Supplementary Information

Broad vertebrate conservation of Treg-associated molecular components constrains models of regulatory T-cell evolution

Nabanita Ghosh[1,*]
Krishnendu Sinha[2]

[1]Department of Zoology, Maulana Azad College, Kolkata, West Bengal, India
[2]Department of Zoology, Siksha Bhavana, Visva-Bharati, Santiniketan, West Bengal, India
[*]nabanitaghosh89@gmail.com

1 Purpose and organization

This Supplementary Information accompanies the manuscript and follows the same evidence order: study design, molecular sets, duplication-aware orthology, genomic representation and specificity, mammalian FOXP3 stage dependence, comparative cellular feasibility, and promoter analysis. It supplies details needed to reproduce or audit the work without repeating the main narrative. The 34 machine-readable tables cited below are included alongside this PDF in the reproducibility package prepared for Zenodo DOI 10.5281/zenodo.22735270; public resolution of that DOI must be confirmed before submission.

2 Supplementary Methods

2.1 Prospective design and evidence boundaries

Four evidence layers were kept separate: genomic family availability, source-defined cellular deployment, mammalian FOXP3 dependence, and regulatory sequence or chromatin features. The de novo component-innovation and ancestral-programme capture/individuation hypotheses were both retained; generic immune conservation and regulatory specialization were explicit alternatives. Thresholds, controls, missing-data states, and stopping rules were fixed before each outcome was opened.

Animals, donors, pools, biological libraries, technical libraries, runs, lanes, and cells were recorded separately. Only resolved animals or biological libraries were inferential units. Unknown nesting remained unresolved. Non-observation, reference limitation, search-negative evidence, supported absence, and unresolved synteny remained distinct. Supplementary Table S1 records the prospective decision path. Supplementary Fig. S1 and Table S33 show the prospective validation of the initial candidate programme; Fig. S8 shows its equivalence-margin sensitivity. This initial programme was retired without redefining membership.

2.2 Molecular sets and comparators

The primary core contained eight source genes assigned to six equal-weight suppressive mechanisms: *Ctla4*; *Il10*; *Tgfb1*; *Ebi3* and *Il12a*; *Entpd1* and *Nt5e*; and *Gzmb*. The broader mammalian Treg-dependency set contained 19 source rows that resolved to 17 unique families. A broader sensitivity version of each set never replaced its primary definition. Eight comparator programmes captured activation, interferon, cytotoxic, B-cell, helper-T-cell, stress, cell-cycle, and housekeeping biology. Supplementary Tables S2–S4 provide row-level definitions, evidence and overlap records.

2.3 Duplication-aware orthology

Exact, checksum-recorded proteomes represented 13 focal taxa plus four backbone taxa (*Petromyzon marinus*, *Callorhinchus milii*, *Lepisosteus oculatus*, and *Latimeria chalumnae*). The primary policy selected the longest translated isoform per locus; a source/canonical-isoform policy was retained as sensitivity. OrthoFinder 3.1.4 produced the 17-taxon primary coordinate. All paralogues were retained. Copy number, relationship class, duplication complexity, external agreement, and reference risk were stored separately.

Coordinate comparisons evaluated bounded claim equivalence, not identifier equality. Seventeen review units covering 18 component rows produced five equivalent and 12 equivalent-with-restrictions decisions; none was classified as non-equivalent or unresolved. Restrictions prohibit exact single-copy or duplication-node claims where support is insufficient. Supplementary Tables S5 and S6 give mappings and adjudications; Supplementary Fig. S6 and Tables S12 and S17 show method and coordinate sensitivities.

2.4 Genomic representation, recovery and specificity

Primary occupancy was the fraction of 13 focal taxa with an observed source-compatible member. Families, rather than taxa, were resampled for descriptive bootstrap intervals. Six mechanisms received equal weight. Broad nonmammalian representation required evidence in more than three nonmammalian lineage bins. Leave-one-taxon, leave-one-family, leave-one-mechanism, coordinate, broad-set and complex-family sensitivities tested robustness without treating taxa as independent replicates.

Eligible non-observations were reviewed with annotation searches, sequence/profile searches, translated-genome evidence, locus integrity and local synteny. Supported absence required a callable, eligible reference and all frozen locus conditions; otherwise the case remained restricted or unresolved. Supplementary Figs. S3 and S4 show core and broader-dependency matrices, and Fig. S5 summarizes recovery uncertainty. Row-level states are in Tables S7–S13.

Specificity matching preserved structural features, including copy class, orthology class, reference-quality stratum, eligibility, overlap and paralog complexity. Numerical covariates were binned prospectively; no family could be reused within a panel. The core-versus-broader-dependency comparison had no complete panel. Five immune controls produced 10,000 panels each, but none showed greater core breadth under the joint decision rule. Supplementary Fig. S7 and Tables S14–S19 contain the complete matching and sensitivity outputs.

2.5 Mammalian FOXP3 stage-dependence reanalysis

Establishment/acquisition and mature-maintenance contexts remained separate. A stringent gene required adjusted $P \leq 0.05$, absolute $\log_2$ fold change ≥ 0.5, same-direction evidence in at least two eligible contexts and no strong opposite context. The frozen sets contained 67 positive and 17 negative establishment/acquisition-associated genes and one negative mature-maintenance-associated gene. Contexts were not pooled as interchangeable experiments. Supplementary Fig. S9 shows effect concordance; Fig. S10 provides descriptive promoter-occupancy context; Table S34 supplies the verification set. The analysis confirms stage dependence reported in the source study but does not define direct targets or evolutionary timing.

2.6 Comparative cellular-data feasibility

Expression resources were audited processed-first before project-gene expression was opened. Eligibility required at least three resolved animals or biological libraries per state, compatible count semantics, source-native labels, a defensible gene-family crosswalk and sufficient cells or depth within each biological unit. No pre-amniote resource passed. Chicken PRJEB88203 resolved four paired animals but lacked the deposited Salmon count object. Human GSE201333 resolved three donors but contained nonexchangeable 10X UMI and Smart-seq2 read-count layers; a target stratum failed the 50-cell rule within either platform. Supplementary Fig. S2 and Table S20 record the gate. No cross-platform pooling or evolutionary cellular analysis was performed.

2.7 Promoter sequence and accessibility

Gene-anchored promoter windows at ±1, ±2 and ±5 kb allowed feature summaries without asserting nucleotide homology. Direct comparisons additionally required reciprocal same-family mapping, unambiguous correspondence and local synteny. JASPAR 2026 CORE vertebrate profiles were scanned after masking at frozen relative-score thresholds 0.80 and 0.85. Motif matches indicate sequence preference, not binding or activity. Structural matching used the same no-reuse logic described above and required at least 1,000 valid panels; none of the primary promoter comparisons met this requirement.

Human GSE118189 used three donors and a donor-blocked DESeq2 model. Mouse GSE297597 was analyzed separately at day 3 (four mice per genotype) and day 7 (three per genotype). A directional promoter required absolute $\log_2$ fold change ≥ 1 and adjusted $P < 0.05$. Complete promoter universes, direct-map records, matching diagnostics, liftover, model results, window/paralog sensitivities and stopping rules are in Supplementary Tables S21–S30.

3 Supplementary Results

The initial candidate FOXP3-independent programme did not validate coherently across two independent bulk resources and a mixed-chimera experiment (Supplementary Fig. S1; Table S33), and no equivalence-margin sensitivity changed that adjudication (Fig. S8). The programme was therefore excluded from evolutionary projection without post-outcome repair.

Eight core families had median occupancy 0.923 and mean occupancy 0.952; all six mechanisms had defensible nonmammalian representation. Seventeen broader-dependency families had median occupancy 1.000 and mean occupancy 0.973. Detailed matrices and recovery records retain every uncertainty and paralogue (Supplementary Figs. S3–S6; Tables S7–S13). Genomic control comparisons did not show greater breadth of the core set (Fig. S7; Tables S14–S19).

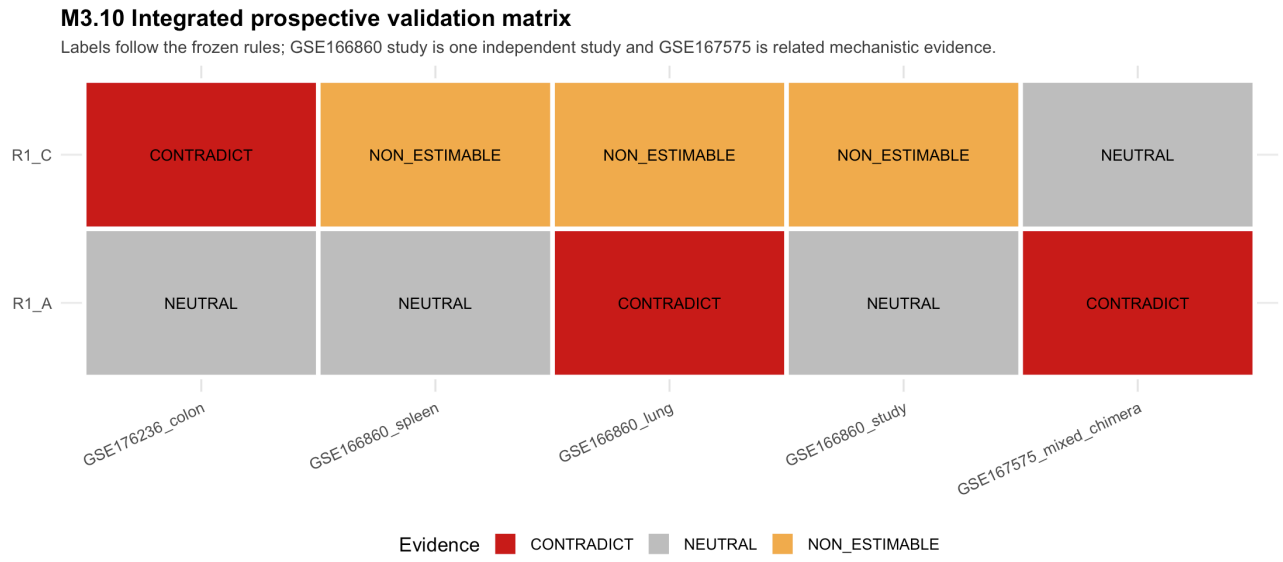
The confirmatory FOXP3 reanalysis contained 84 establishment/acquisition-associated genes versus one mature-maintenance-associated gene. Effect concordance varied by context, and promoter occupancy did not define direct targets (Supplementary Figs. S9 and S10; Table S34).

No cellular resource met every biological-replication and input-compatibility condition (Supplementary Fig. S2; Table S20). Direct homologous promoter evidence was sparse, and zero strict matched panels were obtained at any window. In human promoter accessibility, 581 of 582 programme promoters were nondirectional and one was negative-directional; in mouse, none of 961 tested promoters was directional at either time point (Tables S21–S30). Integrated genomic and regulatory classifications are in Tables S31 and S32; both preserve the conclusion that available evidence constrains but does not resolve the evolutionary mechanism.

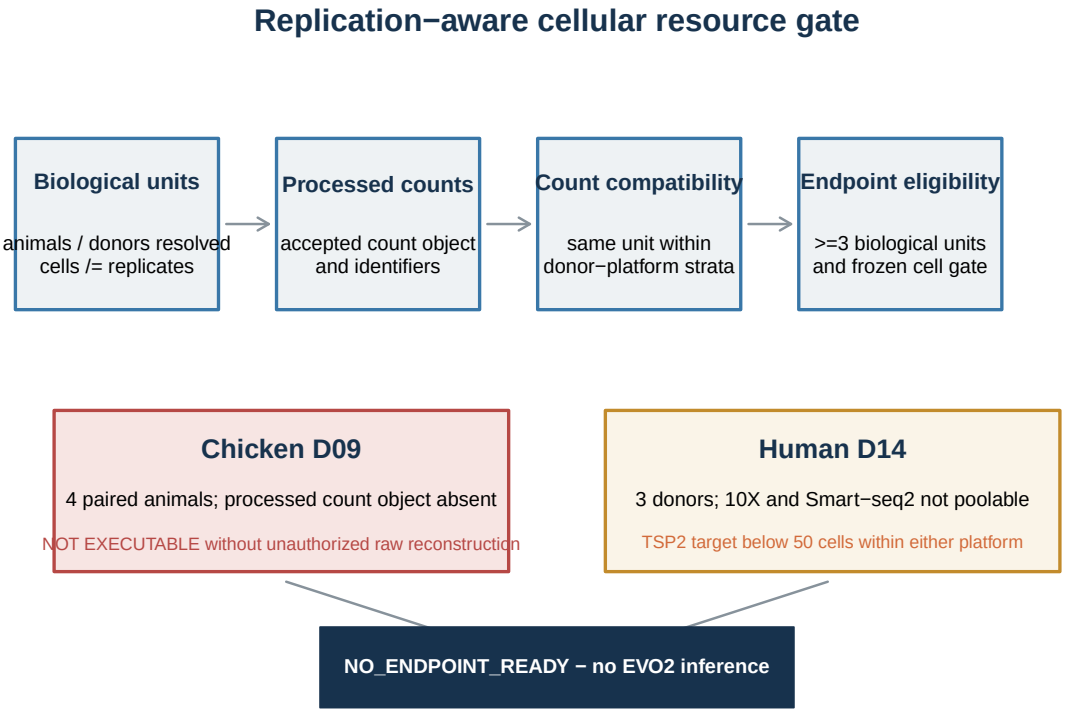
4 Internal code glossary

Code	**Plain-language meaning**
C0-A / C0-B	Primary eight-gene core suppressive-effector set / broader sensitivity set.
C1-A / C1-B	Primary broader mammalian Treg-dependency set / broader sensitivity set.
S0–S3	Four frozen orthology coordinates differing in software generation, taxon roster or representative-protein policy.
P1 / P2	Gene-anchored promoter evidence / direct nucleotide-homology-qualified promoter evidence.
R1-A / R1-C	Two frozen candidate FOXP3-independent programmes tested prospectively and then retired.
M9-D / M10-B4	Internal genomic and regulatory classifications meaning that the preregistered evolutionary alternatives were not discriminated.
REK2	Prespecified stop because fewer than 1,000 valid matched promoter panels were available.
NO_ENDPOINT_READY	Neither restricted cellular resource met every input and replication requirement.

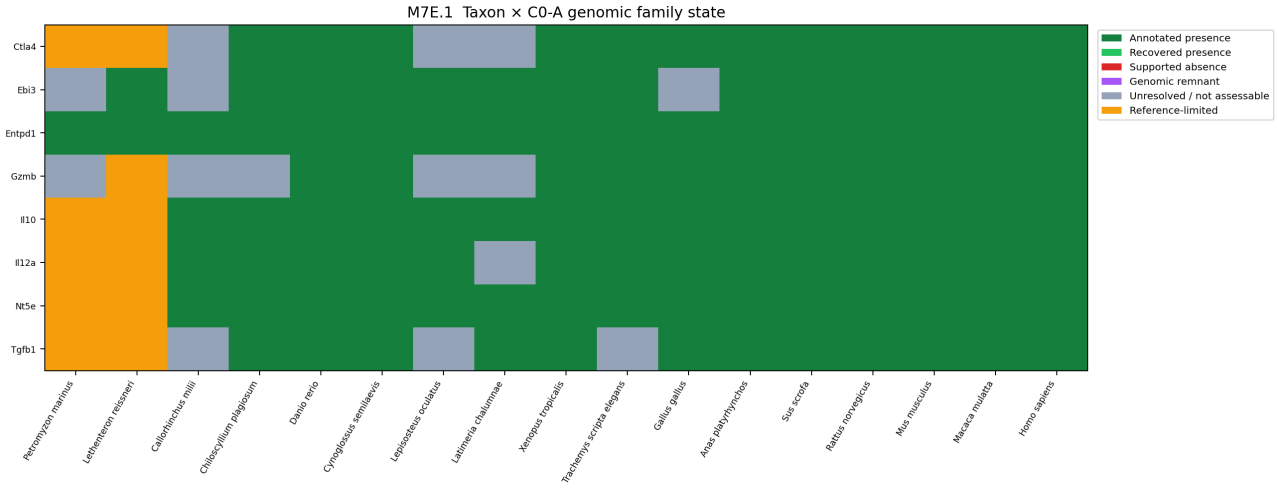
5 Supplementary figure legends



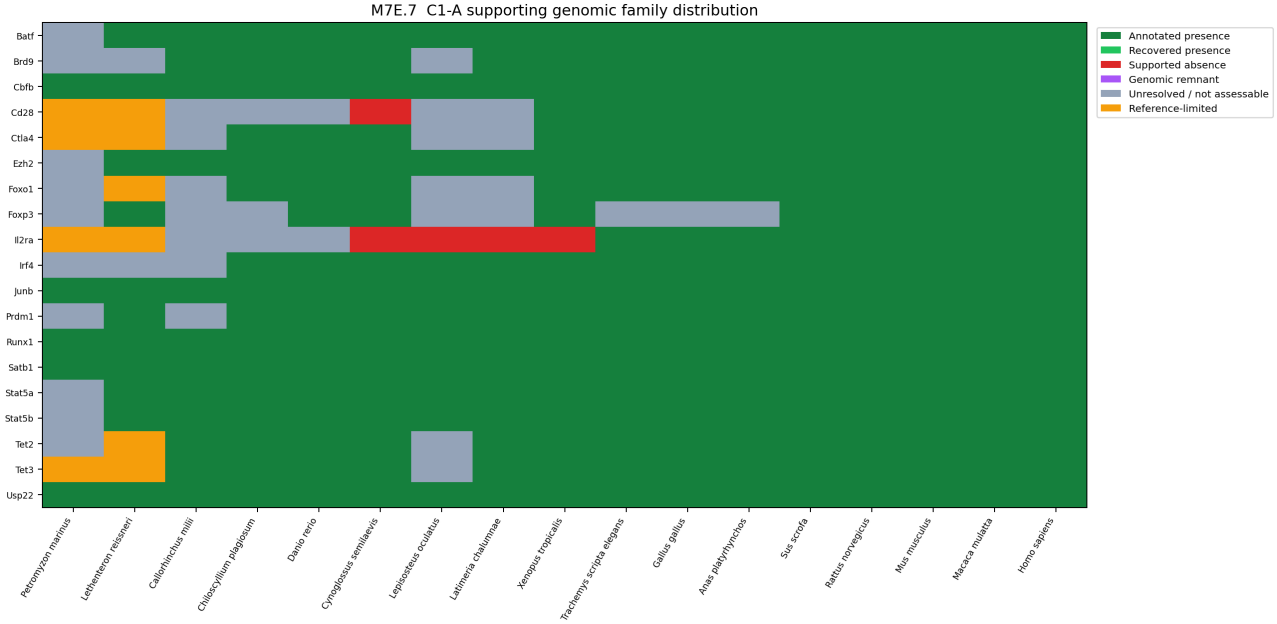
Supplementary Figure S1. Prospective validation of the initial candidate programme. Support, neutral, contradiction and non-estimability remain distinct across independent resources.



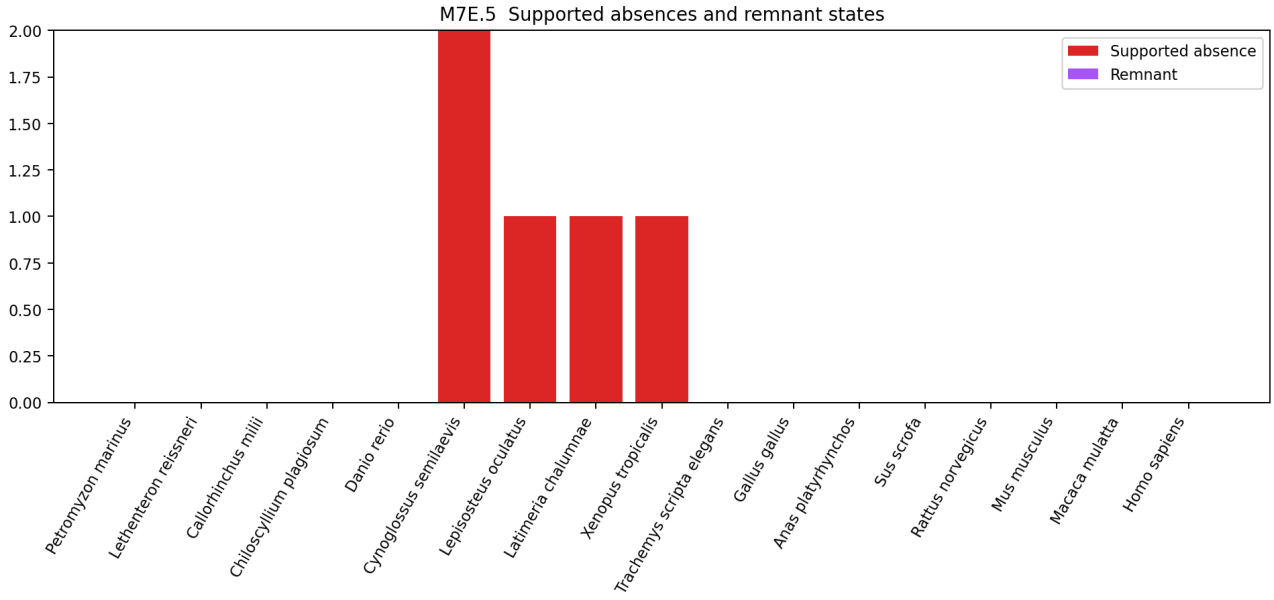
Supplementary Figure S2. Biological-replication and input-compatibility gate. The chicken and human near-term resources fail for different, explicitly recorded reasons.



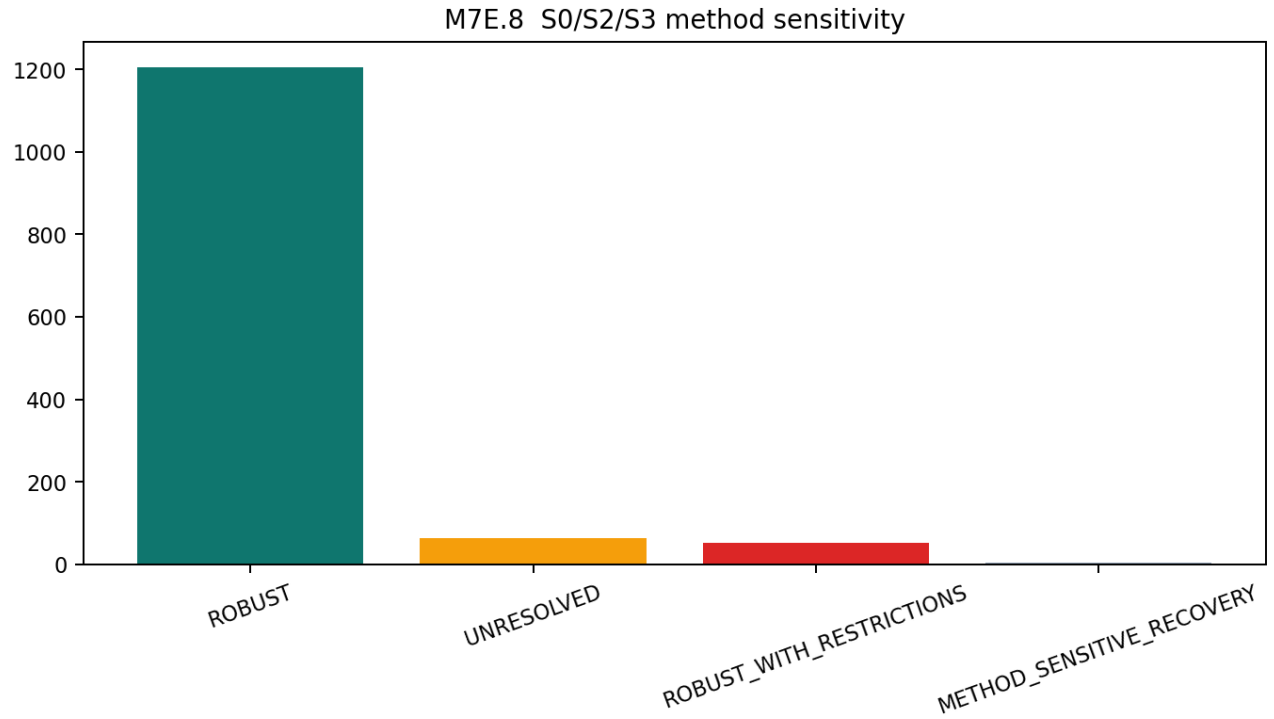
Supplementary Figure S3. Core family-by-taxon evidence matrix. Cells are genomic evidence records, not biological replicates.



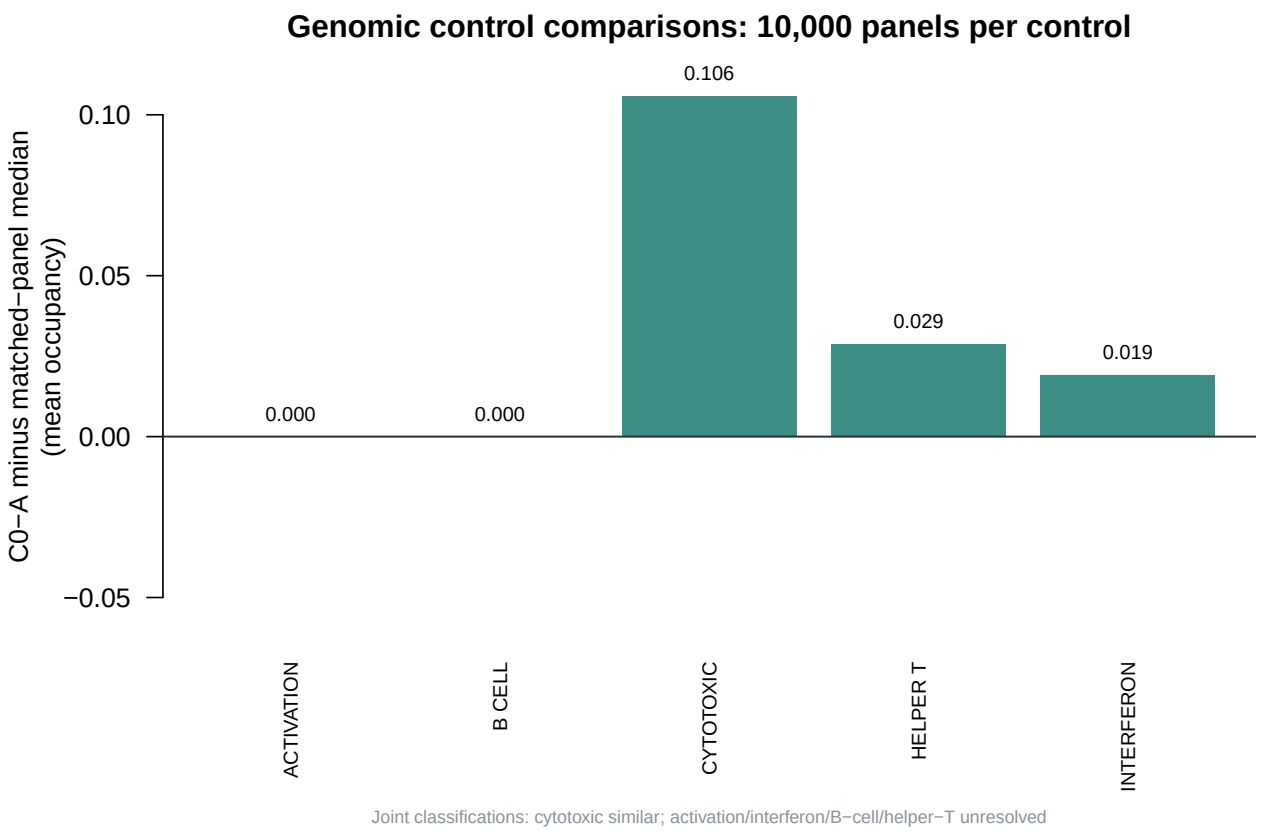
Supplementary Figure S4. Broader Treg-dependency family-by-taxon matrix. Detailed restrictions are retained in the machine-readable tables.



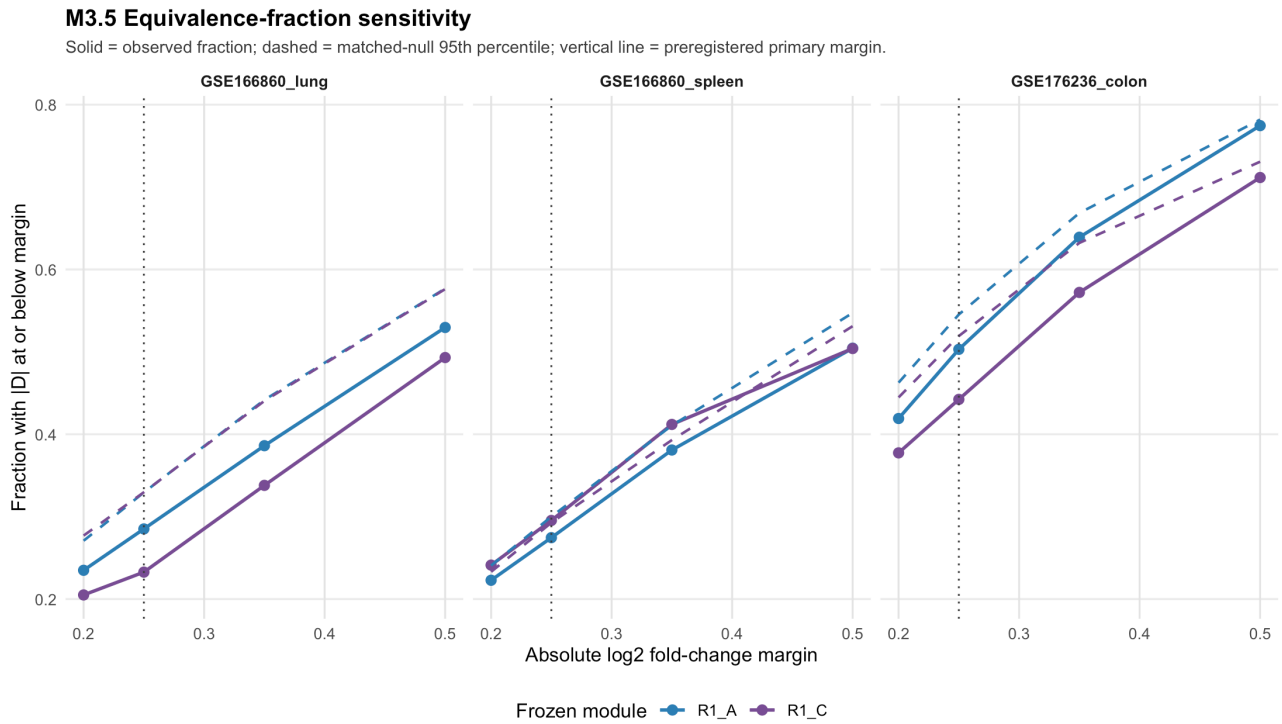
Supplementary Figure S5. Whole-genome recovery and uncertainty. Annotated presence, supported absence, restricted evidence and unresolved synteny remain separate.



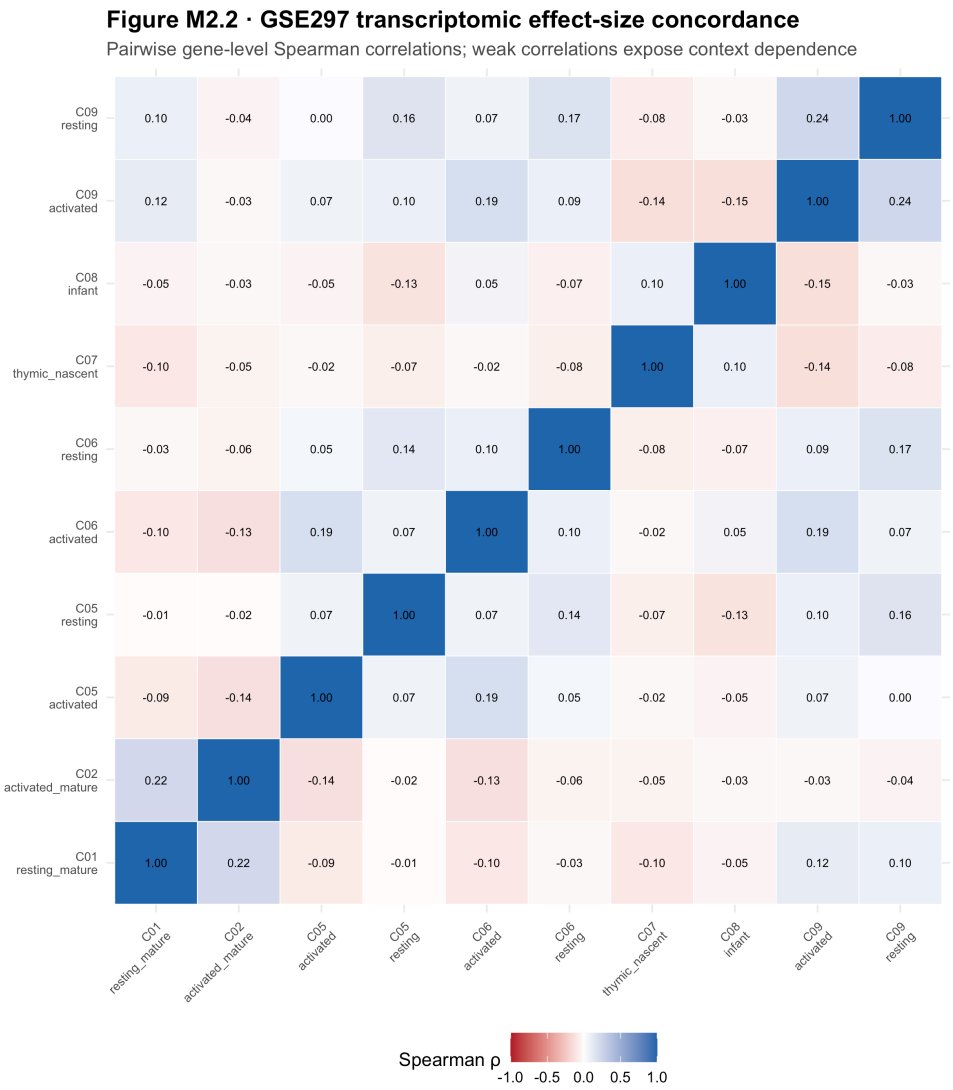
Supplementary Figure S6. Orthology-coordinate sensitivity. Family-state stability is shown across software, taxon-roster and representative-protein choices.



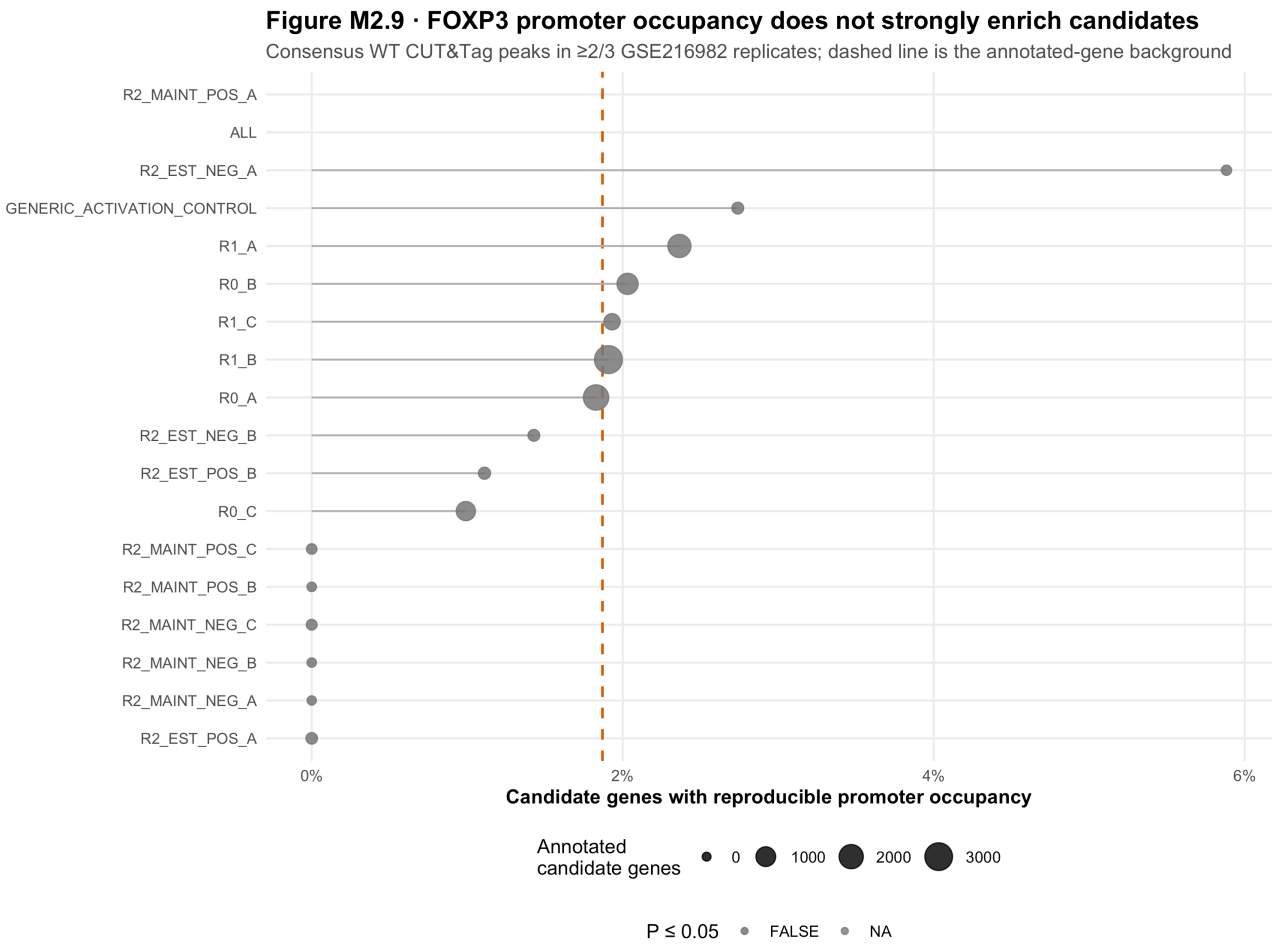
Supplementary Figure S7. Genomic immune-control comparisons. Five controls produced complete panels; none supported greater core-family breadth.



Supplementary Figure S8. Candidate-programme equivalence sensitivity. Equivalence fractions are compared with matched-null 95th percentiles across frozen margins.



Supplementary Figure S9. FOXP3 perturbation effect-size concordance. Heterogeneity across ten contexts motivates the stringent same-direction rule.



Supplementary Figure S10. Descriptive FOXP3 promoter-occupancy context. Overlap with reproducible peaks does not define direct targets or evolutionary timing.

6 Supplementary table index

Table	File	Content
S1	S01_prospective_decisions.tsv	Prospective decisions, gates and claim ceilings.
S2	S02_C0_component_definition.tsv	Core suppressive-effector definitions and evidence.
S3	S03_C1_component_definition.tsv	Broader mammalian Treg-dependency definitions.
S4	S04_control_programmes.tsv	Comparator-programme definitions.
S5	S05_component_HOG_map.tsv	Duplication-aware component-to-family map.
S6	S06_claim_equivalence_adjudication.tsv	Cross-coordinate claim-equivalence review.
S7	S07_family_taxon_recovery.tsv	Full genomic recovery evidence.
S8	S08_C0A_family_state_matrix.tsv	Core family-by-taxon states.
S9	S09_C1A_family_state_matrix.tsv	Broader-dependency family-by-taxon states.
S10	S10_C0_mechanism_taxon_state.tsv	Mechanism-level taxon states.
S11	S11_genomic_uncertainty_audit.tsv	Genomic uncertainty records.
S12	S12_genomic_method_sensitivity.tsv	Genomic method sensitivities.
S13	S13_layer_summary_S2.tsv	Primary genomic-layer summary.
S14	S14_C0_vs_C1_genomic_matching.tsv	Core-versus-broader-dependency matching.
S15	S15_C0_vs_genomic_controls.tsv	Genomic comparator results.
S16	S16_control_mapping.tsv	Comparator-to-family mappings.
S17	S17_orthology_method_sensitivity.tsv	Orthology coordinate sensitivity.
S18	S18_family_complexity_sensitivity.tsv	Complex-family sensitivity.
S19	S19_taxon_sensitivity.tsv	Leave-one-taxon sensitivity.
S20	S20_cellular_resource_gate.tsv	Replication and input-compatibility audit.
S21	S21_promoter_universe.tsv	Gene-anchored promoter universe.
S22	S22_P2_mapping.tsv.gz	Direct-homology-qualified mapping records.
S23	S23_promoter_matching_summary.tsv	Promoter matching yield.
S24	S24_human_liftover_qc.tsv	Human peak-coordinate conversion quality.
S25	S25_human_promoter_DA.tsv	Human promoter-accessibility effects.
S26	S26_mouse_day3_promoter_DA.tsv	Mouse day-3 promoter-accessibility effects.
S27	S27_mouse_day7_promoter_DA.tsv	Mouse day-7 promoter-accessibility effects.
S28	S28_regulatory_window_sensitivity.tsv	Promoter-window sensitivities.
S29	S29_reference_paralog_sensitivity.tsv	Reference/paralogue sensitivities.
S30	S30_regulatory_kill_rules.tsv	Prespecified regulatory stopping rules.
S31	S31_M9_integrated_classification.tsv	Integrated genomic classification.
S32	S32_M10_integrated_classification.tsv	Integrated regulatory classification.
S33	S33_R1_adjudication.tsv	Initial candidate-programme validation.
S34	S34_FOXP3_ establishment_maintenance.tsv	FOXP3 stage-dependence verification.

7 Reproducibility note

The accompanying archive contains LaTeX sources, figures, tables, frozen metadata/configuration, nonrestricted derived results, scripts, environment records and SHA-256 checksums. Public raw reads, genomes, proteomes, annotations, chain files and third-party motif libraries are not redistributed where size or provider terms require independent retrieval. Accessions, versions and acquisition instructions are supplied. The archive does not add any analysis beyond the manuscript and SI.