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.

# Gut microbial structural variation improves disease discrimination and reveals a strain-level regulatory mechanism in rheumatoid arthritis

**Yang Ou**
Nanjing Drum Tower Hospital
**Shengyan Zhao**
West China Hospital of Sichuan University
**Jiachun Weng**
Nanjing Drum Tower Hospital
**Yichen Long**
Duke-NUS Medical School
**Huamei Li**
Nanjing Drum Tower Hospital
**Yue Hou**
Nanjing Drum Tower Hospital
**Qing Xiong**
Nanjing Drum Tower Hospital
**Sha Liu**
Nanjing Drum Tower Hospital
**Zhikang Wang**
Nanjing Drum Tower Hospital
**Guangxiang Huang**
Ruijin Hospital
**Hui Pan**
West China Hospital of Sichuan University
**Yuhui Xu**
Peking Union Medical College Hospital
**Huayong Zhang**
huayong.zhang@nju.edu.cn

Nanjing Drum Tower Hospital
**Lingyun Sun**
Nanjing Drum Tower Hospital

**Research Article**

**Keywords:** Structural variants, rheumatoid arthritis, gut microbiome

**Posted Date:** September 28th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10852285/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

# Background

Species-level microbiome profiling may overlook strain-level genomic variation relevant to rheumatoid arthritis (RA). We investigated whether microbial structural variants (SVs) provide additional information beyond microbial taxonomy.

# Methods

We analyzed shotgun metagenomic data from four independent RA cohorts comprising 491 samples. Microbial SVs were identified using SGV-Finder and quantified across cohorts. Species- and SV-level associations were evaluated by cross-cohort meta-analysis. Meta-analysis-prioritized features were subjected to LASSO selection and random forest classification using combined training cohorts, followed by validation in an independent cohort. Selected SV-associated loci were functionally characterized by genomic annotation, motif analysis, structural modeling, and biolayer interferometry.

# Results

We identified 4,105 SVs, including 2,860 variable and 1,245 deletion SVs. Meta-analysis identified 49 species and 59 SVs associated with RA. Most RA-associated SVs were not accompanied by differential species abundance, with only 7 variable and 4 deletion SVs overlapping species-level associations. Incorporating SVs improved RA discrimination, with AUCs of 0.82 versus 0.71 in training cohorts and 0.62 versus 0.48 in the validation cohort. Functional characterization identified strain-level loss of a retron-associated reverse transcriptase locus in *Faecalibacterium prausnitzii* and an XRE-family transcriptional regulator locus in *Agathobacter rectalis*. Biolayer interferometry confirmed sequence-specific DNA binding by the XRE regulator, which was attenuated by motif mutation.

# Conclusions

Microbial SVs capture RA-associated strain-level genomic variation beyond species abundance and improve disease discrimination, providing an additional genomic layer for microbiome-based characterization of RA.

## Introduction

Rheumatoid arthritis (RA) is a chronic systemic autoimmune disease characterized by persistent synovial inflammation and progressive joint destruction[1]. Its pathogenesis involves complex interactions between genetic susceptibility and environmental factors[2]. Increasing evidence has implicated the gut

microbiome in RA susceptibility and immune dysregulation[3]. Alterations in intestinal microbial composition and function have been associated with impaired immune tolerance, systemic inflammation, and autoantibody-associated immune responses, including anti-citrullinated protein antibodies (ACPAs), which may precede clinical disease onset [4,5].

Although microbiome-targeted interventions such as fecal microbiota transplantation (FMT)[6,7] and probiotic supplementation[8,9] aim to restore microbial homeostasis, their variable efficacy highlights host–microbiome complexity and heterogeneity. A major limitation of current microbiome studies is that conventional taxonomic profiling often fails to capture functional variation within bacterial species. Thus, identifying high-resolution metagenomic biomarkers beyond species-level abundance remains an important challenge.

Recent advances in metagenomics have revealed that microbial structural variants (SVs), including deletion structural variants (dSVs) and variable structural variants (vSVs), represent an important source of strain-level functional diversity[10]. These regions can harbor genes involved in microbial adaptation and host–microbe interactions, providing functional resolution beyond taxonomy. Previous studies have demonstrated that microbial SVs are associated with host metabolic traits and circulating metabolomic profiles, suggesting their potential to capture biologically relevant microbiome variation that is not reflected by taxonomic composition alone[10,11].

Despite the established association between gut microbial dysbiosis and RA, the contribution of microbial SVs to RA remains largely unexplored. In particular, it remains unknown whether strain-level genomic variation provides additional discriminatory information beyond species-level taxonomy and whether RA-associated SVs are linked to specific regulatory mechanisms influencing microbial functions. Addressing these questions may reveal an additional functional layer connecting microbial genomic variation with immune-relevant processes in RA.

To address these gaps, we performed a cross-cohort meta-analysis integrating gut metagenomic datasets from four independent RA cohorts. Beyond conventional taxonomic profiling, we systematically characterized microbial SV landscapes and evaluated their contribution to RA discrimination beyond species-level abundance. Furthermore, we identified a key SV-associated transcriptional regulator and investigated its potential regulatory mechanism through motif analysis and experimental DNA-binding validation. Together, our findings establish microbial structural variation as an additional layer of the gut microbiome associated with RA and highlight its potential for disease discrimination and functional interpretation of strain-level variation.

## Results

# Microbial Diversity and Clinical Profiles Across Cohorts

Clinical characteristics were largely comparable across the four cohorts (Table S1). RA patients were predominantly female (67.5%–82.9%) with median ages ranging from 51 to 60 years. In Cohort D, untreated RA patients exhibited higher inflammatory activity, reflected by elevated erythrocyte sedimentation rate (ESR) and C-reactive protein (CRP) levels compared with treated patients and healthy controls. High ACPA seropositivity was consistently observed across RA groups, particularly among untreated individuals.

Microbial alpha diversity was generally comparable between RA and control groups, with reduced observed richness detected only in Cohort A (Table S2, Figure S1). Other diversity metrics, including Shannon diversity, Simpson index, and Pielou's evenness, showed no significant differences across groups. Beta diversity analysis based on Bray–Curtis dissimilarity revealed cohort-dependent differences in microbial community composition. Significant compositional shifts were observed in Cohorts A and C, whereas no significant differences were detected in Cohorts B and D (Table S3, Figures S1–S4). These findings indicate variability in community-level microbial profiles across RA cohorts.

## Landscape of Gut Microbial Structural Variants Across Cohorts

To investigate genomic variation beyond species-level composition, we characterized microbial structural variants (SVs) across the four cohorts. A total of 4,105 SVs were identified, including 2,860 variable SVs (vSVs) and 1,245 deletion SVs (dSVs) (Fig. 2A, B). SVs displayed strong species specificity, with *Blautia wexlerae*, *Roseburia intestinalis*, and *Agathobacter rectalis* harboring the highest numbers of detected SVs.

The distribution of SV-harboring species varied across cohorts, with *Bacteroides uniformis* and *Faecalibacterium prausnitzii* identified in the largest number of SV-detected samples (Fig. 2C). In addition, species detected with SVs constituted a major fraction of the total microbial community (median cumulative abundance ~ 35%), exhibiting remarkably consistent density distributions across the four cohorts (Fig. 2D).

## Meta-analysis Reveals Consensus Microbial Taxonomic and Structural Signatures

To identify robust RA-associated microbial features across populations, we performed a meta-analysis integrating four independent cohorts. This analysis identified 49 species with consistent differential abundance (Table S4, $P_{meta} < 0.05$, $P_{heterogeneity} > 0.01$) (Fig. 3A). At the genomic level, we further identified 34 vSVs and 25 dSVs significantly associated with RA (Table S5, S6 and Fig. 3B, C).

Interestingly, the vast majority of significant SVs did not reside within differentially abundant species, with only 7 vSVs and 4 dSVs overlapping with species-level abundance changes (Fig. 3D, E). This

decoupling indicates that microbial structural variations provide additional disease-associated information beyond conventional taxonomic profiles.

## Machine Learning Performance and Cross-Cohort Validation of Consistently Prioritized Structural Variants

To evaluate whether microbial SVs improve classification performance for RA while avoiding discovery bias, Cohort A was reserved strictly for candidate SV identification, with model training performed on independent datasets (Cohort B + C for training, Cohort D for external validation). Incorporating SVs markedly enhanced classification performance over species abundance alone across all cohorts (Fig. 4A–F, Table S7). Specifically, Random Forest classifiers subsequently trained using features selected by LASSO under the λ1se criterion achieved areas under the receiver operating characteristic curve (AUCs) of 0.82 in the training cohort. (Cohort B + C; vs. 0.71 for Species-only, $P < 0.01$; Fig. 4A), 0.62 in the independent test cohort (Cohort D; vs. 0.48, DeLong's test $P = 0.04$; Fig. 4B, Table S7), and 0.78 in the discovery cohort (Cohort A; vs. 0.65, $P = 0.02$; Fig. 4C). Under the $\lambda_{min}$ parameterization, similar gains were observed, with AUCs reaching 0.83 in training (vs. 0.77, $P < 0.01$; Fig. 4D), 0.60 in test (vs. 0.45, DeLong's test $P < 0.01$; Fig. 4E, Table S7), and 0.79 in discovery (vs. 0.71, $P = 0.04$; Fig. 4F).

SHapley Additive exPlanations (SHAP) feature importance analysis in the training cohort confirmed that SV features contributed substantially to model predictions, accounting for 40.50% of the cumulative feature importance (and 47.70% in the test cohort; Fig. 4G). Top global features ranked by training magnitude highlighted key specific SVs across cohorts (Fig. 4H). Cross-cohort SHAP profile alignment (Fig. 5A, B) and quantile-scaled SHAP summary distributions (Figure S5) prioritized robust RA-associated features based on consensus criteria (Table S8), including *Alistipes shahii*: 2252_2255;2255_2257;2257_2259;2259_2262;2262_2264;2265_2266, *Bacteroides uniformis*: 3880_3883, and *Bacteroides fragilis*: 4781_4788 (Spearman r = 0.69 to 0.91, pooled ORs = 1.86 to 4.18; Table S8, Figures S6, 7). Conversely, persistent protective factors such as *Faecalibacterium prausnitzii*: 537_538, *Faecalibacterium prausnitzii*: 781_783, *Agathobacter rectalis*: 662_663, and *Ruminococcus bromii*: 97_103 were identified (Spearman r = -0.67 to -0.89, pooled ORs = 0.34 to 0.53; Table S8, Figures S6, 7). Notably, leave-one-out sensitivity analyses confirmed that two key SVs maintained robust cross-cohort stability (Figures S8, 9). In subgroup analyses, *Agathobacter rectalis*: 662_663 showed significant frequency differences in both osteoarthritis (OA) vs. RA and Untreated vs. Treated RA comparisons ($P = 0.04$ and $P = 0.03$, respectively; Fig. 5C, D), whereas *Faecalibacterium prausnitzii*: 781_783 showed no significant differences across these subgroups ($P = 1.00$; Figure S10).

## Functional Annotation and Structural Characterization of Consistently Prioritized Deletion SVs

To explore the strain-level functional impacts of consistently prioritized deletion SVs, we analyzed their genomic locus structures, conserved functional domains, and locus-specific regulatory potential based

on sequence alignments and structural predictions.

For *Faecalibacterium prausnitzii*: 781_783, domain significance analysis revealed that this deletion region encodes a bacterial retron-associated reverse transcriptase element (HMPREF9436_01829) dominated by cd03487: RT_Bac_retron_II (-$\log_{10}$ E-value = 54.32) and NF038233: retron_St85_RT (-$\log_{10}$ E-value = 36.51) (Figure S11). Detailed site profiling mapped conserved catalytic centers, including putative active sites, NTP-binding sites, and nucleic acid-binding sites (Figure S11). Strain-level phylogenetic mapping illustrated the structural loss of this target gene across specific *Faecalibacterium prausnitzii* strains (Figure S12). Retron-associated reverse transcriptases have been implicated in bacterial antiphage defense through the production of multicopy single-stranded DNA (msDNA). Thus, loss of this locus may potentially affect strain-level defense capacity and ecological fitness. Given the role of *Faecalibacterium prausnitzii* as a major butyrate-producing commensal involved in gut mucosal integrity and immune homeostasis, structural loss of this retron-associated locus may reflect altered strain-level bacterial defense capacity or ecological fitness in RA-associated microbial communities.

For *Agathobacter rectalis*: 662_663, domain architecture profiling demonstrated that this SV region uniquely harbors a bacterial transcriptional regulator, *EUR_RS03225* (with the eggNOG-assigned preferred gene symbol *XK27_07105*, encoding NCBI RefSeq protein WP_015515948.1; Table S10), belonging to the Xenobiotic Response Element (XRE) family (Fig. 6A, B). It is characterized by COG1476: XRE (-$\log_{10}$ E-value = 23.56) and smart00530: helix-turn-helix (HTH)_XRE (-$\log_{10}$ E-value = 13.56), containing sequence-specific DNA-binding sites and a salt bridge (Fig. 6A, B). Strain-level phylogenetic mapping illustrated the differential presence/absence of this locus across *Agathobacter rectalis* isolates (Fig. 6C). Hypothesizing that XRE-family proteins commonly mediate negative feedback autorepression through binding to their own promoters[12], we sought to systematically decipher the consensus DNA-binding preference of EUR_RS03225 (Preferred Name XK27_07105). We identified homologous proteins of EUR_RS03225 (Preferred Name XK27_07105) across *Agathobacter rectalis* and related strains using BLASTP, extracted the 500-bp upstream promoter regions of their encoding genes, and performed de novo motif discovery via the MEME suite. This analysis identified a highly conserved consensus DNA-binding motif enriched within these regulatory regions (top motif E-value = $7.20 \times 10^{-88}$) (Figure S13). Notably, the predominant motif exhibited a characteristic conserved inverted-repeat motif, consistent with the dimeric DNA-binding mode of XRE-family regulators, supporting the presence of a conserved autoregulatory binding motif that may contribute to regulation of *EUR_RS03225 (XK27_07105)* expression.

To identify potential genomic targets of EUR_RS03225 (XK27_07105) and investigate the consequences of core motif disruption, we performed genome-wide motif scanning and identified candidate loci with paired palindromic motif matches within putative regulatory regions (Table S9). Subsequently, loci were prioritized based on motif significance and functional relevance to essential metabolic processes (Figure S14, Table S10). We identified paired palindromic motif occurrences across the reference genome and selected three representative loci for further characterization: the *EUR_RS03225 (XK27_07105)* locus

itself, encoding the XRE-family regulator, the lipid and energy metabolism-associated gene *EUR_RS01730 (phbA)*, and the carbohydrate degradation-associated gene *EUR_RS13145 (xsa)*. Each of these loci contained conserved paired inverted-repeat motif occurrences within their upstream regulatory regions (Fig. 6D, Tables S9 and S10).

To explore the structural basis of motif recognition, we performed three-dimensional complex modeling and comparative interface analysis of the EUR_RS03225 (XK27_07105) dimer bound to wild-type (WT) and core-mutated regulatory DNA duplexes using AlphaFold3 (Figure S15, Table S12). Both WT and mutant complexes were predicted to retain canonical helix-turn-helix (HTH)-mediated orientations toward the DNA major groove (ipTM, 0.55–0.62; pTM, 0.65–0.69; Figure S15). Comparative interface analysis following OpenMM energy minimization suggested altered protein–DNA interaction patterns upon core motif mutation, accompanied by a reduction in the number of predicted polar contacts from 56 to 20 for *EUR_RS03225 (XK27_07105)*, from 69 to 32 for *EUR_RS01730 (phbA)*, and from 56 to 34 for *EUR_RS13145 (xsa)* (WT versus mutant; Figure S16). Consistently, structural mapping identified losses of several predicted hydrogen-bonding interactions involving residues within or adjacent to the DNA-recognition helices (Figure S17).

Rosetta-based comparative scoring further indicated less favorable modeled protein–DNA interfaces following core motif mutation. Across all three complexes, mutant structures showed higher total Rosetta scores than their corresponding WT complexes across all three target loci (Table S11). These differences were accompanied by increased steric repulsion terms (fa_rep) and altered electrostatic (fa_elec) and sidechain hydrogen-bonding (hbond_sc) contributions, collectively suggesting that mutation of the conserved motif may perturb the modeled protein–DNA interface.

To experimentally assess these predicted interactions, recombinant EUR_RS03225 (XK27_07105) protein was expressed and purified to homogeneity (Figure S18, Table S13). Real-time bio-layer interferometry (BLI) assays demonstrated robust binding of EUR_RS03225 (XK27_07105) to its own WT promoter duplex ($K_D$ = 1.58 $\times$ $10^{-08}$ M, $k_{on}$ = 5.00 $\times$ $10^{05}$ $M^{-1}s^{-1}$, $k_{off}$ = 7.90 $\times$ $10^{-03}$ $s^{-1}$), whereas binding to the core-mutated duplex was markedly attenuated ($K_D$ = 2.17 $\times$ $10^{-05}$ M; Fig. 6E). Direct binding was also observed for the regulatory sequences associated with the candidate metabolic genes, with micromolar to sub-micromolar affinities toward *EUR_RS01730 (phbA)* WT ($K_D$ = 4.75 $\times$ $10^{-06}$ M) and *EUR_RS13145 (xsa)* WT ($K_D$ = 5.10 $\times$ $10^{-07}$ M). Core motif mutation reduced binding affinity for both sequences ($K_D$ = 2.22 $\times$ $10^{-05}$ M and 1.15 $\times$ $10^{-06}$ M, respectively; Figure S19). The strongest motif-dependent reduction in binding was observed for the EUR_RS03225 (XK27_07105) autoregulatory promoter, whereas motif mutation produced more modest reductions in binding affinity at the *EUR_RS01730 (phbA)* and *EUR_RS13145 (xsa)* regulatory sequences. Collectively, these genomic, structural, and biophysical analyses support sequence-specific recognition of conserved inverted-repeat motifs by EUR_RS03225 (XK27_07105) and demonstrate direct binding of the regulator to its own promoter and to regulatory regions associated with candidate metabolic genes. Disruption of the motif attenuated these interactions, indicating that EUR_RS03225 (XK27_07105) can directly recognize these

regulatory sequences and may regulate its own locus and selected metabolic genes through sequence-specific DNA binding.

## Discussions

In this study, we systematically characterized gut microbial structural variation in rheumatoid arthritis by integrating metagenomic data from four independent cohorts. Beyond conventional species-level profiling, we demonstrated that microbial SVs represent an additional layer of microbiome variation associated with RA and capture disease-related genomic heterogeneity that is not reflected by taxonomic abundance alone. Cross-cohort meta-analysis identified reproducible RA-associated microbial signatures at both species and SV levels, while machine learning analyses showed that incorporation of SV features improved discrimination between RA patients and controls beyond species-level profiles. Furthermore, functional characterization of representative SVs revealed strain-level genomic alterations affecting microbial regulatory and defense-related loci, providing potential biological insights into how microbial genomic variation may contribute to RA-associated microbiome states. Together, these findings highlight microbial structural variation as an underexplored dimension of the gut microbiome with potential relevance for biomarker discovery and functional interpretation in RA.

Previous studies have established an association between gut microbiome alterations and RA, with recurrent observations including the expansion of *Prevotella copri*[13] and increasing evidence supporting immunomodulatory roles of commensal taxa such as *Faecalibacterium prausnitzii* in regulating inflammatory responses and intestinal immune homeostasis[14]. However, species-level abundance profiling provides only a partial representation of microbial variation, as bacterial species frequently comprise genetically diverse populations with substantial strain-level heterogeneity[15]. Strains belonging to the same species can differ extensively in their accessory genomes, including gene gain and loss events, resulting in functional divergence in metabolic capacity, ecological adaptation, and host–microbe interactions[10]. Therefore, changes in species abundance may capture ecological shifts within microbial communities while overlooking genomic heterogeneity among closely related bacterial populations. Microbial SVs, which represent strain-level differences in genome organization and gene content, provide an additional genomic layer capable of capturing functional diversity beyond conventional taxonomic profiling[10]. By resolving microbial variation at higher genomic resolution, SV-based approaches may reveal disease-associated microbial features that remain undetected when analyses are restricted to species-level abundance[10].

Beyond characterizing disease-associated microbial alterations, metagenomic studies have increasingly explored the potential of microbiome-derived features as biomarkers for complex diseases. Previous studies have demonstrated that gut microbial signatures can discriminate individuals with metabolic disorders, inflammatory diseases, and other clinical phenotypes, highlighting the potential of microbiome-based classifiers for disease stratification[16,17]. However, the translational application of microbiome biomarkers remains challenging due to substantial inter-individual variability, population

heterogeneity, and limited reproducibility across independent cohorts[18,19]. Moreover, most existing microbiome classifiers rely primarily on taxonomic abundance profiles, which may overlook functionally relevant genomic differences among strains within the same species. In this study, we found that incorporation of microbial SV features improved classification performance beyond species-level abundance profiles and identified reproducible SV-associated signatures across independent RA cohorts[15]. These findings suggest that microbial genomic variation may complement conventional taxonomic features and provide additional resolution for microbiome-based biomarker development in RA.

Beyond their potential biomarker value, microbial SVs may represent biologically meaningful genomic alterations rather than merely statistical features. Emerging studies have demonstrated that bacterial SVs can capture strain-level functional variation associated with diverse host phenotypes. For example, microbial SVs have been linked to host metabolic traits through strain-specific gene content variation, revealing functional interactions between host genetic background and microbial genomic diversity[20,21]. In addition, disease-associated microbial SV signatures have been identified in children with autism spectrum disorder and were shown to improve disease classification beyond species-level abundance profiles[22]. Furthermore, microbial SVs have been associated with response to immune checkpoint inhibitor therapy, suggesting that strain-level genomic variation may influence host immune-related phenotypes[23]. Together, these findings indicate that microbial SVs represent an emerging functional layer of the microbiome that extends beyond conventional taxonomic composition. However, whether microbial SVs contribute to autoimmune diseases such as RA remains largely unexplored.

In our study, functional annotation of RA-associated SVs revealed alterations affecting diverse microbial functional modules. For example, deletion of a retron-associated reverse transcriptase locus in *Faecalibacterium prausnitzii* suggests that disease-associated SVs may influence bacterial defense-related functions, highlighting the potential functional consequences of strain-level genomic variation[24]. More notably, deletion of the XRE-family transcriptional regulator cassette in *Agathobacter rectalis* provided an example of how microbial SVs may affect bacterial regulatory architecture. XRE-family regulators are sequence-specific DNA-binding proteins involved in transcriptional regulation, and loss of such regulatory elements may alter bacterial gene regulatory networks[25]. In our study, motif analysis, structural modeling, and DNA-binding validation supported sequence-specific recognition of a conserved inverted-repeat motif by the XRE-family regulator. This motif was identified not only in the upstream regulatory region of the regulator itself but also in candidate metabolic genes, including *phbA (EUR_RS01730)*, encoding a thiolase involved in acetyl-CoA metabolism, and *xsa (EUR_RS13145)*, encoding an α-L-arabinofuranosidase involved in carbohydrate degradation. Biolayer interferometry further confirmed direct binding of the regulator to these regulatory sequences, while motif mutation attenuated binding, supporting the potential involvement of this SV-associated regulator in strain-level transcriptional regulation. Although the downstream metabolic consequences require further investigation, the location of the predicted binding motif immediately upstream of the translation

initiation site of a metabolic gene provides a plausible framework linking microbial genomic variation with altered regulatory potential and host-relevant microbial functions[26].

Beyond disease-associated microbial variation, we further explored whether therapeutic intervention may influence this strain-level genomic landscape. Previous studies have demonstrated that methotrexate (MTX) can directly reshape the gut microbiome and alter bacterial metabolic pathways in RA patients, with strain-dependent differences in bacterial responses to treatment[27]. In line with these observations, our strain-resolved analysis further suggests that treatment-associated microbiome remodeling extends beyond taxonomic changes to microbial structural variation, highlighting SVs as an additional genomic layer underlying therapeutic-induced microbial adaptation.

Several limitations should be acknowledged. First, although we integrated four independent cohorts and applied a standardized analytical framework, the retrospective nature of publicly available metagenomic datasets limits the ability to establish temporal relationships between microbial SV alterations and RA development. Prospective longitudinal studies will be required to determine whether specific SV signatures precede disease onset or change during disease progression and treatment. Second, while functional analyses and molecular validation of the XRE-family regulator support its sequence-specific DNA-binding activity and the biological relevance of this SV-associated regulatory locus, future studies evaluating its downstream metabolic consequences, including potential effects on butyrate-associated metabolism, will be important to further clarify its functional impact. Finally, given the complexity and diversity of microbial genomes, the SVs identified in this study likely represent only a subset of functionally relevant microbial variations associated with RA. Limited sample size, microbial strain heterogeneity, and statistical power may have restricted the detection of additional disease-associated SVs, including those with weaker effects or lower prevalence. Therefore, the identified functional SV examples should be considered representative instances supporting the broader concept that microbial structural variation constitutes an important and underexplored layer of host–microbiome interactions.

In conclusion, our study expands the current understanding of RA-associated microbiome alterations by demonstrating that microbial structural variation provides information beyond species-level taxonomy. These strain-level genomic features improve characterization of disease-associated microbial states and reveal functional genomic changes that may otherwise remain undetected by conventional abundance-based approaches. By establishing microbial SVs as a complementary layer of microbiome variation, our findings provide a framework for future studies investigating microbial genomic heterogeneity, biomarker development, and host–microbiome interactions in rheumatoid arthritis.

## Methods

# Data Acquisition and Cohort Integration

This study utilized publicly available gut metagenomic sequencing data from four clinical cohorts of RA patients, totaling 491 samples. Cohort A (Cheng et al., 2022) focuses on the comparison between RA

patients, osteoarthritis (OA) patients, and healthy controls (HC)[28], Cohorts B and C (He et al., 2022) investigate the differences between RA patients and healthy individuals[29], and Cohort D (Zhang et al., 2015) provides a comprehensive comparison between healthy controls and RA patients, the latter of which includes both disease-modifying antirheumatic drug (DMARD)-treated and untreated individuals[30]. Raw metagenomic reads were obtained from the Genome Sequence Archive (GSA) and the European Bioinformatics Institute (EBI) database, while clinical metadata were retrieved from the supplementary materials of the original publications (Table S1). To ensure cross-cohort consistency, all sequence data were reprocessed using a standardized metagenomic workflow based on unified host-read removal, taxonomic profiling, and structural variation identification procedures, with comprehensive baseline clinical data detailed in Fig. 1.

# Metagenomic Preprocessing and Quality Control

The raw metagenomic sequencing data were subjected to quality control and host-read depletion before downstream microbial profiling. Adapter sequences and low-quality reads were trimmed using Trimmomatic (v0.39)[31] with the parameters LEADING:3, TRAILING:3, SLIDINGWINDOW:4:20, and MINLEN:50. Host-derived reads were identified and removed by mapping the processed reads to the human reference genome (GRCh37) using Bowtie2 (v2.3.5.1)[32] in --very-sensitive mode. The preprocessing workflow was performed using the KneadData (v0.6.1) wrapper[33], generating host-depleted microbial reads for subsequent taxonomic profiling and structural variation analysis.

# Identification of Microbial Structural Variations

Microbial SVs, including dSVs and vSVs, were identified using SGV-Finder based on the Iterative Coverage-based Read Assignment (ICRA) algorithm to resolve ambiguous read mappings[10]. Microbial genomes from the proGenomes database[34] were partitioned into 1-kbp bins based on concatenated genome scaffolds for coverage-based SV identification. SV discovery was initially performed in Cohort A (n = 122) using the --min_samp_cutoff parameter set to 12, corresponding to approximately 10% of the cohort size, to identify prevalent structural variants while minimizing rare sample-specific signals. For replication analyses, the SV status of candidate SV regions identified in Cohort A[28] was quantified in Cohorts B, C[29], and D[30] using SGV-Finder with the --by-orig parameter. These SVs were subsequently evaluated for disease associations through cohort-specific statistical analyses and meta-analysis.

# Taxonomic Quantification and Diversity Analysis

Taxonomic profiles were generated by aligning quality-controlled metagenomic reads to the proGenomes database[34] using Bowtie2 (default settings)[32], followed by processing with Samtools[35] and CoverM[36] to generate species-level abundance profiles, including raw read counts, covered fraction, mean coverage depth, TPM, and RPKM values. To minimize noise, microbial features present in fewer than five samples within a cohort were excluded from downstream analysis. Raw read counts were used

for differential abundance analysis, whereas normalized abundance profiles were used for microbial diversity analyses.

Alpha diversity indices, including Shannon diversity, Simpson index, observed richness, and Pielou's evenness, were calculated based on filtered abundance profiles. Beta diversity was assessed using Bray–Curtis dissimilarity calculated from relative abundance data via total sum scaling (TSS) and visualized using Non-metric Multidimensional Scaling (NMDS) and Principal Coordinates Analysis (PCoA), with statistical significance evaluated by PERMANOVA.

# Statistical Analysis and Consensus Meta-Analysis

Differential abundance analysis was performed using the edgeR quasi-likelihood (QL) framework[37] based on raw read counts with Trimmed Mean of M-values (TMM) normalization to account for differences in library size across samples.

To identify consistent microbial signatures across cohorts, random-effects meta-analyses were performed using the metafor[38] and meta[39] packages. Species-level differential abundance results were integrated using restricted maximum likelihood (REML)-based models. Binary dSV associations were summarized as odds ratios using Mantel–Haenszel (MH) random-effects meta-analysis. Continuous vSVs were analyzed as standardized mean differences (SMDs) and subsequently converted to log odds ratios using the formula LogOR = d $\bullet \pi / \sqrt{3}$[40]. Consensus significance was defined by a three-tier filtering system requiring a pooled meta-analysis P < = 0.05, a heterogeneity test P > 0.01, and replication in at least two independent cohorts with a nominal P < = 0.2 and a consistent effect direction[23]. For candidate SV features prioritized by the LASSO models, leave-one-cohort-out sensitivity analyses were conducted by iteratively recalculating pooled effect sizes and heterogeneity metrics after excluding one cohort at a time to confirm cross-cohort stability. For clinical subgroup analyses based on original cohort annotations (e.g., OA vs. RA in Cohort A, and Untreated vs. Treated RA in Cohort D), differences in categorical variables (such as SV frequencies) between two groups were evaluated using two-sided Fisher's exact tests. Differences in continuous variables between two groups were assessed using the Wilcoxon rank-sum test.

# Machine Learning Model Construction and Feature Interpretation

To evaluate whether microbial structural variants provided additional discriminatory information beyond species-level abundance, machine learning models were constructed using meta-analysis-derived microbial features. Cohort A was used for initial microbial SV discovery and calling. The identified SVs were subsequently quantified across all four cohorts and evaluated through cohort-level association analyses and cross-cohort meta-analysis. Features meeting the predefined meta-analysis criteria were subsequently subjected to LASSO-based feature selection using the combined Cohorts B and C, which served as the training dataset. Random forest classifiers were then constructed using the LASSO-

selected features, while Cohort D was reserved as an independent validation dataset. Cohort A was additionally evaluated using the models trained on Cohorts B and C.

Prior to model construction, feature preprocessing was performed using the training dataset only. To reduce redundancy between genomic and taxonomic features, pairwise Spearman correlations were calculated between all SV features and all species abundance features; any SV feature showing a strong correlation (Spearman $|r| > 0.7$) with any species abundance feature was removed. Among the remaining SV features showing high inter-feature collinearity (Spearman $|r| > 0.8$), the feature with lower univariate significance was excluded.

Feature selection was performed using least absolute shrinkage and selection operator (LASSO) regression implemented in the glmnet package[41,42]. Ten-fold cross-validation was applied to determine optimal regularization parameters, and features were selected using both $\lambda_{1se}$ and $\lambda_{min}$ criteria. Separate feature selection procedures were performed for species-only models and combined species-plus-SV models.

Random forest classifiers were subsequently constructed using the ranger package with 1,000 trees and probability estimation enabled[43,44]. Model performance was evaluated using receiver operating characteristic (ROC) curves and area under the curve (AUC). Differences between paired ROC curves were assessed using DeLong's test[45] implemented in the pROC package[46].

Model interpretation was performed using SHAP analysis[47] implemented in the treeshap package. Feature importance was quantified using mean absolute SHAP values. Cross-cohort consistency of feature contributions was evaluated by calculating Spearman rank correlation coefficients across cohort-specific feature-level SHAP profiles. Robust features summarized in Table S8 were prioritized based on consensus criteria requiring: (1) inclusion in the union of the top 20 structural variant (SV) features ranked by mean absolute SHAP values (|SHAP|) in either the training cohort (Cohort B + C) or the independent validation cohort (Cohort D); (2) statistically significant and strong monotonic feature-attribution correlations in both cohorts ($|\text{Spearman } r_s| > 0.60$ and $P < 0.05$); and (3) strictly concordant effect directionality across cohorts ($r_{s,train} \times r_{s,test} > 0$). Directional contributions of prioritized features were evaluated using Spearman's rank correlation between raw feature abundance and corresponding SHAP values, classifying features as persistent risk-promoting factors ($r_s > 0$) or protective factors ($r_s < 0$).

# Functional Annotation of SV-Associated Genes

To investigate the potential functional consequences of disease-associated structural variations, genes located within SV-associated genomic bins were identified based on the corresponding reference genome annotations. Only SV regions containing annotated protein-coding genes were selected for downstream functional characterization. Conserved protein domains and functional features of SV-associated genes were annotated using the NCBI Conserved Domain Database (CD-Search)[48].

Significant domain assignments, conserved catalytic residues, and predicted functional sites were obtained directly from the NCBI CD-Search output.

# Strain-Level Phylogenetic Analysis and Gene Distribution Analysis

To evaluate the strain-level distribution of SV-associated genes, complete chromosome-level genome assemblies were retrieved from the NCBI Assembly database for the corresponding bacterial species. Reference genome sequences were screened using BLASTN against SV-associated gene sequences to determine the presence or absence of the corresponding SV-associated genes across strains. Genome annotations were generated using Prokka[49]. Pan-genome analysis and core genome alignment generation were performed using Panaroo[50] in strict cleaning mode with a 50% core gene threshold (--core_threshold 0.5). Phylogenetic reconstruction was performed using FastTree based on the core genome alignment using the Generalized Time-Reversible (GTR) nucleotide substitution model (-nt -gtr)[51]. The presence or absence patterns of SV-associated genes across strains were visualized together with the phylogenetic tree.

# Conserved Motif Identification and Binding Site Prediction

To investigate whether the SV-associated XRE-family transcriptional regulator mediates autoregulatory feedback and downstream gene repression through conserved promoter binding, homologous protein sequences were identified across related bacterial genomes using BLASTP. For each homologous protein, the 500-bp upstream regions of the corresponding genes were extracted for motif discovery.

Conserved DNA-binding motifs were identified using the MEME Suite[52]. MEME analysis was performed in DNA mode with the following parameters: zero or one occurrence per sequence (zoops) model, five candidate motifs, motif widths ranging from 12 to 20 bp, classic objective function, reverse-complement strand search, and a zero-order Markov background model. The identified top conserved motif was subsequently scanned against the complete reference genome sequence using FIMO with a significance threshold of $P < 1.0 \cdot 10^{-04}$. Complementary conserved inverted-repeat motif on positive (+) and negative (-) strands within < = 2 bp distance were paired to identify candidate adjacent inverted-repeat binding sites, assigned identical Pair_ID identifiers, and ranked in ascending order of their P-values (Table S9). Motif occurrences were mapped onto the annotated genome using GenomicRanges to delineate regulatory contexts, where promoter regions were defined as -200 bp to + 50 bp relative to the translation initiation site (TIS/start codon), alongside coding sequences (CDSs) and intergenic regions. Candidate target genes regulated by or harboring these paired binding sites were functionally profiled using the NCBI RefSeq annotation and eggNOG-mapper (v2), assigning Clusters of Orthologous Groups (COG) functional categories, Enzyme Commission (EC) numbers, and Pfam domain architectures (Table S10). KEGG pathway enrichment analysis of candidate target genes was performed using a hypergeometric test, with significance evaluated via $-\log_{10}$(P-value) and false discovery rates adjusted by the Benjamini-Hochberg procedure.

# Structural Modeling of Transcriptional Regulator Dimer–DNA Complexes

To evaluate the structural basis of DNA recognition and validate downstream transcriptional targets, candidate target genes were prioritized based on a false discovery rate (FDR) threshold of $q < 0.25$ for their paired promoter-binding motifs, combined with their annotated roles in essential bacterial metabolic and autoregulatory pathways (Table S9 and Table S10). Specifically, *EUR_RS03225* (Preferred Name *XK27_07105*, autoregulatory repressor locus), *EUR_RS01730* (Preferred Name *phbA*, belonged to thiolase, involved in lipid and energy metabolism), and *EUR_RS13145* (Preferred *Name xsa*, encoding alpha-L-arabinofuranosidase involved in carbohydrate metabolism) were selected. The three-dimensional structures of the transcriptional regulator dimer bound to the promoter DNA sequences of these prioritized target genes were predicted using AlphaFold3[53]. Both wild-type (WT) and mutant DNA duplexes (with mutated core inverted-repeat binding sites) were modeled in complex with the regulator dimer to assess predicted protein–DNA interaction interfaces and hydrogen-bonding networks (Table S12). Although the predicted local distance difference test (pLDDT) for the DNA regions were relatively low, which is consistent with the known limitations of current AI-based protein–nucleic acid complex prediction, the overall architecture of the predicted complexes was biologically plausible. In particular, the HTH DNA-binding motif was positioned toward the major groove of the target dsDNA, consistent with the canonical DNA recognition mode of XRE-family transcriptional regulators. The predicted complexes exhibited moderate interface confidence, with interface predicted template modeling score (ipTM) values ranging from 0.55 to 0.62 and predicted template modeling score (pTM) values ranging from 0.65 to 0.69.

# Molecular Mechanics Energy Minimization and Rosetta Energy Analysis

To reduce potential steric clashes and locally relax the predicted structures, molecular mechanics energy minimization was performed using OpenMM[54] with the AMBER14 force field[55], including amber14-all.xml and amber14/tip3pfb.xml. Prior to minimization, terminal phosphate groups (P, OP1, OP2, and OP3) at the 5' ends of DNA strands were removed and converted to 5'-hydroxyl groups to ensure compatibility with AMBER topology definitions. Missing hydrogen atoms were subsequently added using OpenMM Modeller.

All systems underwent 1,000 steps of energy minimization using the L-BFGS algorithm until convergence (energy tolerance: 10.0 kJ/mol/nm) to obtain locally relaxed structural models. Following minimization, interface energetic characteristics were evaluated using Shirosetta based on the PyRosetta framework[56]. The Rosetta score function was used to assess relative energetic contributions at the protein–DNA interface, including van der Waals interactions, solvation effects, electrostatic interactions, and hydrogen-bonding-related terms. These scores were interpreted as comparative interface energy profiles rather than experimental binding free energies.

# Quantitative Interface Analysis and Molecular Visualization

Protein–DNA interfacial interactions were quantitatively analyzed from the minimized structures using PyMOL (The PyMOL Molecular Graphics System, Version 2.5, Schrödinger, LLC) with customized scripts. Putative polar interactions, including hydrogen bond-like contacts and salt bridge interactions, were identified based on donor–acceptor heavy atom distances within 3.5 Å.

To investigate both global interface characteristics and sequence-specific recognition features, a dual-scale analysis strategy was applied. First, all potential polar contacts between the transcriptional regulator dimer and DNA molecules were evaluated across the entire interface. Subsequently, conserved residues within the predicted DNA-binding region were identified based on domain annotation using the NCBI CD-Search. Structural analyses were further focused on six conserved residues located within the XRE-family HTH DNA-binding motif (residues 18, 19, 29, 33, 37, and 38), which are predicted to contribute to DNA recognition.

Specific interactions involving these core recognition residues were further characterized using a stringent heavy-atom distance cutoff of 3.3 Å. Structural comparisons between WT and mutant complexes were performed to evaluate potential alterations in DNA recognition patterns and interfacial interaction features associated with the mutation.

# Binding Affinity Measurement of EUR_RS03225 to Target Promoter DNAs by BLI

Recombinant EUR_RS03225 protein (assigned the preferred gene symbol XK27_07105) was heterologously expressed in Escherichia coli BL21(DE3) chemically competent cells and purified using Ni-NTA affinity chromatography (Table S13). Biotin-labeled sense strands and unmodified antisense strands were annealed by heating to 95°C for 5 min, followed by snap cooling to room temperature to form 60-bp double-stranded DNA (dsDNA) duplexes (Table S12).

The binding kinetics and affinity of EUR_RS03225 to target dsDNAs were determined via BLI on a Gator instrument (Gator Bio; Table S13). Streptavidin (SA) biosensors were pre-hydrated in running buffer (HBS-EP+, pH 7.4; Table S13). Biotinylated dsDNA probes were diluted to 100 nmol/L and immobilized onto the SA biosensors. Following baseline stabilization (120 s in running buffer), sensors were exposed to serially diluted EUR_RS03225 protein solutions for association (120 s), followed by dissociation in running buffer (120 s). Between measurement cycles, biosensors were regenerated using commercial regeneration buffer (Gator Bio; Table S13) to remove bound protein. All binding assays were conducted at 30°C. Kinetic sensorgrams were fitted to a 1:1 Langmuir binding model using Gator Bio analysis software (version 2.18.7.0718) to calculate the association rate constant ($k_{on}$), dissociation rate constant ($k_{off}$), and equilibrium dissociation constant ($K_D$).

## Abbreviations

ACPA: anti-citrullinated protein antibody

AUC: area under the receiver operating characteristic curve

BLI: bio-layer interferometry

CDD: Conserved Domain Database

CDS: coding sequence

COG: Clusters of Orthologous Groups

CRP: C-reactive protein

DMARD: disease-modifying antirheumatic drug

dsDNA: double-stranded DNA

dSV: deletion structural variant

EBI: European Bioinformatics Institute

EC: Enzyme Commission

ESR: erythrocyte sedimentation rate

fa_elec: electrostatic interactions score (Rosetta)

fa_rep: Lennard-Jones repulsive interactions score (Rosetta)

FMT: fecal microbiota transplantation

GSA: Genome Sequence Archive

hbond_sc: sidechain-sidechain hydrogen-bonding score (Rosetta)

HTH: helix-turn-helix

ICRA: Iterative Coverage-based Read Assignment

ipTM: interface predicted template modeling score

$K_D$: equilibrium dissociation constant

$k_{off}$: dissociation rate constant

$k_{on}$: association rate constant

LASSO: least absolute shrinkage and selection operator

MH: Mantel–Haenszel

MTX: methotrexate

NMDS: non-metric multidimensional scaling

OA: osteoarthritis

OR: odds ratio

PCoA: principal coordinates analysis

PERMANOVA: permutational multivariate analysis of variance

pLDDT: predicted local distance difference test

pTM: predicted template modeling score

QL: quasi-likelihood

RA: rheumatoid arthritis

REML: restricted maximum likelihood

ROC: receiver operating characteristic

SA: streptavidin

SCFA: short-chain fatty acid

SHAP: SHapley Additive exPlanations

SMD: standardized mean difference

SV: structural variant

TIS: translation initiation site

TMM: trimmed mean of M-values

TSS: total sum scaling

vSV: variable structural variant

WT: wild-type

XRE: Xenobiotic Response Element

# Declarations

**Ethics approval and consent to participate:**

Not applicable.

**Consent for publication:**

Not applicable.

**Availability of data and materials:**

The metagenomic sequencing datasets analyzed in this study are publicly available from the Genome Sequence Archive (GSA) and the European Nucleotide Archive (ENA) under the accession numbers reported in the original publications. The processed structural variation (SV) matrices, associated metadata, and all scripts used for statistical analyses and machine learning are available from Figshare (DOI: 10.6084/m9.figshare.32576595).

**Competing interests:**

The authors declare that they have no competing interests.

**Funding:**

Not applicable.

**Statement:**

During the preparation of this work the authors used Gemini 3.2 in order to refine the English language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

**Authors' contributions:**

Y.O., S.Z., J.W., Y.L., and H.L. contributed equally as co-first authors. Y.O. conceptualized the study, designed the methodology, and wrote the original manuscript draft. S.Z. was responsible for validation, laboratory investigation, and executing the biophysical surface plasmon resonance (SPR) assays. J.W. performed the cross-cohort meta-analysis, led data curation, and generated the primary data visualizations. Y.L. conducted the machine learning modeling, executed the genetics-related bioinformatics pipelines, and performed formal statistical analyses. H.L. contributed to methodology development, bioinformatics workflows, and computational AlphaFold modeling.

Y.H., S.L., G.H., Y.X., H.P., and Z.W. contributed to the clinical interpretation of the genomic findings, provided translational medicine insights, and participated in the conceptual design of the study. Q.X. contributed to big data platform management, multi-cohort dataset infrastructure, and preliminary data processing. L.S. and H.Z. served as co-corresponding authors, jointly supervising the research project. They acquired funding, provided overall scientific supervision, and were responsible for manuscript review, editing, and final approval. All authors reviewed the manuscript drafts and approved the final version for submission.

**Acknowledgements:**

We would like to express our gratitude to the researchers and consortia that made their data publicly available, enabling this work. We are also deeply grateful to the Medical Big Data Analysis Platform at Nanjing Drum Tower Hospital for providing the computational infrastructure and technical support. Additionally, we would like to thank all staff members of the Department of Rheumatology and Immunology at Nanjing Drum Tower Hospital for their continuous support and valuable insights throughout this study.

# References

1. Di Matteo, A., Bathon, J. M. & Emery, P. Rheumatoid arthritis. *Lancet* **402**, 2019-2033 (2023). https://doi.org:10.1016/s0140-6736(23)01525-8
2. Ha, E., Bae, S. C. & Kim, K. Large-scale meta-analysis across East Asian and European populations updated genetic architecture and variant-driven biology of rheumatoid arthritis, identifying 11 novel susceptibility loci. *Ann Rheum Dis* **80**, 558-565 (2021). https://doi.org:10.1136/annrheumdis-2020-219065
3. Wang, J. & Jia, H. Metagenome-wide association studies: fine-mining the microbiome. *Nat Rev Microbiol* **14**, 508-522 (2016). https://doi.org:10.1038/nrmicro.2016.83
4. Maeda, Y. *et al.* Dysbiosis Contributes to Arthritis Development via Activation of Autoreactive T Cells in the Intestine. *Arthritis Rheumatol* **68**, 2646-2661 (2016). https://doi.org:10.1002/art.39783
5. Pianta, A. *et al.* Evidence of the Immune Relevance of Prevotella copri, a Gut Microbe, in Patients With Rheumatoid Arthritis. *Arthritis Rheumatol* **69**, 964-975 (2017). https://doi.org:10.1002/art.40003
6. Liu, X. *et al.* Fecal microbiota transplantation for the management of autoimmune diseases: Potential mechanisms and challenges. *J Autoimmun* **141**, 103109 (2023). https://doi.org:10.1016/j.jaut.2023.103109
7. Zeng, L. *et al.* Efficacy and safety of gut microbiota-based therapies in autoimmune and rheumatic diseases: a systematic review and meta-analysis of 80 randomized controlled trials. *BMC Med* **22**, 110 (2024). https://doi.org:10.1186/s12916-024-03303-4
8. Vaghef-Mehrabany, E. *et al.* Probiotic supplementation improves inflammatory status in patients with rheumatoid arthritis. *Nutrition* **30**, 430-435 (2014). https://doi.org:10.1016/j.nut.2013.09.007

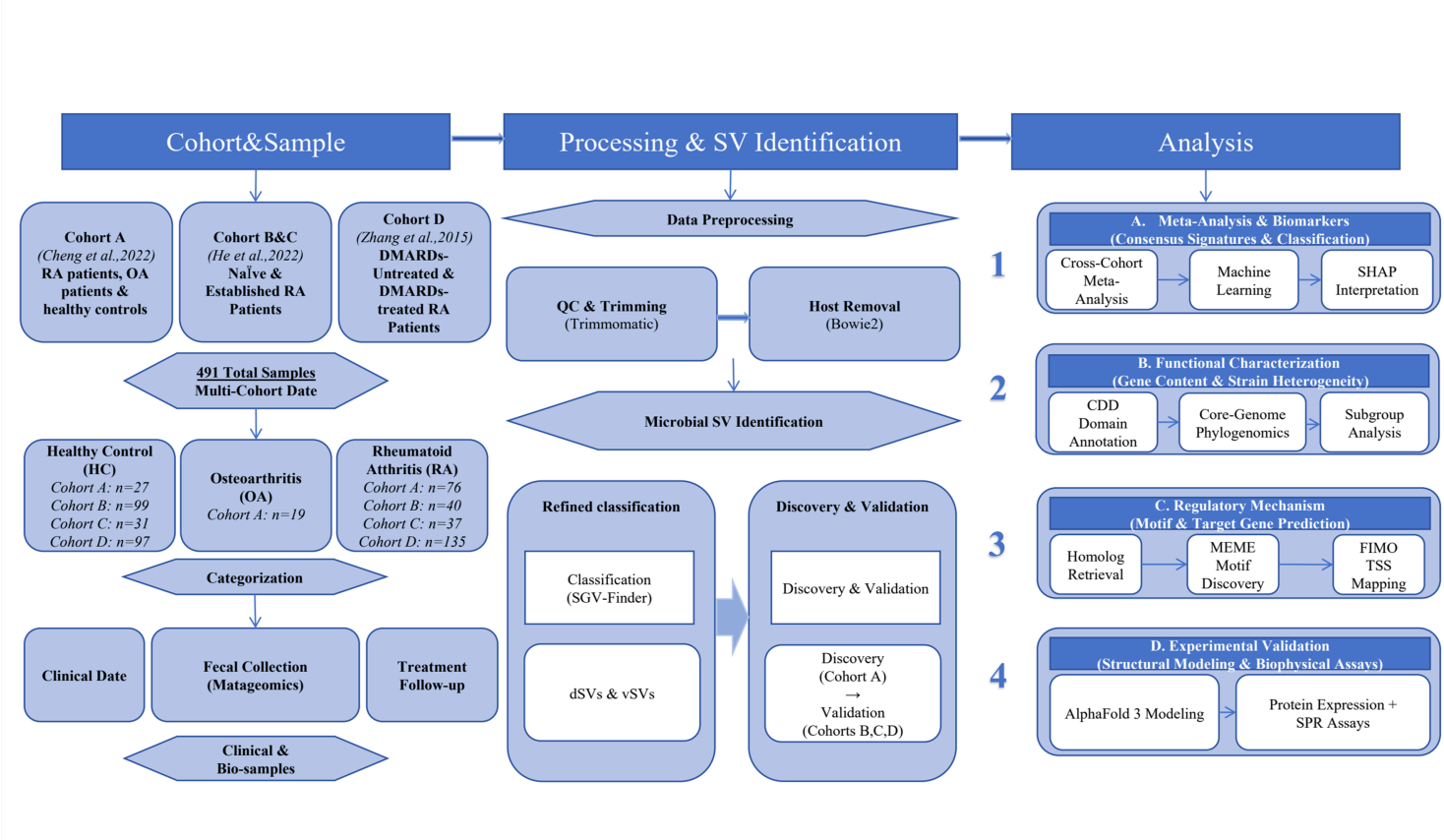
9. Pineda Mde, L. *et al.* A randomized, double-blinded, placebo-controlled pilot study of probiotics in active rheumatoid arthritis. *Med Sci Monit* **17**, Cr347-354 (2011). https://doi.org:10.12659/msm.881808
10. Zeevi, D. *et al.* Structural variation in the gut microbiome associates with host health. *Nature* **568**, 43-48 (2019). https://doi.org:10.1038/s41586-019-1065-y
11. Chen, L. *et al.* Influence of the microbiome, diet and genetics on inter-individual variation in the human plasma metabolome. *Nat Med* **28**, 2333-2343 (2022). https://doi.org:10.1038/s41591-022-02014-8
12. McDonnell, G. E. & McConnell, D. J. Overproduction, isolation, and DNA-binding characteristics of Xre, the repressor protein from the Bacillus subtilis defective prophage PBSX. *J Bacteriol* **176**, 5831-5834 (1994). https://doi.org:10.1128/jb.176.18.5831-5834.1994
13. Scher, J. U. *et al.* Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. *Elife* **2**, e01202 (2013). https://doi.org:10.7554/eLife.01202
14. Maeda, Y. *et al.* Dysbiosis Contributes to Arthritis Development via Activation of Autoreactive T Cells in the Intestine. *Arthritis & Rheumatology* **68**, 2646-2661 (2016). https://doi.org:https://doi.org/10.1002/art.39783
15. Van Rossum, T., Ferretti, P., Maistrenko, O. M. & Bork, P. Diversity within species: interpreting strains in microbiomes. *Nat Rev Microbiol* **18**, 491-506 (2020). https://doi.org:10.1038/s41579-020-0368-1
16. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55-60 (2012). https://doi.org:10.1038/nature11450
17. Pasolli, E., Truong, D. T., Malik, F., Waldron, L. & Segata, N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLoS Comput Biol* **12**, e1004977 (2016). https://doi.org:10.1371/journal.pcbi.1004977
18. Wirbel, J. *et al.* Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat Med* **25**, 679-689 (2019). https://doi.org:10.1038/s41591-019-0406-6
19. Vujkovic-Cvijin, I. *et al.* Host variables confound gut microbiota studies of human disease. *Nature* **587**, 448-454 (2020). https://doi.org:10.1038/s41586-020-2881-9
20. Wang, D. *et al.* Characterization of gut microbial structural variations as determinants of human bile acid metabolism. *Cell Host Microbe* **29**, 1802-1814.e1805 (2021). https://doi.org:10.1016/j.chom.2021.11.003
21. Zhernakova, D. V. *et al.* Host genetic regulation of human gut microbial structural variation. *Nature* **625**, 813-821 (2024). https://doi.org:10.1038/s41586-023-06893-w
22. Liu, W. *et al.* Bacterial genomic structural variations in children with autism serve as diagnostic biomarkers. *Gut* **75**, 937-948 (2026). https://doi.org:10.1136/gutjnl-2025-337280
23. Liu, R. *et al.* Gut microbial structural variation associates with immune checkpoint inhibitor response. *Nat Commun* **14**, 7421 (2023). https://doi.org:10.1038/s41467-023-42997-7

24. Millman, A. *et al.* An expanded arsenal of immune systems that protect bacteria from phages. *Cell Host Microbe* **30**, 1556-1569.e1555 (2022). https://doi.org:10.1016/j.chom.2022.09.017
25. Trouillon, J., Ragno, M., Simon, V., Attrée, I. & Elsen, S. Transcription Inhibitors with XRE DNA-Binding and Cupin Signal-Sensing Domains Drive Metabolic Diversification in Pseudomonas. *mSystems* **6**, 10.1128/msystems.00753-00720 (2021). https://doi.org:10.1128/msystems.00753-20
26. Koh, A., De Vadder, F., Kovatcheva-Datchary, P. & Bäckhed, F. From Dietary Fiber to Host Physiology: Short-Chain Fatty Acids as Key Bacterial Metabolites. *Cell* **165**, 1332-1345 (2016). https://doi.org:10.1016/j.cell.2016.05.041
27. Nayak, R. R. *et al.* Methotrexate impacts conserved pathways in diverse human gut bacteria leading to decreased host immune activation. *Cell Host Microbe* **29**, 362-377.e311 (2021). https://doi.org:10.1016/j.chom.2020.12.008
28. Cheng, M. *et al.* Stage-specific roles of microbial dysbiosis and metabolic disorders in rheumatoid arthritis. *Ann Rheum Dis* **81**, 1669-1677 (2022). https://doi.org:10.1136/ard-2022-222871
29. He, J. *et al.* Intestinal butyrate-metabolizing species contribute to autoantibody production and bone erosion in rheumatoid arthritis. *Sci Adv* **8**, eabm1511 (2022). https://doi.org:10.1126/sciadv.abm1511
30. Zhang, X. *et al.* The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat Med* **21**, 895-905 (2015). https://doi.org:10.1038/nm.3914
31. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014). https://doi.org:10.1093/bioinformatics/btu170
32. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012). https://doi.org:10.1038/nmeth.1923
33. Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* **10** (2021). https://doi.org:10.7554/eLife.65088
34. Mende, D. R. *et al.* proGenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes. *Nucleic Acids Res* **45**, D529-d534 (2017). https://doi.org:10.1093/nar/gkw989
35. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009). https://doi.org:10.1093/bioinformatics/btp352
36. Aroney, S. T. N. *et al.* CoverM: read alignment statistics for metagenomics. *Bioinformatics* **41** (2025). https://doi.org:10.1093/bioinformatics/btaf147
37. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010). https://doi.org:10.1093/bioinformatics/btp616
38. Viechtbauer, W. Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software* **36**, 1 - 48 (2010). https://doi.org:10.18637/jss.v036.i03
39. Balduzzi, S., Rücker, G. & Schwarzer, G. How to perform a meta-analysis with R: a practical tutorial. *Evid Based Ment Health* **22**, 153-160 (2019). https://doi.org:10.1136/ebmental-2019-300117

40. Chinn, S. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Stat Med* **19**, 3127-3131 (2000). https://doi.org:10.1002/1097-0258(20001130)19:22<3127::aid-sim784>3.0.co;2-m
41. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267-288 (1996). https://doi.org:10.1111/j.2517-6161.1996.tb02080.x
42. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33**, 1-22 (2010).
43. Breiman, L. Random Forests. *Machine Learning* **45**, 5-32 (2001). https://doi.org:10.1023/A:1010933404324
44. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* **77**, 1 - 17 (2017). https://doi.org:10.18637/jss.v077.i01
45. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837-845 (1988).
46. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011). https://doi.org:10.1186/1471-2105-12-77
47. Lundberg, S. & Lee, S.-I. *A Unified Approach to Interpreting Model Predictions*. (2017).
48. Lu, S. *et al.* CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res* **48**, D265-d268 (2020). https://doi.org:10.1093/nar/gkz991
49. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069 (2014). https://doi.org:10.1093/bioinformatics/btu153
50. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol* **21**, 180 (2020). https://doi.org:10.1186/s13059-020-02090-4
51. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010). https://doi.org:10.1371/journal.pone.0009490
52. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202-208 (2009). https://doi.org:10.1093/nar/gkp335
53. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493-500 (2024). https://doi.org:10.1038/s41586-024-07487-w
54. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol* **13**, e1005659 (2017). https://doi.org:10.1371/journal.pcbi.1005659
55. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **11**, 3696-3713 (2015). https://doi.org:10.1021/acs.jctc.5b00255

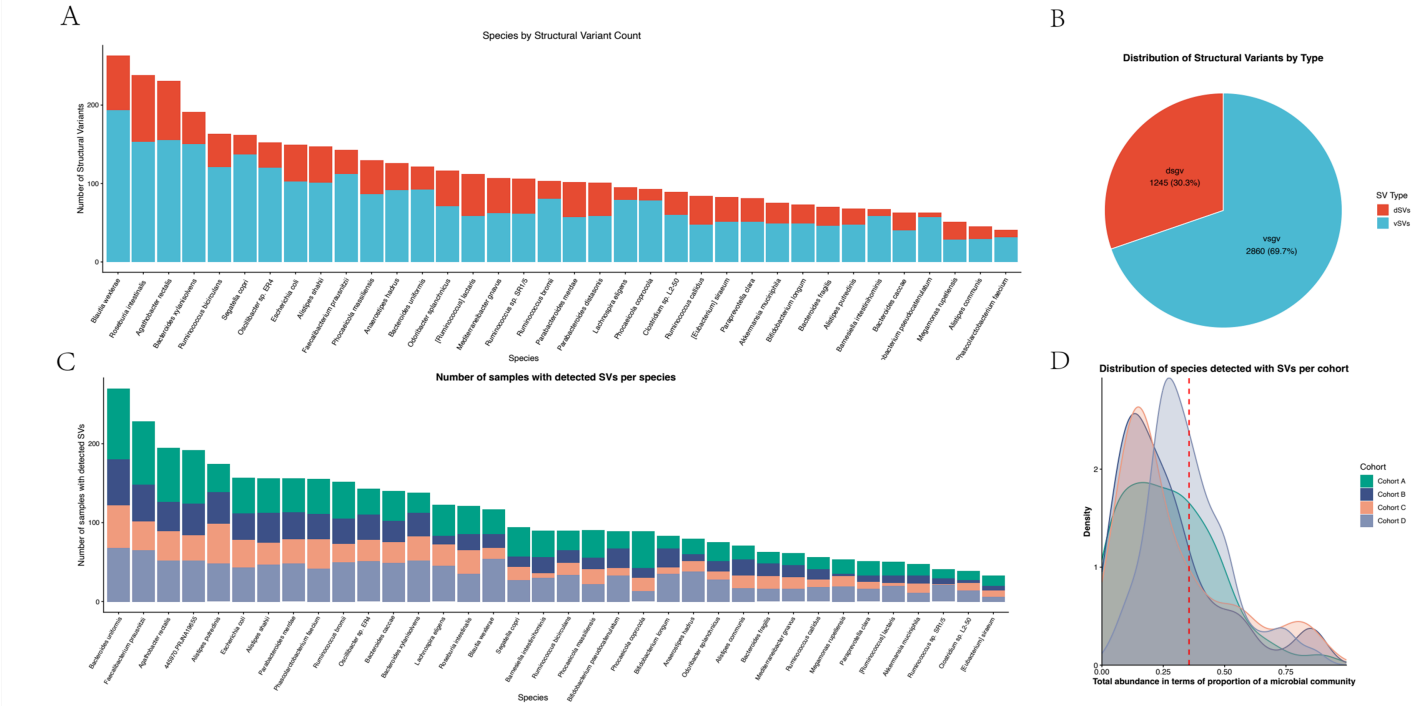
56. Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics***26**, 689-691 (2010). https://doi.org:10.1093/bioinformatics/btq007

# Figures




## Figure 1

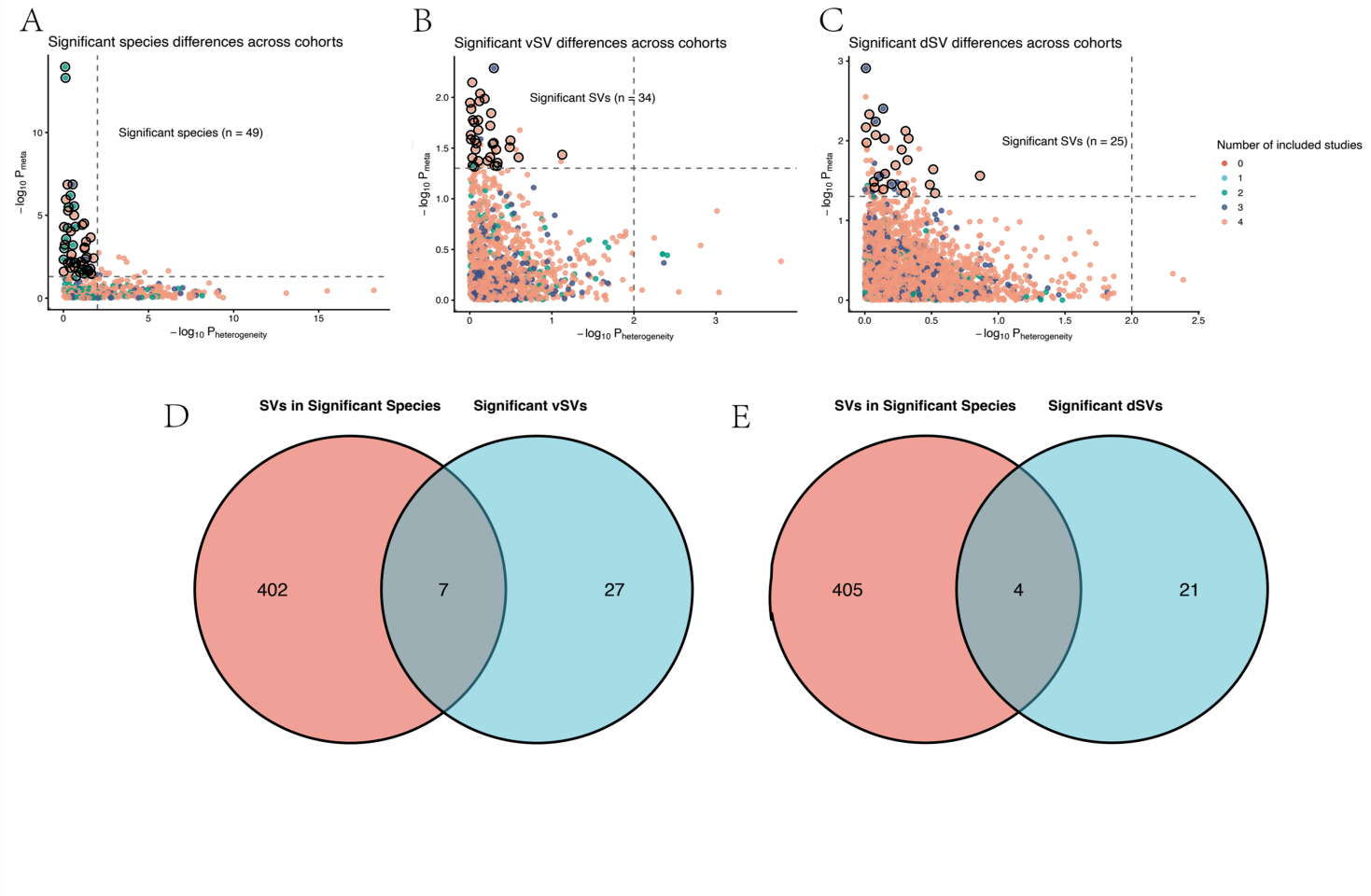
**Study workflow and analytical framework for profiling microbial structural variants (SVs) across multi-center cohorts.**




## Figure 2

**Landscape, distribution, and cohort-level prevalence of gut microbial structural variants (SVs).**

(A) Distribution of detected SV counts across bacterial species, color-coded by SV classification. (B) Proportional distribution of identified SV types, highlighting deletion SVs (dSVs, 30.3%) and variable SVs (vSVs, 69.7%). (C) Number of human stool samples harboring detected SVs per species, stratified across independent study cohorts. (D) Relative abundance density distributions of bacterial species carrying detected SVs within microbial communities across cohorts.




## Figure 3

**Meta-analytic discovery and overlap of species-level and structural variant-level differences across multi-center cohorts.**

(A–C) Volcano plots displaying random-effects meta-analysis results for species-level abundance differences (A), variable SVs (vSVs) (B), and deletion SVs (dSVs) (C) across cohorts. Horizontal dashed lines denote significance thresholds ($P_{meta}$ < 0.05) and vertical dashed lines mark heterogeneity limits ($P_{heterogeneity}$). Point sizes indicate the number of included studies. (D, E) Venn diagrams illustrating the overlap between SVs located within significantly altered bacterial species and significantly differentiated vSVs (D) or dSVs (E).

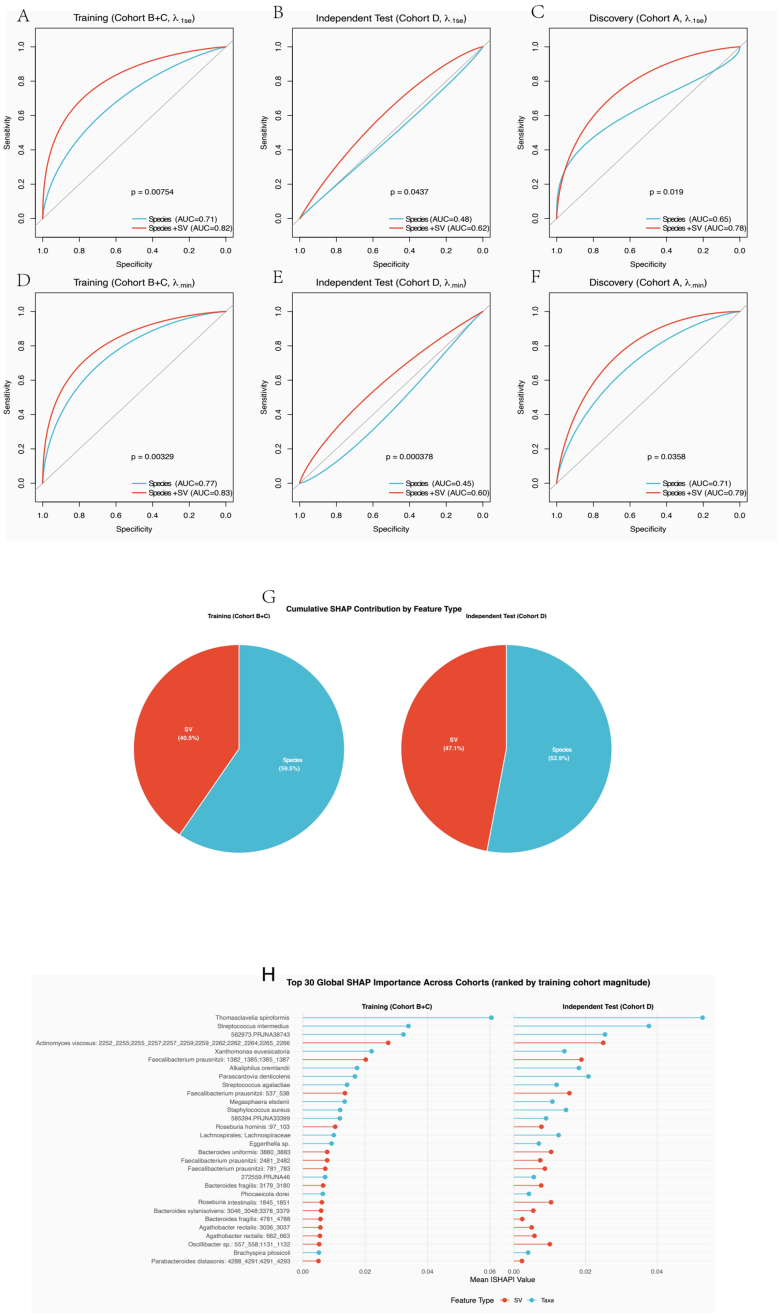



Figure 4

## Figure 4

**Cross-cohort performance, classification discrimination, and global feature importance of microbial SV machine learning models.**

(A–F) Receiver operating characteristic (ROC) curves evaluating disease classification performance using species-only models (blue) versus combined species-plus-SV models (red) across Training Cohort

B+C (A, D), Independent Test Cohort D (B, E), and Discovery Cohort A (C, F) evaluated under alternate regularization parameters ($\lambda_{1se}$ and $\lambda_{min}$). Area Under the Curve (AUC) values and DeLong test P-values are indicated. (G) Cumulative SHAP contribution pie charts comparing the overall predictive weight of SV features versus species features in training and test cohorts. (H) Top 30 global SHAP importance ranking of predictive features across cohorts, ranked by mean absolute SHAP value in the training cohort (B + C).

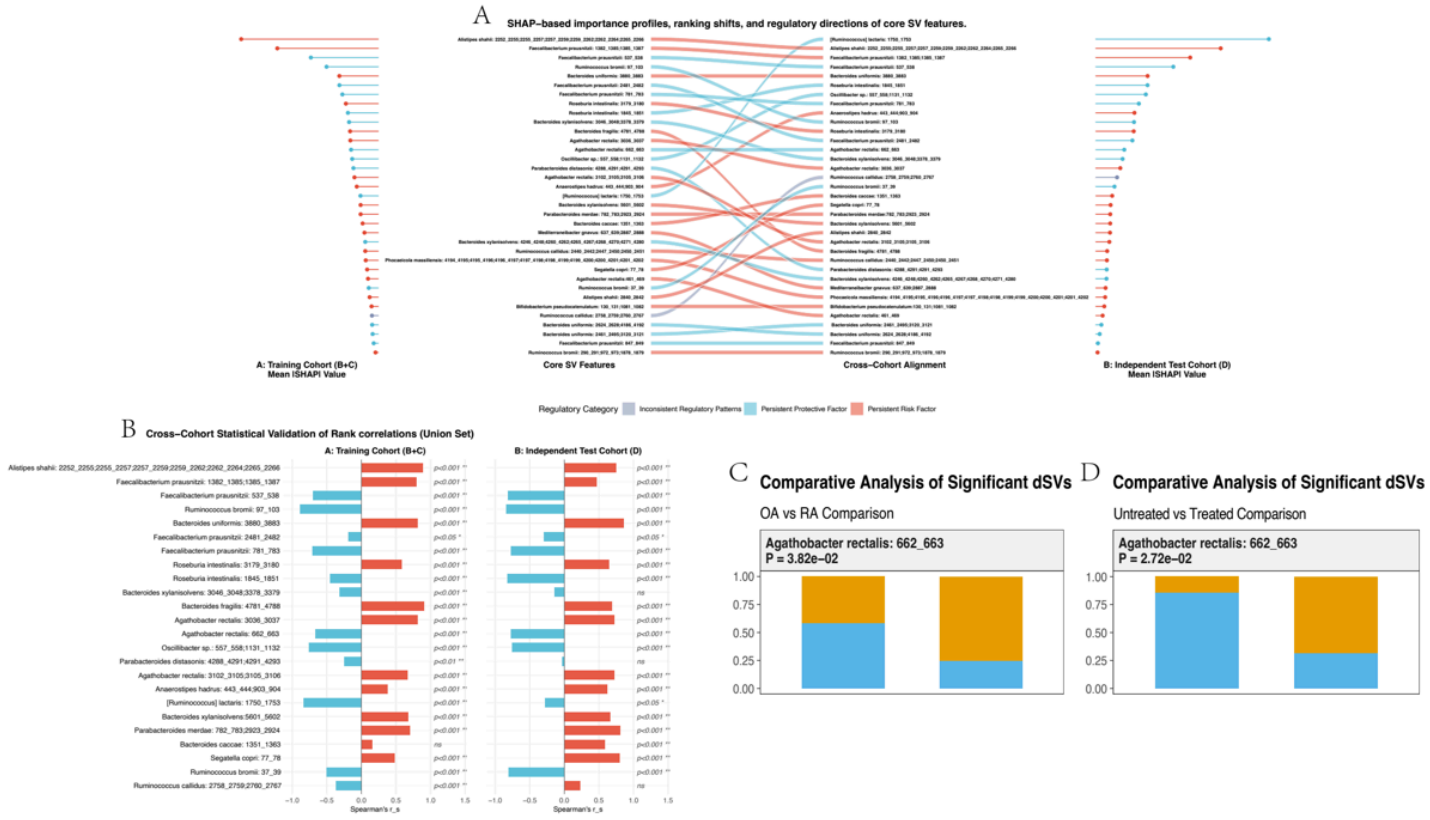
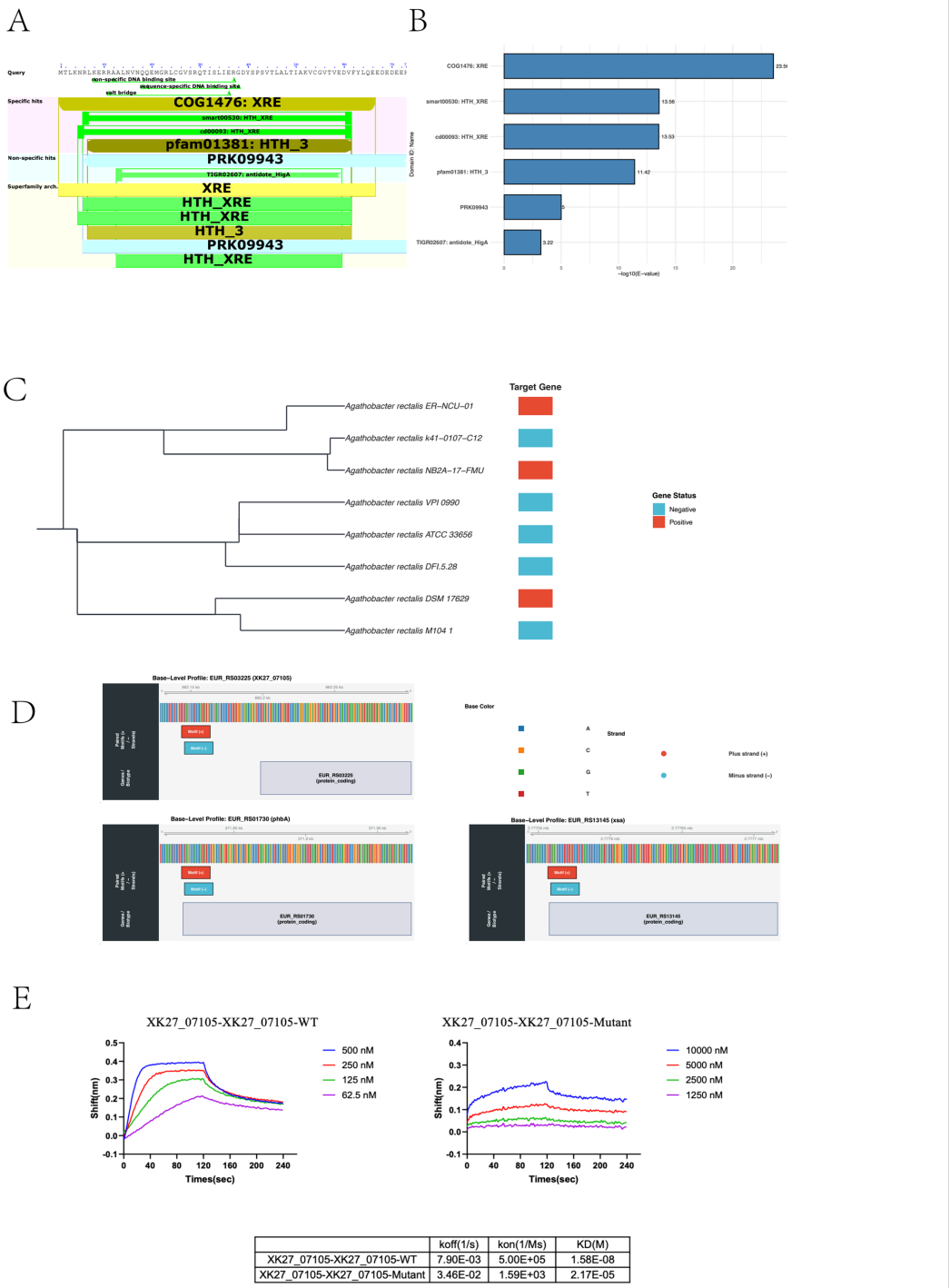



**Figure 5**

**Cross-cohort alignment, ranking shifts, and regulatory consistency of consistently prioritized SV features.**

(A) Cross-cohort statistical validation and alluvial plot mapping rank shifts and mean |SHAP| importance values of top SV features (union of top 20 ranked features in Training Cohort B+C or Independent Test Cohort D). Features are color-coded by regulatory directionality, highlighting persistent protective factors, persistent risk-associated features, and inconsistent regulatory patterns. (B) Spearman correlation matrices illustrating cross-cohort consistency and alignment of feature importance rankings. (C) Comparative prevalence analysis of representative dSVs (*Agathobacter rectalis*: 662_663) between disease subgroups (OA vs. RA, left; Untreated vs. Treated RA, right).




| | koff(1/s) | kon(1/Ms) | KD(M) |
|---|---|---|---|
| XK27_07105-XK27_07105-WT | 7.90E-03 | 5.00E+05 | 1.58E-08 |
| XK27_07105-XK27_07105-Mutant | 3.46E-02 | 1.59E+03 | 2.17E-05 |

Figure 6

## Figure 6

**Functional domain annotation, catalytic profiling, and genomic architecture of the SV-associated transcriptional regulator EUR_RS03225.**

(A) Conserved domain architecture and sequence site profiling of *EUR_RS03225*, highlighting DNA-binding domains, salt bridges, and specific contact sites. (B) Domain significance analysis indicating

primary hits dominated by COG1476: XRE (-$\log_{10}$ E-value = 23.56) and smart00530: HTH_XRE (-$\log_{10}$ E-value = 13.56). (C) Strain-level phylogenetic distribution showing presence (red) or absence (blue) of the target regulator locus across *Agathobacter rectalis* isolates. (D) Base-level genomic track profiling of paired palindromic binding motifs located within the promoter and regulatory regions of candidate target genes: the autoregulatory locus *EUR_RS03225 (XK27_07105)*, the lipid metabolism-associated gene *EUR_RS01730 (phbA)*, and the carbohydrate degradation gene EUR_RS13145 (xsa). Genomic tracks show nucleotide coordinates, four-color single-base sequence tracks (A, blue; C, orange; G, green; T, red), paired reverse-complement motifs on positive (red) and negative (blue) strands, and annotated gene models (grey boxes). (E) Bio-layer interferometry (BLI) binding sensorgrams and corresponding kinetic parameter summary table for recombinant EUR_RS03225/XK27_07105 protein binding to immobilized 60-bp biotinylated promoter DNA duplexes of its own gene locus (*EUR_RS03225/XK27_07105*). Left: Wild-type (WT) promoter duplex assayed across two-fold dilutions (62.5 nM to 500 nM; KD = 1.58 $^{-08}$ M). Right: Core-mutated (Mutant) duplex substituting the central 20-bp motif with a GC-rich sequence, tested at concentrations from 1,250 nM to 10,000 nM (KD = 2.17 $^{-05}$ M). Association rate constants (kon), dissociation rate constants (koff), and equilibrium dissociation constants (KD) derived from a 1:1 Langmuir binding model confirm high-affinity autoregulatory feedback that is abolished upon binding site mutation.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigureSMerged.pdf
- supplementarytables.pdf