LOCUSGATE: OPERATOR-NEUTRAL GRAPH WORKSPACES FOR COMPOSABLE PANGENOME QUERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-stage pangenome analyses revisit one locus for sequence membership, path extraction, and neighborhood inspection, yet independently localized stages can expose different graph evidence. We make locality an explicit operator-neutral value with LocusGate, which materializes $\mathcal{W}(q) = \langle G', \chi \rangle$ before dispatch from complete path-window anchors and deterministic budgeted expansion. The certificate χ binds visible graph content, construction settings, canonical ordering, and the answer domain of each consumer. This interface separates three guarantees: exact membership over declared embedded paths for $|q| \geq w+k-1$, deterministic extraction complete within G', and bounded neighborhood inspection over the certified traversal domain. We prove conditioned exactness for membership and show that certificate equality makes handoff and deterministic rematerialization produce identical normalized outputs for a fixed deterministic operator. By making region identity, guarantee scope, and immutable state travel together, LocusGate gives multi-stage pangenome queries one named body of graph evidence rather than operator-private locality choices.

1 INTRODUCTION

A locus in a pangenome workflow is rarely queried only once. A pipeline may first test a 41-mer, then extract the paths crossing a 10 kb interval, and finally inspect nearby alleles or branches for graph-aware mapping, genotyping, or biological inference (Paten et al., 2017; Eizenga et al., 2020; Dilthey et al., 2015; Garrison et al., 2018; Ebler et al., 2022; Liao et al., 2023). Today these stages usually localize through separate operator state: the membership index follows its keys, the extractor follows its seeds, and the inspector selects its own local view. Consequently, requests that name the same locus can expose different graph content, making both the answer domain and the measured cost depend on an implicit region choice.

Task-specific graph indexes and local-view tools solve important localization problems, but their reusable state remains tied to an operator or selected within its interface. Multi-stage analysis instead needs a query-scoped graph value whose content is fixed before a stage is chosen and whose identity records why that content is present. Such a value must also carry the answer domain of each consumer: exact presence over declared paths is a different promise from complete path enumeration inside a bounded subgraph or recovery under a finite inspection budget.

We introduce LocusGate to provide this interface. For canonical query key q, complete path-window anchors identify mandatory candidates and deterministic budgeted expansion materializes the bounded oriented subgraph $G' = (V', E')$. The accompanying certificate χ binds graph and query digests, anchors, traversal budgets, ordering conventions, implementation version, and the canonical content digest, yielding the operator-neutral workspace $\mathcal{W}(q) = \langle G', \chi \rangle$. Full-interval verification gives exact membership over declared embedded paths for $|q| \geq w+k-1$; stable enumeration is sound, deterministic, and complete for qualifying walks in G'; and a fixed deterministic rule chooses the inspection budget before dispatch. Region identity, guarantee domain, and reusable decoded state therefore travel together.

The central technical question is whether a local graph value can cross operator boundaries without allowing an operator to redefine the visible evidence. LocusGate admits membership-to-extraction,

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

membership-to-inspection, extraction-to-inspection, and inspection-to-extraction as compositions over the same certified workspace. Certificate equality fixes the graph and query digests, anchor order, traversal budget, construction version, and content digest, so a deterministic receiver obtains the same normalized value whether it consumes a handoff or rematerializes the workspace.

The contribution is a content-certified graph value that makes locality an explicit compositional interface rather than hidden operator state. Its deterministic construction couples complete path-window localization to a guarantee hierarchy for membership, extraction, and inspection. The paper formalizes the object identity and dispatch boundary, gives a complete materialization algorithm, proves conditioned membership exactness, and proves deterministic handoff equivalence. Together, these elements show how heterogeneous pangenome queries can reason over one named body of local graph evidence while retaining a precise answer-domain contract.

2 Related Work

Pangenome construction and inference systems make population-scale graph analysis practical (Paten et al., 2017; Eizenga et al., 2020; Li et al., 2020; Liao et al., 2023; Hickey et al., 2024; Ebler et al., 2022), but their reusable query objects have different lifecycles. VG, GCSA2, GBWT, and GraphAligner retain mapper- or index-specific state (Garrison et al., 2018; Sirén, 2017; Sirén et al., 2020; Rautiainen & Marschall, 2020). That state accelerates its associated operation, whereas a query-scoped region is not presented as the immutable common input to heterogeneous operators. ODGI and VRPG can expose local views for inspection (Guarracino et al., 2022; Miao & Yue, 2025); the view boundary is selected for that interface rather than certified before an unknown downstream task. BIGSI, Mantis, Bifrost, SeqOthello, and SBT provide exact presence retrieval over an index key space (Solomon & Kingsford, 2016; Bradley et al., 2019; Pandey et al., 2018; Holley & Melsted, 2020; Yu et al., 2018), but that reusable object serves one answer family rather than membership, path extraction, and neighborhood inspection together.

General systems abstractions make reuse explicit in different ways: memoization retains results by function input (Michie, 1968), materialized views support query answering through precomputed relations (Halevy, 2001), and immutable dataflow objects preserve reusable intermediates and lineage across stages (Zaharia et al., 2012; Akidau et al., 2015). These mechanisms identify a computation or stored value, but they do not define pangenome-local graph evidence visible to several query families or the guarantee each family has under bounded traversal. LocusGate contributes that domain contract by combining pre-dispatch region selection, content-derived identity, and operator-specific answer semantics over the identical G'. The relevant choices are fields of χ and properties of operators over G', so changing cache residency cannot change the graph value. Appendix A and Table 2 compare these lifecycle and invariant differences directly.

3 Method

For a directed sequence graph $G = (V, E)$ and canonical query key q, LocusGate constructs an operator-neutral workspace $\mathcal{W}(q) = \langle G', \chi \rangle$ before selecting a downstream task. The immutable $G' = (V', E')$ contains the visible oriented graph. Certificate χ binds graph and query digests, canonical anchors, finite traversal bounds, ordering rules, materializer version, and the canonical digest of G'. A standalone membership key is its normalized sequence. A locus key additionally identifies a declared path, orientation, source coordinate, sequence context, and requested interval; this form lets heterogeneous operators name one locus without changing workspace identity. Operator settings and output-format choices remain outside the key. Membership, extraction, and neighborhood inspection consume the same $\mathcal{W}(q)$, so region identity, answer scope, and reusable read-only state are fixed at the pre-dispatch boundary. Figure 1 presents that boundary.

Figure 1 and Table 1 define the workspace type: G' fixes visible graph content, χ fixes its identity and construction, and each operator declares its answer domain over that value. This coupling is the technical distinction from reusing cached bytes alone: operator-specific indexes can accelerate a computation, but they cannot alter the certified region or attached guarantee domain after dispatch.

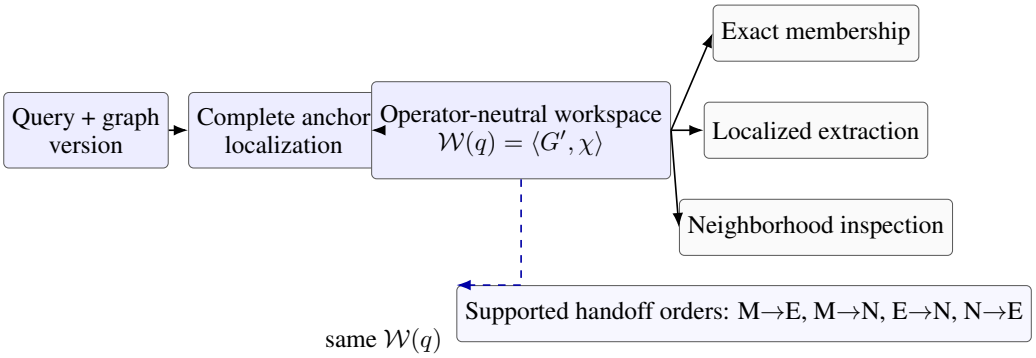


Figure 1: Operator-neutral materialization. LocusGate localizes a query and certifies one immutable workspace $\mathcal{W}(q)$ before dispatch. A receiver consumes the identical value, keeping region selection stable across supported operator orders.

Table 1: Answer semantics carried by the operator-neutral workspace. Each contract is stated over the certified workspace rather than over an operator-private region.

Operator	Guarantee	Certified answer domain	Required condition
Membership	Exact	Literal occurrences on declared embedded paths for $\|q\| \geq w+k-1$	Complete path-window indexing and full-interval verification
Localized extraction	Sound, deterministic, complete within G'	Qualifying oriented walks from the requested start projection in the certified subgraph	Fixed span and stable enumeration
Neighborhood inspection	Bounded	Path-supported alleles reached under the certified traversal budget	Budget selected before dispatch

3.1 CANDIDATE LOCALIZATION

Each node v has an oriented handle $x = (v, \sigma)$, spelling s_v for $\sigma = +$ and its reverse complement for $\sigma = -$. The finite admissible universe $\mathcal{P}^{\pm}$ contains every declared embedded haplotype path and its reverse-complement path. A path may revisit nodes, so recorded cycles retain their path coordinates; duplicate paths keep distinct identifiers. Graph edges alone do not introduce additional recombinant walks into the membership domain. Sequences are normalized to uppercase IUPAC symbols, with literal equality and the standard complement mapping; an ambiguous symbol matches the same symbol rather than acting as a wildcard.

LocusGate enumerates every length-k substring of every path spelling, including substrings crossing any number of short-node boundaries. Each block of w consecutive k-mers selects the minimum canonical key $\kappa(z) = \min\{z, \mathrm{rc}(z)\}$ under a total order of collision-resolved hash, exact string, and leftmost path coordinate (Schleimer et al., 2003; Roberts et al., 2004). Each selected occurrence enters a compressed posting list with path ID, spelling coordinate, initial handle and offset, strand, and the ordered node slices that spell it. Overlapping windows store an identical path-coordinate occurrence once; distinct paths, visits, and orientations remain distinct.

Queries use the identical normalization, canonicalization, and tie rule. For a query minimizer at offset a and an oriented posting at path coordinate b, candidate initialization sets $t = b - a$ and retains valid intervals $[t, t + |q|)$. The verifier reconstructs every path slice intersecting that interval, so a candidate region contains the complete query occurrence even when it spans more than two nodes or a recorded cycle. Candidate regions form a mandatory core that is never shortened by the contextual traversal bounds. Canonical-key collisions and repetitive anchors can add candidates, but exact key resolution and interval verification determine the Boolean answer.

3.2 BOUNDED MATERIALIZATION

An anchor hit (j, x, o) identifies query-minimizer position j, oriented handle x, and boundary offset o. Exact duplicates are removed. Anchor support of a frontier handle is the number of distinct query-minimizer positions with a posting on its underlying node; duplicate paths and postings contribute at most once per query position. Successors are ordered by decreasing support, stable node ID, and orientation ($+$ before $-$), and the first B are retained independently at each frontier state. Thus B is a per-state expansion cap, not a global workspace-degree cap.

A state (u, h, d) records an admitted handle, the edges traversed from its seed, and spelled bases from the anchor coordinate through the exit of u. Forward traversal from (x, o) starts with $d = |s_x| - o$ and reverse traversal with $d = o$. A successor v is charged in full before admission: $d' = d + |s_v|$, and it is excluded when $d' > L$. The finite hop bound H, length bound L, and per-state successor bound B are determined before expansion by the fixed policy π and recorded in χ.

The queue is strictly FIFO: seeds are appended in canonical node/orientation/offset order and children in ranked order. The Pareto set $\mathcal{D}[u]$ suppresses duplicate or dominated visits; every cycle increases hop count and cannot improve length, ensuring termination. The returned E' is the oriented subgraph induced by V', preserving cross-frontier and cycle-closing edges whose endpoints are visible. The contextual expansion is finite under finite $\mathcal{A}^\star$, B, and H, while mandatory candidate intervals contribute a finite verification core.

Certificate χ records immutable graph and query digests, canonical anchors, (k, w, H, L, B), the complete specification and version of π, orientation and ranking conventions, adjacency and materializer versions, and the canonical G' digest. Cache entries are immutable and reusable only under complete certificate equality. Vertices and edges are serialized canonically; concurrent requests receive logical sequence numbers that determine recency and eviction. A downstream operator consumes the workspace identifier directly and cannot mutate G'. These conditions make cache retrieval byte-equivalent to deterministic rematerialization.

3.3 DETERMINISTIC PRE-DISPATCH BUDGET POLICY

A deterministic inspection policy π computes a branch-pressure statistic from canonical anchor support and ordered successor lists, then maps that statistic to B in a finite, totally ordered budget set. Its input definition, thresholds, ordering, and version are fixed before expansion and become fields of

Algorithm 1 Bounded workspace materialization

procedure MATERIALIZEWORKSPACE($G, q, \mathcal{A}, k, w, H, L, \pi$)
  $\mathcal{A}^\star \leftarrow$ deduplicated hits in canonical order
  $B \leftarrow \pi(G, q, \mathcal{A}^\star)$ from a fixed finite ordered budget set
  $V' \leftarrow$ handles in all mandatory candidate regions
  $\mathcal{Q} \leftarrow$ empty FIFO queue; $\mathcal{D}[u] \leftarrow \emptyset$
  for each distinct (x, o) in $\mathcal{A}^\star$ **do**
    enqueue $(x, 0, |s_x| - o)$ and $(\bar{x}, 0, o)$
  while $\mathcal{Q} \neq \emptyset$ **do**
    $(u, h, d) \leftarrow \text{pop_front}(\mathcal{Q})$
    if some $(h_0, d_0) \in \mathcal{D}[u]$ has $h_0 \leq h$ and $d_0 \leq d$ **then**
      CONTINUE
    remove states in $\mathcal{D}[u]$ dominated by (h, d); insert (h, d)
    $V' \leftarrow V' \cup \{u\}$
    if $h < H$ **then**
      $\mathcal{N} \leftarrow$ successors sorted by $(-\text{support}, \text{ID}, \text{orientation})$
      $\mathcal{N} \leftarrow$ first $\min(B, |\mathcal{N}|)$ entries
      for each $v \in \mathcal{N}$ in order **do**
        $d' \leftarrow d + |s_v|$
        if $d' \leq L$ **then**
          $\text{push_back}(\mathcal{Q}, (v, h + 1, d'))$
  $E' \leftarrow \{(u, v) \in E : u \in V' \wedge v \in V'\}$
  $G' \leftarrow (V', E')$
  $\chi \leftarrow \text{Certify}(G, q, \mathcal{A}^\star, k, w, H, L, B, \pi, \text{digest}(G'))$
  return $\mathcal{W}(q) = \langle G', \chi \rangle$

χ. Consequently, a downstream operator cannot select a different traversal bound after receiving the workspace. The policy changes the bounded inspection domain only through the certified budget; it does not change the exact-membership domain or the mandatory verification core.

3.4 CROSS-OPERATOR HANDOFF

A source locus $\ell = (p, c, \sigma)$ maps its sequence context and interval to one canonical key. Membership, extraction, and inspection therefore differ only in the operator applied to $\mathcal{W}(q_\ell)$. Handoff transfers immutable decoded node, edge, and path-projection state together with χ; lookup, certificate validation, and acquisition cannot modify the object. If T_i^V and T_i^E are the distinct workspace nodes and edges decoded or traversed by operator i, the corresponding reuse ratios are $|T_1^V \cap T_2^V|/|T_2^V|$ and $|T_1^E \cap T_2^E|/|T_2^E|$. These definitions distinguish transferring the complete certified workspace from reusing only the state touched by a prior operator.

Proposition 1 (Handoff equivalence)**.** *Fix an operator, its configuration, and its implementation version. Consuming a handed-off $\mathcal{W}(q)$ and consuming a deterministically rematerialized workspace with the same certificate χ produce identical normalized outputs.*

The result follows from canonical materialization and immutable operator input; Appendix D gives the proof.

3.5 GUARANTEE HIERARCHY

For sequence membership, the operator returns `True` exactly when q is a literal substring of an admissible spelling $S(p)$ for some $p \in \mathcal{P}^{\pm}$. When $|q| \geq w + k - 1$, complete path-window indexing ensures that every true occurrence contributes a selected query minimizer and a candidate interval covering the full occurrence. Acceptance requires equality with that complete interval. Appendix C separates candidate recall from verifier soundness and treats orientations, repeated visits, duplicates, and ambiguous bases explicitly.

For localized extraction, an oriented start projection and a fixed positive span define qualifying walks in G'. Walks start at that base, follow E', and spell the requested span with a clipped terminal slice.

Repeated handles remain distinct visits; only identical handle-order/endpoint tuples are deduplicated. Stable depth-first order, positive-length node labels, fixed span, and no output-count cap make enumeration deterministic, terminating, and complete within G'.

Neighborhood inspection reports path-supported alleles and branches by breadth-first traversal of the fixed workspace selected by the pre-dispatch policy. Its answer is exact for that certified (B, H, L) domain. The three contracts are deliberately distinct: exactness is over declared paths, extraction completeness is over G', and inspection is bounded by the certified traversal domain.

3.6 Cross-system output normalization

Membership outputs normalize to one Boolean per query. Extraction outputs normalize to oriented node and edge sets plus spelled intervals. Neighborhood outputs normalize to path-supported allele identifiers. These schemas make answer equality precise for a fixed task-level obligation while preserving each operator's native representation. Handoff equivalence is stated over these normalized outputs; no schema permits an operator to redefine G' or χ after dispatch.

4 Formal Guarantees

The contribution is organized around the guarantees that remain attached to a certified workspace after dispatch. These guarantees have distinct semantic scopes: membership is exact over the declared path universe under explicit indexing assumptions, extraction is complete within the fixed visible subgraph, and inspection is bounded by a certificate-recorded traversal domain.

4.1 Exact membership over declared paths

The membership contract is deliberately narrower and stronger than a graph-reachability statement. It quantifies only over the finite universe $\mathcal{P}^{\pm}$ of declared embedded paths and their reverse complements. Complete path-window indexing supplies a candidate for every occurrence of a query of sufficient length, and full-interval literal verification rejects every non-occurrence. Lemmas 1 and 2 establish the two directions separately. Their combination yields the following result.

Proposition 2 (Conditioned membership exactness)**.** *Assume an immutable graph and declared path collection, complete selected-occurrence indexing over $\mathcal{P}^{\pm}$, no-false-negative lookup and compression, identical query/index normalization, literal IUPAC equality, and candidate regions exempt from contextual traversal bounds. For every $|q| \geq w + k - 1$, membership returns* `True` *if and only if q occurs in some $S(p)$ for $p \in \mathcal{P}^{\pm}$.*

The proposition makes the boundary of exactness auditable. It includes paths, orientations, repeated visits, and node-boundary crossings in the domain; it excludes recombinant walks introduced only by graph topology. Contextual exploration can improve the workspace available to other operators, but it cannot remove a mandatory verification region or alter this membership result.

4.2 Deterministic compositionality

The second guarantee concerns composition. Once χ is fixed, it binds the graph and query identities, ordered anchor set, finite bounds, construction rules, implementation version, and canonical content digest. Deterministic materialization therefore produces one immutable G'. A receiving operator either consumes this value directly or obtains the same value by rematerialization; neither path permits a receiver to modify the visible graph evidence.

Proposition 1 formalizes this equivalence for a fixed deterministic operator and normalization rule. The result gives the handoff interface a semantic role beyond caching: a downstream computation can require equality of the certified input object rather than infer equality from cache residency, request timing, or a task-specific index key.

4.3 Guarantee-preserving operator contracts

Localized extraction starts from a specified projection and enumerates qualifying walks in G' under stable order. It is sound, deterministic, and complete with respect to that certified subgraph, including

visible cycle-closing edges. Neighborhood inspection instead reports path-supported alleles and branches under the selected finite budget. Its scope is exact for the certified domain, while its boundedness remains explicit in χ.

This hierarchy makes the method useful precisely where a single informal notion of "local graph" is inadequate. A consumer can distinguish exact membership from workspace-complete extraction and from budget-bounded inspection, while still receiving one named graph value. The appendix gives the candidate-recall, verifier-soundness, and handoff-equivalence proofs that connect these contracts to the construction algorithm.

5 Discussion

The operator-neutral workspace turns task-local bounded searches into composition over a stable graph value: region identity is fixed once, χ travels with visible content, and each operator interprets its answer over that value. This is a useful separation for pangenome pipelines because the same locus can require distinct answer domains without allowing each stage to silently choose different graph evidence.

The method also separates three questions that are often conflated. Complete path-window localization and full-interval verification establish exact membership over declared paths under stated conditions. Stable enumeration establishes what extraction means inside the certified subgraph. A deterministic pre-dispatch policy establishes the finite domain for inspection. The certificate records the choice that separates these contracts, so a receiver can reason about the scope of an answer without reconstructing the region-selection decision.

Immutable identity makes handoff compositional. Certificate equality fixes the inputs to deterministic materialization and identifies the same canonical G', while normalized output schemas make equality meaningful for each operator. The resulting interface supplies a principled graph-evidence boundary for systems that combine sequence, path, and neighborhood queries.

The formal scope covers declared embedded paths, deterministic construction, fixed operator configuration, and finite traversal bounds. Within that scope, the paper provides a precise foundation for carrying local pangenome evidence across operator boundaries.

6 Conclusion

LocusGate fixes graph locality as an operator-neutral value before dispatch. Its content certificate binds the visible subgraph, construction rules, and answer-domain contract, supporting exact membership over declared paths, workspace-complete extraction, bounded inspection, and deterministic transfer of immutable graph state. The formal guarantees show how multi-stage pangenome queries can compose heterogeneous operations over one named body of graph evidence while retaining an auditable account of what each answer means.

Reproducibility Statement

The method definition, certificate fields, deterministic materialization algorithm, answer-domain contracts, and proofs are specified in Sections 3–4 and Appendices B–D.

AI Use Statement

Generative AI tools assisted manuscript writing, editing, and presentation. The authors reviewed the edits and remain responsible for the final content.

References

Tyler Akidau, Robert Bradshaw, Craig Chambers, Slava Chernyak, Rafael Fernández-Moctezuma, Reuven Lax, Sam McVeety, Daniel Mills, Frances Perry, Eric Schmidt, and Sam Whittle. The dataflow model: A practical approach to balancing correctness, latency, and cost in massive-scale,

unbounded, out-of-order data processing. *Proceedings of the VLDB Endowment*, 8(12):1792–1803, 2015. doi: 10.14778/2824032.2824076.

Phelim Bradley, Henk C. den Bakker, Eduardo P. C. Rocha, Gil McVean, and Zamin Iqbal. Ultrafast search of all deposited bacterial and viral genomic data. *Nature Biotechnology*, 37(2):152–159, 2019. doi: 10.1038/s41587-018-0010-1.

Alexander Dilthey, Charles Cox, Zamin Iqbal, Matthew R. Nelson, and Gil McVean. Improved genome inference in the MHC using a population reference graph. *Nature Genetics*, 47(6):682–688, 2015. doi: 10.1038/ng.3257.

Jana Ebler, Peter Ebert, Wayne E. Clarke, Tobias Rausch, Peter A. Audano, Torsten Houwaart, et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nature Genetics*, 54(4):518–525, 2022. doi: 10.1038/s41588-022-01043-w.

Robert C. Edgar. Syncmers are more sensitive than minimizers for selecting conserved k-mers in biological sequences. *PeerJ*, 9:e10805, 2021. doi: 10.7717/peerj.10805.

Jordan M. Eizenga, Adam M. Novak, Jonas A. Sibbesen, Simon Heumos, Ali Ghaffaari, Glenn Hickey, et al. Pangenome graphs. *Annual Review of Genomics and Human Genetics*, 21:139–162, 2020. doi: 10.1146/annurev-genom-120219-080406.

Erik Garrison, Jouni Sirén, Adam M. Novak, Glenn Hickey, Jordan M. Eizenga, Eric T. Dawson, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36(9):875–879, 2018. doi: 10.1038/nbt.4227.

Andrea Guarracino, Simon Heumos, Sven Nahnsen, et al. ODGI: understanding pangenome graphs. *Bioinformatics*, 38(13):3319–3326, 2022. doi: 10.1093/bioinformatics/btac308.

Alon Y. Halevy. Answering queries using views: A survey. *The VLDB Journal*, 10(4):270–294, 2001. doi: 10.1007/s007780100054.

Glenn Hickey, Jean Monlong, Jana Ebler, Adam M. Novak, Jordan M. Eizenga, Yan Gao, et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nature Biotechnology*, 42(4):663–673, 2024. doi: 10.1038/s41587-023-01793-w.

Guillaume Holley and Páll Melsted. Bifrost: highly parallel construction and indexing of colored and compacted de Bruijn graphs. *Genome Biology*, 21(1):249, 2020. doi: 10.1186/s13059-020-02135-8.

Chirag Jain, Arang Rhie, Haowen Zhang, Claudia Chu, Brian P. Walenz, Sergey Koren, and Adam M. Phillippy. Weighted minimizer sampling improves long read mapping. *Bioinformatics*, 36 (Supplement_1):i111–i118, 2020. doi: 10.1093/bioinformatics/btaa435.

Marek Kokot, Maciej Długosz, and Sebastian Deorowicz. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics*, 33(17):2759–2761, 2017. doi: 10.1093/bioinformatics/btx304.

Heng Li, Xiaowen Feng, and Chong Chu. The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, 21(1):265, 2020. doi: 10.1186/s13059-020-02168-z.

Wen-Wei Liao, Mobin Asri, Jana Ebler, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, Julian K. Lucas, Jean Monlong, Haley J. Abel, et al. A draft human pangenome reference. *Nature*, 617(7960):312–324, 2023. doi: 10.1038/s41586-023-05896-x.

Guillaume Marçais and Carl Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, 2011. doi: 10.1093/bioinformatics/btr011.

Zepu Miao and Jia-Xing Yue. Interactive visualization and interpretation of pangenome graphs by linear reference–based coordinate projection and annotation integration. *Genome Research*, 35(2): 296–310, 2025. doi: 10.1101/gr.279461.124.

Donald Michie. "Memo" Functions and Machine Learning. *Nature*, 218:19–22, 1968. doi: 10.1038/218019a0.

Brian D. Ondov, Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 17(1):132, 2016. doi: 10.1186/s13059-016-0997-x.

Prashant Pandey, Fatemeh Almodaresi, Michael A. Bender, Michael Ferdman, Rob Johnson, and Rob Patro. Mantis: a fast, small, and exact large-scale sequence-search index. *Cell Systems*, 7(2): 201–207.e4, 2018. doi: 10.1016/j.cels.2018.05.021.

Benedict Paten, Adam M. Novak, Jordan M. Eizenga, and Erik Garrison. Genome graphs and the evolution of genome inference. *Genome Research*, 27(5):665–676, 2017. doi: 10.1101/gr.214155.116.

Mikko Rautiainen and Tobias Marschall. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biology*, 21(1):253, 2020. doi: 10.1186/s13059-020-02157-2.

Michael Roberts, Wayne Hayes, Brian R. Hunt, Stephen M. Mount, and James A. Yorke. Reducing storage requirements for biological sequence comparison. *Bioinformatics*, 20(18):3363–3369, 2004. doi: 10.1093/bioinformatics/bth408.

Saul Schleimer, Daniel S. Wilkerson, and Alex Aiken. Winnowing: local algorithms for document fingerprinting. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, pp. 76–85, 2003. doi: 10.1145/872757.872770.

Jouni Sirén. Indexing variation graphs. In *Proceedings of the Nineteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pp. 13–27, 2017. doi: 10.1137/1.9781611974768.2.

Jouni Sirén, Erik Garrison, Adam M. Novak, Benedict Paten, and Richard Durbin. Haplotype-aware graph indexes. *Bioinformatics*, 36(2):400–407, 2020. doi: 10.1093/bioinformatics/btz575.

Brad Solomon and Carl Kingsford. Fast search of thousands of short-read sequencing experiments. *Nature Biotechnology*, 34(3):300–302, 2016. doi: 10.1038/nbt.3442.

Ye Yu, Jinpeng Liu, Xinan Liu, Yi Zhang, Eamonn Magner, Erik Lehnert, Chen Qian, and Jinze Liu. SeqOthello: querying RNA-seq experiments at scale. *Genome Biology*, 19(1):167, 2018. doi: 10.1186/s13059-018-1535-9.

Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J. Franklin, Scott Shenker, and Ion Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*, pp. 15–28. USENIX Association, 2012. URL `https://www.usenix.org/conference/nsdi12/technical-sessions/presentation/zaharia`.

A Extended Related Work

Pangenome graphs support construction, mapping, genotyping, visualization, and downstream inference (Paten et al., 2017; Eizenga et al., 2020; Liao et al., 2023). Minigraph and Minigraph-Cactus represent population variation at construction time (Li et al., 2020; Hickey et al., 2024). LocusGate addresses a different lifecycle point: it materializes and certifies one query-scoped region before downstream task selection, then preserves that exact value across heterogeneous operators.

VG, GCSA2, GBWT, and GraphAligner localize sequence evidence inside task-specific mapping and indexing machinery (Garrison et al., 2018; Sirén, 2017; Sirén et al., 2020; Rautiainen & Marschall, 2020). ODGI and VRPG expose local graph views selected for inspection (Guarracino et al., 2022; Miao & Yue, 2025). BIGSI, Mantis, Bifrost, SeqOthello, and SBT optimize sequence-presence retrieval (Solomon & Kingsford, 2016; Bradley et al., 2019; Pandey et al., 2018; Holley & Melsted, 2020; Yu et al., 2018). These systems offer useful persistent indexes or exportable views; Table 2 asks the narrower capability question relevant here: when a local region is fixed, what identifies it, which answer semantics attach to it, and can several operator families receive the identical value?

Minimizers, winnowing, compact posting lists, and high-throughput k-mer processing supply the indexing substrate (Schleimer et al., 2003; Roberts et al., 2004; Ondov et al., 2016; Marçais &

Table 2: Capability comparison at operator dispatch. The panels separate interface lifecycle from declared answer semantics.

Family	Region fixed	Persistent identity	Cross-operator input
VG/GCSA2/GBWT (Garrison et al., 2018; Sirén, 2017; Sirén et al., 2020)	Inside task execution	Index/version state	Managed within associated interfaces
ODGI/VRPG (Guarracino et al., 2022; Miao & Yue, 2025)	During view selection	View or export identity	Reopened as a selected view
Presence indexes (Bradley et al., 2019; Pandey et al., 2018; Holley & Melsted, 2020; Yu et al., 2018)	By indexed key space	Index/version state	Membership answer family
General reuse systems (Michie, 1968; Halevy, 2001; Zaharia et al., 2012; Akidau et al., 2015)	At function, view, or stage definition	Arguments, view definition, or lineage	Reuse under the corresponding abstraction
LocusGate	Before task selection	Query, construction, and content certificate	Identical value handed among operator families

Family	Operator coverage	Guarantee scope	Declared object
VG/GCSA2/GBWT	Several interfaces	Task-specific	Index-managed state
ODGI/VRPG	Local views	View-specific	Selected view
Presence indexes	Membership	Exact presence	Index query result
General reuse systems	General computation	Abstraction-specific	Function, view, or lineage value
LocusGate	Membership, extraction, inspection	Exact; workspace-complete; bounded	Certified workspace

Kingsford, 2011; Kokot et al., 2017). Syncmers and weighted minimizers offer complementary sensitivity trade-offs (Edgar, 2021; Jain et al., 2020). In LocusGate, complete path-window indexing feeds deterministic certified materialization, making the resulting graph slice stable enough to serve as a compositional operator input.

Systems work provides a second comparison axis. Memoization keys reusable function results (Michie, 1968); materialized-view rewriting relates a new query to stored relational answers (Halevy, 2001); and RDDs and dataflow models expose immutable, lineage-bearing intermediates across computations (Zaharia et al., 2012; Akidau et al., 2015). LocusGate adopts their principle that intermediate state should have explicit identity, then specializes it to bounded graph queries: χ names visible graph content and construction, while the operator contract states what membership, extraction, or inspection means over the same G'.

B Adaptive and Transition Definitions

B.1 Pre-dispatch branch pressure

Let $\mathcal{F}_{B_0}$ be the canonically ordered reached states with more than B_0 successors, counting each repeated state once by its nondominated (u, h, d) representative. For distinct-query-position support

$s_q(v)$, a branch-pressure statistic is

$$\rho_{B_0}(q) = \frac{\sum_{u \in \mathcal{F}_{B_0}} \sum_{\mathrm{rank}_u(v) > B_0} (1 + s_q(v))}{\sum_{u \in \mathcal{F}_{B_0}} \sum_{v \in N^+(u)} (1 + s_q(v))},$$

with $\rho_{B_0}(q) = 0$ when the denominator is zero. A deterministic policy maps $\rho_{B_0}(q)$ to a member of a finite ordered budget set. Its thresholds, ordering, version, and selected budget are recorded in χ before workspace materialization. The choice is therefore part of the certified inspection domain rather than a decision made by a downstream operator.

B.2 Cross-operator identity and state

Each source locus is $\ell = (p, c, \sigma)$. Its canonical key contains sequence context and an interval, allowing membership, extraction, and inspection to name the same $\mathcal{W}(q_\ell)$ while retaining their distinct answer contracts. Handoff supplies immutable decoded G'—node sequences, oriented edges, and path-coordinate projections—and χ; operator-private output buffers are excluded. The receiver validates certificate equality and the content digest before acquiring the read-only object.

Let T_i^V and T_i^E be the distinct workspace nodes and edges decoded or traversed by operator i. Node reuse is $|T_1^V \cap T_2^V|/|T_2^V|$, and edge reuse is $|T_1^E \cap T_2^E|/|T_2^E|$. The handoff makes all certified workspace content available, while these ratios describe only the overlap between states actually touched by two operators.

C Exactness Support for Membership

Fix the finite oriented embedded-path universe $\mathcal{P}^{\pm}$ and literal IUPAC semantics defined in Section 3. For $p \in \mathcal{P}^{\pm}$, let $S(p)$ be its spelling. The index processes every block of w consecutive length-k substrings of $S(p)$ and stores the deterministically selected minimizer with its path coordinate and complete node-slice map.

Lemma 1 (Candidate recall)**.** *Let $|q| \geq w + k - 1$. If $q = S(p)[t : t + |q|]$ for some admissible p and valid t, candidate generation emits (p, t) and a verification region covering the complete occurrence.*

Proof. Choose any block of w consecutive query k-mers, which exists under the length condition, and let its selected minimizer start at query offset a. The identical path block starts at t and has the same literal symbols, canonical keys, and tie position. Complete path-window indexing therefore stores that occurrence at $b = t + a$. Lookup retrieves it, exact-oriented key resolution retains it, and candidate initialization computes $b - a = t$. The region contains every path slice intersecting $[t, t + |q|)$, including matches spanning multiple short nodes and recorded cycle visits. Explicit reverse-complement paths cover reverse-strand occurrences; distinct duplicate paths yield distinct valid candidates. □

Lemma 2 (Verifier soundness)**.** *If exact verification accepts (p, t), then q occurs on admissible path p at coordinate t.*

Proof. Acceptance requires literal equality between q and the ordered concatenation of every slice in $S(p)[t : t + |q|]$. Those slices follow the recorded oriented path edges. Canonicalization, duplicate postings, repetitive anchors, ambiguous literal symbols, and hash collisions cannot change the answer because none bypasses this equality test. □

Lemma 3 (Conditioned exactness)**.** *Assume an immutable graph and path collection, complete selected-occurrence indexing over $\mathcal{P}^{\pm}$, no-false-negative lookup and compression, identical query/index normalization, literal IUPAC equality, and candidate regions exempt from contextual traversal budgets. For every $|q| \geq w + k - 1$, membership returns `True` if and only if q occurs in some $S(p)$ for $p \in \mathcal{P}^{\pm}$.*

Proof. Lemma 1 supplies a complete candidate for every true occurrence, and exact equality accepts it. Conversely, every accepted candidate witnesses an admissible occurrence by Lemma 2. This proves completeness and soundness over the stated universe. □

D HANDOFF EQUIVALENCE

Proof of Proposition 1. Certificate equality fixes the graph and query digests, canonical anchors, traversal bounds, ordering conventions, adjacency and materializer versions, and canonical digest of G'. Deterministic materialization and canonical serialization therefore produce the same immutable workspace bytes. A fixed deterministic operator with the same configuration and implementation version receives identical input, and deterministic normalization maps its result to the same output. The handed-off and rematerialized executions consequently agree. □

E FORMAL SCOPE AND INVARIANTS

The construction operates over a finite declared path universe and an immutable directed sequence graph. Canonical ordering determines selected anchors, queue order, successor order, serialization, and the certificate digest. Finite traversal bounds and dominance pruning make contextual expansion terminate, while mandatory candidate intervals remain independent of those bounds. These invariants are sufficient for the membership and handoff results stated in the main text.

The operator contracts have intentionally different domains. Membership reasons about literal sequence occurrences on declared embedded paths. Extraction reasons about qualifying walks in the certified subgraph. Inspection reasons about path-supported content reached under the recorded finite budget. Recording all boundary-defining choices in χ lets a downstream consumer distinguish these domains while using one immutable graph value.