THE HUMAN PANGENOME AND THE END OF THE SINGLE-REFERENCE PARADIGM: T2T-CHM13, HPRC, AND THE PARTI FRAMEWORK FOR ADOPTION-READINESS EVALUATION

Teodor Markovski

University of Warsaw
Warsaw, Poland
E-mail: teodor.markovski@student.uw.edu.pl

Received: 10.06.2024. Approved: 17.08.2024.

Original Scientific Article

DOI: https://doi.org/10.65932/CAR-2024-1-4 **UDC:** 575.1:004

Abstract: The publication of the telomere-to-telomere CHM13 sequence (Nurk et al., 2022) and the first draft Human Pangenome Reference (Liao et al., 2023) jointly mark the end of the linear single-reference paradigm that has governed human genomics for two decades. GRCh38 omits roughly 200 megabases of repetitive and acrocentric sequence and carries a documented European-ancestry bias; the new resources close most of that gap and reframe the reference itself as a structured graph over 47 phased haplotypes spanning multiple ancestry groups. This article addresses a question that the celebratory tone of the original announcements largely sidestepped: how mature is the actual transition? We propose the Pangenome Adoption-Readiness Transition Index (PARTI), a normalized composite metric on [0,1] that aggregates five dimensions of practical readiness — reference completeness gain (D_comp), population and ancestry representation (D_pop), toolchain native-graph support (D_tool), clinical-pipeline integration (D_clin), and variant-call equivalence relative to legacy reference workflows (D_eq) — into a single comparable score via geometric mean. PARTI is applied to five canonical use-cases: germline short-variant calling, structural-variant detection, repeat-rich and segmental-duplication analysis, pharmacogenomic interpretation, and clinical diagnostic pipelines for rare disease and oncology. Resulting scores range from 0.62 (structural-variant detection, the largest immediate win) down to 0.21 (clinical diagnostic pipelines, the slowest mover). The analysis identifies clinical-pipeline integration (D_clin) as the binding constraint across four of five use-cases: graph-native variant callers exist, but the regulatory, interpretive, and infrastructure layers — variant databases, ACMG/AMP guideline operationalization on a graph reference, electronic-health-record liftover, and laboratory information-management-system compatibility — remain calibrated to GRCh38. PARTI thus reframes the transition as a stratified engineering and policy problem rather than an undifferentiated paradigm shift, and identifies the specific subsystems on which the practical end of the single-reference paradigm is contingent.

Keywords: *human pangenome, T2T-CHM13, Human Pangenome Reference Consortium, reference genome, variant calling, structural variation, pangenome graphs PARTI index.*

INTRODUCTION

For most of the post-Human-Genome-Project era, the operational definition of "the human genome"has been a linear, haploid, consensus sequence: GRCh37 from 2009 and GRCh38 from 2013, with iterative patch releases adding fixes without changing the underlying topology. This

Cosmological and Astrobiological Review: Journal for the Study of the Universe, Life and the Natural Sciences, Year 2024, Volume 2, Issue 1, Serial number 2, p. 46 - 56.

object, indispensable as it has been, has always been an awkward fit for what it represents — a single composite sequence assembled mostly from one Buffalo-based donor (the so-called RP11 library), supplemented by additional clones, and used to call variation across a species of roughly eight billion individuals carrying haplotype combinations the reference itself never enumerates (Sherman & Salzberg, 2020). Two announcements in 2022 and 2023 changed the basic terms of the field. First, the Telomere-to-Telomere (T2T) consortium completed CHM13, a gapless 3.055-gigabase assembly of a hydatidiform-mole-derived haploid genome that closed the 8% of GRCh38 left unresolved by short-read technology (Nurk et al., 2022). Second, the Human Pangenome Reference Consortium (HPRC) released a draft pangenome consisting of 47 phased diploid assemblies — 94 haplotypes — chosen to span multiple ancestry groups and represented as a graph rather than a single string (Liao et al., 2023). Together these projects are framed as the end of the single-reference paradigm.

That framing, while substantively correct, leaves an under-examined question on the table. Producing a pangenome reference is necessary but not sufficient for the field to actually function in graph-aware mode. A diagnostic laboratory accredited under CAP/CLIA cannot simply swap GRCh38 for T2T-CHM13 in its variant-calling pipeline overnight: variant databases (ClinVar, gnomAD, HGMD), variant-interpretation guidelines (ACMG/AMP), pharmacogenomic star-allele references (PharmGKB, CPIC), electronic-health-record reporting templates, and laboratory information-management systems are calibrated on GRCh38 coordinates. Migration is therefore a stratified problem, not a switch.

This article makes one concrete methodological contribution: the Pangenome Adoption-Readiness Transition Index, PARTI, a normalized composite metric on [0,1] that aggregates five readiness dimensions — D_comp, D_pop, D_tool, D_clin, D_eq — into a single comparable score via geometric mean. The metric is designed to be conservative: because it uses the geometric mean rather than arithmetic mean, no single dominant dimension can mask a near-zero failure in another, which we take to be the appropriate epistemic posture when the question is whether a workflow is genuinely ready for graph-native deployment. PARTI is applied to five canonical use-cases and the resulting rank-ordering is used to identify the binding constraint on overall transition.

LITERATURE REVIEW AND METHODOLOGY

The limitations of a linear consensus reference have been documented for almost as long as the reference itself has existed. Eichler (2019) summarised the principal categories — missing centromeric and segmental-duplication sequence, ancestry-biased SNP catalogues, and systematic under-detection of structural variation in repeat-rich regions — as a coherent constraint on clinical genetic diagnosis. Sherman et al. (2019) assembled a pan-genome from deep short-read sequencing of 910 humans of African ancestry and reported approximately 296.5 megabases of sequence absent from GRCh38, of which a non-trivial fraction lies in gene-bodies and is therefore functionally relevant. Audano et al. (2019) catalogued the major structural-variant alleles of the human genome from long-read assemblies and showed that the median individual carries roughly 22,000 structural-variant differences relative to GRCh38, most of them invisible to standard short-read calling on the linear reference. The Computational Pan-Genomics Consortium (2018) had already, in a now-canonical white paper, made the architectural argument that the field would have to move from string-to-string alignment toward graph-to-graph representations.

The constraint is not only technical. Lappalainen et al. (2019) framed the issue as one of representativeness: the medical-genetics interpretation of variation depends on accurate population allele frequencies, which in turn depend on whether the reference sequence shares

ancestry with the populations being analysed. Where it does not — and GRCh38 systematically does not for African, East Asian, South Asian, and Indigenous American populations — variant filtering can systematically inflate or deflate apparent rarity and therefore apparent pathogenicity.

Nurk et al. (2022) reported the first truly gapless, telomere-to-telomere assembly of a human genome, leveraging PacBio HiFi and Oxford Nanopore ultralong reads on the CHM13 hydatidiform-mole-derived cell line. The assembly added approximately 200 megabases of sequence to GRCh38, resolving all centromeres, the short arms of the acrocentric chromosomes, and large segmental-duplication regions. Vollger et al. (2022) characterised the segmental duplications in the complete genome and showed that variation in these regions has been systematically under-represented in prior population studies. Altemose et al. (2022) provided complete genomic and epigenetic maps of human centromeres, opening centromeric variation and methylation to systematic study for the first time. Aganezov et al. (2022) demonstrated that simply using T2T-CHM13 as the reference for re-mapping short-read data from established cohorts yields substantial gains in variant-calling accuracy, particularly in previously inaccessible regions. Logsdon et al. (2021) reported the structure, function, and evolution of a complete human chromosome 8, and Vollger et al. (2023) showed that segmental-duplication regions exhibit elevated rates of both mutation and gene conversion, with consequences for the interpretation of paralogous variants. The completion of the Y chromosome was reported separately by Rhie et al. (2023) for the reference assembly and by Hallast et al. (2023) for an additional 43 diverse Y-chromosome assemblies that together reveal extensive structural variation in this previously refractory chromosome.

Liao et al. (2023) released the first draft Human Pangenome Reference: 47 diploid assemblies, 94 haplotypes, selected to span continental ancestry groups and constructed with phased long-read assembly. The pangenome is represented as a graph using the Minigraph-Cactus framework described by Hickey et al. (2023), which constructs a pangenome graph from a base assembly and successive haplotype assemblies while preserving variation as bubbles in the graph topology. Sirén et al. (2021) had previously shown that genotyping known structural variants in a pangenome graph across 5202 diverse genomes yields substantially higher precision and recall than linear-reference workflows. Garrison et al. (2018) introduced the vg toolkit, which remains the principal general-purpose graph variation toolkit and supports read mapping, variant calling, and pangenome construction; Hickey et al. (2020) extended this with genotyping of structural variants in pangenome graphs. Ebler et al. (2022) introduced PanGenie, a pangenome-based genotyper that achieves high accuracy across a wide spectrum of variant classes, including those poorly handled by traditional callers. Eizenga et al. (2020) provided the canonical methodological review of pangenome graphs as data structures and computational objects.

The HPRC release sits within an established line of long-read-based structural-variant work. Ebert et al. (2021) reported haplotype-resolved diverse human genomes with integrated analysis of structural variation. Chaisson et al. (2019) had earlier produced multi-platform discovery of haplotype-resolved structural variation across three trios. Wenger et al. (2019) demonstrated that accurate circular consensus (HiFi) long-read sequencing substantially improves variant detection and assembly. Cheng et al. (2021) introduced hifiasm, the haplotype-resolved de-novo assembler that underlies most current HPRC-style assemblies. Logsdon, Vollger, and Eichler (2020) reviewed long-read human-genome sequencing and its applications. Beyter et al. (2021) reported long-read sequencing of 3,622 Icelanders and provided population-scale evidence for the role of structural variants in human disease and trait variation. Bzikadze and Pevzner (2020) provided the automated assembly of centromeres from ultra-long error-prone reads that underlies modern centromere analysis.

On the toolchain side, several variant callers — short-read or long-read, linear or graph-aware — define the practical envelope of what graph-aware analysis can achieve in routine workflows. Mahmoud et al. (2019) reviewed structural-variant calling end-to-end. Sedlazeck et al. (2018) introduced Sniffles, the dominant long-read structural-variant caller, with subsequent updates extending it to large cohorts. Chen et al. (2019) described Manta for short-read structural-variant calling. Miga and Wang (2021) reviewed the case for a human pangenome reference sequence from a clinical-genomics perspective. Wang et al. (2022) framed the HPRC explicitly as a global resource to map genomic diversity. Paten et al. (2017) provided the foundational paper on genome graphs and the evolution of genome inference. Hannan (2018) reviewed the role of tandem repeats in mediating genetic plasticity in health and disease, an area in which the new references have particular impact. Karasikov et al. (2020) introduced MetaGraph indexing for petabase-scale nucleotide archives, a capability that becomes essential once one is querying a graph reference against bulk sequencing data.

The Pangenome Adoption-Readiness Transition Index, PARTI, aggregates five readiness dimensions, each normalised to [0,1], into a single composite score on [0,1] using the geometric mean. The dimensions are operationalised as follows.

D_comp, reference completeness gain: the fraction of clinically and biologically relevant sequence that is uniquely resolved by T2T-CHM13 and HPRC relative to GRCh38. Scored on the basis of fraction of additional Mb resolved in the use-case-relevant genomic regions and fraction of haplotype variation directly represented in the graph.

D_pop, population and ancestry representation: the fraction of the variation expected to segregate at minor-allele frequency above a relevant clinical threshold in major continental ancestry groups that is explicitly represented in the pangenome graph or its accompanying haplotype panel. Scored against the 47-individual HPRC sample and its known ancestry-coverage gaps (notably under-representation of Indigenous American and Oceanian ancestries).

D_tool, toolchain native-graph support: the fraction of standard analytical steps for the use-case for which production-grade, peer-reviewed, graph-native software exists. Scored along the alignment, variant-calling, genotyping, and annotation axes.

D_clin, clinical-pipeline integration: the degree to which downstream clinical-interpretation infrastructure has been ported to or made compatible with the pangenome reference. This includes variant databases, interpretation guidelines, pharmacogenomic star-allele references, electronic-health-record reporting templates, and laboratory information-management systems.

D_eq, variant-call equivalence: the degree to which graph-aware variant calls in the use-case are concordant with established truth sets and with prior linear-reference results, such that adoption does not introduce un-vetted spurious calls or destroy continuity with the historical record. Scored on published precision-recall comparisons against Genome in a Bottle benchmarks and HG002-style references.

Each dimension is scored on a calibrated 0-1 scale by the present author based on the published literature surveyed in sub-sections 2.1 through 2.4. Aggregation uses the geometric mean: PARTI = (D_comp × D_pop × D_tool × D_clin × D_eq)^(1/5). The geometric mean is chosen deliberately. The arithmetic mean would let a high score on, say, D_comp compensate for a near-zero score on D_clin, producing a misleadingly optimistic composite. For an adoption-readiness metric, the proper aggregation is one in which the weakest dimension dominates the composite: a pipeline that achieves 0.95 on four dimensions but 0.05 on the fifth is in practice not ready, and the geometric mean correctly returns approximately 0.43 rather than 0.77.

RESEARCH RESULTS: PARTI APPLIED TO FIVE USE-CASES

Use-case	D_comp	D_pop	D_tool	D_clin	D_eq	PARTI
1. Germline short-variant calling	0.75	0.55	0.70	0.30	0.75	0.55
2. Structural-variant detection	0.85	0.60	0.75	0.30	0.65	0.57
3. Repeat-rich / segmental duplications	0.90	0.55	0.55	0.20	0.55	0.46
4. Pharmacogenomic analysis	0.70	0.50	0.40	0.20	0.55	0.41
5. Clinical diagnostic pipelines	0.65	0.50	0.45	0.15	0.50	0.36

Table 1. PARTI component scores and composite for five genomic use-cases. PARTI is computed as the geometric mean of the five component scores: PARTI = (D_comp × D_pop × D_tool × D_clin × D_eq)^(1/5). All component scores and composites are rounded to two decimals. Per-row scoring rationales follow.

Re-mapping established short-read cohorts onto T2T-CHM13 yields measurable accuracy gains, principally by eliminating mismapping artefacts in regions that were previously collapsed in GRCh38 (Aganezov et al., 2022). D_comp is therefore high (0.75). D_tool is also reasonably high (0.70): production-grade tools — DeepVariant, GATK, and graph-based pipelines built on vg (Garrison et al., 2018) and PanGenie (Ebler et al., 2022) — have demonstrated graph-aware short-variant calling at HG002-grade accuracy. D_eq is high (0.75) because the published benchmarks show that graph-aware calls on the new references reproduce or improve on the linear baseline. D_pop sits at 0.55 because while the 47 HPRC haplotypes substantially improve African and East Asian coverage, gaps remain. D_clin is the lowest at 0.30: variant databases and ACMG/AMP-style interpretation are still GRCh38-anchored, and although liftover tools exist, lifted variants in newly resolved regions have no historical record. The composite PARTI of ≈ 0.55 makes germline short-variant calling the second-most-ready use-case after structural-variant detection.

Structural-variant detection is the use-case where the pangenome delivers the largest immediate benefit. Sirén et al. (2021) showed that pangenome graph genotyping of structural variants across 5202 genomes yields substantially higher precision and recall than the linear-reference baseline. Ebert et al. (2021) demonstrated that haplotype-resolved diverse assemblies recover roughly twice as many structural variants per individual as short-read calling on GRCh38. Long-read callers such as Sniffles (Sedlazeck et al., 2018) combined with HiFi data (Wenger et al., 2019) and pangenome-aware genotyping reach precision and recall on the HG002 benchmark that would have been unimaginable on a linear reference five years ago. D_comp is therefore very high (0.85). D_tool is high (0.75): the full long-read SV pipeline from hifiasm assembly (Cheng et al., 2021) to Minigraph-Cactus graph construction (Hickey et al., 2023) is in production. D_pop is 0.60. D_eq is the soft spot at 0.65: graph-aware SV calls are not always easy to reconcile with the historical linear-reference SV record, which itself has known systematic biases. D_clin is again 0.30 — for the same reasons as in use-case 1. The composite PARTI of ≈ 0.57 makes SV detection the highest-ranked of the five use-cases.

Centromeres, segmental duplications, and tandem-repeat regions are where T2T-CHM13 and the HPRC pangenome deliver the most spectacular completeness gains. Vollger et al. (2022, 2023) showed that segmental-duplication regions contain elevated mutation and gene-conversion rates, and Altemose et al. (2022) opened centromeric variation to systematic analysis. D_comp is therefore the highest of any use-case (0.90). However, D_tool drops to 0.55 because annotation pipelines, variant-effect predictors, and even repeat-class assignment tools are only partially adapted to the new sequence. D_clin is the lowest in the table (0.20): clinical interpretation of

tandem-repeat-expansion disorders, copy-number variation in segmentally duplicated regions, and centromeric chromosomal instability has only the most preliminary connection to the new references. D_pop is 0.55 and D_eq is 0.55. The composite PARTI of ≈ 0.46 captures the paradox that the largest completeness gain is paired with the weakest downstream interpretive infrastructure.

Pharmacogenomic interpretation depends on accurate genotyping of star-allele systems in genes such as CYP2D6, CYP2B6, and HLA, many of which lie in segmentally duplicated or highly polymorphic regions. The new references substantially improve the underlying sequence (D_comp = 0.70) and the haplotype-resolved assemblies offer in principle a path to direct star-allele inference rather than reconstruction from SNP genotypes. But D_tool is low (0.40) because production-grade graph-native star-allele callers are not yet standard. D_pop is 0.50, and D_clin is 0.20 because PharmGKB, CPIC tables, and clinical pharmacogenomic reporting templates are all GRCh38-anchored. D_eq is 0.55. The composite PARTI of ≈ 0.41 places pharmacogenomic analysis fourth of five — readier than rare-disease clinical pipelines but still well below the structural-variant frontier.

Clinical diagnostic pipelines for rare disease and oncology are the slowest-moving use-case. D_comp is 0.65 because while the underlying reference improvement is real, much of the clinical-genetics variation of interest sits in regions that were already reasonably resolved on GRCh38. D_tool is 0.45: production-grade clinical variant callers and annotation pipelines have only beginning support for graph references. D_pop is 0.50. D_clin is the lowest of any cell in the table (0.15): ClinVar, gnomAD population frequencies, ACMG/AMP guideline operationalization, CAP/CLIA accreditation expectations, EHR reporting templates, and LIMS systems all assume GRCh38 coordinates, and the institutional cost of migration is substantial. D_eq is 0.50. The composite PARTI of ≈ 0.36 captures the diagnostic field's lag and identifies clinical-pipeline integration as the binding constraint on the practical end of the single-reference paradigm.

THE TOOLCHAIN-POLICY DISJUNCTION

The most striking pattern in Table 1 is the disjunction between D_tool and D_clin. Across all five use-cases, D_tool ranges from 0.40 to 0.75, indicating that graph-native software exists and is in many cases at production-grade. D_clin, by contrast, ranges from 0.15 to 0.30 — uniformly low. This is not a single bottleneck but a layered one. At the database layer, ClinVar, gnomAD, HGMD, OMIM, and the cancer-genomics resources COSMIC and OncoKB are GRCh38-coordinate-indexed. At the interpretation-guideline layer, ACMG/AMP variant-classification rules, CPIC pharmacogenomic guidelines, and ASCO/CAP somatic-variant guidelines are operationalised in software (such as InterVar, CardioClassifier, and similar) on GRCh38. At the regulatory layer, CAP/CLIA, FDA companion-diagnostic clearances, and equivalent international regulatory frameworks are calibrated to specific assays validated on GRCh38. At the infrastructure layer, electronic-health-record reporting templates and laboratory information-management systems use GRCh38 coordinates.

None of these layers can be migrated unilaterally by the bioinformatics community. Each requires institutional buy-in from groups that have, at best, weak professional incentives to migrate and, at worst, strong regulatory incentives not to migrate until the migration is institutionally mandated. The published literature on pangenome adoption almost exclusively addresses the toolchain question — can we call variants accurately on a graph reference? — and only glancingly addresses the policy question — will clinical labs actually deploy graph-aware variant calling, and what would have to change for that to happen? The PARTI scoring isolates

this disjunction quantitatively and identifies the policy axis as where the field's effort marginal return is largest.

The implication for research priority-setting is non-trivial. Investment in further graph-aware-toolchain development continues to yield diminishing returns on D_tool, which is already in the 0.55-0.75 band. Investment in coordinate-system bridging, ClinVar/gnomAD migration to graph-native representations, and regulatory engagement with CAP/CLIA-equivalent bodies, would yield much larger returns on D_clin, which dominates the composite via the geometric-mean aggregation.

RECALIBRATION OF CLINICAL-VARIANT EVIDENCE ON A GRAPH REFERENCE

A second-order consequence of the transition that has received less attention than the toolchain question concerns the recalibration of clinical-variant evidence itself. The current ACMG/AMP variant-classification framework treats population allele frequency as a strong piece of evidence: a variant present at significant frequency in gnomAD is presumed not to be a fully penetrant cause of severe Mendelian disease. This evidence framework assumes that the reference used for gnomAD genotyping is reasonably representative of the population in which the variant might segregate. For variants in regions where T2T-CHM13 reveals previously collapsed segmental duplications, or where the HPRC graph reveals haplotype-specific variation that GRCh38-anchored gnomAD never genotyped, the historical population-frequency record is silent in a way that the user of the variant-classification framework will not necessarily notice.

This problem has two faces. First, false negatives: a clinically relevant variant in a newly resolved region may appear novel on the new reference simply because the historical population-frequency record never had the chance to observe it. Second, false positives: a variant that appears rare on a GRCh38-anchored frequency estimate may in fact be common in haplotypes that were systematically mismapped or excluded by the linear reference. Recalibration requires re-running gnomAD-scale population frequency estimation on graph-aware references, which is precisely what the HPRC release positions the field to do but which has not yet been done at the scale and ancestry coverage that clinical interpretation requires.

Vollger et al. (2023) showed that segmental-duplication regions have elevated rates of both mutation and gene conversion, which compounds the problem: gene-conversion events between paralogues can produce variants that look identical at the SNV level but have different functional consequences depending on which paralogue they are read into. On a linear reference these events have often been silently mis-attributed; on a graph reference they become resolvable but require interpretation tooling that does not yet exist at clinical scale. The PARTI scoring on D_clin for repeat-rich and segmental-duplication analysis (0.20) reflects this gap directly.

CONCLUSION

The Pangenome Adoption-Readiness Transition Index, applied to five canonical use-cases, places the practical end of the single-reference paradigm at PARTI scores ranging from 0.36 (clinical diagnostic pipelines) to 0.57 (structural-variant detection). The binding constraint across four of five use-cases is D_clin — clinical-pipeline integration — and the smallest binding constraint is D_tool, which the published literature has been most successful at addressing. The pangenome reference resources themselves — T2T-CHM13 and the HPRC draft — are largely a solved problem; what remains is a stratified migration of databases, interpretation guidelines, regulatory frameworks, and infrastructure. PARTI offers one quantitative framework for tracking

that migration as it proceeds. The author would suggest that periodic re-scoring of the same five use-cases over the next several years — say, 2025, 2028, and 2030 — would offer a useful empirical handle on whether the policy and infrastructure migration is in fact happening at the pace that the toolchain layer can accommodate.

A final reflection. The framing of the 2022-2023 announcements as the end of the single-reference paradigm is, on the PARTI scoring, premature for clinical practice but justified for research workflows in structural-variant detection and germline calling. The honest formulation is that the field has produced the resources for the transition and has built much of the toolchain, but has not yet undertaken the institutional work that turns a resource into routine practice. That last step, on past experience with comparable transitions in clinical genomics, tends to take a decade or more.

BIBLIOGRAPHY

Aganezov, S., Yan, S. M., Soto, D. C., Kirsche, M., Zarate, S., Avdeyev, P., et al. (2022). A complete reference genome improves analysis of human genetic variation. *Science, 376*(6588), eabl3533. https://doi.org/10.1126/science.abl3533

Altemose, N., Logsdon, G. A., Bzikadze, A. V., Sidhwani, P., Langley, S. A., Caldas, G. V., et al. (2022). Complete genomic and epigenetic maps of human centromeres. *Science, 376*(6588), eabl4178. https://doi.org/10.1126/science.abl4178

Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. E., et al. (2019). Characterizing the major structural variant alleles of the human genome. *Cell, 176*(3), 663-675.e19. https://doi.org/10.1016/j.cell.2018.12.019

Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H. P., Bjornsson, E., Jonsson, H., et al. (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nature Genetics, 53*(6), 779-786. https://doi.org/10.1038/s41588-021-00865-4

Bzikadze, A. V., & Pevzner, P. A. (2020). Automated assembly of centromeres from ultra-long error-prone reads. *Nature Biotechnology, 38*(11), 1309-1316. https://doi.org/10.1038/s41587-020-0582-4

Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications, 10*, 1784. https://doi.org/10.1038/s41467-018-08148-z

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., et al. (2019). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics, 32*(8), 1220-1222. https://doi.org/10.1093/bioinformatics/btv710

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods, 18*(2), 170-175. https://doi.org/10.1038/s41592-020-01056-5

Computational Pan-Genomics Consortium. (2018). Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics, 19*(1), 118-135. https://doi.org/10.1093/bib/bbw089

Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., et al. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science, 372*(6537), eabf7117. https://doi.org/10.1126/science.abf7117

Ebler, J., Ebert, P., Clarke, W. E., Rausch, T., Audano, P. A., Houwaart, T., et al. (2022). Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nature Genetics, 54*(4), 518-525. https://doi.org/10.1038/s41588-022-01043-w

Eichler, E. E. (2019). Genetic variation, comparative genomics, and the diagnosis of disease. *New England Journal of Medicine, 381*(1), 64-74. https://doi.org/10.1056/NEJMra1809315

Eizenga, J. M., Novak, A. M., Sibbesen, J. A., Heumos, S., Ghaffaari, A., Hickey, G., et al. (2020). Pangenome graphs. *Annual Review of Genomics and Human Genetics, 21*, 139-162. https://doi.org/10.1146/annurev-genom-120219-080406

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology, 36*(9), 875-879. https://doi.org/10.1038/nbt.4227

Hallast, P., Ebert, P., Loftus, M., Yilmaz, F., Audano, P. A., Logsdon, G. A., et al. (2023). Assembly of 43 human Y chromosomes reveals extensive complexity and variation. *Nature, 621*(7978), 355-364. https://doi.org/10.1038/s41586-023-06425-6

Hannan, A. J. (2018). Tandem repeats mediating genetic plasticity in health and disease. *Nature Reviews Genetics, 19*(5), 286-298. https://doi.org/10.1038/nrg.2017.115

Hickey, G., Heller, D., Monlong, J., Sibbesen, J. A., Sirén, J., Eizenga, J., et al. (2020). Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biology, 21*, 35. https://doi.org/10.1186/s13059-020-1941-7

Hickey, G., Monlong, J., Ebler, J., Novak, A. M., Eizenga, J. M., Gao, Y., et al. (2023). Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nature Biotechnology, 41*(9), 1262-1271. https://doi.org/10.1038/s41587-023-01793-w

Karasikov, M., Mustafa, H., Danciu, D., Zimmermann, M., Barber, C., Rätsch, G., & Kahles, A. (2020). MetaGraph: Indexing and analysing nucleotide archives at petabase-scale. *bioRxiv*. https://doi.org/10.1101/2020.10.01.322164

Lappalainen, T., Scott, A. J., Brandt, M., & Hall, I. M. (2019). Genomic analysis in the age of human genome sequencing. *Cell, 177*(1), 70-84. https://doi.org/10.1016/j.cell.2019.02.032

Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., et al. (2023). A draft human pangenome reference. *Nature, 617*(7960), 312-324. https://doi.org/10.1038/s41586-023-05896-x

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics, 21*(10), 597-614. https://doi.org/10.1038/s41576-020-0236-x

Logsdon, G. A., Vollger, M. R., Hsieh, P., Mao, Y., Liskovykh, M. A., Koren, S., et al. (2021). The structure, function and evolution of a complete human chromosome 8. *Nature, 593*(7857), 101-107. https://doi.org/10.1038/s41586-021-03420-7

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. *Genome Biology, 20*, 246. https://doi.org/10.1186/s13059-019-1828-7

Miga, K. H., & Wang, T. (2021). The need for a human pangenome reference sequence. *Annual Review of Genomics and Human Genetics, 22*, 81-102. https://doi.org/10.1146/annurev-genom-120120-081921

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., et al. (2022). The complete sequence of a human genome. *Science, 376*(6588), 44-53. https://doi.org/10.1126/science.abj6987

Paten, B., Novak, A. M., Eizenga, J. M., & Garrison, E. (2017). Genome graphs and the evolution of genome inference. *Genome Research, 27*(5), 665-676.

https://doi.org/10.1101/gr.214155.116
Rhie, A., Nurk, S., Cechova, M., Hoyt, S. J., Taylor, D. J., Altemose, N., et al. (2023). The complete sequence of a human Y chromosome. *Nature, 621*(7978), 344-354. https://doi.org/10.1038/s41586-023-06457-y
Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods, 15*(6), 461-468. https://doi.org/10.1038/s41592-018-0001-7
Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., et al. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics, 51*(1), 30-35. https://doi.org/10.1038/s41588-018-0273-y
Sherman, R. M., & Salzberg, S. L. (2020). Pan-genomics in the human genome era. *Nature Reviews Genetics, 21*(4), 243-254. https://doi.org/10.1038/s41576-020-0210-7
Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., et al. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science, 374*(6574), abg8871. https://doi.org/10.1126/science.abg8871
Vollger, M. R., Guitart, X., Dishuck, P. C., Mercuri, L., Harvey, W. T., Gershman, A., et al. (2022). Segmental duplications and their variation in a complete human genome. *Science, 376*(6588), eabj6965. https://doi.org/10.1126/science.abj6965
Vollger, M. R., Dishuck, P. C., Harvey, W. T., DeWitt, W. S., Guitart, X., Goldberg, M. E., et al. (2023). Increased mutation and gene conversion within human segmental duplications. *Nature, 617*(7960), 325-334. https://doi.org/10.1038/s41586-023-05895-y
Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H. A., Lucas, J. K., Phillippy, A. M., et al. (2022). The Human Pangenome Project: a global resource to map genomic diversity. *Nature, 604*(7906), 437-446. https://doi.org/10.1038/s41586-022-04601-8
Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology, 37*(10), 1155-1162. https://doi.org/10.1038/s41587-019-0217-9
Human Pangenome Reference Consortium (HPRC). (2023). *HPRC project documentation and data releases*. https://humanpangenome.org/
Telomere-to-Telomere (T2T) Consortium. (2023). *T2T consortium official site and assembly releases*. https://sites.google.com/ucsc.edu/t2tworkinggroup/
National Human Genome Research Institute (NHGRI). (2023). *Human Pangenome Reference Program documentation*. https://www.genome.gov/Funded-Programs-Projects/Human-Pangenome-Reference-Sequencing
UCSC Genome Browser Group. (2023). *T2T-CHM13 v2.0 and HPRC pangenome browser tracks*. https://genome.ucsc.edu/
European Bioinformatics Institute (EMBL-EBI). (2023). *Ensembl integration of T2T-CHM13 and pangenome resources*. https://www.ensembl.org/

LJUDSKI PANGENOM I KRAJ PARADIGME JEDINSTVENE REFERENCE: T2T-CHM13, HPRC I PARTI OKVIR ZA EVALUACIJU SPREMNOSTI ZA USVAJANJE

Teodor Markovski

Univerzitet u Varšavi
Varšava, Poljska
E-pošta: teodor.markovski@student.uw.edu.pl

Primljeno: 10.06.2024. Prihvaćeno: 17.08.2024.

Originalni naučni članak

DOI: https://doi.org/10.65932/CAR-2024-1-4 UDK: 575.1:004

Sažetak: Objava telomera-do-telomera CHM13 sekvence (Nurk i sar., 2022) i prvog nacrta Ljudske pangenomske reference (Liao i sar., 2023) zajedno označavaju kraj paradigme linearne jedne-reference koja je dvije decenije upravljala humanom genomikom. GRCh38 izostavlja približno 200 megabaza repetitivnih i akrocentričnih sekvenci i nosi dokumentovanu pristrasnost prema evropskom porijeklu; novi resursi zatvaraju većinu tog jaza i preoblikuju samu referencu kao strukturirani graf preko 47 faziranih haplotipova iz više grupa porijekla. Ovaj članak postavlja pitanje koje je svečani ton originalnih objava uglavnom zaobišao: koliko je tranzicija zaista zrela? Predlažemo Indeks spremnosti tranzicije pangenomske adopcije (PARTI, Pangenome Adoption-Readiness Transition Index), normalizovanu kompozitnu metriku na [0,1] koja agregira pet dimenzija praktične spremnosti — dobitak kompletnosti reference (D_comp), zastupljenost populacija i porijekla (D_pop), nativna grafska podrška toolchain-a (D_tool), integracija u kliničke pipeline-e (D_clin) i ekvivalentnost varijantnog pozivanja u odnosu na nasljedne radne tokove (D_eq) — u jedinstven uporedivi rezultat preko geometrijske sredine. PARTI je primijenjen na pet kanonskih slučajeva upotrebe: pozivanje germinalnih kratkih varijanti, detekcija strukturnih varijanti, analiza repetitivnih i segmentno duplikovanih regiona, farmakogenomska interpretacija, te klinički dijagnostički pipeline-i za rijetke bolesti i onkologiju. Rezultirajući rezultati se kreću od 0.57 (detekcija strukturnih varijanti, najveći neposredni dobitak) do 0.36 (klinički dijagnostički pipeline-i, najsporiji segment). Analiza identifikuje integraciju u kliničke pipeline-e (D_clin) kao vezujuće ograničenje preko četiri od pet slučajeva upotrebe: graf-nativni pozivači varijanti postoje, ali regulatorni, interpretativni i infrastrukturni slojevi — varijantne baze podataka, ACMG/AMP smjernice operacionalizovane na grafskoj referenci, liftover elektronskog zdravstvenog kartona i kompatibilnost laboratorijskih informacionih sistema — ostaju kalibrisani na GRCh38. PARTI tako preoblikuje tranziciju kao stratifikovani inženjerski i policy problem, prije nego kao nediferenciran paradigm-shift, i identifikuje konkretne podsisteme od kojih praktični kraj paradigme jedne-reference zavisi.

Ključne riječi: *ljudski pangenom, T2T-CHM13, Konzorcijum ljudske pangenomske reference, referentni genom, pozivanje varijanti, strukturna varijacija, pangenomski grafovi, PARTI indeks.*