Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Reference choice affects positional enrichment of AlphaGenome variant rankings at APOL1-MYH9

Christopher Lawrence

Christopher.lawrence3@nhs.net

Short Report

Keywords: APOL1, AlphaGenome, DNase-seq, chromatin accessibility, variant prioritisation, reference selection, benchmarking

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11116147/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Objective.

We examined how the choice of measured chromatin-accessibility reference set affects interpretation of AlphaGenome variant rankings at the APOL1-MYH9 locus. This exploratory evaluation arose from a regulatory-variant prioritisation project and included a beta-globin comparison locus.

Results.

Among 1,930 common alternate alleles at 1,921 positions, the top 200 variants ranked by the absolute podocyte-labelled differential accessibility score were enriched in whole-kidney peaks (odds ratio 5.98, 95% CI 4.18–8.52) and in a podocyte-inclusive kidney union (8.76, 6.30–12.20). The podocyte estimate exceeded the strongest non-renal estimate numerically under both references, 5.98 against 5.46 for lung and 8.76 against 5.61, with greater separation under the union. We detected no change in the podocyte-minus-comparator enrichment contrast for any of six comparators after partitioning that union, under a conditional peak bootstrap. At the comparison locus the corresponding estimates were 1.44 (0.16–6.56) and 7.54 (3.66–15.31). These observations illustrate reference-dependent interpretation of positional enrichment. They do not establish tissue-specific variant-effect accuracy, equivalent performance between loci, or generalisation beyond training-overlapping data.

Introduction

Two coding haplotypes at APOL1, G1 and G2, account for much of the excess risk of non-diabetic kidney disease in people of recent African genetic ancestry [1], an association first mapped to the adjacent gene MYH9 [2]. Penetrance is incomplete, and transgenic expression of the risk variants in podocytes causes dose-dependent kidney disease in mice [3], so regulatory modifiers of APOL1 expression are a plausible contributor to who develops disease [4]. The locus is overwhelmingly non-coding and largely uncharacterised. Sequence-to-function models predict regulatory activity from DNA and are proposed for prioritising such variants [5, 6], though earlier models of this family perform less well in regions that distinguish cell types [7].

We set out to prioritise regulatory candidates here. That attempt did not yield a defensible candidate list, but in the course of it we found that the apparent answer to a simple question, whether a tissue-labelled ranking concentrates variants in measured accessible chromatin, changed with the reference set used to evaluate it. We report that sensitivity, because the reference sets involved are all defensible choices and the differences between them are easy to overlook.

Methods

Variants. The interval spanning APOL3 to MYH9, chr22:36,140,330 – 36,388,018 (GRCh38, 247,688 bp) was evaluated. Variants were retrieved on 13 September 2026 from the gnomAD v4 [8] GraphQL application programming interface (API) [9] (dataset gnomad_r4, GRCh38 region query), taking allele counts from the afr block of the genome callset with exome fallback; the raw response is deposited. Retaining single-nucleotide substitutions at African-ancestry allele frequency of at least 1% gave **1,930 distinct alternate alleles at 1,921 positions**. The unit of analysis is a unique build, chromosome, position, REF and ALT combination; nine positions are multi-allelic, contributing two alleles each. A deduplication ledger is deposited.

Predictions. Predictions came from the AlphaGenome public API [6] using the ALL_FOLDS model version, the default when none is specified, with the recommended DNase scorer. In client 0.9.0 that is CenterMaskScorer(requested_output=DNASE, width = 501, aggregation_type=DIFF_LOG2_SUM), defined in that client as log2(sum(ALT)) - log2(sum(REF)) over a 501 bp window centred on the variant. Scoring is performed server-side and the public documentation describes the same score with a pseudocount, so we report the pinned client's definition and do not assert the server's internal constant. Of the 305 DNase tracks returned, each ranking selects those whose biosample_name matches exactly and averages them; exactly one matched for every ranking, so no averaging occurred. The track manifest is deposited. Each variant was scored in a 1,048,576 bp context centred on it (start = max(0, pos − 524288), pos the 1-based gnomAD coordinate). Predictions were obtained on 13 and 15 September 2026. Rankings use the **absolute value of that score**, sorted descending, with no quantile calibration; "raw" here means uncalibrated, not untransformed. Ties are broken arbitrarily, but no tie spans a reported cutoff at any depth for any ranking, so tie order changes no result. Seven tissue outputs were ranked: podocyte (glomerular visceral epithelial cell), whole kidney, hepatocyte, liver, lung, brain and stomach, a different list from the six tissues used below to define shared peaks. Because the scorer is differential, the endpoint tests whether predictions concentrate at accessible positions, not which allele changes it.

This evaluation is not independent of model training. AlphaGenome is trained on ENCODE DNase-seq coverage, and every kidney biosample used here corresponds to a named training experiment (ENCSR206OJJ, ENCSR785BDQ, ENCSR000EPW, ENCSR000EOL, ENCSR000EOK, ENCSR175IWT). The served ALL_FOLDS version is described as distilled from teachers trained across all folds, so no genomic region is held out from it; the API also exposes FOLD_0 to FOLD_3. We did not establish which released checkpoint, if any, holds this locus out, and make no claim about a held-out evaluation. What follows are consistency checks between predictions and the data underlying them.

Measured reference sets. ENCODE DNase-seq peak files (GRCh38, released, peak output) [10] were merged over each locus. Coordinates are GRCh38 BED half-open intervals and touching intervals merge (next start not greater than current end). Peaks are classified unclipped, so one merged peak of the podocyte-inclusive union at APOL1-MYH9 extends 447 bp beyond the interval (369 bp of it also present in the whole-kidney reference); coverage percentages are computed from the intersection with the evaluated interval, and no peak overhangs at the comparison locus. Six locus and reference combinations were evaluated (Table 1): whole kidney alone (term_name=kidney); a **podocyte-inclusive**

kidney union adding glomerular visceral epithelial cell ; that union partitioned into **kidney-restricted** and **shared**; and both whole-kidney and podocyte-inclusive definitions at the comparison locus. A merged kidney peak is classified **shared** if a peak from any of six comparison tissues (liver, lung, heart left ventricle, stomach, spleen, thyroid gland; four files per tissue) **covers its midpoint**, and **kidney-restricted** otherwise, so part of a kidney-restricted peak may still overlap a comparison peak. Partitions are mutually exclusive and sum to the union in peaks, base pairs and alleles, asserted in code. Within each locus the eligible allele set and each tissue ranking were held fixed across reference definitions; the loci have different allele sets. A peak manifest listing every file, term and contribution is deposited in the analysis archive [11].

Table 1

Positional enrichment at depth 200 for each locus and measured reference set. The podocyte ranking is held fixed across reference definitions; the comparator column names the strongest non-renal ranking evaluated at the same locus, reference and depth. Odds ratios are conditional maximum likelihood with exact 95% intervals. Full contingency cells for every endpoint and depth are deposited.

Locus	Reference set	Peaks	bp (% locus)	n	Podocyte OR (95% CI)	Strongest non-renal OR (95% CI)
APOL1-MYH9	Whole kidney	67	29989 (12.1)	1930	5.98 (4.18–8.52)	lung 5.46 (3.81–7.79)
APOL1-MYH9	Podocyte-inclusive union	110	44129 (17.8)	1930	8.76 (6.30–12.20)	lung 5.61 (4.02–7.81)
APOL1-MYH9	Union, kidney-restricted	39	14686 (5.9)	1930	4.30 (2.53–7.15)	lung 2.49 (1.36–4.36)
APOL1-MYH9	Union, shared	71	29443 (11.9)	1930	7.94 (5.55–11.33)	lung 6.02 (4.18–8.62)
beta-globin	Whole kidney	10	2598 (1.0)	1929	1.44 (0.16–6.56)	not evaluated
beta-globin	Podocyte-inclusive union	18	5851 (2.4)	1929	7.54 (3.66–15.31)	not evaluated

Coverage percentages are of the 247,688 bp evaluated at each locus, computed from the intersection of merged peaks with that interval. Each estimate is specific to the endpoint its row defines.

Statistics. Enrichment of the top-ranked alleles was tested against the remaining common alleles at the same locus by Fisher exact test, with conditional maximum-likelihood odds ratios and exact intervals; all four contingency cells and the denominator are deposited for every endpoint. Depth 200 is primary; 25, 50 and 100 are nested sensitivity analyses. These intervals assume independence between alleles, which clustering within peaks violates, so they are descriptive, not tests of ranking differences.

Resampling. Each of 500 bootstrap replicates draws, with replacement, 27 kidney-restricted merged peaks, these being the peaks of the 39 that contain at least one eligible allele. A peak drawn k times contributes its allele records k times, and a peak not drawn contributes none. All eligible records outside

kidney-restricted peaks are appended once. Each tissue ranking is then recomputed over that whole replicate pool and its top 200 records taken, so membership of the top set is a consequence of reranking and an allele in the observed top 200 need not survive it. The fixed part of the pool is fixed in its records and scores, not in its contingency membership. One list of draws is generated (seed 20260921, numpy PCG64) and every comparator and both reference definitions are evaluated on each draw, so the six estimates are paired; alternate alleles at one position always travel together. The observed point estimate and every replicate use the same estimator, a Haldane-corrected log cross-product ratio, which is deliberately not the conditional maximum-likelihood odds ratio of Table 1 and is not interchangeable with it. Intervals are the 2.5th and 97.5th percentiles, unadjusted across six comparators, and conditional on the fixed background. Nine further configurations (three replicate counts by three seeds), yielding 54 comparator-specific intervals, are deposited, with pseudocode.

Comparison locus. The beta-globin cluster, chr11:5,150,000–5,397,688, matched on length and paralogue-cluster structure and **approximately count-matched** by subsampling stratified on African-ancestry allele frequency in eight bins: 1,929 alleles against 1,930, because in the 0.10 to 0.20 stratum the test locus required 463 and the control held 462.

Results

Whole-kidney reference. Against whole-kidney peaks alone (67 peaks, 29,989 bp within the locus, 12.1%) the podocyte ranking gave an odds ratio of 5.98 (95% CI 4.18–8.52) at depth 200, numerically above the strongest non-renal ranking at that depth, lung at 5.46 (3.81–7.79), by a small margin. In a nested sensitivity analysis at depth 100 that ordering reversed: podocyte 5.25 and hepatocyte 6.72.

Podocyte-inclusive union. Against the union (110 peaks, 44,129 bp within the locus, 17.8%) the podocyte ranking gave 8.76 (6.30–12.20) against 5.61 (4.02–7.81) for lung (Fig. 1). The separation is larger than under the whole-kidney reference: a difference in the size of a numerical gap, not a demonstration that discrimination is absent under one reference and established under the other.

Partitions. The union partitions into 39 kidney-restricted peaks (14,686 bp) and 71 shared (29,443 bp within the locus), containing 84 and 193 of the 277 alleles in the union. The podocyte ranking led in both, 4.30 (2.53–7.15) and 7.94 (5.55–11.33). The change in the podocyte-minus-comparator contrast on partitioning ranged from − 0.07 to + 0.42 in natural-log odds ratio and no interval excluded zero for any of six comparators (Fig. 2). We detected no change in relative enrichment under this resampling scheme. That is not evidence of equivalence, and it answers a different question from the whole-kidney to union comparison above.

Comparison locus. Against whole-kidney peaks the beta-globin cluster gave 1.44 (0.16–6.56, one-sided p = 0.43; 10 peaks, 2,598 bp). Against the podocyte-inclusive union it gave 7.54 (3.66–15.31; 18 peaks, 5,851 bp). Positional enrichment is therefore **also observed at the comparison locus** under the union definition. We do not read that as equivalent performance: the loci differ in the architecture evaluated, with union coverage of 17.8% at APOL1-MYH9 against 2.4% here, and overlapping intervals are not an

equivalence test. Equally, the wide whole-kidney interval is too imprecise to establish an absence of enrichment.

Discussion

Reference definition changed the magnitude and the interpretation of positional enrichment at both loci. At APOL1-MYH9 the podocyte estimate was numerically above the strongest non-renal estimate under both references, with greater separation under the podocyte-inclusive one. Adding podocyte files changes coverage and peak composition as well as cell-type representation, and these analyses do not isolate those contributions.

The beta-globin comparison also showed enrichment under that definition. That does not establish equivalent performance between the loci, and the imprecise whole-kidney estimate does not establish its absence. Likewise the conditional partition analysis detected no change in the relative enrichment contrast without establishing equivalence.

Three practical safeguards follow. Specify the measured reference and the predicted output together, since neither alone fixes the endpoint. Verify in code that partitions reconstruct their parent in peaks, base pairs and alleles before interpreting any difference between them; an earlier version of this analysis compared partitions against a pooled set built on a different reference. Apply consistent reference definitions across loci. Because the evaluation overlaps model training and tests position rather than allelic effect, it does not validate tissue-specific variant-effect accuracy or identify an APOL1 regulatory modifier.

Limitations

This is an exploratory analysis of one primary locus and one comparison locus, with no pre-registered plan, and the comparisons reported were arrived at iteratively.

The evaluation is not independent of model training, which is the binding limitation: the tracks are named training experiments and the served model is distilled from teachers trained across all folds. We identified no released checkpoint holding this locus out, so these results speak to internal consistency, not generalisation.

The endpoint is positional. A differential score ranked against peak membership does not test which allele changes accessibility, nor the direction or size of any effect, and we did not compare it against a reference-accessibility baseline, so we cannot say what the differential scorer adds over ranking by predicted accessibility alone.

The bootstrap resamples kidney-restricted peaks and holds everything else fixed, including alleles in shared peaks, so its intervals are conditional on that background rather than a general estimate of sampling uncertainty. They are unadjusted for six comparisons. Resampling whole peaks preserves

within-peak clustering but does not account for dependence between separate peaks. The deposited configurations support the stability of zero inclusion; they do not establish precise endpoint stability. A non-significant interaction is not evidence of no interaction, and we specify no margin below which a change would be negligible.

"Kidney-restricted" is defined by a midpoint rule against six comparison tissues and would change with a different panel or rule.

The prioritisation exercise that prompted this work yielded no candidate list reportable under its specified procedure of linkage and prior-literature screening; those intermediate analyses are deposited, and nothing here establishes or excludes an APOL1 regulatory modifier.

Conclusion

Reference choice affected the observed positional enrichment and its interpretation at the two evaluated loci. Consistent reference definitions and reconciled partitions are necessary for interpretable comparisons. These training-overlapping evaluations do not establish allele-specific regulatory accuracy or an APOL1 modifier.

Abbreviations

API
application programming interface
bp
base pairs
CI
confidence interval
DNase
deoxyribonuclease
ENCODE
Encyclopedia of DNA Elements
gnomAD
Genome Aggregation Database
GRCh38
Genome Reference Consortium Human Build 38
OR
odds ratio
REF/ALT
reference/alternate allele.

Declarations

Ethics approval and consent to participate. Not applicable. No new participant data were collected; the study analysed existing publicly available ENCODE and gnomAD data together with model predictions.

Consent for publication. Not applicable.

Availability of data and materials. Variant frequencies are from gnomAD v4 [8, 9] and peak calls from the ENCODE portal [10]. All analysis code, the frozen allele ledger, per-variant predictions, contingency counts for every endpoint, the ENCODE peak manifest, the track manifest, partition labels, the ranking provenance record, the six change-in-contrast estimates with intervals, and the bootstrap stability table are archived at Zenodo [11]. Model predictions were obtained under AlphaGenome's non-commercial research terms and are redistributed subject to them; model weights were not accessed.

Competing interests. The author declares no competing interests.

Funding. No funding was received.

Authors' contributions. CL is the sole author and conceived the study, wrote the code, performed the analyses and wrote the manuscript.

Acknowledgements. During the preparation of this work the author used Claude (Opus 5, Anthropic) to search and verify the published literature, to check quoted figures and citations against primary sources, and to draft and revise the text and analysis code. After using this tool the author reviewed and edited the content as needed and takes full responsibility for the content of the publication. The tool is not listed as an author.

References

1. Genovese G, Friedman DJ, Ross MD, Lecordier L, Uzureau P, Freedman BI, et al. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. Science. 2010;329:841–5.
2. Kopp JB, Smith MW, Nelson GW, Johnson RC, Freedman BI, Bowden DW, et al. MYH9 is a major-effect risk gene for focal segmental glomerulosclerosis. Nat Genet. 2008;40:1175–84.
3. Beckerman P, Bi-Karchin J, Park AS, Qiu C, Dummer PD, Soomro I, et al. Transgenic expression of human APOL1 risk variants in podocytes induces kidney disease in mice. Nat Med. 2017;23:429–38.
4. Ojo AO, Adu D, Bramham K, Freedman BI, Gbadegesin RA, Ilori TO et al. APOL1 kidney disease: conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. Kidney Int. 2025;108:763 – 79.
5. Linder J, Srivastava D, Yuan H, Agarwal V, Kelley DR. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. Nat Genet. 2025;57:949–61.
6. Avsec Ž, Latysheva N, Cheng J, Novati G, Taylor KR, Ward T, et al. Advancing regulatory variant effect prediction with AlphaGenome. Nature. 2026;649:1206–18.

7. Kathail P, Shuai RW, Chung R, Ye CJ, Loeb GB, Ioannidis NM. Current genomic deep learning models display decreased performance in cell type-specific accessible regions. Genome Biol. 2024;25:202.
8. Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genomic mutational constraint map using variation in 76,156 human genomes. Nature. 2024;625:92–100.
9. Genome Aggregation Database. gnomAD v4, GraphQL API dataset identifier gnomad_r4, GRCh38 region query, afr genetic-ancestry group, genome callset with exome fallback. https://gnomad.broadinstitute.org/api (accessed 13 September 2026). The raw API response is deposited with the analysis archive.
10. ENCODE Project Consortium. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583:699–710. Individual experiment and file accessions are listed in the deposited peak manifest.
11. Lawrence C. apol1-alphagenome: analysis code and frozen results, version v1.0-bmcrn. Zenodo; 2026. doi:10.5281/zenodo.22875737. Also at https://github.com/PatientThread/apol1-alphagenome.

Figures

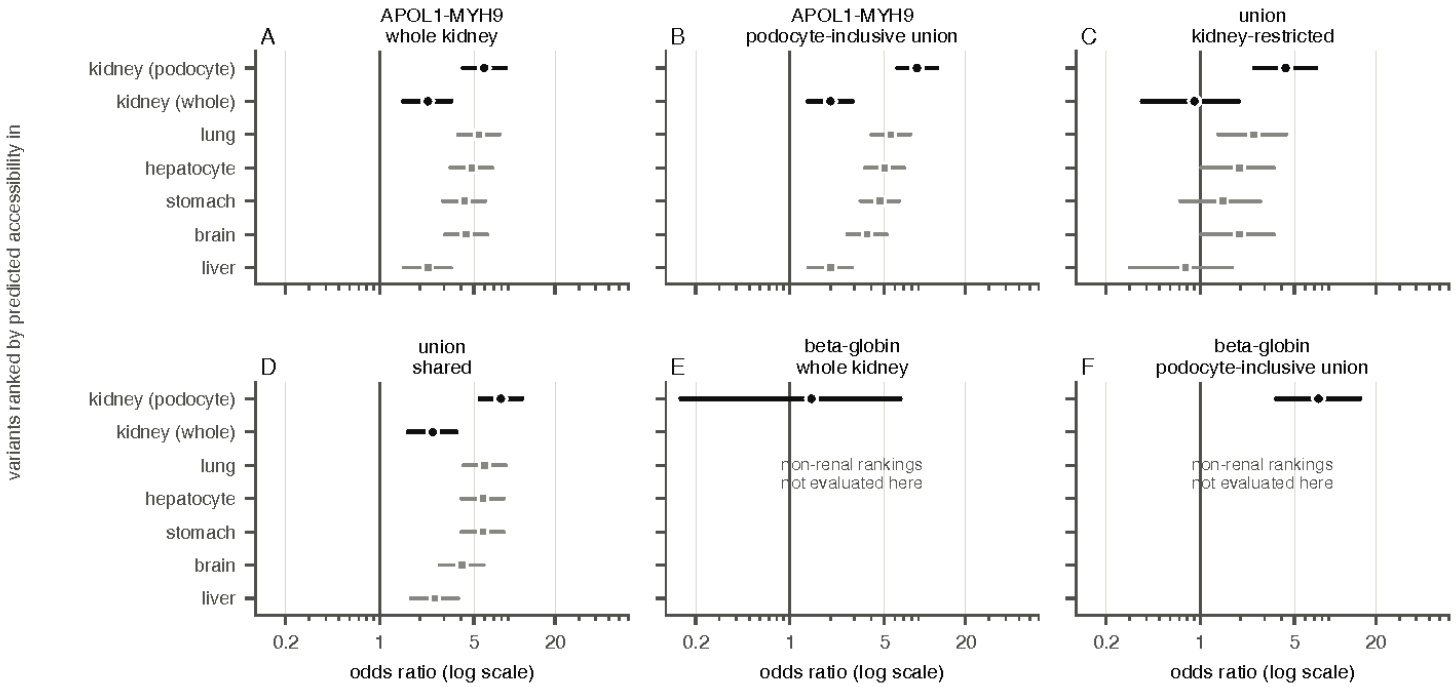


Figure 1

Positional enrichment of the top 200 alleles by the tissue output used to rank, across the six locus and reference combinations of Table 1. Panels: (A) APOL1-MYH9, whole kidney; (B) APOL1-MYH9, podocyte-inclusive union; (C) union, kidney-restricted; (D) union, shared; (E) beta-globin, whole kidney; (F) beta-globin, podocyte-inclusive union. Conditional maximum-likelihood odds ratios with exact 95% intervals

against the remaining common alleles at the same locus, all at depth 200; eligible n = 1,930 at APOL1-MYH9 and 1,929 at beta-globin. Common log scale with OR = 1 marked; tissue order identical in every panel; podocyte and whole kidney labelled separately and drawn in black, non-renal rankings in grey. Only the podocyte ranking was evaluated at the comparison locus, so panels E and F show that estimate alone. Intervals assume independence between alleles.

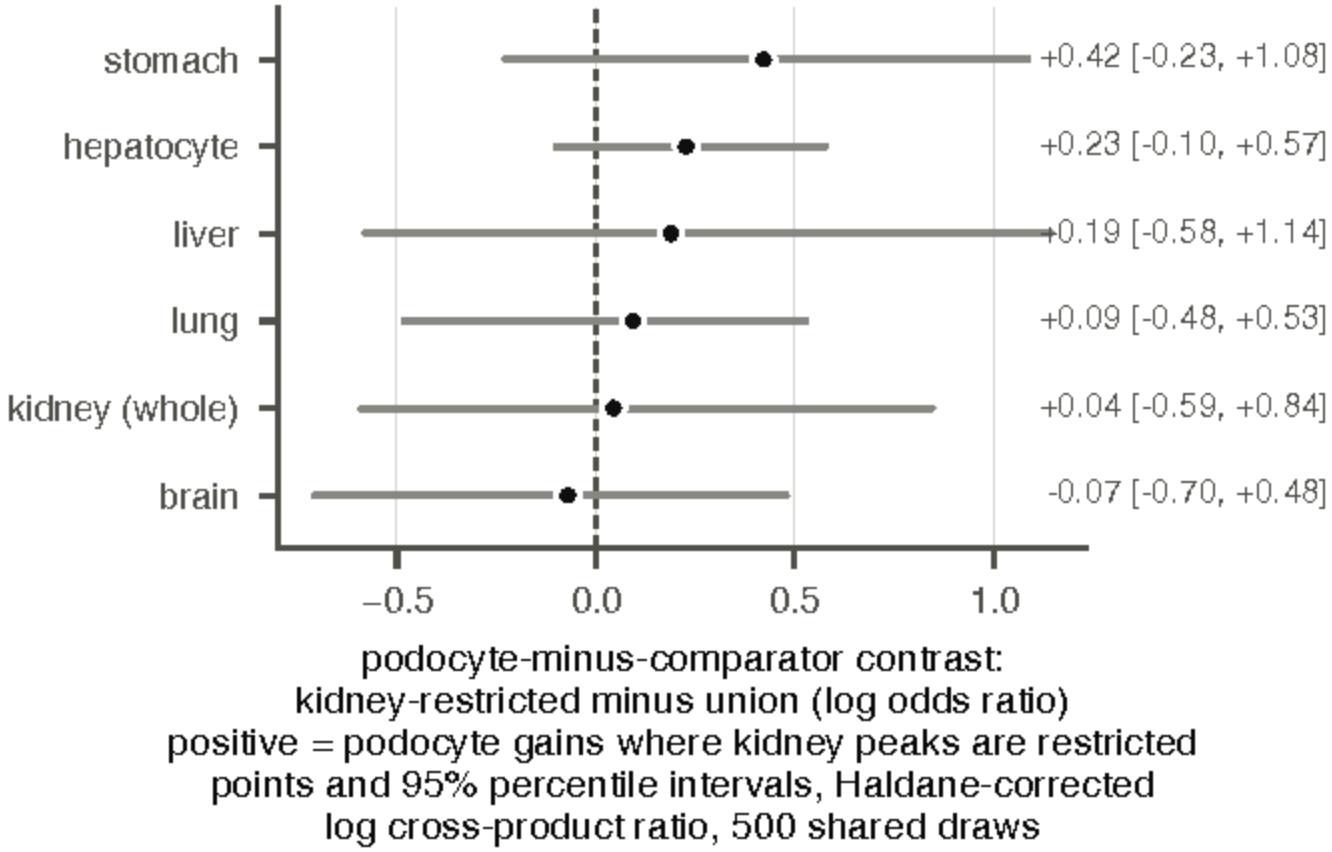


Figure 2

Change in tissue contrast on partitioning. For each of six comparator rankings, points show the podocyte-minus-comparator contrast in kidney-restricted regions minus the corresponding contrast in the podocyte-inclusive union, on the natural-log odds-ratio scale. Positive values indicate a larger relative podocyte advantage in restricted regions. Each of 500 paired bootstrap replicates sampled 27 allele-containing restricted peaks with replacement, appended the fixed non-restricted records once, and reselected the top 200 records for each ranking. Observed points and replicate contrasts use Haldane-corrected log cross-product ratios, whereas the estimates in Table 1 are conditional maximum-likelihood odds ratios. Bars show conditional 95% percentile intervals (2.5th to 97.5th percentiles), unadjusted for six comparisons, with numerical limits beside each point. No interval excludes zero; this does not establish equivalence.