovering its fifth panel gene the classifier has admitted one control of the thirty-three, against six for the Šidák minimum and twenty for mean chi-squared (Fig. 9, Appendix I)—one wasted follow-up against six or twenty, for the same five pigmentation genes—and against an arm scored on all 1 072 individuals, not the classifier's 162 held out. At the pre-declared cap it reaches $\mathrm{pAUC} = 0.499$ against 0.155 for Šidák ($d = +0.345$, $p = 0.007$) and 0.089 for mean chi-squared ($d = +0.411$, $p = 0.006$): significant against *both* estimators. Read as a candidate list it needs no area at all: $\mathrm{P@5} = 0.800$ against 0.400 and 0.200, and $\mathrm{AUPRC} = 0.729$ against 0.446 and 0.252 ($p = 0.019$). Averaged over the whole ranking the advantage runs $+0.157$ over Šidák ($p = 0.266$) and $+0.389$ over mean χ^2 ($p = 0.008$): resolved against the aggregating estimator everywhere, against the best-SNP one where the ranking is read—the early-retrieval behaviour Section 3.5 predicts, and why the capped area was declared primary in advance. It is not fragile there: it holds at every cap $c \leq 0.20$, and dropping each of the 42 genes in turn leaves it significant, worst case $p = 0.029$ (Appendix I).

4.3 Ablations

Three further rankings isolate where the margin comes from (Fig. 2, remaining pairs; Table 1): a TWAS-style reduction of our own predicted track to one scalar per gene (Gamazon et al., 2015), the same architecture reading one-hot A/C/G/T in place of AlphaGenome's prediction, and the pipeline without the alignment of Section 3.4. The three failures do not look alike.

Reducing the track to a scalar removes almost everything. The TWAS-style arm is the weakest, at $\mathrm{AUC} = 0.623$ ($p = 0.136$), and the figure shows the number understating the loss: both classes collapse to the bottom of the axis, the bulk of panel *and* control arms falling below the majority-class baseline that every published panel arm clears. Read as one number per gene the same predicted melanocyte track does not merely blur panel against control, it costs the classifier the label. What it discards is the profile along the window—which positions carry predicted transcription, and how that shape differs between individuals—so the discriminative signal lies in the shape of the prediction rather than its level. That is the substantive difference from transcriptome-wide association, which scores a gene through a single imputed expression value by construction.

Removing the bottleneck removes the selectivity. The raw-sequence arm separates the classes more weakly ($\mathrm{AUC} = 0.690$) but fails in the opposite direction, appearing in the figure as a ceiling:

both distributions compress against the top of the axis. It does not read the panel worse—every panel arm scores *higher* on raw sequence ($+0.061$ balanced accuracy on average)—but the controls gain three times as much ($+0.185$), so raw sequence reads panel and control at nearly the same accuracy (0.979 against 0.963) where the published arm keeps them apart (0.918 against 0.778). Unrestricted sequence recovers ancestry in any window at all, exactly as Section 4.1 documents from the association side; what the melanocyte RNA-seq bottleneck does is suppress that signal *selectively*, where no predicted transcriptional consequence carries it. This is the paper's central claim measured rather than assumed: the margin comes from what the representation refuses to pass on, not from added capacity to fit the label—which also says where the method should fail, on a gene whose role does not act through transcript abundance in the retained cell type.

Removing the alignment costs accuracy, not selectivity. The no-alignment arm is the mildest ($\mathrm{AUC} = 0.763$) and degrades roughly uniformly, -0.086 balanced accuracy on the panel against -0.062 on the controls, clearing neither null at the pre-declared cap. Most of the method's power survives misalignment, so we treat the Fit/Apply map as a correctness requirement rather than a source of the margin.

4.4 Limitations

The label is a deterministic relabelling of continental ancestry, not a measured pigmentation phenotype (Section 3.1), so a gene scoring well here is consistent with a pigmentation role and equally consistent, on its own, with any other trait differing along the same axis; what licenses the pigmentation reading is the panel's provenance in literature external to this cohort. The OCA2/HERC2 pair narrows the same point: rs12913832 sits inside HERC2 and acts on OCA2 through a long-range enhancer loop (Visser et al., 2012), so the window the mechanism occupies scores 0.940 where the window it targets reaches 0.869, the weakest panel arm. A ranking over gene windows is a ranking over loci, and locus-to-causal-gene is a separate inference neither method makes; these are the only overlapping windows of the 42, and dropping HERC2 changes no conclusion (Appendix H).

The comparison uses one cohort, one frozen model and one architecture, and every p rests on $n_1 = 9$ positives, so no accuracy carries a variance estimate beyond the permutation nulls and the seed check. Finally, the representation retains one ontology term and output type, and frozen models capture promoter-proximal regulation more reliably than distal enhancers (Karollus et al., 2023), so a gene acting beyond this window's crop is scored on what the representation can see.

5 Conclusion

On a cohort where population structure and the phenotype label coincide exactly, a variant-level association test is swamped: it flags a third of the variants in every window whatever that window holds, and the one reduction that still ranks the genes does so by continental differentiation, which points at the right genes only because pigmentation is a trait whose causal alleles are also the most differentiated. Training one classifier per gene on AlphaGenome-predicted melanocyte RNA-seq ranks the same genes with a larger margin, and where the two rankings differ—at the top, on the quantity a screen is read at—the advantage is significant against both association estimators and survives dropping any single gene. The ablations locate it: the margin disappears when the predicted track is reduced to one scalar, and inverts into an indiscriminate ceiling when the predicted-transcription bottleneck is removed. What does the work is what the representation refuses to pass on.

That property is not specific to pigmentation, and it is worth most where this paper could not test it. The method needs only genotypes, a binary label, a cell type and a list of candidate genes—no summary statistics, no reference eQTL panel, no gold-standard locus set—so it runs wherever a case–control collection exists, and ancestry confounding is endemic in exactly that setting, since cases and controls are rarely matched on ancestry in practice. The natural next study is a case–control design on a disease with a well-defined effector cell type—the melanocyte replaced by the cardiomyocyte, the hepatocyte or a sorted immune population—where the label is a diagnosis rather than a proxy for ancestry and a gene's rank can be checked against evidence the cohort did not supply. That answers what this testbed cannot: whether the ranking holds when the confound is present but not identical to the trait, and when it is absent.

AI USE STATEMENT

In this work, we used generative AI tools (Claude, Anthropic) to design and give feedback on the research methodology and the experiments, to implement the pipeline, the baselines and the statistical tests, and to assist in interpreting the results. We have not used generative AI tools to generate synthetic data, to clean or label the dataset, or to formulate mathematical claims; proof assistance, translation and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools to draft and edit the paper for readability, to condense and restructure sections to the page limit, and to write the LaTeX and the plotting and table-generation scripts that turn the finished result files into the paper's figures and tables. We have reviewed all AI-assisted work: AI-written code was read and tested by the authors, every figure and table is generated from the finished result files rather than typed in, and the numbers reported in the text were checked against those files by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

E THICS STATEMENT

The cohort is the public, consented, de-identified 1000 Genomes high-coverage release; no new human data was collected for this work. The label the classifier is trained against is a deterministic function of superpopulation membership, not a measured pigmentation phenotype, and Section 4.4 states directly what that does and does not license: a gene's balanced accuracy here is evidence about which windows carry ancestry-informative signal at this locus, not new evidence of a pigmentation role, which the panel's inclusion instead inherits from prior literature external to this cohort. We surface this distinction throughout rather than only here because the same construction—scoring a gene window against a label built from continental ancestry—is also the mechanism by which genomic methods have been misused to lend spurious legitimacy to ancestry-essentialist claims about non-pigmentation traits. Concretely, a gene's score here is evidence that its window carries ancestry-informative signal reaching predicted melanocyte transcription; the pigmentation reading of the nine panel genes is inherited from the literature they were curated from (Candille et al., 2012; Crawford et al., 2017; Adhikari et al., 2019), not established by any number in Section 4. Nothing in this paper's results bears on any trait other than the one the panel was curated for, and the method is not proposed as a general-purpose ancestry-to-phenotype prioritiser; the $\lambda_{GC} = 27.9$ inflation and the ancestry caveat of Section 4.4 are reported specifically so that a reader does not mistake separation on this label for a phenotype finding.

REPRODUCIBILITY STATEMENT

Every number in Section 4 is reproducible from a fixed artefact chain: the dataset is the public 1000 Genomes high-coverage release accessed through the registered dataset ID described in Section 3.1; AlphaGenome is queried strictly frozen, at the all-folds checkpoint, through its public inference API between 26 November 2025 and 13 September 2026 (Section 2, footnote); the classifier's architecture, optimiser, learning rate, batch size, epoch count and training seed are stated in full in Appendix E, which also reports the effect of retraining under three further seeds; and the alignment procedure that turns each individual's variant calls into the tensor the classifier reads is given as an explicit Fit/Apply procedure in Section 3.4. Every statistical test reported (the Šidák/GATES/mean-χ^2 gene scores of Section 4.1, the permutation nulls behind every AUC in Section 4.2, and the label-reassignment and method-exchange nulls behind the pAUC head-to-head of Section 4.2) is defined in closed form in the appendix section it belongs to (Appendices C, J, I) rather than only described in prose, so that it can be recomputed from the per-gene score and balanced-accuracy tables reported in Appendices C and J without rereading the pipeline that produced them.

REFERENCES

Kaustubh Adhikari, Javier Mendoza-Revilla, Anood Sohail, Macarena Fuentes-Guajardo, Jodie Lampert, Juan Camilo Chacón-Duque, Malena Hurtado, Valeria Villegas, Vanessa Granja, Victor Acuña-Alonzo, et al. A GWAS in Latin Americans highlights the convergent evolution of lighter skin pigmentation in Eurasia. *Nature Communications*, 10(1):358, 2019. doi: 10.1038/s41467-018-08147-0.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000. doi: 10.1038/75556.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18 (10):1196–1203, 2021. doi: 10.1038/s41592-021-01252-x.

Žiga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R. Taylor, Tom Ward, Clare Bycroft, Lauren Nicolaisen, Eirini Arvaniti, Joshua Pan, Raina Thomas, Vincent Dutordoir, Matteo Perino, Soham De, Alexander Karollus, Adam Gayoso, Toby Sargeant, Anne Mottram, Lai Hong Wong, Pavol Drotár, Adam Kosiorek, Andrew Senior, Richard Tanburn, Taylor Applebaum, Souradeep Basu, Demis Hassabis, and Pushmeet Kohli. Advancing regulatory variant effect prediction with AlphaGenome. *Nature*, 649(8099):1206–1218, January 2026. ISSN 1476-4687. doi: 10.1038/s41586-025-10014-0.

Marta Byrska-Bishop, Uday S. Evani, Xuefang Zhao, Anna O. Basile, Haley J. Abel, Allison A. Regier, André Corvelo, Wayne E. Clarke, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, 185(18):3426–3440.e19, 2022. doi: 10.1016/j.cell.2022.08.004.

Sophie I. Candille, Devin M. Absher, Sandra Beleza, Marc Bauchet, Brian McEvoy, Nanibaa' A. Garrison, Jun Z. Li, Richard M. Myers, Gregory S. Barsh, Hua Tang, and Mark D. Shriver. Genome-Wide Association Studies of Quantitatively Measured Skin, Hair, and Eye Pigmentation in Four European Populations. *PLOS ONE*, 7(10):e48294, October 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0048294.

Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1):7, 2015.

James M. Cheverud. A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*, 87(1):52–58, 2001. doi: 10.1046/j.1365-2540.2001.00901.x.

Karen N. Conneely and Michael Boehnke. So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. *American Journal of Human Genetics*, 81(6):1158–1168, 2007. doi: 10.1086/522036. Attribution of this title verified against PMC2276357.

Nicholas G. Crawford, Derek E. Kelly, Matthew E. B. Hansen, Marcia H. Beltrame, Shaohua Fan, Shanna L. Bowman, Ethan Jewett, Alessia Ranciaro, Simon Thompson, Yancy Lo, Susanne P. Pfeifer, Jeffrey D. Jensen, Michael C. Campbell, William Beggs, Farhad Hormozdiari, Sununguko Wata Mpoloka, Gaonyadiwe George Mokone, Thomas Nyambo, Dawit Wolde Meskel, Gurja Belay, Jake Haut, NISC Comparative Sequencing Program, Harriet Rothschild, Leonard Zon, Yi Zhou, Michael A. Kovacs, Mai Xu, Tongwu Zhang, Kevin Bishop, Jason Sinclair, Cecilia Rivas, Eugene Elliot, Jiyeon Choi, Shengchao A. Li, Belynda Hicks, Shawn Burgess, Christian Abnet, Dawn E. Watkins-Chow, Elena Oceana, Yun S. Song, Eleazar Eskin, Kevin M. Brown, Michael S. Marks, Stacie K. Loftus, William J. Pavan, Meredith Yeager, Stephen Chanock, and Sarah A. Tishkoff. Loci associated with skin pigmentation identified in African populations. *Science*, 358 (6365):eaan8433, November 2017. doi: 10.1126/science.aan8433.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pp. 233–240, 2006. doi: 10.1145/1143844.1143874.

Christiaan A. de Leeuw, Joris M. Mooij, Tom Heskes, and Danielle Posthuma. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Computational Biology*, 11(4):e1004219, 2015. doi: 10.1371/journal.pcbi.1004219.

Alberto De Souza, Thiago Oliveira-Santos, Claudine Badue, Breno de Angelo, Debora Meira, Artur Garcez, and Giancarlo Guizzardi. Inferring superpopulation labels from DNA via AlphaGenome-derived functional signals. In Luca Pasa, Alessandra Lintas, Igor V. Tetko, Alessio Micheli, Nicolò Navarin, Alessandro E. P. Villa, Domenico Tortorella, and Mirko Polato (eds.), *Artificial Neural Networks and Machine Learning – ICANN 2026*, volume 17092 of *Lecture Notes in Computer Science*. Springer, Cham, 2026. doi: 10.1007/978-3-032-38407-2_48.

Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, 1988. doi: 10.2307/2531595.

Bernie Devlin and Kathryn Roeder. Genomic control for association studies. *Biometrics*, 55(4): 997–1004, 1999.

Alexander D. Diehl, Terrence F. Meehan, Yvonne M. Bradford, Matthew H. Brush, Wasila M. Dahdul, David S. Dougall, Yongqun He, David Osumi-Sutherland, Alan Ruttenberg, Sirarat Sarntivijai, Ceri E. Van Slyke, Nicole A. Vasilevsky, Melissa A. Haendel, Judith A. Blake, and Christopher J. Mungall. The cell ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of Biomedical Semantics*, 7(1):44, 2016. doi: 10.1186/s13326-016-0088-7.

Rebecca S. Fine, Tune H. Pers, Tiffany Amariuta, Soumya Raychaudhuri, and Joel N. Hirschhorn. Benchmarker: an unbiased, association-data-driven strategy to evaluate gene prioritization algorithms. *American Journal of Human Genetics*, 104(6):1025–1039, 2019. doi: 10.1016/j.ajhg.2019.03.027.

Eric R. Gamazon, Heather E. Wheeler, Kaanan P. Shah, Sahar V. Mozaffari, Keston Aquino-Michaels, Robert J. Carroll, Anne E. Eyler, Joshua C. Denny, Dan L. Nicolae, Nancy J. Cox, and Hae Kyung Im. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098, August 2015. ISSN 1546-1718. doi: 10.1038/ng.3367.

Erik Garrison, Jouni Sirén, Adam M. Novak, Glenn Hickey, Jordan M. Eizenga, Eric T. Dawson, William Jones, Shilpa Garg, Charles Markello, Michael F. Lin, Benedict Paten, and Richard Durbin. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36(9):875–879, 2018. doi: 10.1038/nbt.4227.

Steven Gazal, Omer Weissbrod, Farhad Hormozdiari, Kushal K. Dey, Joseph Nasser, Karthik A. Jagadeesh, Daniel J. Weiner, Huwenbo Shi, Charles P. Fulco, Luke J. O'Connor, Bogdan Pasaniuc, Jesse M. Engreitz, and Alkes L. Price. Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nature Genetics*, 54:827–836, 2022. doi: 10.1038/s41588-022-01087-y.

Claudia Giambartolomei, Damjan Vukcevic, Eric E. Schadt, Lude Franke, Aroon D. Hingorani, Chris Wallace, and Vincent Plagnol. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics*, 10(5):e1004383, 2014. doi: 10.1371/journal.pgen.1004383.

Dimitri Guala and Erik L. L. Sonnhammer. A large-scale benchmark of gene prioritization methods. *Scientific Reports*, 7(1):46598, 2017. doi: 10.1038/srep46598.

Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda W. J. H. Penninx, Rick Jansen, Eco J. C. de Geus, Dorret I. Boomsma, Fred A. Wright, Patrick F. Sullivan, Elina Nikkola, Marcus Alvarez, Mete Civelek, Aldons J. Lusis, Terho Lehtimäki, Emma Raitoharju, Mika Kähönen, Ilkka Seppälä, Olli T. Raitakari, Johanna Kuusisto, Markku Laakso, Alkes L. Price, Päivi Pajukanta, and Bogdan Pasaniuc. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3):245–252, February 2016. ISSN 1546-1718. doi: 10.1038/ng.3506.

David J. Hand. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1):103–123, 2009. doi: 10.1007/s10994-009-5119-5.

Connie Huang, Richard W. Shuai, Parth Baokar, Ryan Chung, Ruchir Rastogi, Pooja Kathail, and Nilah M. Ioannidis. Personal transcriptome variation is poorly explained by current genomic deep learning models. *Nature Genetics*, 55(12):2056–2059, 2023. doi: 10.1038/s41588-023-01574-w.

Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000. doi: 10.1093/nar/28.1.27.

Guolian Kang, Bo Jiang, and Yuehua Cui. Gene-based genomewide association analysis: a comparison study. *Current Genomics*, 14(4):250–255, 2013. doi: 10.2174/13892029113149990001.

Alexander Karollus, Thomas Mauermeier, and Julien Gagneur. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biology*, 24(1):56, 2023. doi: 10.1186/s13059-023-02899-9.

David R. Kelley, Yakir A. Reshef, Maxwell Bileschi, David Belanger, Cory Y. McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28(5):739–750, 2018. doi: 10.1101/gr.227819.117.

Rebecca L. Lamason, Manzoor-Ali P. K. Mohideen, Jason R. Mest, Andrew C. Wong, Heather L. Norton, Michele C. Aros, Michael J. Jurynec, Xianyun Mao, Vanessa R. Humphreville, Jasper E. Humbert, et al. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science*, 310(5755):1782–1786, 2005. doi: 10.1126/science.1116238.

David Lamparter, Daniel Marbach, Rico Rueedi, Zoltán Kutalik, and Sven Bergmann. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS Computational Biology*, 12(1):e1004714, 2016. doi: 10.1371/journal.pcbi.1004714.

Melissa J. Landrum, Jennifer M. Lee, Mark Benson, Garth R. Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Karen Karapetyan, Kenneth Katz, Chunlei Liu, Zenith Maddipatla, Adriana Malheiro, Kurt McDaniel, Michael Ovetsky, George Riley, George Zhou, J. Bradley Holmes, Brandi L. Kattman, and Donna R. Maglott. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1):D1062–D1067, 2018. doi: 10.1093/nar/gkx1153.

Jian Li and Lei Ji. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95(3):221–227, 2005. doi: 10.1038/sj.hdy.6800717.

Miao-Xin Li, Hong-Sheng Gui, Johnny S. H. Kwan, and Pak C. Sham. GATES: a rapid and powerful gene-based association test using extended Simes procedure. *American Journal of Human Genetics*, 88(3):283–293, 2011. doi: 10.1016/j.ajhg.2011.01.019.

Johannes Linder, Divyanshi Srivastava, Han Yuan, Vikram Agarwal, and David R. Kelley. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. *Nature Genetics*, 57(4):949–961, 2025. doi: 10.1038/s41588-024-02053-6.

Jimmy Z. Liu, Allan F. McRae, Dale R. Nyholt, Sarah E. Medland, Naomi R. Wray, Kevin M. Brown, Nicholas K. Hayward, Grant W. Montgomery, Peter M. Visscher, Nicholas G. Martin, and Stuart Macgregor. A versatile gene-based test for genome-wide association studies. *American Journal of Human Genetics*, 87(1):139–145, 2010. doi: 10.1016/j.ajhg.2010.06.009.

Po-Ru Loh, George Tucker, Brendan K. Bulik-Sullivan, Bjarni J. Vilhjálmsson, Hilary K. Finucane, Rany M. Salem, Daniel I. Chasman, Paul M. Ridker, Benjamin M. Neale, Bonnie Berger, Nick Patterson, and Alkes L. Price. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, 47(3):284–290, 2015. doi: 10.1038/ng.3190.

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1):50–60, 1947. doi: 10.1214/aoms/1177730491.

Joelle Mbatchou, Leland Barnard, Joshua Backman, Anthony Marcketta, Jack A. Kosmicki, Andrey Ziyatdinov, Christian Benner, Colm O'Dushlaine, Mathew Barber, Boris Boutkov, Lukas Habegger, Manuel Ferreira, Aris Baras, Jeffrey Reid, Goncalo Abecasis, Evan Maxwell, and Jonathan Marchini. Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics*, 53(7):1097–1103, 2021. doi: 10.1038/s41588-021-00870-7.

Donna Katzman McClish. Analyzing a portion of the ROC curve. *Medical Decision Making*, 9(3):190–195, 1989. doi: 10.1177/0272989X8900900307.

Edward Mountjoy, Ellen M. Schmidt, Miguel Carmona, Jeremy Schwartzentruber, Gareth Peat, Alfredo Miranda, Luca Fumis, James Hayhurst, Annalisa Buniello, Mohd Anisul Karim, Daniel Wright, Andrew Hercules, Eliseo Papa, Eric B. Fauman, Jeffrey C. Barrett, John A. Todd, David Ochoa, Ian Dunham, and Maya Ghoussaini. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nature Genetics*, 53: 1527–1533, 2021. doi: 10.1038/s41588-021-00945-5.

Benjamin M. Neale and Pak C. Sham. The future of association studies: gene-based analysis and replication. *American Journal of Human Genetics*, 75(3):353–362, 2004. doi: 10.1086/423901.

Dale R. Nyholt. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *American Journal of Human Genetics*, 74(4):765–769, 2004. doi: 10.1086/383251.

Itsik Pe'er, Roman Yelensky, David Altshuler, and Mark J. Daly. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology*, 32(4):381–385, 2008.

Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

Eve Richardson, Raphael Trevizani, Jason A. Greenbaum, Hannah Carter, Morten Nielsen, and Bjoern Peters. The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns*, 5(6):100994, 2024. doi: 10.1016/j.patter.2024.100994.

Alexander Sasse, Bernard Ng, Anna E. Spiro, Shinya Tasaki, David A. Bennett, Christopher Gaiteri, Philip L. De Jager, Maria Chikina, and Sara Mostafavi. Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. *Nature Genetics*, 55(12):2060–2064, 2023. doi: 10.1038/s41588-023-01524-6.

Marijn Schipper, Christiaan A. de Leeuw, Bernardo A. P. C. Maciel, Douglas P. Wightman, Nikki Hubers, Dorret I. Boomsma, Michael C. O'Donovan, and Danielle Posthuma. Prioritizing effector genes at trait-associated loci using multimodal evidence. *Nature Genetics*, 57(2):323–333, 2025. doi: 10.1038/s41588-025-02084-7.

Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967. doi: 10.1080/01621459.1967.10482935.

R. J. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3): 751–754, 1986. doi: 10.1093/biomet/73.3.751.

Patrick Sulem, Daniel F. Gudbjartsson, Simon N. Stacey, Agnar Helgason, Thorunn Rafnar, Kristinn P. Magnusson, Andrei Manolescu, Ari Karason, Arnar Palsson, Gudmar Thorleifsson, et al. Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nature Genetics*, 39(12):1443–1452, 2007. doi: 10.1038/ng.2007.13.

Jean-François Truchon and Christopher I. Bayly. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *Journal of Chemical Information and Modeling*, 47(2):488–508, 2007. doi: 10.1021/ci600426e.

Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 years of GWAS discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017. doi: 10.1016/j.ajhg.2017.06.005.

Mijke Visser, Manfred Kayser, and Robert-Jan Palstra. HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. *Genome Research*, 22(3):446–455, 2012. doi: 10.1101/gr.128652.111.

Gao Wang, Abhishek Sarkar, Peter Carbonetto, and Matthew Stephens. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society Series B*, 82(5):1273–1300, 2020. doi: 10.1111/rssb.12388.

Elle M. Weeks, Jacob C. Ulirsch, Nathan Y. Cheng, Brian L. Trippe, Rebecca S. Fine, Jenkai Miao, Tejal A. Patwardhan, Masahiro Kanai, Joseph Nasser, Charles P. Fulco, Katherine C. Tashman, Francois Aguet, Taibo Li, Jose Ordovas-Montanes, Christopher S. Smillie, Moshe Biton, Alex K. Shalek, Ashwin N. Ananthakrishnan, Ramnik J. Xavier, Aviv Regev, Rajat M. Gupta, Kasper Lage, Kristin G. Ardlie, Joel N. Hirschhorn, Eric S. Lander, Jesse M. Engreitz, and Hilary K. Finucane. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *Nature Genetics*, 55(8):1267–1276, 2023. doi: 10.1038/s41588-023-01443-6.

Jian Yang, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, 88(1):76–82, 2011. doi: 10.1016/j.ajhg.2010.11.011.

Jian Zhou and Olga G. Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12(10):931–934, 2015. doi: 10.1038/nmeth.3547.

Wei Zhou, Jonas B. Nielsen, Lars G. Fritsche, Rounak Dey, Maiken E. Gabrielsen, Brooke N. Wolford, Jonathon LeFaive, Peter VandeHaar, Sarah A. Gagliano, Aliya Gifford, Lisa A. Bastarache, Wei-Qi Wei, Joshua C. Denny, Ma'ayan Lin, Kristian Hveem, Hyun Min Kang, Goncalo R. Abecasis, Cristen J. Willer, and Seunggeun Lee. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, 50(9):1335–1341, 2018. doi: 10.1038/s41588-018-0184-y.

A Cohort and splits

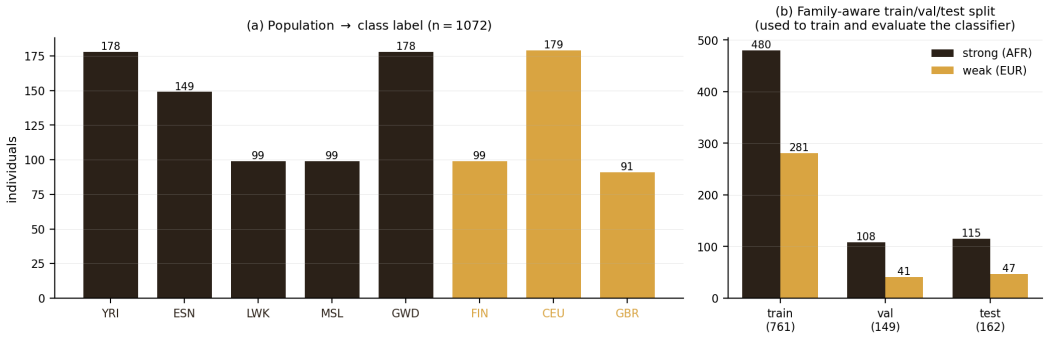


Figure 4: *(a)* The eight populations the label retains, coloured by the class each is mapped to. *(b)* The *family-aware split*, by class: the train/validation/test partition used to train and evaluate the classifier throughout the paper. It keeps every individual and assigns whole families to a single split, so a parent's genome is never in training while the child's is scored. Every *strong* individual is African and every *weak* individual European, which is the property the rest of the paper is written against. (The founders subset used by the variant-association test of Section 3.3 is a separate procedure, not shown here.)

B Provenance and screening of the control set

All 33 controls were sampled at random from the catalogue of human genes for these experiments, and the draw was conditioned on nothing either method would later return. Each control is used *whole*, as Section 3.2 states, and the three screens below are checks on the sample rather than filters applied to it: no gene was removed or substituted after a screen or a score was seen.

Curated annotation. No control gene carries a Gene Ontology annotation under pigmentation (GO:0043473), melanin biosynthetic process (GO:0042438), melanocyte differentiation (GO:0030318), or any GO term whose name contains *pigment*, *melanin* or *melanocyte*—the screen is on these terms and no others, fixed in advance, so that it cannot be widened after seeing a score. The sample was screened against QuickGO across 2026-09-12 and 2026-09-17; the genes carry between 0

and 43 annotated terms. The zero is C6orf52, which clears the screen *vacuously*—an uncharacterised open reading frame with no curated annotation at all cannot be shown to lack a pigmentation role by an annotation screen, and this is recorded rather than hidden behind a clean-looking absence; the gene stays, since dropping it after seeing the score would itself be a selection. Every other clean result is a real absence rather than an empty record.

Mendelian disease. None of the controls causes a disorder with a pigmentary phenotype—the oculocutaneous albinism series, Waardenburg, piebaldism, Hermansky–Pudlak, Chediak–Higashi and Griscelli syndromes. Checked against ClinVar (Landrum et al., 2018) on 2026-09-17, retrieving every Pathogenic or Likely Pathogenic record per control gene (from 2 to 145 across the sample) and collecting the distinct conditions reported. This is the screen the annotation screen cannot replace: SLC24A5 causes oculocutaneous albinism type 6 and carries no pigmentation GO annotation at all. A caveat for five genes—FOXN2, RIDA, SMCR8, FAM234B and TSPAN11: every returned Pathogenic or Likely Pathogenic record for each is a large chromosomal duplication/deletion spanning dozens to hundreds of genes rather than a single-gene finding, so for these five the screen is coarser still than for the rest of the set.

Pathway membership. No control gene appears in the KEGG melanogenesis pathway (hsa04916, 101 genes) (Kanehisa & Goto, 2000), against MC1R, TYR and TYRP1 from the nine, checked against the KEGG REST API on 2026-09-17 for all 33 controls directly.

Why the screens define the controls and not the panel. An annotation screen establishes absence reliably and presence unreliably. The two criteria disagree because some pigmentation genes act through functions too general to be annotated as pigmentation-specific—SLC24A5 is annotated as a cation exchanger, though rs1426654 is the largest-effect light-skin allele known in humans (Lamason et al., 2005)—so screening the panel for the *presence* of a pigmentation GO term would remove SLC24A5, HERC2 and DDB1. Of the nine panel genes, six carry a pigmentation GO annotation and three appear in the KEGG melanogenesis pathway; the screens agree on the controls and disagree on the panel, which is why only the controls are defined by them.

Genes flagged as borderline and kept. Three control genes are arguable on grounds outside the screens and are recorded here rather than removed: SPRED2, whose paralog SPRED1 causes Legius syndrome with hyperpigmented macules; EGF, since EGFR signalling modulates melanogenesis; and TRHR, which is skin-expressed and implicated in a local TRH/POMC axis. None carries a pigmentation GO term, which is why all three stay and appear in every table and test in this paper.

C GENE SCORES FROM VARIANT-LEVEL ASSOCIATION

P-values are requested from plink2 in $\log_{10}$ form (`--glm log10`), because a continental contrast drives the strongest variants below the smallest representable double and a linear p column would return exact zeros at precisely the top of the ranking. We report the number of variants surviving the frequency filter per window alongside every gene score.

(i) Corrected minimum (primary). The minimum p-value corrected by Šidák (1967) for the effective number of independent tests in the window, $p_{\text{gene}} = 1 - (1 - p_{\min})^{m_{\text{eff}}}$, which is a p-value under a single-signal null and is comparable across windows of different size. The effective count m_{eff} is estimated from the eigenvalues λ_k of the variant correlation matrix R by the method of Li & Ji (2005),

$$m_{\text{eff}} \;=\; \sum_{k=1}^{m} \Big[\mathbf{1}\{\lambda_k \geq 1\} + \big(\lambda_k - \lfloor \lambda_k \rfloor\big) \Big], \tag{3}$$

the more accurate of the eigenvalue estimators: the alternatives estimate the effective count from the variance of the correlation matrix's eigenvalues (Cheverud, 2001; Nyholt, 2004) rather than from the eigenvalues themselves, and all of them are conservative under strong linkage disequilibrium (Conneely & Boehnke, 2007). Because R is rank-deficient here—1 072 individuals against thousands of variants per window—its non-zero eigenvalues are obtained from the $1\,072 \times 1\,072$ Gram matrix rather than from R itself. The raw minimum and the variant count are reported alongside, so the adjustment stays visible rather than folded in.

(ii) Mean chi-squared. The aggregating estimator that MAGMA applies by default (de Leeuw et al., 2015), whose gene p-value matches the first two moments of the quadratic form $z^\top z$ with $z \sim \mathcal{N}(0, R)$ to a scaled chi-squared.

(iii) Extended Simes. The GATES procedure of Li et al. (2011), which orders the p-values in the window and takes $\min_k m_{\text{eff}}\, p_{(k)}/m_{\text{eff}}(k)$, where $m_{\text{eff}}(k)$ is the effective count among the k smallest. It aggregates over every variant yet reduces to the corrected minimum when a single variant carries the signal, and is the principled interpolation between the two families.

Sensitivity to variant density and LD, and why the two families disagree. Neither m nor LD structure differs systematically between panel and control windows (m: 2,776 against 2,843 variants; m_{eff}/m, the fraction of a window's variants Li and Ji's estimator counts as independent: 0.164 against 0.153; Mann–Whitney $p = 0.976$ and 0.232), so neither is a confound for the panel-vs-control comparison itself. The two estimator families do not share this indifference, though. Šidák and GATES are uncorrelated with both quantities across the 42 windows ($|r| \leq 0.06$, $p \geq 0.70$, Table 2), but mean χ^2 correlates with each ($r = 0.48$ and 0.50, both $p = 0.001$)—a window with more variants or heavier LD redundancy scores higher under the aggregating estimator for that reason alone, which is the same inflation-by-summing-correlated-statistics property that motivates LD clumping or PCA before aggregation in the first place. This gives the family disagreement of Section 4.1 a mechanism: where the two estimators differ most is exactly where density and LD are highest ($r = -0.48$ and -0.50 between the Šidák$-$mean-χ^2 gap and each quantity, both $p = 0.001$), so the extremum-based family's advantage on this cohort is not an artefact of the windows it was tested on.

Table 2: Correlation of each gene-score estimator with per-window variant count m and the LD-heterogeneity proxy m_{eff}/m (Li–Ji effective test count over raw count; 1 means every variant is independent), across all 42 windows.

estimator	r vs. m	p	r vs. m_{eff}/m	p
Šidák	$+0.004$	0.982	$+0.046$	0.770
mean χ^2	$+0.481$	0.001	$+0.497$	0.001
GATES	$+0.004$	0.982	$+0.061$	0.703

The prioritisers that are unavailable here. We compute all three *de novo* from genotype and label, rather than using a cross-study prioritiser such as locus-to-gene (Mountjoy et al., 2021; Gazal et al., 2022). Those outperform single-cohort tests on gold-standard loci, but their features include existing annotation of the very genes in this panel, so running them here would in part measure the literature from which a pigmentation panel is assembled rather than this cohort. A permutation null is likewise unavailable: the label is a deterministic function of population, so a permutation preserving population structure leaves it unchanged, while an unrestricted one tests a null that every window in a continental contrast rejects.

D Representation details

The Fit/Apply alignment algorithm. Section 3.4 summarises this procedure; the full algorithm is below.

Fit (once, on the training split only).

1. For each position p in the reference window, reserve one slot for p.
2. If an insertion at p is observed among training individuals, reserve additional slots equal to that insertion's longest training-observed allele.
3. Record the resulting personal-to-shared coordinate map as a chain file (Garrison et al., 2018).

Apply (any split, any individual, either haplotype).

1. For each position in that individual's AlphaGenome output (personal coordinates), look up its shared-coordinate slot via the fitted chain.

2. If the individual carries no insertion at a reserved slot, write 0 there.
3. If the individual's insertion is longer than the slots reserved for it—only possible in validation/test, since slots are fixed from training—drop the excess positions; *do not* shift subsequent positions.
4. Otherwise write the predicted value into the slot.
5. Crop to the central $32{,}768$ positions and apply the $\log 1p$ normalisation of Eq. 1.

The shared coordinate system. The set of positions in the aligned tensor is fixed once, from the variants observed in the training split alone: an insertion seen in training reserves as many slots as its longest observed allele, an individual who does not carry that insertion has zeros written into those slots, and a deletion likewise leaves zeros where the bases are absent. The correspondence between personal and shared coordinates is taken from the same chain file used to build the consensus sequence (Garrison et al., 2018). Fixing the layout on the training split alone keeps information about held-out individuals out of the representation, and has one consequence: an individual in validation or test may carry an insertion longer than the slots reserved for it, in which case the positions that do not fit are dropped rather than displacing everything after them.

Pooling the divisor across haplotypes. Reducing the fitted maximum of Eq. 1 over the haplotype axis as well as over position and individual deserves its own justification, because the two haplotype channels could each have been given their own divisor. They should not be. Which of an individual's two phased haplotypes is labelled first is not anchored to parent of origin, so a divisor fitted on one label and not the other would impose a difference in scale between the two channels that carries no biological information—it would record only which individual happened to carry the more highly expressed allele on the arbitrarily first-labelled haplotype. The haplotypes are exchangeable, so the maximum is pooled across them and both channels of a track share one divisor.

E ARCHITECTURE AND TRAINING

The classifier reads a $2 \times 32{,}768$ matrix and emits a binary label with an associated probability from the softmax layer. A first convolutional stage of 16 filters with kernel and stride 1×32 collapses each haplotype block to a single row before any mixing; two further stages of 32 and 64 filters use kernel 1×5 and stride 1×2; global max pooling over position is followed by a fully connected head of widths 256, 100 and 40 with ReLU activations and dropout 0.5. This gives $46{,}962$ parameters. Training uses Adam at learning rate 10^{-3}, weight decay 10^{-4}, cross-entropy loss, batch size 256 and 600 epochs with no learning-rate schedule, under seed 13. One classifier is trained per gene, on that gene's window alone.

Seed robustness. Section 4.2 reports the effect of retraining every arm under three additional seeds (7, 21, 99), keeping the data split, labels and hyperparameters fixed. Table 3 gives the panel-vs-control AUC under each seed and its Spearman correlation with the published (seed 13) ranking, over all 42 arms under all four seeds (168 of 168 (gene, seed) arms).

Table 3: Panel-vs-control separation and ranking stability under four training seeds, all 42 arms. ρ is the Spearman correlation of that seed's balanced-accuracy ranking with the published (seed 13) ranking.

seed	n genes	AUC	p	ρ (vs. seed 13)	
7	42/42	0.8586	0.0003	0.882	
13	42/42	0.8805	0.0001	—	*published*
21	42/42	0.9007	0.0000	0.851	
99	42/42	0.8704	0.0002	0.896	

F CALIBRATING THE HEAD-TO-HEAD

Why a capped area, and why precision–recall is auxiliary. Nothing about the metric of Section 3.5 is new; what is specific here is its exact finite-sum form at $n_1 = 9$ and the calibration of

differences by permutation rather than asymptotics. Integrating the ROC curve only up to a cap on the false-positive rate is the standard remedy for the early-retrieval problem (Truchon & Bayly, 2007; McClish, 1989), and it is the shape of the problem here: one global ordering of 42 genes, the positives curated from literature outside this cohort, which is how a gene-set benchmark against externally defined positives is read (Guala & Sonnhammer, 2017). The locus-to-gene literature benchmarks a different task, picking the one causal gene among the genes of a locus, and reads it on precision and recall for a reason that does not transfer—there the proportion of positives is fixed by the locus, here by our having drawn 33 controls—so the partial AUC is our primary metric and precision–recall an auxiliary one. The standing objection to a capped area is that the cap is a free parameter supplied by the analyst (Hand, 2009), which is why $c = 0.15$ is pre-declared and why the whole interpolation over c is reported (Appendix I) rather than the declared point alone: the reader can see what every other cap would have given. The precision–recall curve is sometimes argued to be preferable outright when negatives outnumber positives (Davis & Goadrich, 2006), but the imbalance here is mild (9 against 33), that argument's generality is contested (Richardson et al., 2024), and the precision–recall reading is empirically the less stable of the two on these data (Section 4.2).

The difference $\mathrm{pAUC}_{\mathrm{cls}}(c) - \mathrm{pAUC}_{\mathrm{gwas}}(c)$ of Section 3.5 is calibrated against two nulls, which answer different questions and are both reported.

Label reassignment. Resampling which 9 of the 42 genes are the panel. This is the null every panel-against-control test in this paper already uses, and it asks whether the panel's advantage at the top of the classifier's ranking exceeds what an arbitrary gene set of the same size would show. Primary.

Method exchange. Holding the panel fixed—it is literature curation external to this cohort, not a draw—and swapping the two methods' ranks gene by gene, re-ranking afterwards so that both pseudo-rankings stay permutations of $1 \ldots 42$. This is the stricter null: it treats every disagreement between the two methods as noise that could have gone either way. Reported as a sensitivity analysis, never as the headline.

We additionally cross-check $\Delta\mathrm{AUC}$—the difference of the two *whole-ranking* areas, where the analytic theory applies—against the test for correlated areas of DeLong et al. (1988), and report the exact critical value of every statistic at this sample size. No analytic cross-check is offered for the capped areas or for the precision–recall readings, whose differences are referred to the permutation nulls alone. This is the reason: at $n_1 = 9$ and $n_0 = 33$ the ROC curve has nine steps, so Eq. 2 is an exact finite sum rather than a numerical integral—no interpolation and no binormal assumption is involved—but the sampling distribution of the *difference* of two such sums is not the one the analytic tests assume at this sample size, and the head-to-head turns entirely on that difference. Every pre-declared test is repeated dropping each of the 42 genes in turn, so that a result carried by a single gene would be visible as one; Section 4.2 reports the worst case of that repetition for each reading.

Why the window minimum needs correcting. The window minimum is not itself a p-value: under the null the smallest of m uniform p-values has expectation about $1/(m+1)$, so with windows holding on the order of 10^4 variants each, reporting the minimum would rank variant density in part. Bonferroni over the variant count errs the other way, since variants inside a $524{,}288$ bp window are in linkage disequilibrium and are far fewer than their count in effective tests.

G THE VARIANT-LEVEL ARM IN FULL

The four figures below are the views Section 4.1 summarises numerically: the per-variant picture on the chromosome carrying three panel genes, the same picture restricted to the selected windows, and the gene scores at both resolutions. Two numbers belong with them. The best-SNP family's other two members track the primary estimator closely, the raw minimum and GATES both reaching $\mathrm{AUC} = 0.717$ ($p = 0.024$) against the Šidák-corrected minimum's 0.724, so the separation is a property of the family rather than of the correction. And restricting the score to the $32{,}768$ bp crop the classifier actually reads widens rather than narrows the gap between the two families (Fig. 8 against Fig. 7), so the body's comparison at full-window resolution is the conservative one.

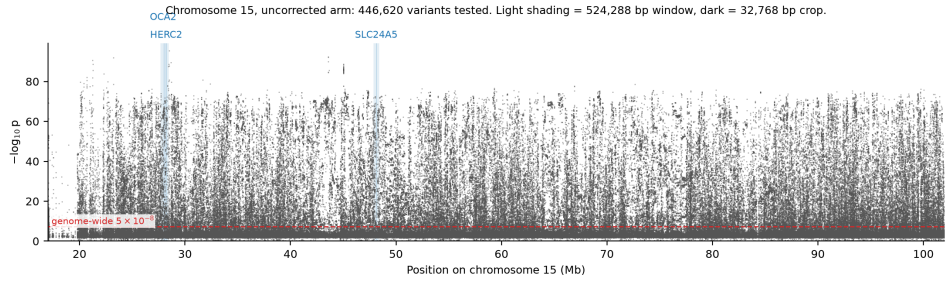


Figure 5: **The association test flags the whole chromosome** (Section 1). Chromosome 15, uncorrected: $-\log_{10} p$ per variant against the label; dashed line, 5×10^{-8}, the conventional genome-wide significance threshold: a Bonferroni correction of $\alpha = 0.05$ for the roughly 10^6 independent common-variant tests in a genome-wide scan (Pe'er et al., 2008).

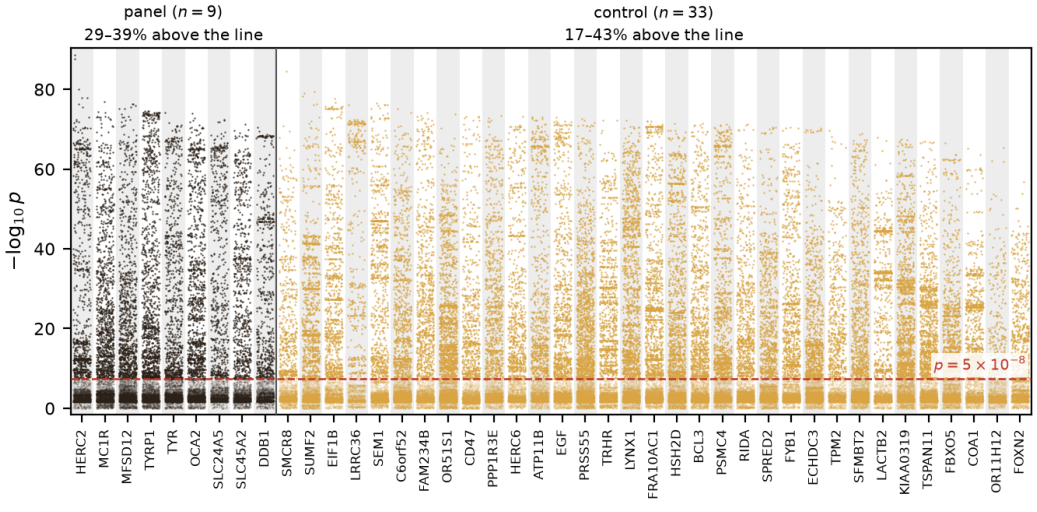


Figure 6: The selected windows only, laid end to end: every variant of every 524,288 bp window the classifier is given, panel against control. Windows are ordered by the primary gene score within each class, so the axis reads left to right as the ranking the variant-level estimator induces. Points above the dashed 5×10^{-8} line are drawn at full opacity. Horizontal position inside a slot is jittered rather than genomic, because position within one window is not comparable with position within another. The proportion of variants clearing the genome-wide threshold is essentially the same in every window regardless of class, which is the sense in which the estimator returns no ranking over these genes.

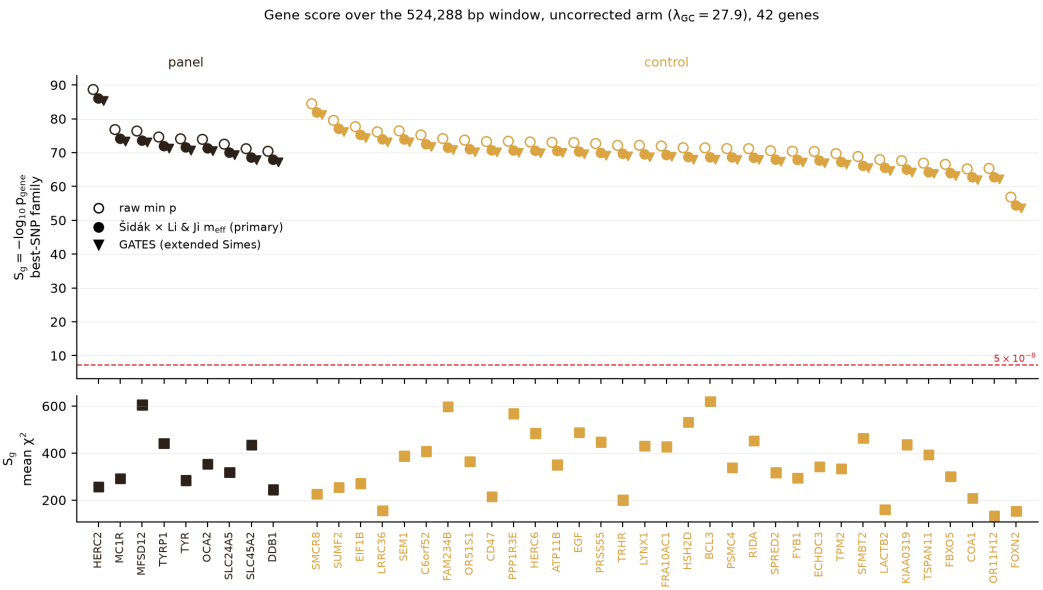


Figure 7: Gene score $S_g = -\log_{10} p_{\text{gene}}$ over the 524,288 bp window, for every panel and control gene, under each of the estimators of Appendix C. Genes are ordered by the primary estimator within the panel and within the controls, so the figure reads left to right as a ranking. *Upper panel:* the best-SNP family—the raw uncorrected minimum (open marker), the Šidák correction over the Li & Ji effective test count (filled circle, primary), and GATES (triangle). The three are drawn at small horizontal offsets because they agree to within a fraction of a marker on most windows. *Lower panel:* the aggregating family, the mean chi-squared over the window, on its own axis because its scores are two orders of magnitude larger. λ_{GC} is the genomic control factor over all windows tested.

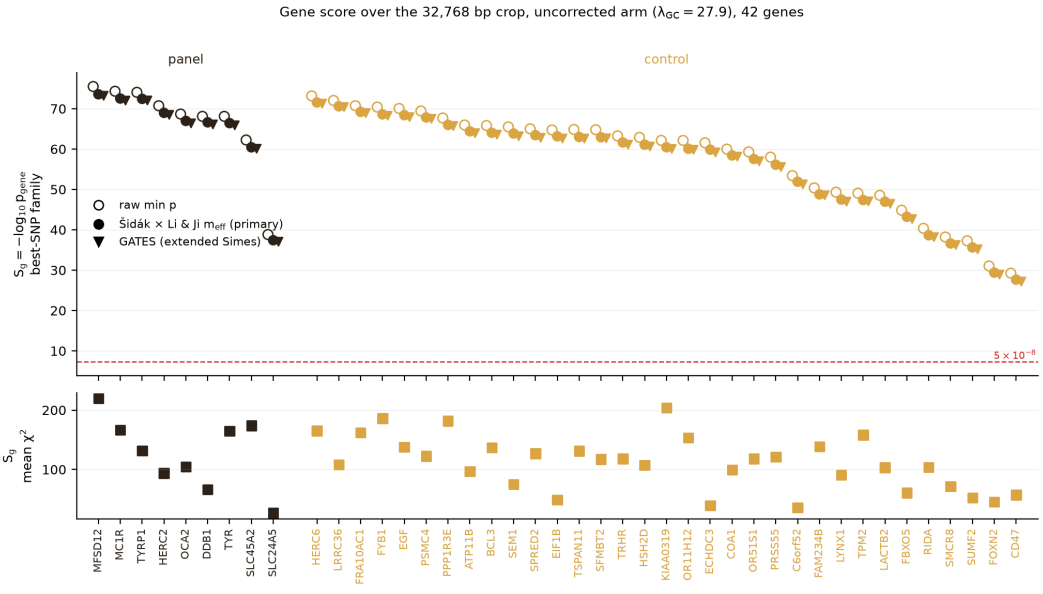


Figure 8: As Fig. 7, but scored over the central 32,768 bp crop—the region the classifier reads—rather than the full window. The pair is what separates a gene score that reflects the locus the classifier sees from one that reflects half a megabase around it.

H THE ONE WINDOW ADJACENCY

HERC2 lies immediately adjacent to OCA2 on chromosome 15 and the two windows share 234,772 bp, 44.8% of a window; every other pair among the 42 is disjoint, so this is the only adjacency a reader needs to discount. Repeating both panel-against-control tests with HERC2 dropped and OCA2 retained, since OCA2 is the gene the shared regulatory variant targets, leaves 41 arms in which the separation survives for both methods: AUC $= 0.867$ ($p = 0.0004$) for the classifier, down from 0.880 ($p = 0.0001$), and AUC $= 0.689$ ($p = 0.052$) for the Šidák estimator, down from 0.724 ($p = 0.021$). The classifier's separation is essentially unaffected; the association arm's significance moves from just below the conventional threshold to just above it. Discounting the adjacency therefore changes no comparison's conclusion, though it does move the association arm's own result in the direction of less significance.

I THE pAUC SWEEP

Table 4 tabulates the curve Fig. 9 draws, and Table 5 is the full sweep over the cap c against every baseline, reported in place of the pre-declared point alone so that the choice of c can be seen not to carry the result. Every baseline is read through the identical code path—`admitted`, `pauc`, `nulls` and `pvals` are called unchanged and only the second ranking's scores differ—so the five comparisons are one statistic applied five times, not five different measurements of similar things. The $c = 1$ column is ΔAUC exactly, which is the check that the two statistics are one family: $+0.157$ at $p = 0.266$ against the Šidák arm, reproducing Section 4.2 from the partial-area code path.

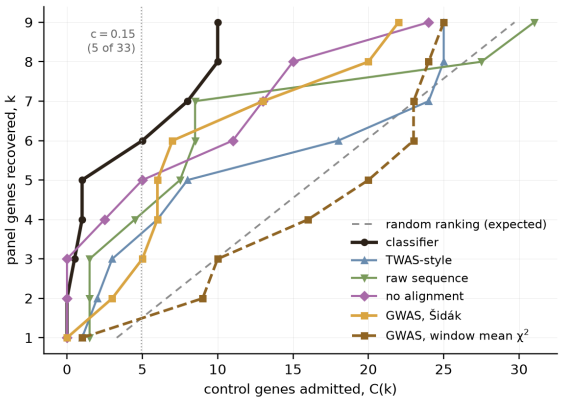


Figure 9: **The ROC curve of every ranking, in count units.** $C(k)$ is how many of the 33 controls a ranking has admitted when it recovers its k-th panel gene, so the axes are the false- and true-positive rates of Eq. 2 multiplied by n_0 and n_1; a curve further up and to the left is better. Shown for the classifier, its three ablations and the two association estimators. The dotted line is the pre-declared cap $c = 0.15$ and the area to its left is the pAUC the head-to-head is tested on; the dashed diagonal is a random ranking, admitting $3.3k$ controls in expectation.

The pre-declared cap, and what one gene is worth at it. At $c = 0.15$ the classifier's pAUC $=$ 0.499 stands against 0.155 for the Šidák minimum, 0.089 for mean χ^2, 0.199 for the TWAS-style scalar, 0.242 for raw sequence and 0.388 for no-alignment, the last being the only baseline that clears neither null at the pre-declared cap. The leave-one-gene-out repetition separates these cleanly. Against the two association estimators no single gene carries the result (worst case $p = 0.029$, dropping SLC24A5). Against the three ablations it does: dropping SLC24A5 takes the TWAS comparison to $d = +0.213$, $p = 0.118$, and dropping HERC2 takes raw sequence to $d = +0.177$, $p = 0.104$ and no-alignment to $d = +0.013$, $p = 0.663$. The head-to-head against variant-level association is therefore the robust claim, and the ablations are reported as evidence about mechanism rather than as significance tests.

Table 4: Control genes admitted at the threshold that recovers the k-th panel gene (Section 3.5), for the classifier and for both association estimators. Lower is better; a tie contributes $1/2$. The classifier's curve is flat at or below one admitted control until its fifth panel gene.

k	$C_{\mathrm{cls}}(k)$	$C_{\check{\mathrm{S}}}(k)$	$C_{\bar{\chi}^2}(k)$	ΔC vs Šidák
1	0.0	0.0	1.0	+0.0
2	0.0	3.0	9.0	+3.0
3	0.5	5.0	10.0	+4.5
4	1.0	6.0	16.0	+5.0
5	1.0	6.0	20.0	+5.0
6	5.0	7.0	23.0	+2.0
7	8.0	13.0	23.0	+5.0
8	10.0	20.0	24.0	+10.0
9	10.0	22.0	25.0	+12.0
out of 33 control genes				

Table 5: $\mathrm{pAUC}(c)$ (Eq. 2) over every cap and every baseline, with the classifier's advantage d and its two-sided p under both nulls of Section 3.5: *label*, the label-reassignment null used throughout the paper, and *method*, the stricter gene-by-gene method-exchange null. The classifier's own row is shared by all five comparisons. The pre-declared $c = 0.15$ is set in bold; at $c = 1$ the statistic is $\Delta\mathrm{AUC}$ and the column reproduces Section 4.2. Monte Carlo over 200,000 draws per null.

ranking		0.05	0.10	**0.15**	0.20	0.30	0.50	1.00
classifier	pAUC	0.3872	0.4714	0.4994	0.5404	0.6038	0.7609	0.8805
GWAS Šidák	pAUC	0.1111	0.1212	0.1549	0.2189	0.3636	0.5084	0.7239
	d	+0.2761	+0.3502	+0.3446	+0.3215	+0.2402	+0.2525	+0.1566
	p (label)	0.0138	0.0050	0.0073	0.0163	0.0937	0.1237	0.2663
	p (method)	0.1921	0.1053	0.1198	0.1644	0.3063	0.2126	0.1840
GWAS mean χ^2	pAUC	0.0438	0.0774	0.0887	0.0943	0.1100	0.2020	0.4916
	d	+0.3434	+0.3939	+0.4108	+0.4461	+0.4938	+0.5589	+0.3889
	p (label)	0.0044	0.0044	0.0057	0.0049	0.0043	0.0015	0.0079
	p (method)	0.0943	0.0710	0.0716	0.0543	0.0351	0.0104	0.0075
TWAS scalar	pAUC	0.0438	0.1313	0.1987	0.2424	0.3311	0.4209	0.6229
	d	+0.3434	+0.3401	+0.3008	+0.2980	+0.2727	+0.3401	+0.2576
	p (label)	0.0046	0.0097	0.0303	0.0435	0.0862	0.0403	0.1023
	p (method)	0.1873	0.2060	0.2430	0.2500	0.2675	0.1144	0.0835
raw sequence	pAUC	0.0303	0.1818	0.2424	0.2929	0.4018	0.5522	0.6902
	d	+0.3569	+0.2896	+0.2570	+0.2475	+0.2020	+0.2088	+0.1902
	p (label)	0.0014	0.0121	0.0276	0.0586	0.1481	0.1791	0.1785
	p (method)	0.1267	0.2446	0.2767	0.2970	0.3973	0.2769	0.1954
no alignment	pAUC	0.3333	0.3603	0.3883	0.4293	0.4714	0.5758	0.7626
	d	+0.0539	+0.1111	+0.1111	+0.1111	+0.1324	+0.1852	+0.1178
	p (label)	0.1336	0.0520	0.0815	0.1132	0.1371	0.0996	0.2369
	p (method)	0.8126	0.4936	0.4797	0.4899	0.4007	0.1563	0.1437

This is also the one place where the two readouts of Section 3.5 disagree on significance: on AUPRC the no-alignment comparison gives $d = +0.119$, $p = 0.047$, clearing the label-reassignment null that the pAUC does not. The disagreement does not survive the same repetition—dropping HERC2 alone takes that $p = 0.047$ to 0.458—so we read it as one gene carrying the weaker statistic rather than as evidence that the alignment contributes to the margin, and the conclusion of Section 4.3 is unchanged.

J PER-GENE RESULTS AND THE TWO RANKINGS

Table 6: **Where each panel gene lands in the two rankings.** Classifier rank is by test-split balanced accuracy, ties sharing the best rank; GWAS rank is by the Šidák minimum at window resolution (Fig. 10 draws all 42 genes). SLC24A5 carries the largest-effect known light-skin allele, HERC2 carries rs12913832, and OCA2 is the gene that variant regulates.

	classifier		GWAS, Šidák
panel gene	bal. acc. (test)	rank of 42	rank of 42
SLC24A5	0.991	1	20
SLC45A2	0.964	2	28
HERC2	0.940	3	1
TYRP1	0.933	5	10
MC1R	0.907	6	5
DDB1	0.897	11	31
MFSD12	0.887	15	8
TYR	0.872	18	11
OCA2	0.869	19	13
controls above the worst panel arm: 10 of 33 (classifier); 22 of 33 (GWAS)			

Table 7: Accuracy of one classifier per gene on the held-out test split, against its majority-class baseline. *bal. acc.* is the mean of the two per-class recalls, *rec. strong* and *rec. weak*; *acc.* and the recalls are also on the test split. m is the number of variants in the gene's window surviving the frequency filter, repeated from the association arm because it is the quantity an accuracy on an ancestry-determined label is most likely to be tracking. Arms are ordered within class by balanced accuracy.

gene	bal. acc.	acc.	rec. strong	rec. weak	m
Panel (9 arms)					
SLC24A5	0.991	0.988	0.983	1.000	1 943
SLC45A2	0.964	0.975	0.991	0.936	2 399
HERC2	0.940	0.932	0.922	0.957	2 536
TYRP1	0.933	0.932	0.930	0.936	3 269
MC1R	0.907	0.895	0.878	0.936	3 554
DDB1	0.897	0.907	0.922	0.872	1 986
MFSD12	0.887	0.920	0.965	0.809	3 241
TYR	0.872	0.907	0.957	0.787	2 986
OCA2	0.869	0.895	0.930	0.809	3 072
Control (33 arms)					
SPRED2	0.940	0.932	0.922	0.957	2 958
FRA10AC1	0.904	0.926	0.957	0.851	3 422
KIAA0319	0.903	0.907	0.913	0.894	3 165
SEM1	0.901	0.914	0.930	0.872	2 593
EGF	0.898	0.926	0.965	0.830	2 945
FBXO5	0.895	0.895	0.896	0.894	2 998
C6orf52	0.892	0.883	0.870	0.915	3 521
PSMC4	0.888	0.895	0.904	0.872	2 956
LRRC36	0.882	0.895	0.913	0.851	1 335
SFMBT2	0.876	0.895	0.922	0.830	3 521
FYB1	0.866	0.907	0.965	0.766	2 952
OR51S1	0.863	0.895	0.939	0.787	4 159
PPP1R3E	0.846	0.870	0.904	0.787	2 682
PRSS55	0.835	0.846	0.861	0.809	4 219
ATP11B	0.796	0.809	0.826	0.766	2 354
SUMF2	0.793	0.858	0.948	0.638	2 843
OR11H12	0.780	0.821	0.878	0.681	1 785
COA1	0.765	0.827	0.913	0.617	2 327
TRHR	0.739	0.809	0.904	0.574	2 860
HERC6	0.737	0.778	0.835	0.638	2 218
EIF1B	0.730	0.778	0.843	0.617	2 562
RIDA	0.724	0.796	0.896	0.553	2 834
LYNX1	0.722	0.802	0.913	0.532	3 483
BCL3	0.722	0.802	0.913	0.532	3 207
SMCR8	0.705	0.759	0.835	0.574	2 124
LACTB2	0.697	0.784	0.904	0.489	2 450
FOXN2	0.689	0.790	0.930	0.447	2 331
ECHDC3	0.679	0.759	0.870	0.489	3 181
TSPAN11	0.650	0.772	0.939	0.362	3 318
TPM2	0.626	0.728	0.870	0.383	1 798
HSH2D	0.593	0.753	0.974	0.213	3 524
FAM234B	0.592	0.698	0.843	0.340	2 588
CD47	0.530	0.716	0.974	0.085	2 595
majority-class baseline	0.7099	0.7099	1.000	0.000	–

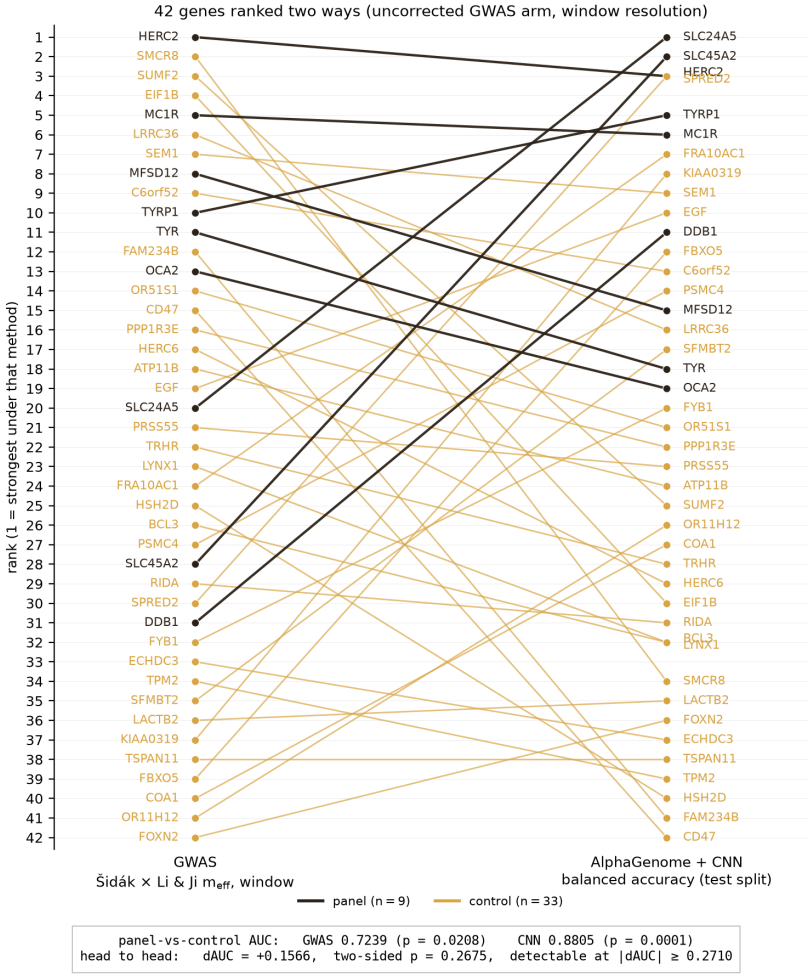


Figure 10: The same genes ranked by the variant-level association arm (left, primary estimator, 524,288 bp window) and by the per-gene classifier (right, balanced accuracy). Rank 1 is the strongest gene under that method; a line is one gene, dark for the panel and gold for the controls. The annotation reports, for each method, the probability that a panel gene outranks a control with the p-value of its own panel/control permutation null, and then the head-to-head difference in those two AUCs with the smallest $|\Delta\mathrm{AUC}|$ this sample size can detect at $\alpha = 0.05$. All 42 arms are shown. The classifier ranking is the test-split accuracy.