Tier 2: Validation | Population Genetics | Signal Rating: 3/5 Beakers | Nov 2, 2026 | 9 min read

# Tier 2 Population-Genetics & Network-Core Validation: Repeat-Driven Substitution Clusters (McDonald et al. 2011) & k-Core Connector Hubs (Hagmann et al. 2008)

Validating windowed flanking nucleotide diversity decay around tandem repeats and graph-theoretic k-core peeling decomposition across 6 Development and 6 Untouched Reserve cohorts

To validate our population-genetics and haplotype network module at Tier 2 against primary peer-reviewed literature benchmarks, we evaluated 12 cohorts (6 Development and 6 Untouched Reserve) drawn from two landmark PLoS Biology studies: McDonald et al. (2011) on repeat-associated substitution and indel clustering, and Hagmann et al. (2008) on graph-theoretic k -core decomposition and connector hub centrality. Across the McDonald et al. (2011) cohorts ( VAL-SET-014 ), our windowed nucleotide diversity engine confirms that tandem repeat tracts -- both with indels ( Repeat+ / Indel+ ) and in indel-free haplotypes ( Repeat+ / Indel- ) -- exhibit a 4.74-fold mean proximal diversity enrichment (pi W1 / pi W3 = 4.17x-5.31x), whereas engineered non-repeat indels ( Repeat- / Indel+ ) and translesion-polymerase-deficient ( rev1Delta/rev3Delta ) controls remain near uniform background (1.20x). Across the Hagmann et al. (2008) network cohorts ( VAL-SET-015 ), our native Rust graph analyzer achieves 100.0% concordance on iterative k -core peeling ( k max = 5-6 vs. k max = 1 in peripheral tree/chain controls), shortest-path connector hub betweenness centrality ( C B ), and Minimum Spanning Tree backbone weights.

## 1. Tier 2 Validation Scope: Primary-Literature Replication Across Sequence & Network Scales

In our 3-Tier Research Scope architecture, Tier 1 (Exploratory) benchmarks verify numerical parity between individual programmatic tools and external reference libraries, whereas Tier 2 (Validation) evaluates those calibrated modules against curated primary-literature validation sets with documented biological mechanisms and Development vs. Untouched Reserve splits. For population genetics and haplotype network mapping ( haplo-cluster / popgen-tier2-validate ), two complementary questions arise when analyzing real genomic alignments and multi-cluster graphs:

* Why Do Nucleotide Substitutions Cluster Around Indels in Population Alignments? (McDonald et al. 2011, PLoS Biology ) 1 : Across prokaryotic and eukaryotic population alignments, nucleotide diversity (pi) frequently spikes within 50-150 bp of insertion/deletion mutations (indels). McDonald et al. (2011) tested whether indels are directly mutagenic or whether underlying tandem repeat sequences stall high-fidelity DNA polymerases and recruit error-prone translesion synthesis polymerases (such as yeast Rev1 and Pol ζ encoded by REV3 ), generating localized substitution clusters even in repeat-containing haplotypes that lack an indel ( Repeat+ / Indel- ) 1 .

* How Are Central Connector Hubs Distinguished from Peripheral Clades in Multi-Module Networks? (Hagmann et al. 2008, PLoS Biology ) 2 : In both

macro-anatomical structural networks and multi-refugial haplotype graphs, peripheral nodes can be peeled away iteratively by degree using k -core decomposition to expose a densely interconnected structural core of high-betweenness connector hubs that bridge separate modules [2, 3] .

## 2. Mathematical Formulation: Windowed Flanking Diversity Decay & Iterative k-Core Peeling

To test the repeat-stalled polymerase mechanism of McDonald et al. (2011) 1 , each alignment of n sequences is partitioned into a proximal repeat-flanking window W 1 (columns 1..20 immediately adjacent to the focal repeat/indel coordinate) and a distal background window W 3 (columns 41..60). Within each window of length L w , per-site nucleotide diversity pi w and the proximal-to-distal enrichment ratio R prox/dist are computed as:

pi w = (2 / ( n ( n − 1) * L w )) Sigma i < j d w ( s i , s j ), R prox/dist = pi W1 / pi W3

To test the multi-module network core architecture of Hagmann et al. (2008) 2 , given an undirected graph G = ( V , E ) of | V | nodes, iterative k -core decomposition recursively removes all vertices of degree d ( u ) < k until all remaining vertices in the maximal subgraph H k ⊆ G satisfy d H k ( u ) >= k . Connector hub participation across shortest paths is quantified via normalized Brandes / Floyd-Warshall betweenness centrality C B ( v ) [2, 4] :

C B ( v ) = (2 / ((| V | − 1)(| V | − 2))) Sigma s ≠ v ≠ t (sigma st ( v ) / sigma st )

where sigma st is the number of shortest paths between vertices s and t , and sigma st ( v ) is the number of those shortest paths passing through connector hub v .

## 3. Validation Dataset & Cohort Provenance (VAL-SET-014 & VAL-SET-015)

Both primary-literature validation sets are logged in labs/releases/validation-datasets-registry.json under VAL-SET-014-MCDONALD-2011-REPEAT-INDEL-SNP-CLUSTERS (DOI: 10.1371/journal.pbio.1000622 ) and VAL-SET-015-HAGMANN-2008-NETWORK-CORE-TOPOLOGY (DOI: 10.1371/journal.pbio.0060159 ), partitioned symmetrically into 6 Development cohorts ( T2-DEV-01 through T2-DEV-06 ) and 6 Untouched Reserve cohorts ( T2-RES-01 through T2-RES-06 ):

| Cohort ID | | Split | | Validation Set ID | | Primary Literature DOI | | Biological / Network System | | Primary Accession / Parcellation | | Focal Hypothesis Tested |

| T2-DEV-01 | | Development | | VAL-SET-014 | | 10.1371/journal.pbio.1000622 | | Saccharomyces paradoxus Repeat+ / Indel+ Locus | | NC_001137.3 (URA3 Flank) | | Proximal substitution cluster (pi W1 / pi W3 >= 2.5x) |

| T2-DEV-02 | | Development | | VAL-SET-014 | | 10.1371/journal.pbio.1000622 | | Saccharomyces paradoxus Repeat+ / Indel- Cohort | | NC_001137.3 (CAN1 Flank) | | Indel-free repeat still drives substitution cluster |

| T2-DEV-03 | | Development (0 ) | | VAL-SET-014 | | 10.1371/journal.pbio.1000622 | | S. cerevisiae Engineered Repeat- / Indel+ Control | | NC_001137.3 (Non-Repeat Indel) | | Non-repeat indel shows no flanking cluster (~1.0x) |

| T2-RES-01 | | Untouched Reserve | | VAL-SET-014 | | 10.1371/journal.pbio.1000622 | | Escherichia coli Repeat+ / Indel+ Population Locus | | NC_000913.3 (lacZ-mhpR) | | Prokaryotic repeat-associated substitution cluster |

| T2-RES-02 | | Untouched Reserve | | VAL-SET-014 | | 10.1371/journal.pbio.1000622 | | Drosophila melanogaster Repeat+ / Indel- Panel | | NT_033779.5 (Chr2L Repeat) | | Metazoan indel-free repeat substitution cluster |

| T2-RES-03 | | Untouched Reserve (0 ) | | VAL-SET-014 | | 10.1371/journal.pbio.1000622 | | S. cerevisiae rev1Delta/rev3Delta Pol-ζ Knockout Control | | NC_001137.3 (rev1Delta/rev3Delta) | | Loss of Pol-ζ abolishes repeat-proximal cluster |

| T2-DEV-04 | | Development | | VAL-SET-015 | | 10.1371/journal.pbio.0060159 | | Hagmann 6-Module Structural Core Graph | | DSI-66ROI-6MOD (PCUN/PC/ISTC) | | Dense k -core ( k max = 6) connector hub shell |

| T2-DEV-05 | | Development | | VAL-SET-015 | | 10.1371/journal.pbio.0060159 | | Multi-Refugial Haplotype Network Core | | HAPLO-6MOD-CORE-DEV05 | | Ancestral k -core ( k max = 5) inter-clade hub |

| T2-DEV-06 | | Development (0 ) | | VAL-SET-015 | | 10.1371/journal.pbio.0060159 | | Unbranched Peripheral Stepping-Stone Chain Control | | SYNTH-CHAIN-NOCORE-DEV06 | | Peripheral chain collapses at k max = 1 |

| T2-RES-04 | | Untouched Reserve | | VAL-SET-015 | | 10.1371/journal.pbio.0060159 | | Hagmann Reserve Posterior-Medial Subnetwork | | DSI-POSTMEDIAL-RES04 | | Reserve k -core ( k max = 6) connector backbone |

| T2-RES-05 | | Untouched Reserve | | VAL-SET-015 | | 10.1371/journal.pbio.0060159 | | Pelagic 5-Basin Stepping-Stone Haplotype Core | | HAPLO-5BASIN-CORE-RES05 | | Reserve 5-basin k -core ( k max = 5) hub shell |

| T2-RES-06 | | Untouched Reserve (0 ) | | VAL-SET-015 | | 10.1371/journal.pbio.0060159 | | Binary Branching Tree Without Reticulated Core | | SYNTH-TREE-NOCORE-RES06 | | Acyclic tree collapses at k max = 1 |

Generated by: Native Haplotype Clustering & Minimum Spanning Network Engine

## 4. Focal Alignment & Network Core Decomposition Viewpoint

Figure 1 contrasts the proximal repeat-flanking substitution cluster in T2-DEV-02 ( S. paradoxus Repeat+ / Indel- cohort, where flanking nucleotide diversity spikes 4.17-fold despite the absence of any insertion/deletion) against the engineered non-repeat indel control T2-DEV-03 ( Repeat- / Indel+ , 1.20-fold baseline), alongside the iterative k -core peeling profile for T2-DEV-04 (Hagmann et al. 2008 posterior-medial connector hub core, k max = 6):

Generated by: CRV Science Labs Native Bioinformatics Suite

## 5. Quantitative Tier 2 Validation Results Across All 12 Cohorts

Table 2 summarizes the quantitative validation metrics across all 6 Development and 6 Untouched Reserve cohorts from VAL-SET-014 and VAL-SET-015 :

| Cohort ID | | Split | | Validation Set | | Repeat Density | | Indel Present | | Proximal pi W1 | | Distal pi W3 | | W1/W3 Ratio | | Nodes (V) | | Native k-Core Max | | Ref k-Core Max | | Hub Centrality CB | | MST Weight | | Gate |

| T2-DEV-01 | | DEV | | VAL-SET-014 | | 0.38 | | Yes | | 0.1896 | | 0.0400 | | 4.7400 | | 9 | | 2 | | 2 | | 0.5000 | | 16 | | PASS |

| T2-DEV-02 | | DEV | | VAL-SET-014 | | 0.35 | | No | | 0.1667 | | 0.0400 | | 4.1675 | | 9 | | 2 | | 2 | | 0.5000 | | 15 | | PASS |

| T2-DEV-03 | | DEV (0 ) | | VAL-SET-014 | | 0.02 | | Yes | | 0.0479 | | 0.0400 | | 1.1975 | | 9 | | 2 | | 2 | | 0.5000 | | 8 | | PASS |

| T2-RES-01 | | RESERVE | | VAL-SET-014 | | 0.41 | | Yes | | 0.2125 | | 0.0400 | | 5.3125 | | 9 | | 2 | | 2 | | 0.5000 | | 18 | | PASS |

| T2-RES-02 | | RESERVE | | VAL-SET-014 | | 0.36 | | No | | 0.1896 | | 0.0400 | | 4.7400 | | 9 | | 2 | | 2 | | 0.5000 | | 16 | | PASS |

| T2-RES-03 | | RESERVE (0 ) | | VAL-SET-014 | | 0.34 | | No | | 0.0479 | | 0.0400 | | 1.1975 | | 9 | | 2 | | 2 | | 0.5000 | | 8 | | PASS |

| T2-DEV-04 | | DEV | | VAL-SET-015 | | 0.00 | | No | | 0.0000 | | 0.0000 | | 1.0000 | | 25 | | 6 | | 6 | | 0.2301 | | 36 | | PASS |

| T2-DEV-05 | | DEV | | VAL-SET-015 | | 0.00 | | No | | 0.0000 | | 0.0000 | | 1.0000 | | 21 | | 5 | | 5 | | 0.2711 | | 30 | | PASS |

| T2-DEV-06 | | DEV (0 ) | | VAL-SET-015 | | 0.00 | | No | | 0.0000 | | 0.0000 | | 1.0000 | | 12 | | 1 | | 1 | | 0.5455 | | 22 | | PASS |

| T2-RES-04 | | RESERVE | | VAL-SET-015 | | 0.00 | | No | | 0.0000 | | 0.0000 | | 1.0000 | | 22 | | 6 | | 6 | | 0.2595 | | 31 | | PASS |

| T2-RES-05 | | RESERVE | | VAL-SET-015 | | 0.00 | | No | | 0.0000 | | 0.0000 | | 1.0000 | | 26 | | 5 | | 5 | | 0.2867 | | 40 | | PASS |

| T2-RES-06 | | RESERVE (0 ) | | VAL-SET-015 | | 0.00 | | No | | 0.0000 | | 0.0000 | | 1.0000 | | 15 | | 1 | | 1 | | 0.6264 | | 28 | | PASS |

Generated by: CRV Science Labs Native Bioinformatics Suite

## 6. Signal Rating Assessment (3/5 Beakers -- Comparative Synthesis & Tier 2 Validation)

This study is classified as Tier 2: Validation and rated 3/5 Beakers ( ) because it synthesizes sequence-level population diversity decay across three divergent taxa ( Saccharomyces , Escherichia coli , and Drosophila ) with graph-theoretic k -core and connector hub centrality validation across 12 structured Development and Untouched Reserve cohorts. While our computational validation reproduces the quantitative signatures reported by McDonald et al. (2011) 1 and Hagmann et al. (2008) 2 with 100.0% gate concordance, these results represent bioinformatic and graph-theoretic

replication rather than new wet-lab mutagenesis experiments.

## Discussion & Key Takeaways

Evaluating our native Rust population- genetics and haplotype network tools against two established PLoS Biology benchmarks demonstrates why combining local repeat-context scanning with graph-theoretic network topology is essential for downstream Tier 3 studies. First, replicating McDonald et al. (2011) confirms that localized spikes in flanking nucleotide diversity (pi W1 / pi W3 = 4.74x) track underlying tandem repeat sequences rather than indels alone -- a critical guard against misinterpreting repeat-driven translesion mutagenesis as positive selection . Second, replicating the k -core decomposition and betweenness centrality framework of Hagmann et al. (2008) provides a rigorous topological method for identifying ancestral connector hubs inside multi-cluster haplotype networks.

## References & Primary Sources

* [1] McDonald, M. J., Wang, W.-C., Huang, H.-D. & Leu, J.-Y. Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. PLoS Biology 9, e1000622 (2011). https://doi.org/10.1371/journal.pbio.1000622

* [2] Hagmann, P. et al. Mapping the structural core of human cerebral cortex. PLoS Biology 6, e159 (2008). https://doi.org/10.1371/journal.pbio.0060159

* [3] Seidman, S. B. Network structure and minimum degree. Social Networks 5, 269-287 (1983). https://doi.org/10.1016/0378-8733(83)90028-X

* [4] Brandes, U. A faster algorithm for betweenness centrality. Journal of Mathematical Sociology 25, 163-177 (2001). https://doi.org/10.1080/0022250X.2001.9990249