Results Comparison Table

Llm Result Id	Peer Result Id	Llm Status	Peer Status	Agreement Status	Notes	N Llm	N Peer	N Itx	Llm Reasoning	Peer Reasoning
R4	R1	SUPPORTED	UNCERTAIN	disagree	Strongly overlapping claims on k-mer PCA/GLM for structure correction (C4,C49,C53,C64–C70,C134–C138). LLM deems demonstrations sufficient; reviewers are uncertain and request head-to-head with SNP-PCA, clearer covariates, and post-adjustment QQ plots.	15	16	15	PCA on a large random subset of k-mers separates YRI and TSI, with interpretability of PCs and clear separation shown (C64–C67, C134–C135). Incorporating PCs and other covariates in logistic models removes k-mer signals as expected under perfect separation, demonstrating adjustment capability and confounder assessment (C53, C68–C70, C136–C137). Simulations further show detectability of associations under confounding control, supporting feasibility (C138), thus validating inference and correction (C4, C49).	Reviewers raised serious concerns about how population structure and other confounders are accounted for and requested thorough demonstration. They asked for a head-to-head comparison with standard genetics PCA, clarity on the logistic regression covariates (gender, sequencing depth), QQ plots after adjustment, and noted potential misinterpretation of PCs (e.g., PC1 correlating with depth). Thus, adequacy of the k-mer PCA/GLM approach to infer and correct structure is not yet established.
R5	R5	SUPPORTED	UNCERTAIN	disagree	E. coli blaTEM validation (C5,C50,C139–C146,C149) is treated as confirmatory by LLM; reviewers remain uncertain pending a head-to-head with traditional variant calling to establish added value.	11	13	11	In E. coli, thousands of significant k-mers assemble into sequences mapping to regions including the blaTEM-1 gene, with strongest signals within or upstream of the causal gene (C139–C145). The QQ and Manhattan plot interpretations are consistent with true signal enrichment around blaTEM-1 (C146), and a conventional mapping/variant-calling pipeline failed to detect genome-wide significant hits (C149). This strongly validates the approach on a known resistance phenotype (C5, C50).	Reviewers considered the E. coli analysis closer to a true association study but asked whether traditional variant calling plus association would recover the same gene and to demonstrate what is added by HAWK. Until a clear head-to-head comparison shows added value, using this as definitive validation of HAWK is premature.
R3	R8	SUPPORTED	UNCERTAIN	disagree	Detection of SVs/non-reference signals (C3,C105–C107,C112–C113,C153–C155). LLM sees sufficient evidence; reviewers call it provisional until SNP-only analyses and QQ calibration demonstrate false-positive control.	15	10	9	HAWK identifies associations spanning indels, structural variants, and CNVs (C105–C107), and assembles sequences longer than the SNP-limited maximum, consistent with multi-variant/structural signals (C106–C108). Numerous associated sequences do not map to the human reference and include BLAST matches to non-reference human sequences or EBV contamination (C112–C114, C116–C118, C153–C155), supporting detection outside the reference genome (C3). These observations collectively substantiate the claim across multiple analyses.	Although the work is proposed to be well suited for structural variants, reviewers emphasized that these harder-to-verify claims should be supported by well-calibrated SNP-only analyses demonstrating false-positive control. Until such calibration is provided, assertions about detecting structural variants and non-reference regions should be viewed as provisional.

Llm Result Id	Peer Result Id	Llm Status	Peer Status	Agreement Status	Notes	N Llm	N Peer	N Itx	Llm Reasoning	Peer Reasoning
R2	R3	SUPPORTED	UNCERTAIN	disagree	1000 Genomes YRI–TSI overlap/'largely agree' claims (C2,C46,C91,C93,C95–C97,C150). LLM supports; reviewers are uncertain and request SNP-restricted tests, genotype dosages, data parity, and QQ plots.	9	14	8	In the YRI–TSI comparison, 80.3% of significant SNPs from genotype-based analysis are covered by HAWK sequences and 95.2% by at least one k-mer (C96–C97), indicating substantial concordance. The study design and counts are reported (C91, C93, C95), and the conclusion that results "largely agree" is consistent with these overlaps (C2, C46, C150). QQ plot interpretation (C94) does not contradict this conclusion.	For the 1000 Genomes comparison, reviewers requested analyses restricted to SNPs, inclusion of genotype dosages, confirmation that both approaches use the same data, and QQ plots to show false-positive control. Without these targeted checks, the claims of 'largely agree' and overlap-based validation are not fully substantiated.
R7	R4	SUPPORTED	UNSUPPORTED	disagree	BEB–TSI allele-frequency differences in CVD-linked genes (C48,C124,C125). LLM supports inter-population differences; reviewers judge unsupported due to using population identity instead of disease status and resultant confounding/interpretability issues.	10	3	3	The BEB–TSI analysis reports statistically significant frequency differences for multiple non-synonymous variants in CVD-linked genes, including ApoB SNPs rs1042034 and rs676210, and other loci such as SH2B3 and TTN (C48, C109, C125, C130–C131). Prior literature contextualizes disease relevance for these SNPs (C126–C129). While no CVD phenotypes were analyzed (C124), the stated result is about inter-population allele frequency differences in CVD-related genes and is supported.	Reviewers stated the CVD study 'seems very odd' because population identity (BEB vs TSI) was used instead of disease status, so detected k-mers likely reflect ancestry rather than disease. They found it almost impossible to interpret whether results reflect population differences or disease association and questioned the focus on genes already linked to CVD, undermining the evidentiary value of these findings.
R6	R2	SUPPORTED	UNCERTAIN	disagree	Modeling/power for count-based tests (overlap on C61,C72). LLM supports Poisson-count approach via simulations; reviewers are uncertain about Poisson assumption, overdispersion, and multiple-testing calibration, requesting NB baselines and QQ plots.	6	7	2	Multiple simulations show Poisson count-based tests have greater power than logistic tests and presence/absence approaches, with specific scenarios where presence/absence fails (e.g., 2 vs 1 copy) (C72–C73, C90). They also observe Poisson-vs-NB p-values are comparable for typical k-mers (C61). Together these support the stated power advantage of using k-mer counts (C37, C151), though the evidence is primarily simulation-based.	Reviewers challenged the Poisson assumption and reliance on Bonferroni, requesting empirical evidence that counts are Poisson, baseline negative binomial tests, and calibration (QQ plots) to rule out overdispersion and confounding. They also noted that stating Bonferroni is conservative is insufficient and asked for discussion of appropriate error measures (e.g., q-values). Consequently, the adequacy of the modeling and multiple-testing strategy remains uncertain.

Llm Result Id	Peer Result Id	Llm Status	Peer Status	Agreement Status	Notes	N Llm	N Peer	N Itx	Llm Reasoning	Peer Reasoning
R1	R2	SUPPORTED	UNCERTAIN	partial	Partial overlap (C52,C59,C60,C62). LLM assesses method existence/implementation as supported; reviewers' uncertainty targets distributional assumptions and error control rather than existence, so scopes differ.	18	7	4	The manuscript clearly describes an alignment-free k-mer based association pipeline and its implementation (C1, C51, C52, C55–C60). Software, parameters, and components (ABySS, Jellyfish, Eigenstrat modifications) are specified and released (C75–C76, C79–C84), supporting that the method exists and is executable. This is a methodological contribution rather than an empirical hypothesis and is adequately documented.	Reviewers challenged the Poisson assumption and reliance on Bonferroni, requesting empirical evidence that counts are Poisson, baseline negative binomial tests, and calibration (QQ plots) to rule out overdispersion and confounding. They also noted that stating Bonferroni is conservative is insufficient and asked for discussion of appropriate error measures (e.g., q-values). Consequently, the adequacy of the modeling and multiple-testing strategy remains uncertain.
R1	R7	SUPPORTED	UNCERTAIN	partial	Minimal overlap (C51). LLM focuses on tool availability/description; reviewers question broader applicability beyond case–control and to other sequencing data, making the comparison scope-mismatched.	18	6	1	The manuscript clearly describes an alignment-free k-mer based association pipeline and its implementation (C1, C51, C52, C55–C60). Software, parameters, and components (ABySS, Jellyfish, Eigenstrat modifications) are specified and released (C75–C76, C79–C84), supporting that the method exists and is executable. This is a methodological contribution rather than an empirical hypothesis and is adequately documented.	Reviewers noted HAWK appears only applicable to case-control traits and requested making this limitation explicit and explaining (ideally demonstrating) how to extend to quantitative traits. They also asked to define the scope beyond generic ' sequencing data' with empirical support (e.g., RNA-seq). Thus, broader applicability and proposed extensions are not yet substantiated.
	R6		UNCERTAIN	partial	Peer-only claims on Cortex vs HAWK scalability and reference use (C27–C29,C40–C42) have no corresponding LLM evaluation.		6	0		Reviewers asked the authors to explain why Cortex is not suited to large numbers of individuals and to provide a comparison with HAWK, as well as to comment on leveraging reference information as highlighted by Iqbal et al. Without these analyses and comparisons, the scalability and reference-use claims remain insufficiently supported.