Computational methods for pangenomics and multiomics integration

Computational methods for pangenomics and multiomics integration

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von
MSc Simon Heumos
aus Aalen

Tübingen
2024

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	28.03.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Sven Nahnsen
2. Berichterstatter:	Prof. Dr. Oliver Kohlbacher
3. Berichterstatter:	Prof. Dr. Sven Rahmann

For Ernie and Bert

License

Eidesstattliche Erklärung

Ich erkläre hiermit, dass ich die zur Promotion eingereichte Arbeit mit dem Titel

"Computational methods for pangenomics and multiomics integration"

selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich erkläre, dass die Richtlinien zur Sicherung guter wissenschaftlicher Praxis der Universität Tübingen (Beschluss des Senats vom 25.5.2000) beachtet wurden. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

Ort, Datum Unterschrift

Abstract

Biomedical research models often simplify complex biological processes, with each focusing on one specific molecular mechanisms. For example, genomics can examine the heritable genotype of an organism, while the phenotype refers to the observable traits or characteristics of an organism resulting from the interaction of its genotype with the environment. However, a single data source is often insufficient to explain complex genotype-phenotype relationships due to analysis bias. To address this, the integration of multiple omics data sources, multiomics, provides a more comprehensive approach. Nevertheless, some omics analysis techniques still rely on reference-based methods, which can introduce reference bias and complicate the discovery of accurate genotype-phenotype relationships. Pangenome models offer a solution by relating a representative set of genomic sequences within a population. Pangenome graphs, in particular, store both the shared and variant regions of a set of genomes in one data structure. The contributions of this thesis lie in two different fields: Multiomics and pangenomics. On the multiomics side this thesis showcases the explorative power of integrative multiomics for genotype-phenotype validation and discovery in cancer immunotherapy. Through cell surface molecule profiling of cancer cell panel data, and integration with transcriptomics and proteomics data, I identified potential cancer-specific markers. I validated biomarker candidates using public data to highlight the importance of comprehensive multiomics analysis and data integration for discovering and validating cancer-specific biomarkers. On the pangenomics side this thesis explores two main research questions. First, to overcome the reference bias and implementation limitations of existing pangenome graph construction pipelines, I developed a cluster-efficient, reference-free pipeline to build pangenome graphs, enabling comprehensive genomic diversity studies. The second research question addressed the need to efficiently visualize and analyze pangenome graphs. Therefore, I developed a new layout algorithm that enables efficient visualization of pangenome graphs at the gigabase scale. Additionally, I implemented methods for detecting complex regions, manipulating structure, annotating, and performing exploratory analysis, which allow for comprehensive analysis of these graphs at the same scale. This enables researchers to examine the genotype-phenotype relationships encoded in gigabase-scale pangenome graphs in an unbiased manner. The results of this work show that integrating data from different biological origins improves interpretation and uncovers relationships that single data sources cannot, effectively mitigating analysis bias. The models proposed and the results presented in this doctoral thesis contribute to advancing current knowledge towards improved genotype-phenotype discovery in biomedical research.

Kurzfassung

In der biomedizinischen Forschung nutzen Wissenschaftler häufig vereinfachte Modelle, die sich auf spezifische molekulare Mechanismen konzentrieren und von den tatsächlichen biologischen Prozessen abweichen. So untersucht die Genomik den vererbbaren Genotyp, während der Phänotyp die durch die Umwelt beeinflussten Merkmale beschreibt. Einzelne Datenquellen reichen oft nicht aus, um komplexe Genotyp-Phänotyp-Beziehungen zu erklären, da sie Verzerrungen verursachen können. Die Integration mehrerer Omics-Datenquellen, Multiomics, kann dies mildern, wobei referenzbasierte Methoden jedoch weiterhin zu Verzerrungen führen können. Pangenom-Modelle bieten eine Lösung, indem sie genomische Variationen innerhalb einer Population abbilden. Pangenom-Graphen speichern dabei sowohl gemeinsame als auch variable Regionen von Genomen in einer Datenstruktur. Diese Dissertation leistet Beiträge in den Bereichen Multiomics und Pangenomics. Im Bereich Multiomics wird das Potenzial der integrativen Analyse zur Entdeckung und Validierung von Genotyp-Phänotyp-Beziehungen in der Krebsimmuntherapie aufgezeigt. Ich identifizierte krebspezifische Zelloberflächenmoleküle durch die Integration von Zelloberflächenprofilen mit Transkriptomik- und Proteomikdaten und validierte Biomarkerkandidaten mit öffentlichen Daten. Dies unterstreicht die Bedeutung der Multiomics-Analyse für die Entdeckung und Validierung von Biomarkern in der Krebsforschung. In der Pangenomics behandle ich zwei Hauptfragen: Erstens entwickelte ich eine cluster-effiziente, referenzfreie Pipeline zur Konstruktion von Pangenom-Graphen, um Referenzverzerrungen und Implementierungsgrenzen zu überwinden. Diese Pipeline ermöglicht umfassende Studien zur genomischen Vielfalt einer Population. Zweitens befasste ich mich mit der Notwendigkeit, Pangenom-Graphen effizient zu visualisieren und zu analysieren. Daher entwickelte ich einen neuen Layout-Algorithmus, der eine effiziente Visualisierung von Pangenom-Graphen im Gigabasenmaßstab ermöglicht. Zudem implementierte ich Methoden zur Erkennung komplexer Regionen, Strukturmanipulation, Annotation und explorativen Analyse, die eine umfassende Analyse dieser Graphen auf demselben Maßstab ermöglichen. Dies erlaubt Forschern, die Genotyp-Phänotyp-Beziehungen in Gigabasen-Skalierung Pangenom-Graphen unvoreingenommen zu untersuchen. Die Ergebnisse dieser Arbeit zeigen, dass die Integration von Daten aus verschiedenen biologischen Quellen die Interpretation verbessert und Beziehungen aufdeckt, die einzelne Quellen nicht offenbaren können. Dadurch werden Analyseverzerrungen effektiv gemildert. Die entwickelten Modelle und Ergebnisse tragen zur Verbesserung des Wissens über die Entdeckung von Genotyp-Phänotyp-Beziehungen in der biomedizinischen Forschung bei.

Acknowledgements

I am especially grateful to my advisor Sven Nahnsen, for the years of guidance and support at the Quantitative Biology Center. His encouragement and advice have been instrumental throughout my research journey. He allowed me the freedom to explore my interests, while always being there when things became challenging. He showed me to always stay calm in the most difficult of situations.

My heartfelt thanks goes to Erik Garrison, who took me under his pangenomic wings right after I started my PhD. Without him, I would not have taken this fascinating journey through the world of pangenomics. His creativity and mentorship have profoundly shaped my research. Erik often said that research is like an endless marathon, and his perspective helped me stay motivated and focused throughout this journey. One of the highlights of my PhD was the hackathon where we managed to recruit Andrea Guarracino – who quickly became not just a colleague but a friend. I thoroughly enjoyed working with him, and I learned a great deal, especially how *not* to be Italian! Our collaboration enriched my PhD experience, both scientifically and personally. I also greatly appreciate Pjotr Prins's dry humor, which added a unique and enjoyable element to our collaborations. Flavia Villani was always there with a listening ear and is an excellent cook – her hospitality and culinary skills made our time together even more enjoyable. Vincenza Colonna was a fantastic host in Naples and Lavello, making those visits truly memorable. Our teamwork enriched my PhD experience, both scientifically and personally. I also cherish the resulting friendships and collaborations.

I would also like to extend my sincere gratitude to Oliver Kohlbacher for his insightful feedback and thoughtful critique, which greatly enhanced the quality of my thesis. His expertise and meticulous attention to detail were deeply appreciated.

Special thanks go to Michael Krone, whose support and valuable feedback helped me refine my ideas and strengthen my research. His guidance and constructive input were invaluable.

I'm grateful to Jörg Hagmann, who has been an invaluable sounding board and conversational partner throughout this journey. Our discussions, whether scientific or otherwise, were always thought-provoking, and I hold those moments in high regard. Thanks also to Sebastian Schultheiss and Computomics for their early support, which helped me to begin my PhD.

I want to extend a special mention to Friederike Hanssen and Caroline Schwitalla from QBIC. The many walks with Friederike provided me with fresh perspectives and a welcome break from work. Caroline, thank you for introducing me to so many volleyball groups – those experiences helped keep me balanced during this intense pro-

cess. I also want to thank Sven Fillinger who always had an open ear for any work-related problems. Additionally, the camaraderie and support from my other lab mates: Francesca Barletta, Jonas Scheid, Shraddha Pawar, Daniel Straub, Steffen Lemke, Josua Stadelmaier, Matthias Seybold, Sabrina Krakau, Laurence Kuhlburger, Morgana Oquendo, John Francis Maria Joseph, Stefan Czemmel, Mark Polster, Jannik Seidel, Aline Breitinger, Andreas Friedrich, Tobias Koch, Steffen Greiner, Oskar Wacker, Till Englert, Marissa Dubbelaar, Luis Kuhn, Júlia Mir Pedrol, Mahnaz Azimi Asiabar, Jun-Hoe Lee, Gisela Gabernet, and Susanne Jodoin – made this journey more enjoyable and enriched my experience. I am thankful for the administrative support from Katrin Leichtle throughout the PhD process, and for the undergraduate researchers I have had the pleasure of mentoring.

I also want to extend my thanks to the following individuals and research groups for their support and collaboration: Jerven Bollemann, Toshiyuki T. Yokoyama, Jan-Niklas M. Schmelzle, Jiajie Li, Zhiru Zhang, Michael L. Heuer, Daniel Dörr, Sandra Dehn, Michael Schindler, Xian Xeng, Jean Monlong, Karen Miga, Matthias Hörtenhuber, Maxime Garcia, Adam Talbot, and the nf-core community. Their contributions were invaluable in shaping the work presented in this thesis.

My heartfelt thanks go to Jordan Eizenga, who gave critical feedback on my dissertation, helping to strengthen its content.

I also want to thank Philipp Ehmele for his friendship and support outside of science. His presence in my life has been a source of personal encouragement and camaraderie, in addition to his valuable scientific contributions.

I also want to highlight my friendship with Günter Jäger, Alexander Peltzer, and Markus List, with whom I've shared many mountain biking adventures – these experiences were a great way to unwind and recharge. A special mention goes to Florian Battke, who always found the flaw in the system during our many discussions. Special thanks to Jakob Admard and Nicolai Wahn for the fun barbecue sessions we had.

Finally, I would like to thank my family. My parents for their unwavering support, my sister for her encouraging nature, and my brother Lukas for his invaluable feedback. Our thermal bath bet – where he wagered he would get more citations than me within two months but ultimately needed nine months, leading to several free thermal bath visits for me! – provided both a challenge and a welcome distraction. Our gaming sessions also kept my spirits high during tougher times.

To everyone who has supported me throughout this journey, I extend my deepest thanks. This thesis is not just the product of data, analyses, and words, but of the kindness, patience, and support of the remarkable people who surrounded me.

Contents

List of Figures

List Of Abbreviations

1D	1 Dimension
2D	2 Dimensions
BED	Browser Extensible Data
BEDPE	Browser Extensible Data Paired-End
CCS	Circulare Consensus Sequencing
cDNA	Complementary DNA
CHM13	Complete Hydatidiform Mole 13
CPC	Chinese Pangenome Consortium
DAG	Directed Acyclic Graph
DFTD	Devil Facial Tumor Disease
DNA	DeoxyriboNucleic Acid
DP	Dynamic Programming
ESI	ElectroSpray Ionization
FACS	Fluorescence-Activated Cell Sorting
FAIR	Findable Accessible Interoperable Reusable
FPGA	Field Programmable Gate Arrays
GFA	Graphical Fragment Assembly
GFF	General Feature Format
GINA	Genetic Information Nondiscrimination Act
GP	Genotype-Phenotype
GTF	General Transfer Format
GTEx	Genotype-Tissue Expression
GFA	Graphical Fragment Assembly
H3ABioNet	Pan-African Bioinformatics Network for H3Africa
H3Africa	Human Heredity and Health in Africa
HGNC	HUGO Gene Nomenclature Committee
HGP	Human Genome Project
HGSVC	Human Genome Structural Variation Consortium
HPRC	Human Pangenome Reference Consortium
HUGO	Human Genome Organization
ICI	Immune Checkpoint Inhibitors
IGGSy	International Genome Graph Symposium
IHC	Immunohistochemistry
LC	Liquid Chromatography

LC-MS	Liquid Chromatography-Mass Spectrometry
M3	Malignom Metabolome Microbiome
MCIA	Multiple Co-Inertia Analysis
MFI	Mean Fluorescence Intensity
MHC	Major Histocompatibility Complex
MISC	Miscellaneous
MOFA2	Multi-Omics Factor Analysis 2
MSA	Multiple Sequence Alignment
M/Z	Mass-To-Charge-Ratio
NGS	Next-Generation Sequencing
ODGI	Optimized Dynamic Genome/Graph Implementation
ONT	Oxford Nanopore Technologies
PAF	Pairwise Alignment Format
PacBio	Pacific Biosciences
PANGAIA	Pan-European Pangenome Consortium
PGGB	PanGenome Graph Builder
PG-SGD	Path-Guided Stochastic Gradient Descent
POA	Partial Order Alignment
abPOA	Adaptive Band Partial Order Alignment
SPOA	SIMD Partial Order Alignment
Px	Proteomics
RPPA	Reverse Phase Protein Array
RNA	RiboNucleic Acid
RNA-Seq	RNA Sequencing
SGD	Stochastic Gradient Descent
SIMD	Single Instruction Multiple Execution
SMRT	Single-Molecule Real-Time
T2T	Telomere-to-Telomere
TSV	Tab-Separated Values
Tx	Transcriptomics
VCF	Variant Call Format
VG	Variation Graph
VG msga	Variation Graph multiple sequence/graph aligner

1 Scientific Contributions

1.1 Core Publications

This cumulative doctoral thesis is based on the manuscripts listed below. The order of the listed publications is determined by topic rather than chronological order. Since each work was a collaboration of several scientists, the following pages are dedicated to indicate my personal contributions.

1. **Simon Heumos***, Sandra Dehn*, Konstantin Bräutigam, Marius C. Codrea, Christian M. Schürch, Ulrich M. Lauer, Sven Nahnsen, Michael Schindler. Multiomics surface receptor profiling of the NCI-60 tumor cell panel uncovers novel theranostics for cancer immunotherapy. *Cancer Cell Int* 22, 311 (2022).
   `https://doi.org/10.1186/s12935-022-02710-y`
2. **Simon Heumos***, Andrea Guarracino*, Jan-Niklas M. Schmelzle, Jiajie Li, Zhiru Zhang, Jörg Hagmann, Sven Nahnsen, Pjotr Prins, Erik Garrison. Pangenome graph layout by Path-Guided Stochastic Gradient Descent. *Bioinformatics* 40, 7 (2024), btae363.
   `https://doi.org/10.1093/bioinformatics/btae363`
3. **Simon Heumos**, Michael F. Heuer, Friederike Hanssen, Lukas Heumos, Andrea Guarracino, Peter Heringer, Philipp Ehmele, Pjotr Prins, Erik Garrison, Sven Nahnsen. Cluster efficient pangenome graph construction with nf-core/pangenome. *bioRxiv* 2024.
   `https://doi.org/10.1101/2024.05.13.593871`
   Under review at *Bioinformatics*.
4. Andrea Guarracino*, **Simon Heumos***, Sven Nahnsen, Pjotr Prins, Erik Garrison. ODGI: understanding pangenome graphs. *Bioinformatics* 38, 13 (2022), 3319–3326.
   `https://doi.org/10.1093/bioinformatics/btac308`

* indicates equal contribution

Contributions

1. **Simon Heumos***, Sandra Dehn*, Konstantin Bräutigam, Marius C. Codrea, Christian M. Schürch, Ulrich M. Lauer, Sven Nahnsen, Michael Schindler. Multiomics surface receptor profiling of the NCI-60 tumor cell panel uncovers novel theranostics for cancer immunotherapy. *Cancer Cell Int* 22, 311 (2022).
`https://doi.org/10.1186/s12935-022-02710-y`

   My contributions: I did data curation and quality control, performed the MCIA, did the RNAseq analysis and TCPA data exploration, wrote methods sections of the software tools and steps I applied, generated visualizations for Figures 1-3, and edited the manuscript.

   Author contributions: SH did data curation and quality control, performed the MCIA with support from MCC and SN and did the RNAseq analysis and TCPA data exploration. SD cultured the NCI-60 panel and analyzed all cell lines by flow cytometry, and well assisted in data analyses. KB and CMS analyzed the HPA data. UML provided resources and assisted in data interpretation. SN provided infrastructure and assisted in data interpretation and analyses. MS planned, designed and supervised the overall study, provided resources, analyzed data and wrote the manuscript. All authors edited the manuscript draft together to its final form. All authors read and approved the final manuscript.

2. **Simon Heumos***, Andrea Guarracino*, Jan-Niklas M. Schmelzle, Jiajie Li, Zhiru Zhang, Jörg Hagmann, Sven Nahnsen, Pjotr Prins, Erik Garrison. Pangenome graph layout by Path-Guided Stochastic Gradient Descent. *Bioinformatics* 40, 7 (2024), btae363.
`https://doi.org/10.1093/bioinformatics/btae363`

   My contributions:
   I led and executed the algorithm implementation, co-wrote the software tests and documentation, designed and conducted the experiments, and wrote the manuscript.

   Author contributions: The algorithm was implemented by SH, AG, EG, JNMS, and LJ. Software tests were written by SH, AG, PP, and EG. Documentation was created by SH, AG, and EG. The experiments were designed and conducted by SH with feedback from AG, SN, PP, ZZ, JH and EG. SH wrote the initial draft manuscript. SH and AG wrote the manuscript with support from SN, and EG. All authors edited the manuscript draft together to its final form.

* indicates equal contribution

3. **Simon Heumos**, Michael F. Heuer, Friederike Hanssen, Lukas Heumos, Andrea Guarracino, Peter Heringer, Philipp Ehmele, Pjotr Prins, Erik Garrison, Sven Nahnsen. Cluster scalable pangenome graph construction with nf-core/pangenome. *bioRxiv* 2024.
`https://doi.org/10.1101/2024.05.13.593871`
Under review at *Bioinformatics*.

   My contributions: I concepted and implemented the pipeline, wrote the software tests and documentation, designed and conducted the experiments, and wrote the manuscript.

   Author contributions: Conception of the pipeline was acquired by SH, MLH, SN. The tool was developed by SH, MLH, LH, AG, PH, PE. Software was tested by SH, MLH, FH. SH wrote the software documentation. Experimental design and execution was conducted by SH. Project guidance was given by PP, EG, SN. The manuscript was written by SH with feedback from LH, FH, AG, SN.

4. Andrea Guarracino*, **Simon Heumos***, Sven Nahnsen, Pjotr Prins, Erik Garrison. ODGI: understanding pangenome graphs. *Bioinformatics* 38, 13 (2022), 3319–3326.
`https://doi.org/10.1093/bioinformatics/btac308`

   My contributions: I co-developed the tool, co-wrote software tests and documentation, co-executed the experiments, and co-wrote the manuscript.

   Author contributions: The tool was developed by AG, SH, and EG. Software tests were written by SH, AG, PP, and EG. Documentation was created by SH, AG, and EG. The experiments were conducted by AG and SH with feedback from SN and EG. EG designed the original software. AG and SH wrote the manuscript with support from SN, PP, and EG. All authors edited the manuscript draft together to its final form.

* indicates equal contribution

1.2 Associated Publications

1. Erik Garrison*, Andrea Guarracino*, **Simon Heumos**, Flavia Villani, Zhigui Bao, Lorenzo Tattini, Jörg Hagmann, Sebastian Vorbrugg, Santiago Marco-Sola, Christian Kubica, David G. Ashbrook, Kaisa Thorell, Rachel L. Rusholme-Pilcher, Gianni Liti, Emilio Rudbeck, Sven Nahnsen, Zuyu Yang, Mwaniki N Moses, Franklin L Nobrega, Yi Wu, Hao Chen, Joep de Ligt, Peter H. Sudmant, Nicole Soranzo, Vincenza Colonna, Robert W. Williams, Pjotr Prins. Building pangenome graphs. *bioRxiv* 2023.
`https://doi.org/10.1101/2023.04.05.535718`
Accepted at *Nature Methods*.

   My contributions: I co-developed the pipeline, co-wrote tests and documentation, helped with testing, contributed to Figure 1, wrote Section A1, made Figure A1, and contributed to paper writing and editing.

   Author contributions: Project conception was done by EG. Project guidance was given by EG, SN, NS, VC, RWW, and PP. The tool was developed by EG, AG, SH, SMS, and MNM. EG, AG, SH, FV, ZB, LT, JH, SV, CK, KT, RLRP, AAG, SN, ZY, MNM, FLN, HC, JL, and PHS tested the software. Quality evaluation was conducted by EG, AG, LT. Experiments were designed by EG. Experiments were executed by AG. Documentation was written by AG and SH. Parameter settings was evaluated by SV. Algorithmic development was done by SMS. High Performance Computing environments were managed by PP. In the following the contributors of each pangenome: *Mus musculus*, *Rattus Norvegicus*: FV, DGA, HC, VC; *Tomato*: ZB; *S. cerevisiae*, *S. paradoxus*: LT, GL; *Soy G. max*: JH; *A. thaliana*: SV, CK, ZB, DW; *Helicobacter pylori*: KT, ER; *Neisseria mingitidis*: JL; *SARS-CoV-2*: MNM; *E. coli*, *Coliphages*: FLN, YW; *Primates*: PHS. EG, AG, SH, VC, RWW, and PP wrote and edited the manuscript draft together to its final form.

2. Wen-Wei Liao*, Mobin Asri*, Jana Eble*, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, Julian K. Lucas, Jean Monlong, Haley J. Abel, Silvia Buonaiuto, Xian H. Chang, Haoyu Cheng, Justin Chu, Vincenza Colonna, Jordan M. Eizenga, Xiaowen Feng, Christian Fischer, Robert S. Fulton, Shilpa Garg, Cristian Groza, Andrea Guarracino, William T. Harvey, **Simon Heumos**, Kerstin Howe, Miten Jain, Tsung-Yu Lu, Charles Markello, Fergal J. Martin, Matthew W. Mitchell, Katherine M. Munson, Moses Njagi Mwaniki, Adam M. Novak, Hugh E. Olsen, Trevor Pesout, David Porubsky, Pjotr Prins, Jonas A. Sibbesen, Jouni Sirén, Chad Tomlinson, Flavia Villani, Mitchell R. Vollger, Lucinda L. Antonacci-Fulton, Gunjan Baid, Carl A. Baker, Anastasiya Belyaeva, Konstantinos Billis, Andrew Car-

* indicates equal contribution

roll, Pi-Chuan Chang, Sarah Cody, Daniel E. Cook, Robert M. Cook-Deegan, Omar E. Cornejo, Mark Diekhans, Peter Ebert, Susan Fairley, Olivier Fedrigo, Adam L. Felsenfeld, Giulio Formenti, Adam Frankish, Yan Gao, Nanibaa' A. Garrison, Carlos Garcia Giron, Richard E. Green, Leanne Haggerty, Kendra Hoekzema, Thibaut Hourlier, Hanlee P. Ji, Eimear E. Kenny, Barbara A. Koenig, Alexey Kolesnikov, Jan O. Korbel, Jennifer Kordosky, Sergey Koren, HoJoon Lee, Alexandra P. Lewis, Hugo Magalhães, Santiago Marco-Sola, Pierre Marijon, Ann McCartney, Jennifer McDaniel, Jacquelyn Mountcastle, Maria Nattestad, Sergey Nurk, Nathan D. Olson, Alice B. Popejoy, Daniela Puiu, Mikko Rautiainen, Allison A. Regier, Arang Rhie, Samuel Sacco, Ashley D. Sanders, Valerie A. Schneider, Baergen I. Schultz, Kishwar Shafin, Michael W. Smith, Heidi J. Sofia, Ahmad N. Abou Tayoun, Françoise Thibaud-Nissen, Francesca Floriana Tricomi, Justin Wagner, Brian Walenz, Jonathan M. D. Wood, Aleksey V. Zimin, Guillaume Bourque, Mark J. P. Chaisson, Paul Flicek, Adam M. Phillippy, Justin M. Zook, Evan E. Eichler, David Haussler, Ting Wang, Erich D. Jarvis, Karen H. Miga, Erik Garrison, Tobias Marschall, Ira M. Hall, Heng Li, Benedict Paten. A draft human pangenome reference. *Nature* 617 (2023), 312–324.
`https://doi.org/10.1038/s41586-023-05896-x`

My contributions: I contributed to the development of algorithms and software in PGGB, and to the pangenome graph construction with PGGB.

Author contributions: Pangenome empirical analysis and pangenome quality control: W-WL, DD, MH, GH, CM, J Monlong, HJA, JMZ, EEE, TM, IMH, PM, JW. Paper writing: W-WL, MA, JE, DD, MH, GH, SL, J Monlong, RSF, SG, T-YL, MWM, AMN, HEO, TP, JAS, MRV, G Bourque, KHM, EG, TM, IMH, BP, REG and LH. Paper editing: W-WL, DD, MH, GH, XHC, HC, AG, AMN, PP, AMP, EEE, EDJ, KHM, EG, EEK, TM, IMH, HL, BP, OEC, PE, GF, ANAT, AVZ. Assembly creation: MA, JKL, HC, AMP, HL, D Puiu, AAR, AVZ. Assembly quality control and assembly reliability analysis: MA, JKL, HC, JC, SG, K Howe, TP, D Porubsky, CT, MRV, AMP, JMZ, EEE, KHM, HL, REG, SK, J McDaniel, SN, NDO, D Puiu, MR, AAR, AR, VAS, KS, FT-N, JW, BW, JM DW, ABP. Pangenome applications (structural variants): JE, GH, HJA, WTH, PP, EEE, TM, HPJ, HM. Pangenome graph creation: DD, GH, AG, SH, MNM, FV, EG, YG, SM-S. Data coordination and management: MH, JKL, RSF, WTH, MJ, CT, AMP, EDJ, KHM, TW, LLA-F, SC, MD, SF, REG. Transcriptome and annotation: MH, JME, FJM, MRV, MJPC, KB, MD, AF, CGG, LH, TH, FFT. Pangenome applications (small variants): GH, J Monlong, CM, AMN, PP, JMZ, G Baid, AB, AC, P-CC, DEC, HPJ, AK, MN, KS, JW. Pangenome visualization and complex loci analysis: SL, JC, CF, AG. Population genetic analysis: SB, AG, VC. Sample selection: XF, KMM, AMP, EEK, EEE, KHM, SF, JOK. Sequencing: RSF, MJ, MWM, KMM, HEO, AMP, EEE, EDJ, KHM, CAB, OF, REG, K Hoekzema, JOK, JK, APL, J

Mountcastle, SS, ADS. Pangenome applications (ChIP-seq analysis): CG and G Bourque. Principal investigator and laboratory organizer within the HPRC: MJ, MWM, MJPC, PF, EEK, AMP, EEE, DH, EDJ, KHM, TW, TM, BP, REG, VAS. Pangenome applications (VNTR analysis): T-YL and MJPC. Pangenome applications (RNA-seq analysis): JAS and JME. Development of algorithms and software: JS, HPJ, HL, GH, AMN, PP, AG, SH, DD, MNM, FV, EG, YG, SM-S, HL, JME, JAS, BP, TM, XHC. Ethical, legal and social implications: RMC-D, NAG, BAK, AM, ABP. Programme organization: ALF, BIS, MWS, HJS.

3. Jordan M. Eizenga, Adam M. Novak, Jonas A. Sibbesen, **Simon Heumos**, Ali Ghaffaari, Glenn Hickey, Xian Chang, Josiah D. Seaman, Robin Rounthwaite, Jana Ebler, Mikko Rautiainen, Shilpa Garg, Benedict Paten, Tobias Marschall, Jouni Sirén, Erik Garrison. Pangenome Graphs. *Annual Review of Genomics and Human Genetics* 21, 1 (2020), 139-162.
`https://doi.org/10.1146/annurev-genom-120219-080406`

   My contributions:
   I made Table 1 and contributed to Sections 4.4 and 6.1 and Figure 2.

   Author contributions: JME wrote Sections 4.5, 6.2, 6.3, and 8.1; made Table 2; and helped revise the article. AMN wrote Section 6.1, contributed to Section 6.3, and helped revise the article. JAS wrote Sections 8.2 and 8.3 and contributed to Section 6.3. SH made Table 1 and contributed to Sections 4.4 and 6.1 and Figure 2. AG made Figure 1. G.H. contributed to Section 8.1. XC contributed to Section 6.3. JDS contributed to Table 1 and contributed significantly to Section 6.1. RR contributed to Section 4. JE contributed to Section 8.1. MR contributed to Section 8. SG contributed to Section 8.2. BP provided guidance and opinion. TM contributed to Sections 1, 2, and 6.2. JS wrote Section 5. EG organized the work; wrote Sections 1–3, 7, and 9; made Figures 2 and 3; and helped revise the article.

4. Jordan M. Eizenga, Adam M. Novak, Emily Kobayashi, Flavia Villani, Cecilia Cisar, **Simon Heumos**, Glenn Hickey, Vincenza Colonna, Benedict Paten, Erik Garrison. Efficient dynamic variation graphs. *Bioinformatics* 36, 21 (2020), 5139-5144.
`https://doi.org/10.1093/bioinformatics/btaa640`

   My contributions:
   I implemented some ODGI subcommands (pathindex, server, panpos), optimized one (bin), and wrote the whole documentation for ODGI.

1.3 Additional Publications

1. Gisela Gabernet, Susanna Marquez, Robert Bjornson, Alexander Peltzer, Hailong Meng, Edel Aron, Noah Y. Lee, Cole Jensen, David Ladd, Friederike Hanssen, **Simon Heumos**, nf-core community, Gur Yaari, Markus C. Kowarik, Sven Nahnsen, Steven H. Kleinstein. nf-core/airrflow: An adaptive immune receptor repertoire analysis workflow employing the Immcantation framework. *PLOS Computational Biology* 20, 7 (2024).
`https://doi.org/10.1371/journal.pcbi.1012265`

   My contributions:
   When the pipeline development started, I helped to get the initial docker image running and edited the manuscript.

2. Amr Aly, Zsofia I. Laszlo, Sandeep Rajkumar, Tugba Demir, Nicole Hindley, Douglas J. Lamont, Johannes Lehmann, Mira Seidel, Daniel Sommer, Mirita Franz-Wachtel, Francesca Barletta, **Simon Heumos**, Stefan Czemmel, Edor Kabashi, Albert Ludolph, Tobias M. Boeckers, Christopher M. Henstridge, Alberto Catanese. Integrative proteomics highlight presynaptic alterations and c-Jun misactivation as convergent pathomechanisms in ALS. *Acta Neuropathologica* 146 (2023), 451-457.
`https://doi.org/10.1007/s00401-023-02611-y`

   My contributions:
   I performed curation, quality control, and differential analysis of the proteomics and phosphoproteomics data. I edited the manuscript.

3. François Vasseur, Denis Cornet, Grégory Beurier, Julie Messie, Lauriane Rouan, Justine Bresson, Martin Ecarnot, Mark Stahl, **Simon Heumos**, Marianne Gérard, Hans Reijnen, Pascal Tillard, Benoît Lacombe, Amélie Emanuel, Justine Floret, Aurélien Estarague, Stefania Przybylska, Kevin Sartori, Lauren M. Gillespie, Etienne Baron, Elena Kazakou, Denis Vile, Cyrille Violle. A perspective on plant phenomics: coupling deep learning and near-infrared spectroscopy. *Frontiers in Plant Science* 13, 836488 (2022).
`https://doi.org/10.3389/fpls.2022.836488`

   My contributions:
   I managed the experimental design of the several hundred samples of the study. I edited the manuscript.

4. Christoph Ruschil, Gisela Gabernet, Gildas Lepennetier, **Simon Heumos**, Miriam Kaminski, Zsuzsanna Hracsko, Martin Irmler, Johannes Beckers, Ulf Ziemann, Sven Nahnsen, Gregory P. Owens, Jeffrey L. Bennett, Bernhard Hemmer, Markus

C. Kowarik. Specific induction of double negative B cells during protective and pathogenic immune responses. *Frontiers in immunology* 11, 606338 (2020). https://doi.org/10.3389/fimmu.2020.606338

My contributions:
I curated the initial data set. I edited the manuscript.

1.4 Conference Contributions

1.4.1 Papers

1. Andrej Baláz, Travis Gagie, Adrián Goga, **Simon Heumos**, Gonzalo Navarro, Alessia Petescia, Jouni Sirén. Wheeler maps. *LATIN 2024: Theoretical Informatics*. *Lecture Notes in Computer Science* 14578.
   `https://doi.org/10.1007/978-3-031-55598-5_12`

   My contributions:
   I advised on the integration of a wheeler maps implementation with real life pangenome graphs, built and provided initial pangenome graphs for testing the implementation, and made edit suggestions to the manuscript.

2. Jiajie Li, Jan-Niklas Schmelzle, Yixiao Du, **Simon Heumos**, Andrea Guarracino, Giulia Guidi, Pjotr Prins, Erik Garrison, Zhiru Zhang. Rapid GPU-Based Pangenome Graph Layout. *International Conference for High Performance Computing, Networking, Storage, and Analysis (SC)*. 2024.
   DOI pending.

   My contributions:
   I gave guidance on how the current algorithm is implemented and I gave feedback to the cache optimized CPU and GPU implementations. I tested these implementations. I read, criticized and made edit suggestions to the manuscript.

1.4.2 Posters

1. SWAT[4] HCLS Conference 2019 in Edinburgh, Scotland: Conference poster *Semantic Genome Graphs*.
2. ISMB BioViz 2020, virtual: Conference poster *Pantograph - Scalable Interactive Graph Genome Visualization*.
3. ISMB Bio-Ontologies 2020, virtual: Conference poster *Semantic Variation Graphs: Ontologies for Pangenome Graphs*. I won a best poster prize together with Toshiyuki T. Yokoyama.
4. T2T / HPRC Conference 2020, virtual: Conference poster *Graph Layout by Path-Guided Stochastic Gradient Descent*.
5. VCBM 2020, virtual: Conference poster *Graph Layout by Path-Guided Stochastic Gradient Descent*.
6. BoG 2021, virtual: Conference poster *The PanGenome Graph Builder*.
7. TüBiT 2021 in Tübingen, Germany: SSD sessions poster *The PanGenome Graph Builder*.

8. VIZBI 2022, virtual: Conference poster: *Graph Layout by Path-Guided Stochastic Gradient Descent*.
9. QBiC 10 Years and SAB Meeting 2022 in Tübingen, Germany: Symposium poster *Pangenome Graphs*.
10. TüBMI 2023 in Tübingen, Germany: Conference poster *Pangenome Graphs*.

1.4.3 Talks

1. Japan DBCLS Biohackathon 2019 in Fukuoka, Japan: Invited long talk symposium speaker *VG Browser: Interactive Visualization of Genome Variation Graphs*.
2. ISMB BioViz 2020, virtual: Conference talk *Pantograph - Scalable Interactive Graph Genome Visualization*. Josiah Seamann gave the talk, but I heavily contributed to the slides and the content of the talk.
3. ISMB Bio-Ontologies 2020, virtual: Shared conference talk *Semantic Variation Graphs: Ontologies for Pangenome Graphs*. Toshiyuki T. Yokoyama and I shared each one half of the talk.
4. GCB 2021, virtual: Shared 25 minute lecture *ODGI - scalable tools for pangenome graphs*. Andrea Guarracino and I shared each one half of the talk.
5. HPRC Pangenome Working Group August 2021, virtual: Talk *Identifying T2T and Centromere-Assembling Contigs*.
6. Institute for Medical Biometry and Bioinformatics October 2021 in Düsseldorf, Germany: Invited talk *Exploring pangenome graphs and possible applications*.
7. HPRC Pangenome Working Group October 2021, virtual: Talk *Precisely Identifying Assembly Breakpoints Relative to the References*.
8. IGGSy 2022 in Ascona, Switzerland: Conference talk *Graph Layout by Path-Guided Stochastic Gradient Descent*. Since I was involved in a heavy car accident previous to the conference, Erik Garrison took my slides and actually did the talking.
9. TüBMI 2022 in Tübingen, Germany: Conference talk *Exploring Pangenome Graphs*.
10. IBMI PhD Talks 2023 in Tübingen, Germany: Talk *Pangenome Graphs*.
11. MemPanG23 in Memphis, TN, USA: Invited workshop talk *nf-core/pangenome, pangenome growth*.
12. Nextflow Summit Barcelona 2023 in Barcelona, Spain: Conference talk *Cluster scalable pangenome graph construction with nf-core/pangenome*.
13. nf-core bytesize talks 2023, virtual: Bytesized talk *Cluster scalable pangenome graph construction with nf-core/pangenome*.
14. HPRC HUGO24 Workshop in Rome, Italy: Invited workshop talk *Building and analyzing pangenome graphs*.
15. MemPanG24 in Memphis, TN, USA: Invited workshop talk *nf-core/pangenome, pangenome growth*.
16. M3 Workshop 2024 Tübingen, Germany: Workshop talk *Cluster efficient pangenome graph construction with nf-core/pangenome*.

17. HPRC Pangenome Working Group June 2024, virtual: Talk *Cluster efficient pangenome graph construction with nf-core/pangenome*.
18. IGGSy 2024 in Ascona, Switzerland: Conference talk *Cluster efficient pangenome graph construction with nf-core/pangenome*.

2 Introduction

My doctoral research focuses on two interrelated areas, multiomics and pangenomics, each addressing a crucial aspect of reducing methodological biases in understanding biology. The overarching aim is to overcome the limitations of current methods that hinder our ability to fully understand biological systems. By investigating these areas, I seek to enhance the accuracy and depth of our biological insights.

"As was predicted at the beginning of the Human Genome Project, getting the sequence will be the easy part as only technical issues are involved. The hard part will be finding out what it means, because this poses intellectual problems of how to understand the participation of the genes in the functions of living cells."[1] This quote summarizes the duality of the results of the initial Human Genome Project (HGP) [159, 89]: Now we have the first human sequence, but understanding it is the real challenge. Solely reading this genomic code, a single one-dimensional (1D) data type, is insufficient to explain the set of observable characteristics of an individual. For example, mutations on the DeoxyriboNucleic Acid (DNA) level may influence protein expression which in turn can result in loss of functions or biological defects. More information is required to explore such complex biological patterns. Instrumental in this regard are a variety of so-called *omics* technologies which provide a multifaceted approach to characterizing biological systems, encompassing various molecular and functional dimensions: DNA in genomics and epigenomics, RiboNucleic Acid (RNA) in transcriptomics, proteins in proteomics, metabolites in metabolomics, microbes in microbiomics, and quantitative features from medical images in radiomics.

In practice, relying solely on a single data source can oversimplify the analysis and may actually increase noise, sometimes obscuring meaningful connections. Instead, a systems biology perspective that integrates multiple omics layers is crucial for understanding the complex biological functions that unfold across these molecular levels. This holistic approach, often referred to as *multiomics*, provides a more comprehensive view and helps to elucidate the interactions between different molecular systems.

This is crucial when matching molecular and phenotypic factors, especially in disease studies. Diseases often entail complex interactions across multiple biological layers, and a single data type may not capture the full complexity of these interactions. By integrating multiomics data, researchers can better understand the multifaceted nature of diseases, identify key biomarkers, and develop more targeted and effective treatments.

[1] Sydney Brenner (13 January 1927 – 5 April 2019) was a South African nobel prize winning biologist

Despite significant advances in omics research, challenges persist. One notable issue is the reliance on reference genomes in traditional genomics, which introduces potential *bias*. Reference genomes are typically assembled from multiple individuals and may not accurately represent the genomic diversity within a population. This can result in *reference bias*, where discrepancies between the individual's genome and the reference genome lead to mismatches, errors, or difficulties in data alignment [6].

To overcome these limitations, *pangenomics* [170] offers an innovative approach. A pangenome models a representative set of genomic sequences from a population, providing a more comprehensive view of genetic diversity. In contrast to reference-based genomic approaches that compare sequences to a single linear reference genome, pangenomics relates each new sequence to all other sequences within the pangenome. This approach helps to capture a broader range of genetic variation and mitigate some of the biases inherent in traditional genomics.

A *pangenome graph* can compress the shared and variant sequences into one graphical representation. In pangenome graph models, DNA sequences are stored in nodes and edges connect the nodes as they occur in the individual sequences [73]. Genomes are encoded as paths traversing the nodes [59]. However, constructing and analyzing pangenome graphs present their own set of challenges. Existing pipelines often suffer from reference bias or limitations in computational efficiency and scalability. Leveraging the full potential of a pangenomic data set, novel, qualitatively different computational methods and paradigms are needed.

To address these limitations, my doctoral research focuses on two interrelated areas: multiomics and pangenomics, each targeting specific research questions to advance our understanding of genotype-phenotype relationships.

In the multiomics domain, the research aims to explore novel theranostics for cancer immunotherapy. Thus, my first research question investigates how integrating cell surface molecule data from a diverse cancer cell panel with transcriptomics and proteomics profiles can identify unique cell surface markers for specific cancer types. The second question examines the validity of these markers by comparing tumor versus normal samples from RNA-Seq and Reverse Phase Protein Array (RPPA) data. This approach leverages comprehensive multiomics analyses and biodata catalogs to enhance phenotype discovery and validation.

In the pangenomics domain, the research addresses several main questions. The first question focuses on developing a cluster-efficient, reference-free pangenome graph construction pipeline to overcome issues with reference bias and deployment limitations. A pipeline is a series of interconnected processes that automate the workflow for constructing and analyzing data. This approach is essential for efficiently managing complex analyses, ensuring reproducibility, particularly when utilizing containerization for consistent deployment across different computational environments [87]. By integrating multiple steps, a pipeline enhances the overall efficiency of data handling and allows researchers to focus on interpretation rather than manual data processing. My pipeline aims to integrate multiple reference genomes into a single graphical model, facilitating the study

of genomic diversity across populations. Existing algorithms for visualizing and analyzing pangenome graphs often fall short, making it difficult to fully understand these complex structures. My second research question addresses these limitations by developing a new pangenome graph layout algorithm that leverages biological information to enhance visualization. Additionally, my aim is to implement methods for detecting complex regions, adding annotation, and performing exploratory analyses. The goal is to enable researchers to explore genotype-phenotype relationships in large-scale pangenome graphs, accommodating all sequences in the graph and not just a single reference sequence.

By addressing these research questions, my thesis aims to advance our ability to interpret complex biological data and improve the discovery of genotype-phenotype relationships minimizing the bias of reference data, paving the way for more precise and effective biological insights.

2.1 Genotype-Phenotype Relationship

In 1911, the Danish botanist Wilhelm Ludvig Johannsen coined the terms *genotype* and *phenotype* elaborating on their conceptual relationship [82, 25]. The genotype is an organism's complete set of genetic material. It consists of the specific combination of genetic variants or alleles that an individual possesses. Genetic code is inherited from the individual's parents. In polyploid organisms, if all copies of the alleles at a given locus are identical, the genotype is *homozygous*. If there are different alleles present among the multiple sets of chromosomes, the genotype is *heterozygous* [71]. A combination of alleles across multiple adjacent loci or genes within a chromosome is typically called a *haplotype*. Phasing refers to the process of determining which specific alleles are inherited together on the same chromosome, allowing for the accurate reconstruction of haplotypes from sequencing data [15].

A genotype partially influences an organism's characteristics and observable traits, the phenotype of an organism, alongside epigenetic or environmental factors. This is called *genotype-phenotype (GP) relationship*. Organisms sharing the same genotype can have different phenotypes due to environmental influences. A good everyday example are identical twins, who have identical genotypes but diverging phenotypes.

A trait that is solely determined by a genotype usually follows the Mendelian inheritance pattern (Fig. 2.1). In 1866, Gregor Mendel observed that, although the phenotypes of individuals in the parent generation varied, the offspring expressed a single, uniform phenotype [106]. He attributed this phenomenon to the segregation of dominant and recessive alleles. However, the interaction of genotypic and environmental factors gives rise to more complex traits. A classic example are flamingos: Naturally white, their diet's carotenoid content determines their pink plumage [2]. This illustrates how environmental factors can interact with genetic ones to influence complex traits.

Thus, scientists require an arsenal of various *omics* technologies in order to compre-

hend the complex GP landscape of an organism. One single point of observation does not suffice to capture the multifaceted complexities underlying biological systems. In medicine, the integration of GP data is essential for tackling intricate medical issues, but it's not strictly necessary for every condition (e.g a broken leg).

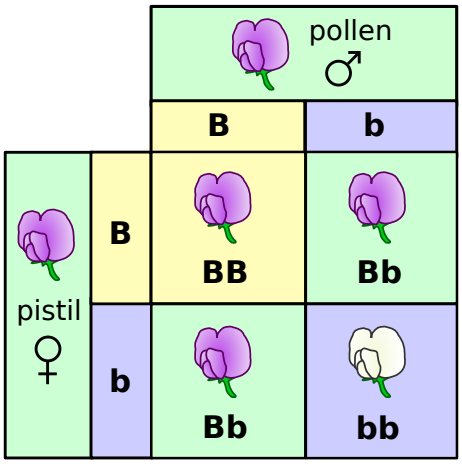

Figure 2.1: Mendelian genotype-phenotype relationship. A Punnett square illustrates the character of petal colour in a pea plant. The letters represent alleles and the pictures show the resultant flowers. The diagram shows the cross between two heterozygous parents where B represents the dominant allele (purple) and b represents the recessive allele (white). Figure and caption modified from *Ball 2007* [5].

2.2 Multiomics

Multiomics (sometimes referred to as panomics) is the integrative study of multiple omics technologies incorporating datasets from genome, proteome, transcriptome, epigenome, radiome, metabolome, as well as assays like fluoresence-activated cell sorting (FACS), and phenotypic or clinical data (Fig. 2.2). Multiomics is an analysis approach to understand and study the complexity of life in an integrated and complementary manner: The integration of information from complementary molecular mechanisms leads to more accurate analysis results compared to making use of only one single technology, effectively reducing analysis bias. In multiomics analysis, goals include unsupervised clustering or supervised classification to uncover sample-specific patterns or predict outcomes, as well as identifying key features for biomarker discovery. Additionally, feature-focused analyses aim to explore relationships across omic layers through integration techniques that reveal shared structures and interactions, often visualized as networks. Multiomics plays a key role in personalized medicine [120, 84], health and disease [18], the discovery of relevant biomarkers, and is capable of refining matching GP relation-

ships [145]. In the following, I will focus my introduction on the subset of multiomics technologies that are important for my thesis.

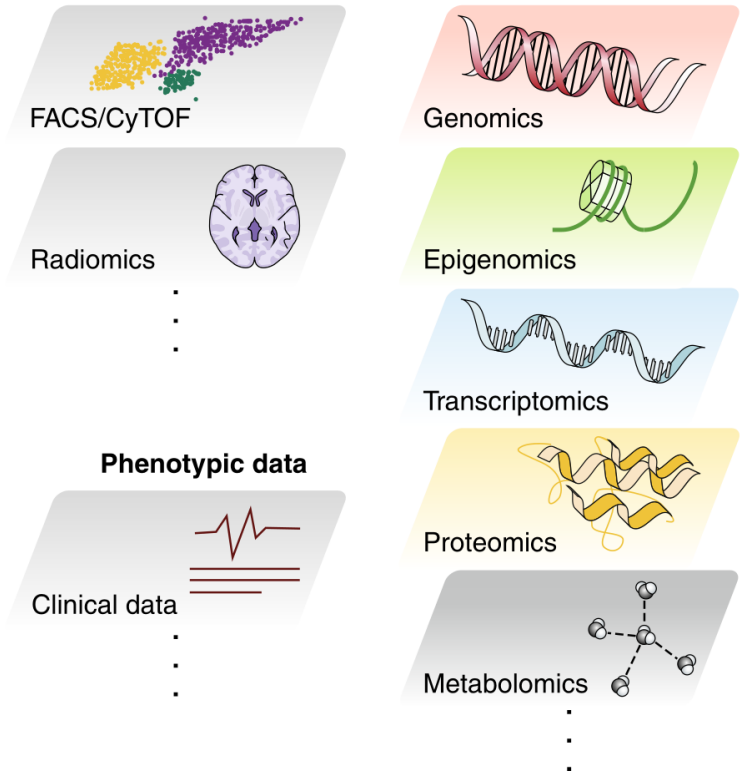

Figure 2.2: Multiomics datasets. Multiomics datasets may be defined by any combination of molecular profiling data modalities, such as genomics, epigenomics, transcriptomics, proteomics or metabolomics, and include other types of high-throughput data such as FACS (fluorescence-activated cell sorting), CyTOF (cytometry by time of flight), or radiomics measurements, as well as phenotypic or clinical co-variates. Figure and caption modified from *Tarazona 2021* [146].

2.2.1 Genomics

Genomics studies the evolution, structure, function, mapping and editing of an organism's genome. Scientists can determine DNA nucleotide sequences from a wide range of sequencing technologies. Sanger Sequencing [122] was the first major sequencing technology. It synthesizes DNA using chain-terminating nucleotides, which allow for the determination of the DNA sequence by producing fragments of varying lengths. The method provides highly accurate sequences with read lengths of up to approximately 1000 base pairs. However, its low throughput limits its application in large-scale genomic studies, making it less practical for comprehensive genomic projects.

Next-Generation Sequencing (NGS) technologies [137] have largely replaced Sanger sequencing due to their high throughput and scalability. Among these, Illumina's se-

quencing platform is predominant [121]. Illumina sequencing measures the incorporation of fluorescently labeled nucleotides during DNA synthesis. The process involves bridge amplification to form clusters of clones from single DNA molecules, followed by synthesizing and sequencing the DNA one base at a time. This massively parallel process produces millions of short reads (36-300 base pairs for a single end read [123]) per sample, offering a high degree of accuracy and efficiency. However, the relatively short read lengths can pose challenges in assembling highly repetitive or complex genomic regions [150].

Third-Generation Sequencing technologies, such as those developed by Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), measure single molecules and offer significant advancements in read length and throughput. ONT [42] employs nanopore sequencing, which measures changes in electrical current as a DNA molecule passes through a protein nanopore. The distinct current fluctuations correspond to different nucleotides, allowing for real-time sequencing of long DNA molecules. ONT can generate extremely long reads, sometimes exceeding 1,000,000 base pairs, which is advantageous for assembling complex genomes and detecting structural variants. However, it has a higher error rate compared to other sequencing technologies, necessitating robust error correction algorithms. PacBio [108] utilizes Single Molecule Real-Time (SMRT) sequencing, which measures the incorporation of fluorescently labeled nucleotides by a DNA polymerase in real-time. The method involves attaching DNA molecules to a SMRT cell, where DNA polymerase is anchored in the zero-mode waveguide of the cell. As the DNA polymerase synthesizes the DNA, the incorporation of fluorescently labeled nucleotides is detected in real-time. While single-pass long reads can reach lengths of 10,000 to 30,000 base pairs, the accuracy of these reads is significantly improved through Circular Consensus Sequencing (CCS). CCS typically produces highly accurate reads with lengths ranging from 10,000 to 20,000 base pairs, offering a trade-off between read length and accuracy. The main limitation of PacBio is its lower throughput and higher cost compared to NGS technologies.

2.2.2 Transcriptomics

Sequencing technologies can provide insights not only into the genomics parts of the organisms, but are also valuable tools when it comes to studying the presence and quantity of the complete set of RNA molecules, the transcriptome of an organism.

Microarray

While microarrays do not involve sequencing, they are commonly used for gene expression analysis and are often compared to sequencing technologies due to their ability to measure the expression levels of multiple genes simultaneously. However, unlike sequencing technologies, microarrays do not provide the full nucleotide sequences of the

target genes, limiting their ability to detect novel transcripts or variations. A microarray chip measures gene expression levels of thousands of genes in parallel by hosting millions of known nucleic acid fragment sequences, called probes, on a solid surface. In a typical experiment, DNA or RNA samples are first labeled with fluorescent dyes. These labeled nucleic acids are then washed over the chip, where they hybridize with complementary probes on the array. The fluorescent dye attached to the nucleic acids emits light when excited by a laser or another light source. The intensity of the fluorescence emitted from each spot on the chip correlates with the amount of hybridized nucleic acid, which indicates the gene expression levels [147]. This method allows for massively parallel gene expression profiling. However, the design of the chip can lead to cross-hybridization artifacts, poor quantification of lowly or highly expressed genes, and is limited to measuring only known transcripts. This is where the RNA sequencing technology steps in.

RNA-Seq

RNA sequencing (RNA-Seq) generates a snapshot of gene expression by determining both the sequence and abundance of transcripts in a given sample [163]. The method converts a pool of RNA molecules into a library of complementary DNA (cDNA), to which adapters are attached at both ends. This library is then sequenced using high-throughput NGS technology. The resulting reads are aligned to a reference genome, and the number of reads corresponding to each transcript is counted. RNA-Seq offers a comprehensive global gene expression profile without being limited to known transcripts. However, the actual proteins present in an organism may differ from what the transcriptomic profile suggests, due to various post-transcriptional regulatory mechanisms. To complement this analysis, proteomics is employed to profile all proteins expressed in the organism, providing insights into the functional output of the genome.

2.2.3 Proteomics

Proteomics is the large-scale study of proteins, particularly their structures and functions. Several advanced techniques are used in proteomics to identify and quantify proteins within a sample. Proteomics can assist in identifying protein biomarkers, understanding disease mechanisms, and discovering new therapeutic targets [1]. In the following an overview is given of the technologies that appear within this thesis.

Mass Spectrometry In Proteomics

Tandem mass spectrometry (MS/MS) is a powerful technique used in proteomics to analyze proteins by measuring the mass-to-charge ratio (m/z) of ions and their subsequent fragmentation patterns. In this approach, proteins are first digested into peptides using an

enzyme such as trypsin. The resulting peptides are then separated by liquid chromatography (LC) and ionized, typically using electrospray ionization (ESI) [47].

Once ionized, the peptides are introduced into the mass spectrometer, where they undergo a first round of mass analysis to determine their m/z ratios. Selected peptides are then fragmented, and the m/z ratios of the resulting fragmentation patterns provide information that allows for the determination of the peptide's amino acid sequence. This combination of liquid chromatography and tandem mass spectrometry (LC-MS/MS) enables comprehensive proteomic analysis but necessitates careful sample preparation and sophisticated data interpretation.

Fluorescence-Activated Cell Sorting

Although fluorescence-activated cell sorting (FACS) does not directly analyze proteins in the same way as mass spectrometry or protein arrays, it can provide information about protein expression on a cellular level. It can be used to study cell (surface) molecules, including protein markers and antigens at the cell surface or within cells rather than providing detailed protein characterization or quantification. While cell surface receptors and antigens are distinct concepts with different primary roles, a receptor can act as an antigen under specific circumstances where it is recognized by the immune system such when it interacts with an antibody or a T cell receptor. However, not all antigens are cell surface receptors. FACS [13, 101] is a specialized technique of flow cytometry to separate and analyze individual cells from a heterogeneous biological sample, one cell at a time. The light scattering and fluorescent characteristics of a cell determine in which container a cell is put. Typically, fluorescently labeled antibodies bind to specific cellular molecules unique to each cell type. The Mean Fluorescent Intensity (MFI) quantifies the expression level of surface or intracellular molecules of a cell population, potentially validating antibody binding. FACS has several limitations, including its dependence on fluorescent labeling, which can affect specificity, and its inability to provide detailed information on intracellular proteins or quantify protein levels accurately. Additionally, while the equipment can be costly, the overall cost per sample is generally lower compared to sequencing methods. FACS data can also be complex to interpret. Alternatives to FACS include mass cytometry and single-cell RNA sequencing, which offer different approaches for analyzing cellular and molecular characteristics.

Reverse Phase Protein Array

Reverse Phase Protein Array (RPPA) is a high-throughput technique that quantifies the expression of hundreds of proteins across many samples simultaneously using antibody-based microarray technology [27]. Compared to LC-MS, it often provides higher reproducibility for specific protein quantification under controlled conditions.

In contrast to forward phase protein arrays, where the antibodies are fixed on a surface and the samples are passed over for binding, in the RPPA process the proteins of the

samples are immobilized onto the surface and a set of specific antibodies is used to probe the samples [139]. This method enables the analysis of multiple protein-signaling pathways and their activity. RPPA's robustness comes from replicates for each antibody, ensuring accuracy. However, its effectiveness is highly dependent on the quality and specificity of the antibodies used. While RPPA can provide reliable pathway analysis, the antibodies often bind unspecifically in complex cell lysates, even if they perform well in Western blot assays. Additionally, RPPA provides less detailed peptide-level information compared to mass spectrometry.

Immunohistochemistry

Immunohistochemistry (IHC) measures semiquantitative protein expression within cells or tissues by detecting specific antigens with antibodies that are linked to detectable markers. Antibodies are conjugated to markers such as enzymes or fluorescent dyes. When antibodies bind to their target proteins in tissue sections, the enzyme converts a substrate into a colorimetric product visible under a light microscope, or the fluorescent dye emits light when excited by a specific wavelength, which is visible under a fluorescence microscope. The intensity of the color or fluorescence indicates the abundance of the target protein. IHC allows for the localization and quantification of proteins within tissue contexts, making it valuable for studying protein distribution and abundance. However, it is semiquantitative and relies on the quality of antibody binding and tissue preservation. IHC has successfully been applied in the diagnosis and validation of metastatic carcinomas [127].

2.2.4 Multiomics Algorithms

Algorithms used in multiomics are diverse and tailored to handle the high-dimensional and complex data sets produced by multiomics experiments. Multiomics is the integration of multiple omics feature-by-sample matrices with optional annotation or other metadata. Analysis methods provide means to cluster samples, discover molecular mechanisms, or predict therapy outcome [11].

One exploratory analysis approach I employed during my thesis is multiple co-inertia analysis (MCIA) [107]. MCIA is an extension of Co-Inertia Analysis (CIA) [39] and works by simultaneously analyzing two or more datasets in a way that maximizes the covariance between them. MCIA uses a covariance optimization criterion to project multiple datasets into a shared dimensional space, aligning diverse feature sets on a common scale. This allows for the extraction of the most variable features from each dataset, aiding in biological interpretation and pathway analysis. The method identifies a set of loading vectors for each dataset, which are used to project the datasets into a common dimensional space. Additionally, it determines a "synthetic" center that represents the combined data from all datasets. By maximizing the sum of squared covariances between the linearly transformed datasets and this synthetic center, MCIA enhances the

alignment of the datasets in the shared space. This optimization process ensures that the most significant patterns and relationships across the different datasets are captured, facilitating more coherent biological interpretation and pathway analysis.

Specifically, MCIA requires omics data matrices where the number of features (rows) exceeds the number of measurements (columns). A prerequisite is that either features or measurements are matched and have equal weights. Given an omics data table $M = [m_{ij}]$ with $1 \leq i \leq n$ and $1 \leq j \leq q$. M is a $(n\ x\ q)$ matrix with row index i and column index j. The row sum is m_{i+}, the column sum is m_{j+}, and the total sum of the matrix is m_{++}. The relative contribution of i to the variation of the dataset is $r_i = \frac{m_{i+}}{m_{++}}$, and of j it is $c_j = \frac{m_{j+}}{m_{++}}$. The contribution of each individual element to the total variation is given by $p_{ij} = \frac{m_{ij}}{m_{++}}$. The key mathematical steps of an MCIA according to *Meng et al. 2014* [107] involve:

- **Table ordination method:** A Principle Component Analysis (PCA) or comparable method is applied to each dataset separately, so that only the most informative components, those that capture the largest variance, from each dataset are considered when computing the co-inertia.
- **Data centering:** We derive a new matrix X by normalizing each data point:

  $$x_{ij} = \frac{p_{ij}}{r_i} - c_j$$

  This yields the centered row profile x_{ij}.
- **Maximizing sum of squared covariance:** This step is a generalization of the CIA. It simultaneously analyzes a set of statistical triplet (X_k, Q_k, D) where $k = 1, ..., k, K$ and X_k is a set of transformed matrices. Let Q_k be a $q_k\ x\ q_k$ matrix where the diagonal elements r_{ij} represent the feature metrics in the hyperspace. Let D be an $n\ x\ n$ identity matrix, reflecting equal weights across all columns of the tables involved. In this context, MCIA aims to maximize the sum of the squared covariance between scores of each table with synthetic axes v:

  $$f(u_1, ..., u_k, ...u_K, v) = \sum_{k=1}^{K} w_k\ cov^2(X_k Q_k u_k, v)$$

  where cov^2 denotes the square of the covariance of the quantities within the parentheses. The term w_k represents the weight assigned to each table. The vector v indicates the reference structure or synthetic center while u_k are the auxiliary axes. The score for each individual table can be expressed as:

  $$X_k = Q_k u_k$$

  Unlike other ordination methods, MCIA determines the solutions for u_k and v sequentially. Multiple matrices X_k can be weighted and concatenated into a single

matrix X represented as:

$$X = [w_1^{1/2} X_1 | ... w_K^{1/2} X_K]$$

Similarly, the individual feature metrics Q_k can be concatenated into a single feature metric Q as follows:

$$Q = [Q_1 | ... | Q_k]$$

Next we want to find the principal components:

- **First order solutions (first principal component):** First, we calculate the first principal component, which is essentially the "main direction" where the data varies the most across all the tables. This is done using the following equation:

  $$wXQX^T Dv = \lambda v$$

  where X is the data from your tables, Q represents how we measure the importance of the features in the data, D is the identity matrix, v is the first "synthetic" axis (or direction) that represents the shared structure across all tables, and λ is an eigenvalue, which helps to measure how important that direction v is. What we are doing here is finding a "direction" (or axis) v that captures the most important shared patterns across the data tables.
- **Calculating auxiliary axes:** Now that we have v, we can calculate the first auxiliary axis u_{k1} for each table. This axis shows how much each individual table follows the shared pattern v.

  $$u_{k1} = \frac{X^T Dv_1}{||X^T Dv_1||_{Q_k}} \ (k = 1, ..., K)$$

  This means we use data X and the direction v to calculate the new axis for each table, where $||\cdot||_{Q_k}$ is a way to normalize this value. This gives us the first direction for each table that aligns with the shared pattern.
- **New directions must be orthogonal:** After finding the first axis v_1, we now want to find the next principal component – the second-most important direction where the data varies. But to make sure that this new direction V_2 doesn't overlap with the first, we introduce a constraint:

  $$v_j^T Dv_s = 0 \quad \text{and} \quad u_j^T Q_k u_s = 0 \quad \text{for} \quad (1 \leq j < s)$$

- **Residual matrix:** To find the next direction, we need to "remove" the effect of the first principal component. This is done by subtracting it from the data using a residual matrix:

  $$X_1^{(2)} = X - XP_{k1}$$

  This removes the influence of the first axis u_{k1} from the data, allowing us to

focus on the remaining variation.

– **Projecting the data:** The matrix P_{k1} is a projection matrix, which represents the first direction we found. It's given by:

$$P_{k1} = (u_{k1} u_{k1}^T Q_k u_{k1}^T)^{-1} u_{k1} Q_k$$

This matrix essentially captures how the first direction u_{k1} interacts with the feature metrics Q_k. We use it to "remove" the first direction and focus on finding new, independent patterns in the data.

The processes **Residual matrix**, **New directions must be orthogonal**, and **Projecting the data** are repeated until we have found the desired number of principal components.

As a result, MCIA produces a joint ordination of both columns (measurements) and rows (features) from multiple tables within a unified hyperspace. Features or measurements that exhibit similar patterns will be projected closely together. A more detailed explanation of MCIA, along with the proof that these axes maximize covariance, can be found in the work of Chessel and Hanafi [20].

2.2.5 Multiomics Resources

There exists a wide range of multiomics data portals: The Cancer Genome Atlas (TCGA) [164], the Human Protein Atlas (HPA) [149], The Cancer Proteome Atlas (TCPA) [154], the Genotype-Tissue Expression (GTEx) [105], and recount2 [29], a resource of processed and summarized expression data. This is by no means a comprehensive list, but it is sufficient for following this thesis. Readers interested in exploring more resource can refer to [31, 141, 113, 155] for additional information.

2.3 The Truth Lies In The Eye Of The Reference

Some of the omics methods that have been presented here, such as the analysis of DNA sequencing and RNA sequencing data, rely on *reference genomes*. These reference genomes are widely used in genomics, serving as a foundation for a variety of analyses, including protein identification, gene annotation, read mapping, and variant detection [134]. However, this reliance can introduce *reference bias*, where discrepancies between the reference genome and individual genomes may lead to inaccuracies in these analyses.

High quality collections of population-wide sequences are becoming more common due to low-cost whole-genome assembly. This offers new opportunities to study genomic variation as never before. However, it is a challenge to simultaneously represent and analyze hundreds of genomes at a gigabase scale [6].

One possible solution can be a pangenome modeled as a graph (Figure 2.3). Compared to a single linear reference genome, pangenome graph models introduce several improvements: (i) The inclusion of variants from multiple individuals better represents a diverse population, (ii) the higher accuracy of read mapping to a pangenome graph especially in highly variable or repetitive regions leads to better variant calling results, (iii) complex regions are better resolved in a pangenome graph, (iv) novel reference-agnostic variants are detected, (v) and annotation of genetic variants across different individuals improve understanding of genetic function and evolutionary processes.

In clinical settings, using a pangenome graph can improve the accuracy of genomic diagnostics and the identification of disease-associated variants which leads to better personalized treatment strategies. In plant and animal breeding, pangenome graphs can help identify beneficial genetic variations that contribute to desirable traits, facilitating the development of improved breeds and cultivars. In summary, pangenomics improves our ability to explain traits and provides an in-depth understanding of the GP relationship, which would not be possible when working with a traditional single linear reference genome.

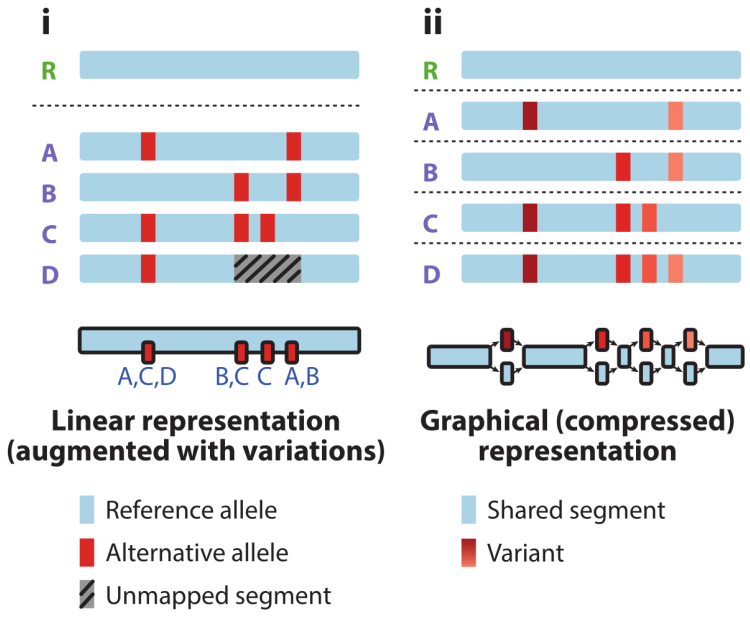

Figure 2.3: Pangenomic models. **(i)** Traditional linear pangenome representation. Regions of some genomes are unalignable against the reference and cannot be represented in a list of variants. **(ii)** A graphical model of the genomes allows a direct all-to-all comparison, capturing all of their sequence relationships. **R** in green is the reference genome. **A,B,C,D** in blue are alternate sequences. Figure and caption modified from *Eizenga et al. 2020* [43].

2.3.1 The Human (Pangenome) Reference Is In Flux

Sequencing the human genome has always been a great challenge. Since its first official *completion* in 2001 [159, 89], there have been many updates and modifications due to an improvement in sequence assembly technologies [24, 79]. The latest release, GRCh38 [125], came really close to a *complete* reference, however, it does not encode highly repetitive and complex regions like centromeres or telomeres. In 2022 the Telomere-to-Telomere (T2T) consortium released a gapless telomere-to-telomere reference genome based on a haploid human cell line – CHM13 [112], adding the sequence of HG002's chromosome Y a year later [117] (the CHM13 original cell line is female) [117]. The complex centromeric regions were studied in detail [104]. But this still left scientists with a single linear reference genome.

This is addressed by the Human Pangenome Reference Consortium (HPRC) initiative. It aims at assembling a comprehensive set of individuals' sequences that encompasses the full spectrum of human genome diversity. Making use of the most recent sequencing technologies, the HPRC sequenced and assembled 47 phased ultra high quality genomes from a genetically diverse background [100]. They demonstrated that over 99% of each assembly, both at the base pair and structural levels, and more than 90% of highly repetitive sequences were structurally accurate, adding a significant number of structural variants compared to GRCh38. The HPRC created a draft human pangenome reference with these assemblies. The final aim is to acquire the haplotypes of 350 diverse individuals.

Since then, other consortia around the globe have published or are actively working on their specific pangenomes: The Chinese Pangenome Consortium (CPC) released a pangenome of 36 populations [53], and the H3ABioNet consortium, the Pan African Bioinformatics Network for the Human Heredity and Health in Africa (H3Africa), assembled 910 human genomes of African origin [131]. The Pan-European Pangenome Consortium (Pangaia) is a project to teach young researchers in the area of pangenomics.

2.3.2 Pangenome Graphs

Variation graphs are a mathematical formalism to represent pangenome graphs [54]. All newly developed tools and algorithms presented in this thesis use this variation graph model.

Definition 2.3.1. In the variation graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{P})$, nodes (or vertices) $\mathcal{V} = v_1 \dots v_{|\mathcal{V}|}$ contain nucleotide sequences. Each node v_i has a unique identifier i and an implicit reverse complement $\bar{v_i}$. The node strand o represents the node orientation. Edges $\mathcal{E} = e_1 \dots e_{|\mathcal{E}|}$ connect ordered pairs of node strands ($e_i = (o_a, o_b)$), defining the graph topology. Paths $\mathcal{P} = p_1 \dots p_{|\mathcal{P}|}$ are series of connected steps s_i that refer to node strands in the graph ($p_i = s_1 \dots s_{|p_i|}$); the paths represent the genomes embedded in the graph.

In order to understand pangenome graphs, we need to see them. Some visualization styles of a simple variation graph are given in Figure 2.4. More advanced and

better-scalable visualizations are given in Figure 2.5. As of 2020, the Bandage force-directed 2D layout algorithm was the best-scaling one up to 100 Mbp [43]. An interactive gigabase-scale and human-pangenome-detail visualization with zoom levels did not exist. During my thesis I bridged the gap working on a more scalable 2D layout algorithm.

2.4 Graphical Pangenomic Analysis

One standard way to create a pangenome is from a Multiple Sequence Alignment (MSA). The most recent algorithms employ Partial Order Alignment (POA) [90]: The iterative alignment to a directed acyclic graph (DAG) through dynamic programming (DP). Effectively, an MSA can be described as a pangenome graph and vice versa (Figure 2.6). However, even the most recent implementations, Single Instruction Multiple Execution (SIMD) Partial Order Alignment (SPOA) [158], which utilizes SIMD operations for parallel processing of data, adaptive banded Partial Order Alignment (abPOA) [52], or POASTA [157] have limits at megabase-length alignments and can't scale to gigabase sequences.

Alternative unbiased pangenome graph construction methods employ *k*-mers [110, 58, 23], but the resulting graphs are often complex and impractical to use for downstream analysis due to their parameter sensitivity, especially *k*-mer length.

Other projects [94, 77] rely on tree or reference-guided approaches [4, 111]. They

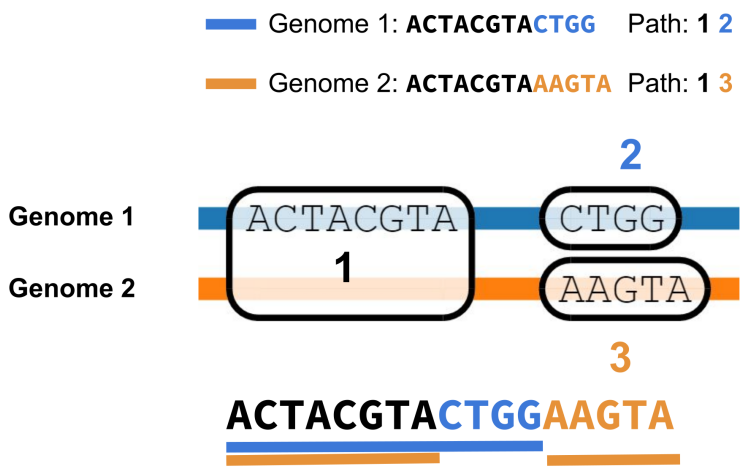

Figure 2.4: Variation graph example. The large numbers **1,2,3** are node identifiers. (Top) The path information of the graph. (Mid) Sequence Tube Map [12] style visualization. Nodes are drawn as rectangles including their respective sequence. Genomic sequences are drawn as colored paths through the nodes. Genomes only visit nodes that contain their respective sequence. (Bottom) *odgi viz* [65] style 1D visualization of the same graph. The pangenomic sequence is arranged from left to right. Colored bars indicate the sequence of the respective genome.

employ a *reference pangenome graph* [94] model: Sequences are iteratively aligned to a linear reference genome, which serves as a fixed positional backbone, as well as to previously aligned sequences within the pangenome graph. This obviously introduces reference bias. It can also lead to an incomplete representation of the input sequences in the resulting pangenome graph [58], especially when pruning complex sequences like centromeres or other satellite sequences [94, 77].

One solution here is to treat all input sequences the same way, building up a pangenome graph from all-vs-all alignments without preferential treatment of a *reference*.

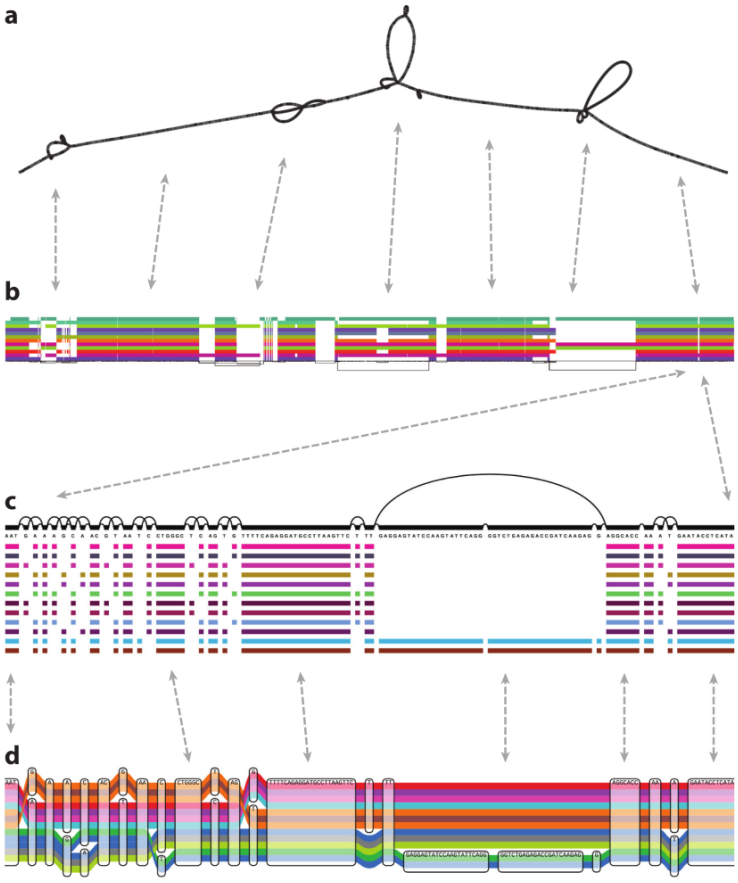

Figure 2.5: Visualizing a graph of GRCh38 and its alternate sequences in the gene HLA-DRB1 built with VG msga (Variation Graph multiple sequence/graph aligner) [59]. (a) Bandage's force-directed layout, revealing large-scale structures [166]. (b) An ODGI viz (Optimized Dynamic Genome Graph Implementation visualization) binned, linearized rendering of the paths (colored bars) versus the sequence and topology of the graph (thin lines below the bars). (c) A fragment of a VG viz (Variation Graph visualization) linearized rendering, showing base-level detail. (d) The same fragment rendered with Sequence Tube Map [12]. Dashed lines show the correspondences between the visualizations. Path colors are assigned independently by each method. Figure and caption modified from *Eizenga et al. 2020* [43].

Such an approach is described in the following paragraphs.

The PanGenome Graph Builder (PGGB) [60] pipeline (Figure 2.7), which constructs pangenome graphs based on the symmetric comparison of all genomes to all others, thereby mitigating reference bias, is such an approach. Unlike competing methods, PGGB allows for multiple reference genomes to be fully embedded in the graph. PGGB is unaffected by genome input order and orientation.

PGGB iteratively refines an all-to-all whole-genome alignment graph, enabling the exploration of sequence conservation and variation, phylogenetic inference, and the identification of recombination events. First, the whole-chromosome pairwise sequence aligner WFMASH [67] generates the all-vs-all alignments. These are squished into a variation graph with SEQWISH [58]. The graph is normalized with SMOOTHXG [60]: SMOOTHXG iteratively applies a local MSA kernel, POA, to refine and compress the pangenome graph. In detail, the graph's nodes are ordered [76] according to their occurrence in the graph's embedded paths and then split into segments on which POA is applied. By default, the SMOOTHXG process is applied 3 times in order to smooth the edge effects at the boundaries of the segments. The final normalization step applies GFAFFIX [100] in order to collapse redundant nodes. Graph statistics and diagnostic 1D and 2D visualizations are summarized in an interactive HTML report.

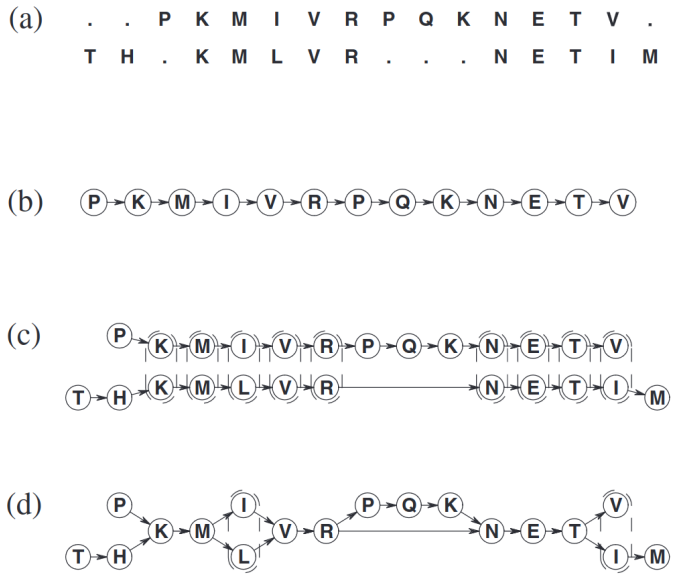

Figure 2.6: MSA in the POA representation. (a) MSA representation of a pairwise protein sequence alignment. Dots indicate no base level alignment. (b) A single sequence visualized as a simple DAG. (c) Two protein sequences in a DAG aligned to each other. Dashed ovals indicate that two nodes are aligned. (d) DAG representation of a pairwise protein sequence alignment. Dashed ovals indicate that two nodes are aligned. Figure and caption modified from *Lee et al. 2002* [90].

Once we have built a reference-unbiased pangenome graph, we want to explore the biology it models. A scientist might want to identify variation, detect complex regions, measure conservation, extract pangenomic loci, detect recombination events, remove artifacts, do exploratory analysis, and infer phylogenetic relationships. This would make the graphs themselves a valuable tool for studying sequence evolution and variation deepening our understanding of complex GP relationships. In 2019, the state-of-the-art software was the vg toolkit [59]. But since then, pangenome graphs became more complex, calling for new software to be implemented to understand reference-unbiased pangenome graphs.

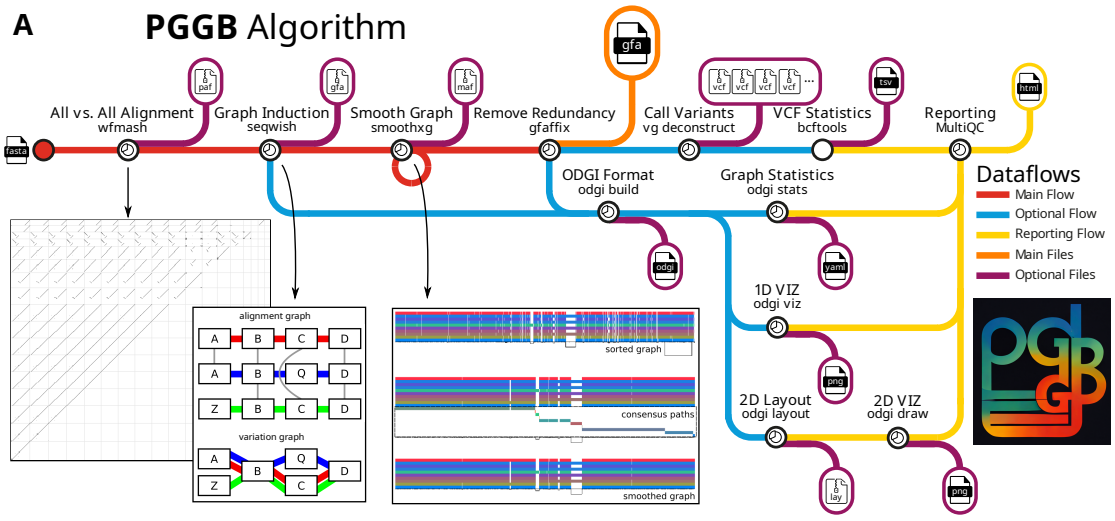

Figure 2.7: PGGB. (A) PGGB's algorithms/data flows. Primary flow (red) proceeds from FASTA to alignment, graph induction, smoothing, to normalization with GFAFFIX [100], ending with the final variation graph (orange). Optional outputs (blue): statistics, variant calls, and 1D/2D graph visualizations. Figure and caption modified from *Garrison et al. 2023* [60].

3 Objectives

The overall goal of my thesis was to apply, develop, and evaluate comprehensive state-of-the-art methods advancing the discovery of genotype-phenotype relationships. The challenge was to integrate data of different biological origins in order to interpret and uncover relationships for which the information from a single data source would not have been sufficient. I specifically focused on two key areas of biological research: multiomics and pangenomics.

The multiomics part of my thesis focuses on utilizing integrative multiomics to explore novel theranostics for cancer immunotherapy. Immunotherapy using immune checkpoint inhibitors (ICI) is untargeted and can lead to unwanted side effects [86]. Ideally, cancer immunotherapy interferes with cancer-specific cell molecules only. We profiled the cell surface molecules of the NCI-60 cancer cell panel [132] using the flow cytometry method Fluorescence-Activated Cell Sorting (FACS). As a first research question, I screened for the most variably expressed cell surface antigens. I integrated the FACS data with previously generated transcriptomics (Tx) and proteomics (Px) NCI-60 profiles [63, 107] using a MCIA [107]. I identified potential cell surface markers unique for a specific cancer entity. The second research question directly arose from the first one: Investigating the validity of the identified cancer entity specific cell surface markers with a tumor to normal comparison. I therefore analyzed tumor versus normal samples of RNA-Seq data of The Cancer Genome Atlas (TCGA) [17] and I explored the Reverse Phase Protein Array (RPPA) data of The Cancer Proteome Atlas (TCPA) [153]. My findings were possible because of (i) the comprehensive multiomics analysis, and (ii) the data mining of biodata catalogs, both emphasizing the value of integrating existing data sources for phenotype discovery and validation.

The pangenomics part of my thesis investigates two main research questions. The first one addresses shortcomings of existing pangenome graph construction pipelines: They are either (i) reference-biased [94, 77], or (ii) their implementation limits their ease of deployment, optimal use of compute resources, and cluster scalability [60]. To address these issues, I aimed to develop a cluster efficient pipeline for building unbiased reference-free pangenome graphs. This enables researchers to combine several reference genomes into one graphical model that offers the opportunity to study the entire genomic diversity of a population. The second research question arose from the need to understand pangenome graphs: We require efficient algorithms to visualize and analyze them. Therefore, I investigated a new pangenome graph layout algorithm which uses

the biological information in the pangenome graph to guide the layout procedure. To provide the possibility to explore the biology of a pangenome graph, I developed methods to detect complex regions, extract loci, remove artifacts, manipulate structure, add annotation, and perform exploratory analysis. Since pangenome graphs can grow taxing in size and complexity [65], I aimed to parallelize the implemented algorithms. This enables researches to explore the genotype-phenotye relationships encoded in gigabase-scale pangenome graphs not only with respect to one reference sequence source, but with respect to all sequences in the graph.

4 Results And Discussion

In this chapter, I will summarize my research towards computational methods mitigating analysis bias for enhancing GP discovery. First, the main ideas and results of each manuscript will be briefly presented and discussed by topic. Section 4.1 focuses on the multiomics part of my thesis by presenting and discussing my research on how integrative multiomics analysis of cell surface molecules of the NCI-60 tumor cell panel can uncover novel theranostics, combining targeted diagnostics and therapies, for cancer immunotherapy. My pangenomics research is presented in Section 4.2. Here I will present and discuss my research regarding the visualization, reference-unbiased construction, and understanding of pangenome graphs. Both sections will relate the work I have done in the respective fields to each other and to the main goal of computational methods mitigating analysis bias for enhancing GP discovery. An integrated discussion of my research is given in Section 4.3. It connects the different directions I explored provides a joint perspective on my research towards computational methods mitigating analysis bias.

During this chapter, I will switch between the subject pronouns *we* and *I*. If the general work published in manuscripts is presented, I will use the plural pronoun since each manuscript has several authors. If my own contributions and views are presented, I will indicate that by using the singular pronoun.

Some texts and figures of Section 4.1 and Section 4.2 are directly copied or modified from my respective manuscripts which are all licensed under a Creative Commons Attribution 4.0 International License (`http://creativecommons.org/licenses/by/4.0/`), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as one gives appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. I am surrounding copied text with enclosing "". A footnote specifies from which manuscript the cited text originates. The manuscripts themselves are reprinted in the Appendix A.8.

4.1 A Showcase Of Integrative Multiomics Mitigating Analysis Bias

As indicated in the introduction, integrative multiomics reduces analysis bias, because it does not rely on one single data source, but it combines the information of different omics

technologies. This is of the utmost importance for the discovery of new GP relationships, and holds especially true for precision medicine. Hence, the research I conducted for manuscript 1 is focused on showcasing the explorative power of integrative multiomics for GP validation and discovery in cancer immunotherapy.

4.1.1 Multiomics surface receptor profiling of the NCI-60 tumor cell panel uncovers novel theranostics for cancer immunotherapy (Manuscript 1)

Immunotherapy using immune checkpoint inhibitors (ICI) that is based on binding of cell surface receptors with therapeutic antibodies or engineered cells, is one of the most promising and disruptive immunotherapeutical developments in the recent years [118]. However, ICIs are untargeted and can lead to unwanted side effects [86]. Cancer immunotherapy targeting specific cell surface molecules is therefore an important focus of current research. Tumor cell panels, which consist of diverse cancer cell lines, are instrumental in this effort by allowing researchers to study drug responses and molecule-specific interactions, thereby advancing the development of targeted therapies that enhance efficacy while minimizing off-target effects.

In manuscript 1, we comprehensively profiled the surface molecules of the NCI-60 tumor cell panel [132] with a set of 332 arrayed antibodies with FACS. I integrated the results with existing proteomic and transcriptomic datasets [63, 107] via MCIA to identify surface accessible biomarkers of the various tumor entities that furthermore represent diagnostic and therapeutic targets, such as novel theranostics [74, 142] (Figure 4.1). The integrated analysis grouped all three datasets into consistent cell tissue clusters (Figure 4.1), positively validating the general outcome of the FACS experiment. This was also visible in the phylogenetic trees I created from the different data sets [74].

I analyzed the most variant surface profiles: Skin, brain, colon, kidney, and bone marrow. I reported several cancer entity specific cell surface molecules for colon cancer[1]: CD15, CD104, CD324, CD326, CD49f, and for renal cancer: CD24, CD26, CD106, EGFR, β3GalT5[2], SSEA-4, TIM1, and TRA-1-60R. I conducted additional validation using RNA-Seq data from The Cancer Genome Atlas (TCGA) [17] to compare normal versus cancerous samples. Data mining of protein expression tissue of the Human Protein Atlas (HPA) [152] strengthened our hypothesis that indeed *VCAM1* and EGFR are potential markers for renal cancer. Analysis of the Cancer Proteome Atlas [153] indicated that EGFR is specifically associated with the "Kidney Chromophobe" (KIRC) cancer subtype, while no data for *VCAM1* was available in this atlas.

Our study significantly advances the field of immunotherapy by applying a multiomics

[1]The HGNC symbols for some of the genes involved have been updated. Current symbols at the time of writing are: CD15 → *FUT4*, CD104 → *ITGB4*, CD324 → *CDH1*, CD326 → *EPCAM*, CD49f → *ITGA6*, CD26 → *DPP4*, CD106 → *VCAM1*, TIM1 → *HAVCR1*.

[2]biosynthesizes SSEA-3 [21]

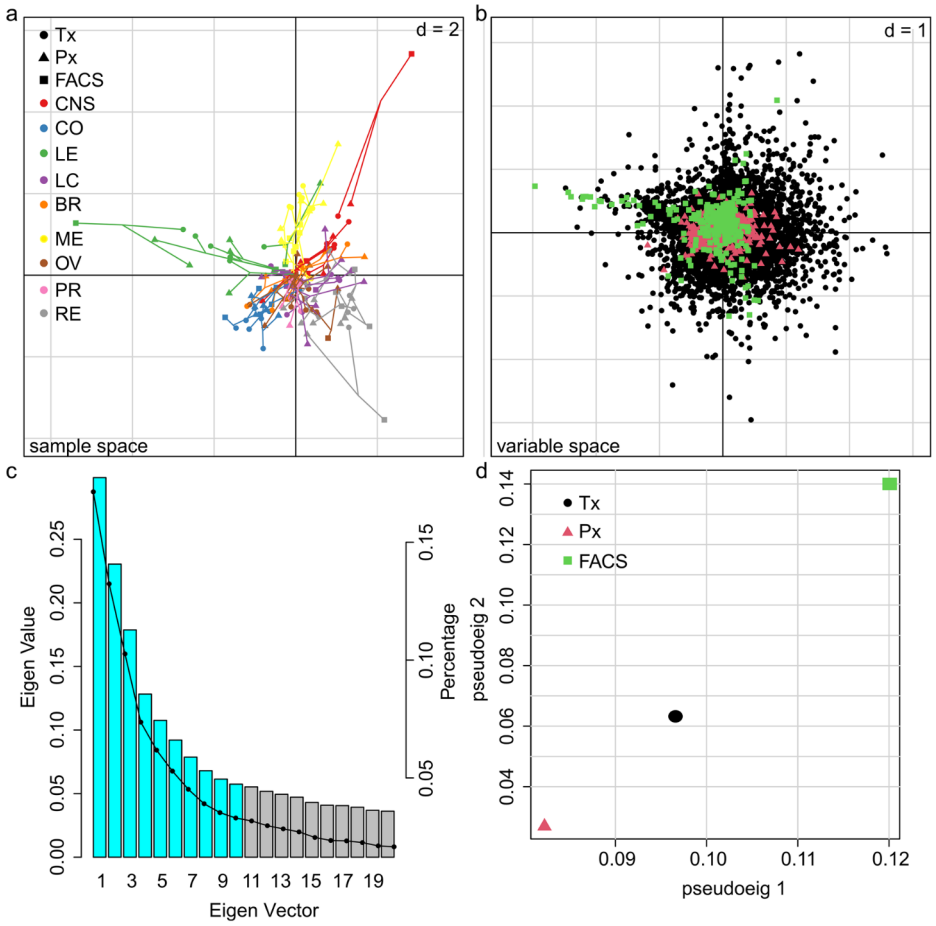

Figure 4.1: MCIA of the NCI-60 panel data. (a) The first two principal components of the MCIA plot show similar trends in Tx, Px, and FACS profiles. The type of shape indicates the respective omics platform. Shapes are connected by lines joining a common point representing the maximized covariance reference structure derived from the MCIA analysis. The length of a line models the divergence between the data from the same tumor cell line. Colors represent the nine NCI-60 different tissues covered by the tumor cell lines. Central nervous system (CNS) and leukemia (LE) cell lines are separated along the first axis (PC1, horizontal). Melanoma (ME) was projected on the positive side of the second axis (PC2, vertical). CO: colon. LC: lung. BR: breast. OV: ovarian. PR: prostate cancer. RE: renal. (b) A tissue specific feature will be projected in the direction of this tissue. The larger the distance from the origin, the more potentially significant a feature is. (c) A scree plot showing the eigenvalues on the y-axis and the number of PCs on the x-axis. Used to rationalize the number of PCs included in the analysis. (d) The pseudo-eigenvalue space of the NCI-60 data sets summarizes the consensus between the platforms, highlighting which omics technique contributes more to the total variance (Tx, black; Px, red; flow cytometry, green). *Figure and caption modified from Heumos et al. 2022* [74].

approach to develop surface-accessible theranostics for tailored, tumor-specific treatments. By utilizing our FACS-expanded multiomics approach, we identified *VCAM1*

and EGFR as key biomarkers that could serve as targets for precise cancer therapies. This approach enables the direct development of targeted treatments specifically aimed at these biomarkers, rather than employing a broad or random approach to gene targeting.

The advantage of using FACS is that it allows for the isolation and characterization of cells based on their surface markers, which helps in developing therapies that are directly relevant to the identified biomarkers. This approach leverages advanced multiomics data and FACS technology to pinpoint specific biomarkers, enabling the development of targeted treatments that are tailored to the unique characteristics of the cancer cells. Our results underscore the importance of integrating data from multiple multiomics sources, demonstrating that this comprehensive analysis offers significant benefits over relying on single data sources, leading to more effective and personalized cancer therapies.

4.1.2 Discussion Of Multiomics Research

With our work published in the article "Multiomics surface receptor profiling of the NCI-60 tumor cell panel uncovers novel theranostics for cancer immunotherapy" (*Heumos et al. 2022* [74]) we showed that by combining the cell surface expression patterns of a newly carried out flow cytometric screen with previously defined Tx and Px datasets of the NCI-60 panel, we were able to identify potential tumor biomarkers for five out of the nine cancer entities.

Since the identification of the found surface molecules is solely based on comparisons within the NCI-60 tumor cell panel, we can't classify them as biomarkers. But by cross-analysis with the TCGA recount2 RNA-Seq data, which contains expression values in healthy and tumor tissue, we were able to narrow down our data revealing tumor specific biomarkers for two tumor entities, i.e., colon and renal cancer. A deeper investigation of the TCGA data revealed that healthy tissue RNA-Seq data for skin, bone marrow, and lymph nodes is missing. Therefore we were not able to validate our potential candidates. Another benefit of the combination of the flow cytometry data with the already available Tx and Px data by MCIA brought us was the strongly enhanced confidence in our initial hits.

Further tumor biomarker candidates were validated using resources HPA and TCPA. While earlier studies, including those summarized in the HPA, identified several promising cancer biomarkers, our research builds on this knowledge by focusing on markers that are accessible on the cell surface. This makes them potential candidates for tumor-specific immunotherapy. In particular, *VCAM1* and EGFR have emerged as promising targets for immunotherapy. Our data show that these biomarkers are highly expressed on the surface of renal cancer cells, suggesting they could be effective targets for more precise and targeted cancer treatments. Recruiting even more omics data from TCPA, we were able to identify a potential tumor specific immunotherapeutic target down to the specific cancer type. This would not have been possible at all, if the additional omics resources would not have been publicly available. However, for e.g. *VCAM1*, no data

was available in the TCPA.

In general, the non-availability of reliable healthy versus tumor data, either on the protein expression or gene expression level, can greatly hinder research. We were not able to answer all open questions our experiments resulted in, due to missing data in existing data sources. This can be a limitation for multiomics studies and any other research work.

While the conducted research lead to the discovery of tumor specific cell surface molecules of 5 cancer entities, there still is the open question why we were not able to find specific cell surface marker expression patterns for lung cancer, breast cancer, ovarian cancer, and prostate cancer. We hypothesize this might be due to the large heterogeneity of these tumor types in our data. Another reason could be that the signals of the 5 cancer entities is so strong that it outshines the ones in the other 4. A possible follow up analysis would be to repeat the MCIA with only these 4 cancer cell lines, gaining deeper knowledge of their cell surface molecular profile.

The results of the first published manuscript demonstrate that one omics data type is not enough for the detection and validation of biomarkers, and phenotype discovery. Some omics methods employed still rely on one single reference genome (e.g. Microarray or RNA-Seq). This can introduce bias in downstream analysis. In the following I will present the results of the second research direction that I explored towards unbiased graphical pangenomics analysis.

4.2 Towards Unbiased Graphical Pangenomics Analysis

Reference-bias is the systematic error that can occur when using a single linear reference genome for bioinformatics analysis. Pangenome graphs try to alleviate this issue. In the following sections, I present my research that was published in the manuscripts 2, 3, and 4. The manuscripts share the same research question. How to leverage the full potential of a pangenomics data set exploring novel, qualitatively different computational methods and paradigms? My research focuses on new bioinformatics techniques for pangenome graphs: human readable low-dimensional layout (manuscript 2), reference-unbiased construction (manuscript 3), and understanding (manuscript 4).

4.2.1 Pangenome Graph Layout By Path-Guided Stochastic Gradient Descent (Manuscript 2)

Since pangenome graphs can grow excessive in size on disk and in RAM (tens to hundreds of gigabytes), a big challenge is to provide scalable and interactive visualizations to better understand and analyze them. The basis of such a visualization can be a human readable graph layout: A low dimensional graph embedding in 1D or 2D. This would allow scientists to (interactively) explore the genotype and, by adding annotation, the phenotype of pangenome graphs in detail.

A graph layout arranges nodes and edges in an N-dimensional space to minimize overlaps, reduce edge crossings, and enhance clarity. Force-directed graph drawing, a common approach [19], uses physical simulations with repulsive and attractive forces to create visually appealing layouts. However, this method can get stuck in local minima. Multi-layer strategies like the Fast Multipole Multilevel Method (FM^3) [69] and Stochastic Gradient Descent (SGD) [174] address this. **"**However, *Zheng et al. 2019*'s SGD algorithm has a quadratic up front cost in the number of nodes to find pairwise distances to guide the layout, making it impossible to apply to pangenome graphs with millions of nodes. Also, existing generic graph layout approaches ignore the biological information inherent in pangenome graphs.**"**[3]

These issues are addressed in manuscript 2: We developed a new pangenome graph layout algorithm: the Path-Guided Stochastic Gradient Descent (PG-SGD) [76] which uses the biological information in the pangenome graph, the genomes encoded as paths as an embedded positional system in the graph, to sample genomic distances between pairs of nodes. This prevents the quadratic computational time of previous implementations of graph drawing by Stochastic Gradient Descent (SGD) [174].

Inspired by *Zheng et al. 2019* [174] our algorithm moves a randomly selected pair of nodes (v_i, v_j) at a time, minimizing the disparity between the layout distance of a node pair and their actual nucleotide distance in the genome. In a 2D layout, nodes have two ends. For each movement of a node pair, we move one end of each node (Figure 4.2).

[3] From *Heumos et al. 2024(b)* [76].

The first node v_i of a pair is a path step s_i uniformly sampled from all steps of $\mathcal{P}$. The second node v_j of a pair is a path step s_j sampled from the same path of s_i by drawing a uniform or Zipfian distribution [177] with equal chance. This balances the global and local layout updates: The Zipf sampling raises the chance that s_i and s_j are close in nucleotide space, refining the layout of nodes a few base pairs apart. The uniform sampling captures long-range distances leading to a globally linear layout.

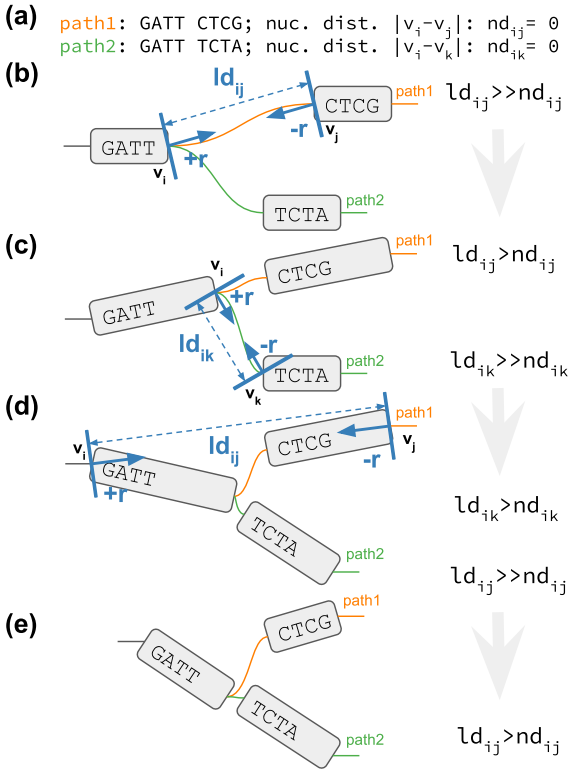

Figure 4.2: 2D PG-SGD update operation sketches. **(a)** The path information of the graph. *path1* and *path2* both visit the same first node. Then their sequence diverges and they visit distinct nodes. **(b-e)** v_i/v_j or v_i/v_k is the current pair of nodes to update. ld_{ij}/ld_{ik} is the current layout distance. $+r, -r$ is the current size of the update. **(b)** Initial graph layout highlighting the future update of the two nodes of *path1*. **(c)** The graph layout after the first update. The nodes appear longer now, because we updated at the end of the nodes. Highlighted is the future update of the two nodes of *path2*. **(d)** The graph layout after the second update. Highlighted is the future update of the two nodes of *path1*. **(e)** Final graph layout after three updates using the 2D PG-SGD. Figure and caption modified from *Heumos et al. 2024(b)* [76].

We implemented the 1D and 2D PG-SGD algorithms in the ODGI [65] toolkit. Utilizing a modified HOGWILD! [116] strategy, the multithreaded implementation uses shared memory for storing the layout. Each layout update of each thread occurs without any thread locking for unrestricted parallel computing and speed optimization.

I applied 2D PG-SGD to each of the chromosomal pangenome graphs of the HPRC

[100]. Each graph consist of 90 whole human haplotypes. On average, a layout was calculated within 50 minutes. The average memory consumption was 29.66 GB RAM. BandageNG, the current state-of-the-art for graph visualization, required 8 times more time while using 2 times more memory to process the same data. To benchmark scalability, I applied the 2D PG-SGD to the full graph with all chromosomes together. Here, BandageNG was unable to produce any layout within 1 week while 2D PG-SGD took around 1 day. I generated 2D visualizations of all the HPRC chromosomal pangenome graphs revealing biological features of the different parts of the chromosome 6 HPRC graph. In Figure 4.3, I show how the 2D PG-SGD's layout reveals biological features of the different parts of chromosome 6 HPRC graph, zooming in on the MHC region.

The 1D and 2D layouts produced by PG-SGD offer an unprecedented high-level view on pangenome graphs. The algorithm can be extended to any number of dimensions. This might open doors to detect and classify variants in the future. The algorithm's 2D layout already is the foundation of upcoming pangenome graph browsers [48, 49] which allow the incorporation of annotation. The PG-SGD layout techniques have been applied to construct (Sections 2.4, 4.2.2) and analyze (Section 4.2.3) the first draft human pangenome reference [100], as well as exploring the heterologous recombination of the human acrocentric chromosomes [66].

Furthermore, they are applied in the understanding of pangenome graphs [65], see Section 4.2.3. Of note, the 1D PG-SGD algorithm is the key step for the construction of pangenome graphs [60, 75], specifically in the pipeline presented in the following Section 4.2.2: PG-SGD orders the nodes according to their occurrence in the graph's embedded paths.

4.2.2 Cluster Efficient Pangenome Graph Construction With nf-core/pangenome (Manuscript 3)

Current pangenome graph construction methods often introduce biases, excluding complex sequences or rely on references. The PGGB pipeline [60] addresses such limitations by iteratively polishing an all-vs-all whole-genome alignment graph. However, PGGB's bash implementation limits its ease of deployment, optimal use of compute resources, and cluster scalability.

With manuscript 3, we aimed to compensate for that: We developed nf-core/pangenome [75], a reference-unbiased approach to construct pangenome graphs (Figure 4.4). Mirroring PGGB, nf-core/pangenome is implemented in Nextflow's [36] latest domain-specific language (DSL2) syntax and follows the nf-core [45] best practice development guidelines. Providing all software dependencies in biocontainers [33] makes the pipeline portable and easy to install on HPC environments.

Unlike PGGB, nf-core/pangenome distributes the quadratic all-to-all base-level alignments across cluster nodes by dividing the approximate alignments into equally sized tasks. I benchmarked the time spent on base-pair level alignments and demonstrated

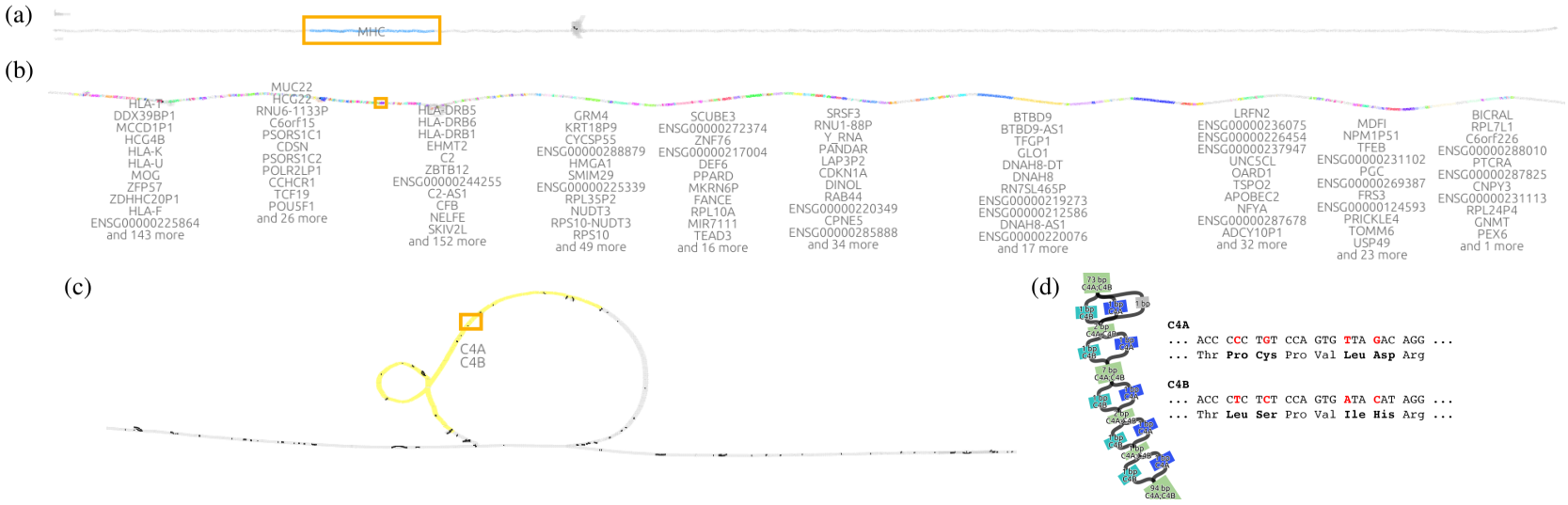

Figure 4.3: 2D visualizations of of the 90 haplotypes chromosomes 6 pangenome graph of the Human Pangenome Reference Consortium (HPRC), the major histocompatibility complex (MHC), and the complement component 4 (C4). **(a)** *gfaestus* screenshot of the chromosome 6 layout. Colored in blue and highlighted in orange is the MHC. The hairball in the middle is the centromere. The black structures in the centromere are edges. **(b)** *gfaestus* screenshot of the MHC. All MHC genes are color annotated and the names of the genes appear as a text overlay. The C4 gene region is highlighted in orange. **(c)** *gfaestus* screenshot of the region around C4, specifically color highlighting genes C4A and C4B. The black lines are the edges of the graph. **(d)** Additionally annotated Bandage layout of the C4 region where C4A and C4B genes differ due to single nucleotide variants leading to changes in the encoded protein sequences. Node labels were annotated by using *odgi position*. Figures and caption **(a)-(c)** modified from *Heumos et al. 2024(b)* [76], and figure and caption **(d)** integrated from Figure 2 from *Guarracino, Heumos et al. 2022* [65].

that it decreases linearly as the number of alignment problem chunks increases. To estimate the carbon dioxide equivalent (CO2e) emissions, I employed the nf-co2footprint Nextflow plugin [88] for all reported results.

As a demonstration case, I built a pangenome graph of 1000 sequences of chromosome 19 of the 1000 Genomes Project (1KGP) [41]. While PGGB finished after 7 days, nf-core/pangenome took 3 days. To assess the scalability of the pipeline, I constructed a pangenome graph of 2146 *E. coli* sequences. Because of wall clock time restrictions on our cluster, PGGB did not finish after 30 days. nf-core/pangenome built the graph within 10 days. Both workflow scalability evaluations showed that nf-core/pangenome is at least two times faster than PGGB while not increasing greenhouse gas emissions.

nf-core/pangenome offers a scalable way to build a comprehensive genotype resource, but it remains unclear what implications this brings for the corresponding phenotypes, and how one can understand the biology encoded in a pangenome graph. The next section describes a tool that was specifically designed for the analysis and interpretation for pangenome variation graphs.

4.2.3 Understanding Pangenome Graphs With ODGI (Manuscript 4)

Analyzing the genotypes of hundreds of gigabase-scale genomes using pangenome graphs is a major challenge. Highly repetitive regions (centromeres, segmental duplications, and acrocentric chromosomes) increase the complexity of the operations performed on graphs. Hence, fast and versatile software that can deal with the sheer size and complexity of such graphs is required. However, this is not well supported by existing tools like e.g., VG [59] or gfatools [94].

In manuscript 4 we present a newly implemented toolset named Optimized Dynamic Genome/Graph Implementation (ODGI) [65]. ODGI implements an efficient variation graph structure in computer memory that can be dynamically updated using multiple CPU cores in parallel. "ODGI includes tools for detecting complex regions, extracting pangenomic loci, removing artifacts, exploratory analysis, manipulation, validation and visualization. Its fast parallel execution facilitates routine pangenomic tasks, as well as pipelines that can quickly answer complex biological questions of gigabase-scale pangenome graphs."[4] ODGI can handle pangenome graphs in the Graphical Fragment Assembly (GFAv1) format. Most of the tools are implemented to be applied together by piping the output from one tool into the next. Currently, ODGI includes more than 40 tools. As a demonstration case (Figure 4.5), I analyzed the metrics of a 90-haplotype human pangenome graph of the exon 1 huntingtin gene (graph name: *HTTexon1*) highlighting some of ODGI's key features:

- **odgi viz, layout & draw**: Pangenome visualization provides convenient insight

[4] From *Guarracino, Heumos et al. 2022* [65].

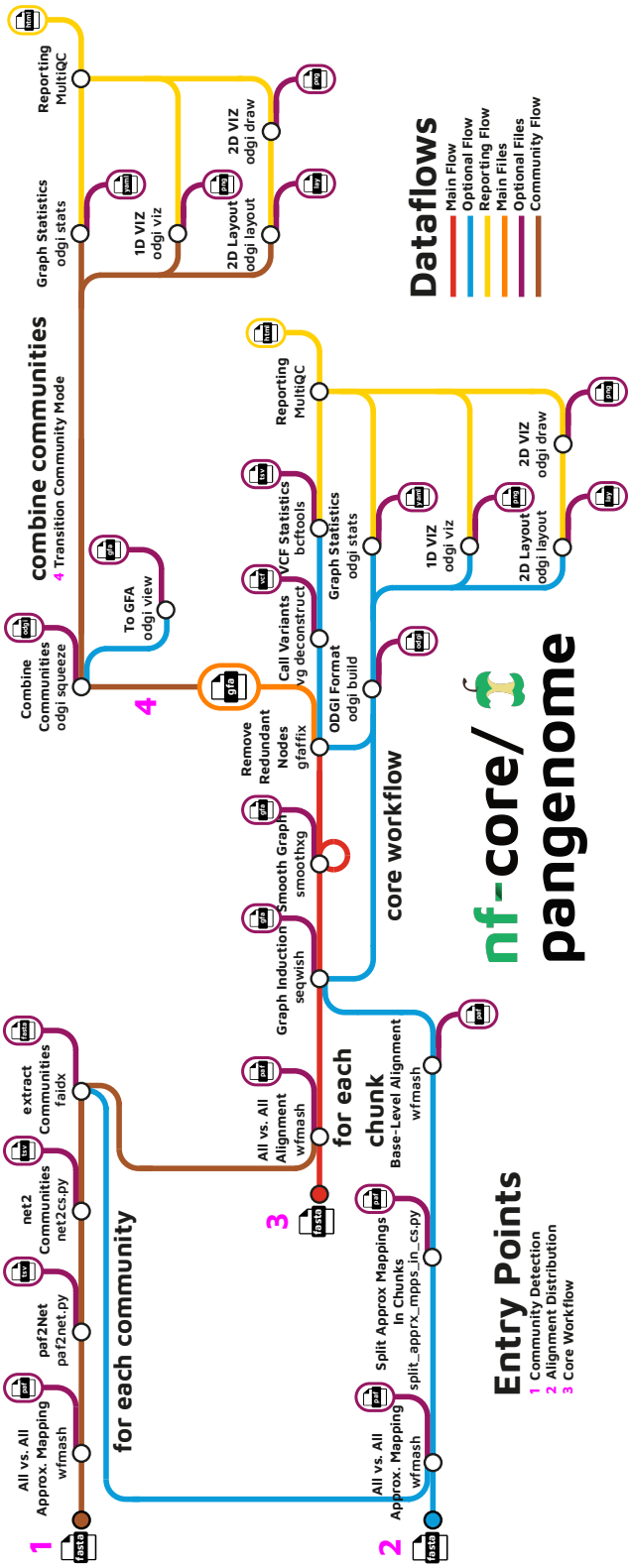

Figure 4.4: Schematic representation of the nf-core/pangenome workflow processes and detailed analysis steps. The input consists of one FASTA file containing all sequences. The pipeline comes with 3 major entry points: Community detection (1), alignment distribution (2), and core workflow (3). Optional community detection (1) is performed on the input sequences. If selected, the heavy all-to-all baise-pair level alignments (2) can be split into problems of equal size. nf-core/pangenome's core workflow (3) is a direct mirror of PGGB. If running in community mode, all communal graphs are combined into one (4) and the subsequent quality control subworkflow is executed. The output is a pangenome graph in GFA format. Figure and caption modified from *Heumos et al. 2024(a)* [75].

into genomic variation. *odgi viz* generates a linearized representation of the pangenome (Figure 4.5d) and is capable of handling full length human chromosomes. *odgi layout* and *odgi draw* extend the visualization in 2D (Figure 4.3).

- **odgi stats, depth & degree**: Graphs statistics provide alternative ways to gain insight into pangenomes complexity. *odgi stats* returns the number of nodes, edges, paths, and graph length which can be interactively explored by the MultiQC [44] ODGI module I implemented (Figure 4.5a). *odgi degree* and *odgi depth* compute node degree and depth as defined by user-provided criteria (Figures 4.5b,4.5c). These methods allow the detection of complex regions generated by highly repetitive sequences (Figure 4.5).

- **odgi squeeze & extract**: Pangenomes can be constructed chromosome-wise. *odgi squeeze* merges multiple graphs into the same file whilst preventing node ID collisions (Figure 4.4). *odgi extract* extracts regions of the graph as defined by certain criteria, allowing downstream processing of smaller subgraphs (Figure 4.5).

- **odgi position**: Pangenome graphs are flexible when it comes to coordinate systems. *odgi position* can use the coordinate system from a contained reference genome to display coordinates and other localized features. Given annotation, I implemented a way so that the positions of the annotation can be projected onto the nodes of the graph (Figure 4.3). This serves as a key bridge between genotype and phenotype in pangenomes.

I evaluated the performance for routine pangenomic tasks of ODGI and VG. I measured the execution time and memory usage of (i) transforming a GFAv1 file into a tool's native format, (ii) the extraction of a subgraph, (iii) the visualization of a pangenome graph and, (iv) the finding of path positions in a pangenome graph. Generalizing across all graph key operation evaluations, ODGI makes comparatively better use of multithreading and requires much less memory. Working with very complex graphs, for example extracting the centromeric subgraph of chromosome 6, ODGI is up to 40 times faster and requires 8 times less memory than VG. With ODGI we implemented a state-of-the-art tool suite that can transform, analyze and visualize pangenome graphs at large scale. Lifting over annotations, and detecting complex graph structures place the suite as the bridge between traditional linear reference genome analysis and pangenome graphs. This makes ODGI an inevitable tool not only for the detailed exploration of the genotype of a pangenome graph, but also for it's linking with annotation and other metadata for the enhanced phenotype discovery.

Already now, the tools are the backbone of my pipelines such as PGGB (Section 2.4) or nf-core/pangenome (Section 4.2.2). Since ODGI's sequence model is alphabet agnostic, it was already applied to RNA and protein [34] sequences. This makes it well prepared for potential multiomics analysis: Exploring biology from different angles using different high-throughput technologies.

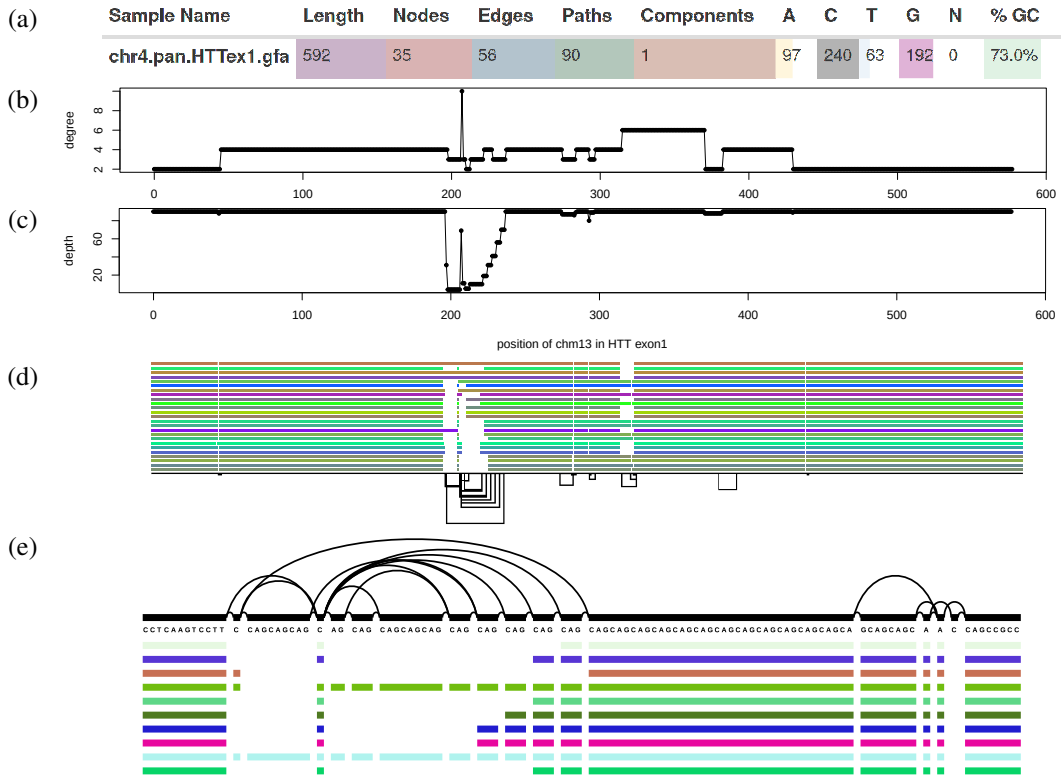

Figure 4.5: Features of a 90-haplotype human pangenome graph of the exon 1 huntingtin gene (graph name: *HTTexon1*): **(a)** Excerpt of vital statistics of the *HTTexon1* graph displayed by MultiQC's ODGI module. **(b)** Per nucleotide node degree distribution of CHM13 in the *HTTexon1* graph. Around position 200 there is a huge variation in node degree. **(c)** Per nucleotide node depth distribution of CHM13 in the *HTTexon1* graph. The alternating depth around position 200 indicates polymorphic variation complementing the above node degree analysis. **(d)** *odgi viz* visualization of the 23 largest gene alleles, CHM13, and GRCh38 of the *HTTexon1* graph. The 1D visualization is a binned binary matrix, where the graph is ordered in 1D across the horizontal axis, with each path represented by a row of the vertical axis. For each path, graph nodes are arranged from left to right, with the colored bars indicating the paths and the nodes they cross. White spaces indicate where paths do not traverse the nodes. Directly consecutive nodes are displayed with no white space between the two. **(e)** *vg viz* nucleotide-level visualization of 10 gene alleles, CHM13, GRCH38 of the *HTTexon1* graph focusing on the CAG variable repeat region. Figure and caption are modified from *Guarracino et al. 2022* [65].

4.2.4 Discussion Of Pangenomics Research

With our work published in the article "Pangenome graph layout by Path-Guided Stochastic Gradient Descent" (*Heumos et al. 2024(b)* [76]), we presented the first layout algorithm for pangenome graphs that leverages the biological information available within the genomes (and therefore genotypes) represented in the graph. Other generic graph layout algorithms, such as the one offered by BandageNG, ignore this additional information. Our implementation efficiently computes the layout of pangenome graphs representing thousands of whole genomes. Our method can be combined with annotation, providing layouts for browsers, and ultimately enables scientists to explore the GP relationship of a collection of sequences in a reference-unbiased manner.

Graph visualization is key for understanding the genotypes inherent of a pangenome graphs, the genome variations. "The layouts produced by PG-SGD offer an unprecedented high-level perspective on pangenome variation. We implemented PG-SGD to generate layouts in 1D and 2D. These graph projections have already been employed in constructing and analyzing the first draft human pangenome reference [100], as well as in the discovery of heterologous recombination of human acrocentric chromosomes [66]. Furthermore, they are applied in the creation and analysis of pangenome graphs for any species [65, 60] (Sections 2.4, 4.2.2). Of note, there still remains a gap in interactive and scalable solutions that merge layouts of large pangenome graphs with annotation. Our algorithm will underpin new pangenome graph browsers for studying graph layouts, the genome variation they represent [49]."[5] Adding annotation or other metadata will shift the analysis frame from a genotype-only to a GP one.

The performance analysis demonstrates that our 2D implementation surpasses BandageNG when handling large and complex pangenome graphs. As especially observed with the chromosome 16 evaluation, our implementation's speed is memory-bound. The current data structure doesn't consider the memory access pattern. Therefore, a CPU cache optimized implementation could make use of the massive parallelized calculations on a GPU. Such an effort is currently in progress [97].

"The classical force-directed layout methods of state of the art generic graph algorithms, such as FM³-based ones, places nodes according to their attractive and repulsive forces. This force can be seen as equivalent to how our 2D PG-SGD moves the nodes' ends in 2D: If the nucleotide distance of the randomly chosen path steps is smaller than the layout distance of the nodes' ends, we move them closer together ("attractive force"), else we move them further away ("repulsive force"). However, the key difference here is that this approach is path-guided: Paths represent biological sequences in pangenome graphs, so it is as if PG-SDG considers a "biological force" for placing the graph nodes. Theoretically, it would be possible to combine our approach with a force-directed one. Combining both methods, we might get the best of both worlds: multi-threadable PG-

[5] From *Heumos et al. 2024(b)* [76].

SGD iteratively applied to different graph-layout-levels. We can imagine that such an approach can lead to a further speedup when calculating the layout. However, for generic graphs, this would only work if path information for each node could be added: We would replace the classical physical simulation approach with our path-guided method. If such information is not available, one could randomly cover the graph with paths. This function is already provided in *odgi cover*. However, this is an NP-hard problem and our preliminary solutions proved inefficient.

With assembly graphs we face the same problem: they usually do not carry path information during each assembly step. One could map the initial assembly reads back against the assembly graph in order to build paths through the graph. This would allow us to obtain a layout using PG-SGD.

PG-SGD can be extended to any number of dimensions. It can be seen as a graph embedding algorithm that converts high-dimensional, sparse pangenome graphs into low-dimensional, dense, and continuous vector spaces, while preserving biologically relevant information. This enables the application of machine learning algorithms that use the graph layout for variant detection and classification. Our future research involves leveraging these graph projections to detect structural variants and to identify and correct assembly errors."[6]

Since the algorithm is sequence agnostic, we expect it works with RNA and protein sequences to support pantranscriptome graphs [133] and panproteome graphs [34]. This makes PG-SGD applicable in a multiomics setting. While this manuscript mostly focuses on the exploration of the genotypes, the algorithm introduced is the foundation for comparative genomics and consequently GP discovery. 1D PG-SGD is a key a step when building pangenome graphs.

With our work published in the article "Cluster efficient pangenome graph construction with nf-core/pangenome" (*Heumos et al. 2024(a)* [75]), we presented an easy-to-install, portable, and cluster-scalable pipeline for the reference-unbiased construction of pangenome variation graphs. We showed how the all-vs-all base pair level alignment process can be distributed across nodes of a cluster reducing the wall clock time of constructing a pangenome graph.

The core workflow steps of nf-core/pangenome have been successfully applied to *Neisseria meningitidis* [171], wild grapes [28], humans [66, 100], grapevines [68], taurines [109], chicken [119], and rats [160]. This highlights the community's efforts to establish a best-practice workflow for generating reference-unbiased and sequence-complete pangenome graphs. "The modular domain-specific language (DSL) 2 pipeline structure eases the exchange of key processes with alternative tools, the extent of the pipeline with new tools, and the integration of parts of the pipeline with other (sub-)workflows.

[6] From *Heumos et al. 2024(b)* [76].

We have shown that we are able to perform all-vs-all base pair level alignments of thousands of sequences. When executed on an HPC, nf-core/pangenome's parallel workflow accelerates graph construction compared to PGGB. PGGB's inability to assign individual computational resources to each pipeline step leads to the allocation of one whole node of an HPC, despite the fact that some processes can only make use of one thread. This blocks valuable CPU cycles for other users working on the same HPC and ultimately can lead to additional costs. In contrast, nf-core/pangenome does not have such limitations: Nextflow's process management enables the optimal workload of given compute resources which can be especially important when running a pipeline in commercial clouds.

Competing pipelines don't use any workflow management system to connect their processes [23], or their workflow language of choice is e.g. Toil [161, 77] which makes them less user-friendly, less cluster efficient, and less portable [168]. nf-core/pangenome is currently the only pangenomics pipeline that is optionally monitoring its CO_2 footprint. The measurements have shown that constructing extensive pangenome graphs, such as the 2146 *E. coli* graph, requires a considerable amount of energy. Therefore, before executing environmentally questionable experiments, we would recommend thoroughly assessing both the rationale and the methodology.

Although we expect our pipeline to scale well for future pangenome graph construction challenges, such as for the next HPRC phase which targets 350 individuals, there still is potential for further optimization: We consider using the IMplicit Pangenome Graph (IMPG) [57], a tool that extracts homologous loci from all genomes mapped to a specific target region. This would allow us to break the whole genome multiple alignments into smaller pieces, construct a pangenome graph for each piece, and lace these together into a full graph with gfalace [56].

We anticipate the pipeline, or its parts, will enhance current single linear reference analysis methods to explore whole population variation instead of focusing on one reference only. Looking ahead, pangenome construction pipelines like nf-core/pangenome will play a pivotal role in studying entire populations, single-cell whole genome sequencing analysis, and constructing personalized (medical) pangenome references [136] shining a light on previously unexplored GP relationships. Once we have built a graph, we want to understand its biology."[7]

In our the article "ODGI: understanding pangenome graphs" (*Guarracino, Heumos et al. 2022* [65]), we introduce ODGI, a novel suite of tools that implements scalable algorithms for detecting complex regions, extracting pangenomic loci, removing artifacts, exploratory analysis, manipulation, validation and visualization. It serves as a bridge between genotype and phenotype and works with reference-centric as well as reference-unbiased pangenome graphs.

[7] From *Heumos et al. 2024(a)* [75].

"Pangenome graphs stand to become a ubiquitous model in genomics thanks to their ability to represent any genetic variant without being affected by reference bias [43]. However, despite this great potential, their spread is impeded by the lack of tools capable of managing and analyzing pangenome graphs easily and efficiently.

By providing a set of standard analysis "verbs" to interact with pangenome graphs, ODGI enables users to explore and discover important biological features captured in this flexible, inclusive model. It provides tools to easily transform, analyze, simplify, validate, and visualize pangenome graphs at large scale. In particular, lifting over annotations and linearizing nested graph structures place the suite as the bridge between traditional linear reference genome analysis and pangenome graphs. With the increased adoption of long read sequencing we expect pangenomic tools to become increasingly common in the genomic studies at different taxonomic levels and in biomedical research. This progression is already afoot, particularly for targets that involve complex variation, such as cancer [30], plant pangenomics [7, 103, 114, 95, 8], and metagenomics [175]. Also, when studying animals like bovines [91, 144, 14].

Currently, bacterial pangenomes are best handled by specialized tools like PPanGGolin [62], PanGraph [111] or PanX [38]. The latter one doesn't build a graphical representation of a pangenome. But, it already has a very developed eco-system, which allows a detailed analysis of bacterial pangenomes using an interactive GUI. Unlike these approaches, which provide a monolithic, integrated solution to understanding pangenomes, ODGI is designed as a low-level toolkit that can work on a generic pangenome graph model frequently used by other existing methods. We hope that this design renders it useful to pangenome analysis pipeline authors. Other pangenome analysis platforms, like PanTools [129] provide access to pangenome analyses at the scales we demonstrate with ODGI, but use specialized de Bruijn graph models to achieve this. In contrast ODGI supports the highly generic variation graph model, which has greater representational power than de Bruijn graphs.

ODGI will facilitate disentangling, describing and analyzing a much larger set of variation than previously was possible with tools that depend on short reads and reference genomes. Furthermore, users can even consider ODGI as a framework, taking advantage of its algorithms to develop new and more advanced tools that work on pangenome graphs, thus expanding the type of possible pangenomic analyses available to the scientific community.

The performance analysis shows that ODGI outperforms VG when handling large, complex pangenome graphs. Across the evaluation of key graph operations, ODGI's memory peak was 10GB. This makes it perfectly suited to be run interactively on a recent laptop. We expect that ODGI will be able to handle the next phase of the HPRC, a pangenome graph constructed from 350 individuals, without any problems.

While ODGI does not construct graphs from scratch nor is it capable of extending them, it is already the backbone of the Pangenome Graph Builder pipeline [60] and nf-core/pangenome [75]. Its static, large-scale 1D and 2D visualizations of the pangenome graphs allow an unprecedented high-level perspective on variation in pangenomes, and

have also been critical in the development of pangenome graph building methods. However, an interactive solution that combines the 1D and 2D layout of a graph with annotation and read mapping information across different zoom levels is still missing. Recent interactive pangenome graph browsers are reference-centric [12, 172], have a limited predefined coordinate system [40], or focus primarily on 2D representations [166, 64]. Our graph sorting and layout algorithms can provide the foundation for future tools of this type. We plan to focus on using these learned models to detect structural variation and assembly errors.

ODGI has allowed us to explore *context mapping* deconvolution of pangenome graph structures via the path jaccard metric. This resolves a major conceptual issue that has strongly guided existing algorithms to construct pangenome graphs. Previously, great efforts have been made to prevent the "collapse" of non-orthologous sequences in the graph topology itself [94]. This has been seen as essential to making these new bioinformatic models interpretable. While our presentation is primarily qualitative, our work demonstrates that we can mitigate this issue by exploiting the pangenome graph not as a static reference, but as a dynamic model of the mutual alignment of many genomic sequences. Because pangenome graphs can contain complete genomes, we are able to query them to summarize the information they contain in easily-interpretable and reusable pairwise formats that are widely supported in bioinformatics. ODGI also projects variation graphs into vector and matrix representations that allow the direct application of machine learning and statistical models to the pangenome. We expect that ODGI will provide a reference interface between pangenomic and genomic approaches for understanding genome variation."[8]

The results of this manuscript showed that one main claim of this thesis holds true: Analyzing reference-unbiased pangenome graphs gives access to GP relationships which would not have been possible when using a single genome as a basis for biomedical analyses.

4.3 Integrated Discussion

The common theme of my research was to explore computational analysis methods mitigating analysis bias for enhancing genotype-phenotype discovery. Exploring the genotype-phenotype relationships in complex biological settings, in particular in healthcare, depends on the utilization of high-dimensional and multi-modal data generated by individual omics technologies. In biological data analyses, if only one data source is available, there are significant limitations. Experts can analyze data extensively, but a single data source can never model complex molecular mechanisms well enough to answer complex questions comprehensively. Multifaceted data analysis has a huge potential to revolutionize the field by integrating diverse biological data sources, enabling a

[8] From *Guarracino, Heumos et al. 2022* [65].

comprehensive understanding of complex molecular mechanisms, and driving advancements in research, in healthcare, and targeted therapies. I argue that the integration of biological data from different origins can act as a bridge to unite experts from diverse research backgrounds, with the potential to positively impact the interdisciplinary research for more precise and informed conclusions in both research and clinical contexts. I approached the viability of computational methods mitigating analysis bias from two different research directions.

The first part of my research exemplified how multiomics approaches integrate diverse data types of various molecular layers, including DNA, RNA, proteins, and cell surface molecules, each contributing to unique and comprehensive insights when understanding complex cancerous processes, ultimately mitigating analysis bias. However, in multiomics research, bias can occur at various stages throughout any research process due to the inherent subjectivity and decision-making of the individuals involved: (i) Selection bias in data acquisition: In genomic studies, improper selection of samples or data points can impact the generalization of the genetic results [46]. (ii) Measurement bias in data generation: The instruments or measurement techniques are inconsistent. For example, in transcriptome studies, varying RNA-Seq protocols can introduce bias [173]. (iii) Analytical bias in data analysis: Wrongly selecting statistical methods like normalization methods for proteomics studies can lead to misinterpretation of protein abundance levels [35].

Despite these biases, the core strength of multiomics lies in its ability to explore and validate molecular mechanisms from multiple perspectives. Its real power, however, emerges from integrating various biological layers, each representing distinct stages of regulation and control. This approach uncovers complex interactions and dynamic networks that cannot be fully understood through a single perspective, reflecting how evolution has leveraged multiple levels of biological information to fine-tune adaptation and specialization. These layers, such as the genome or proteome, are examples of how diverse processes contribute to the overall functionality of organisms. However, this does not mean multiomics is free of limitations.

One critical challenge in multiomics research is the biological variation across different omics layers. Even when measuring the same cell line under identical conditions, results can vary widely between transcriptomics and proteomics data [102]. This biological noise, when integrating both data types, can compound and partially explain the poor correlation often observed between transcriptomics and proteomics measurements. Furthermore, in dynamic systems such as perturbation experiments, there is often a time lag between changes at the transcript level and those observed at the protein level, further complicating data integration. Integrating single-cell RNA sequencing (scRNA-seq) with other omics layers, such as proteomics, offers a promising solution to these challenges by providing high-resolution, cell-type-specific data that reduces biological variability and improves the alignment between transcriptomic and proteomic profiles [140]. This approach allows researchers to more accurately capture the dynamic changes and cellu-

lar heterogeneity, which can mitigate some of the noise and time lag issues inherent in bulk measurements [169].

What I didn't show in my research is that the algorithm choice influences the results and their interpretation. Staying in the algorithm class of dimensionality reduction techniques, an alternative to the applied MCIA is Multi-Omics Factor Analysis 2 (MOFA2) [3] which, in contrast to MCIA, can actually work with missing values by imputation. On the other hand, MOFA2 models Gaussian noise and forces sparsity constraints on latent factors. This makes MOFA2 potentially more powerful, but interpretation harder when analyzing the complex biological interactions of a multiomics data set justifying my choice for using MCIA.

Further algorithmic approaches like data integration algorithms (e.g iCluster [130], Similarity Network Fusion [22]), bayesian approaches (e.g. BNfinder [50]), or machine learning approaches (e.g. CustOmics [10]) come with a comparatively high computational demand and the results are difficult to interpret. In the end, the algorithmic approach must be chosen experiment by experiment. The interpretation potential, computational intensity, availability of training data, and of course the biological question provide guidance here. I opted for MCIA in this study, as it provides a more straightforward interpretation of the biological signals and interactions, which was crucial for our experimental goals. Additionally, MCIA's ability to handle multiple omics layers in a balanced way made it a good fit for identifying novel biomarkers, which we subsequently confirmed with other omics technologies. For a more comprehensive understanding, it would be beneficial to apply some of the aformentioned multiomics methods, where applicable.

It's important to recognize the critical role of data integrity in biological and medical sciences. As was shown, the resources of current public data bases like TCGA are incomplete when it comes to the comparison of tumor versus healthy tissue. Another factor for multiomics analysis is data quality. Low data quality can negatively impact analysis results. Based on my research findings, I strongly support the current movement in the field towards prioritizing data quality as a key focus. But, scientific data should not only be of high quality, but findable, accessible, interoperable, and reusable (FAIR) [167]. Maybe there already exists some protein tissue data for *VCAM1*, but it is just not findable or accessible. The UK biobank is a great step into the right direction. For a sustainable GP discovery, it is inevitable that every scientist lives up to the paradigm *open science with open data*.

What I have discussed so far was about tackling analysis bias utilizing different omics technologies. However, most nucleic acid omics technologies suffer from reference bias. This motivated the second part of the research I conducted towards the overarching goal of my thesis: pangenomics research.

Progress in pangenomics was hindered by selection bias in genomic analyses. Genome inference technologies, like genotype arrays or NGS, were limited to easily detectable

variants leading to a skewed understanding of genomic diversity. Specific variants in more repetitive or complex regions of the genome are not observable with such technologies. Consequently, the identifiable variants are variants that would have been least likely affected by reference bias in the first place.

The situation has changed. Thanks to recent advances in long read sequencing technology, we can now take a look at the *dark parts of a genome*. The most prominent example is the T2T Consortiums's first complete human genome assembly [112], complemented with a fully assembled chromosome Y [117] and entirely assembled human centromeres [104]. This effort is amplified by the Human Genome Structural Variation Consortium (HGSVC) cataloging the structural variants of human. Bringing it all together, the HPRC has recently released the first draft human pangenome reference [100] exemplifying how scientist can tackle reference bias *in vitro* and *in silico* for an advanced genotype representation.

The pangenomic-centric aspect of my research presented in this dissertation aligns with the second phase of pangenomics, which focuses on leveraging recent technological advances to achieve a more comprehensive understanding of genomic diversity. While my work extends this phase slightly, it contributes valuable insights into further enhancing genomic interpretation.

In manuscript 2 I led implementation efforts to project high-dimensional graphs into a low-dimensional (human readable) space which helped to build and visualize a pangenome graph of the HPRC. Manuscript 3 details an approach to build pangenome graphs efficiently on a cluster. Building up on the core algorithm to construct the pangenome graph for the first phase of the HPRC, the pipeline would be a perfect candidate to build the next iteration of a draft human pangenome reference. Finally, in manuscript 4, I contributed toward novel algorithms developed in graphical pangenomics: ODGI, the Swiss Army knife to understand pangenomes bridging pangenomics and traditional genomics which allowed us to deeply explore the biology of the HPRC pangenome graph.

In general, the tools I contributed to and I developed have achieved considerable prominence within the scientific community [151]. However, there are also some currently unresolved challenges ahead. What I often observed during my PhD is that scientists fight over the *one* graphical data model that is best suited for pangenomics in the long term. One prominent dispute is the juxtaposition of the *reference pangenome graph* model of Heng Li [92, 93] versus the *pangenome variation graph* model from Erik Garrison [55]. The first one allowed the HPRC to build graphs against which short and long reads can be mapped efficiently [100]. It is also directly ready for linear reference-based downstream analysis due to its ability to maintain the reference coordinate system (even with existing traditional genomic tools). Reference pangenome-based methods are still reference-biased. Their model requires that for each selected reference genome, one graph is built. In contrast, a pangenome variation graph constructed with PGGB avoids this bias by incorporating data from all genomes equally, without privileging any single reference genome [78]. Furthermore, only a pangenome variation graph is able to model the chromosome-crossing phenomenons like the recombination between heterologous

human acrocentric chromosomes [66].

I was also actively involved in quite some discussions at the International Genome Graph Symposium (IGGSy)[9]. The consensus from these discussions was that no single pangenome graph model is universally applicable. Instead, the choice of data structure, genomes, and algorithms used during and after graph construction should be tailored to the specific use cases and research objectives. I expect for consortia like the HPRC to provide human pangenome graph references for several major use cases. Once the current algorithms have matured and read mapping is possible regardless of the used graph construction algorithm, I envision that the pangenome graph variation model will dominate in the scientific community, because of its flexibility and ability to represent a greater biological range than the reference-biased pangenome graph model ever could.

Not necessarily staying in the graphical world, an alternative way would be to just share the all-vs-all alignments with the community, from which they can build a local, personalized graph with IMPG, potentially masking regions which are not required in downstream analyses. We propose a foundational data structure that is universally applicable, but allow for customization of the pangenome graphs based on specific use cases. This approach offers a balanced solution, providing a common framework while accommodating individual needs for graph construction. This seems the most likely future scenario from my current point of view. This approach would save CO_2 emissions.

Another class of pangenome graph models is De Bruijn graphs, which encode genomic sequences in a k-mer structure, which significantly distinguishes them from the two models presented above. Sophisticated tools for their construction and downstream analysis exist [85, 83], but these graphs are challenging to interpret primarily due to their abstract representation of genomic sequences as overlapping k-mers. This can result in complex and dense graphs that are difficult to visualize and analyze. Additionally, because De Bruijn graphs focus on k-mer overlaps rather than full sequences, they might obscure important biological context and can be cluttered with errors or artifacts from sequencing. Research is ongoing to convert De Bruijn graphs to pangenome variation graphs and vice versa [26], aiming to integrate a wide range of scientific tools for the benefit of pangenomics researchers. However, the future of such concepts remains uncertain.

The general genomics community is still very skeptical about pangenomes. "Switching to these new tools seems such a hassle!" — This is an important observation. Even when the scientific community acknowledges the superiority of a new approach, widespread adoption can take years, if not decades. We can see this in the continued reliance on outdated references like hg19/GRCh37, despite the availability of more accurate references like GRCh38. The adoption of pangenomic tools is likely to follow a similar trajectory. Promoting a phased adoption strategy — where research groups gradually incorporate pangenomic tools alongside existing methods — could ease the transition. The development of interoperable software pipelines that allow for seamless switching between linear and graph-based methods might accelerate acceptance. Addi-

[9] `https://iggsy.org/` (last accessed August 2024)

tionally, investing in benchmarking studies that directly compare pangenomic tools with linear reference-based tools across diverse datasets will help provide the evidence the field needs to justify the transition. One major promise pangenomes have to fulfill is to improve read mapping accuracy while ideally not increasing the computational requirement for this task. Indeed, there has been quite some academic advancements here [124, 37, 59, 54, 115, 94, 135, 16]. However, competitors from the industry are running ahead: As reported by Illumina, their developed DRAGEN [9] tool does not only align short reads best, but also fastest, because of the Field Programmable Gate Arrays (FPGAs) Illumina employs. There have been academic efforts to develop hardware tailored for computational pangenomics, like the PANORAMA [32] project, but this is in progress. I envision that the future is hardware accelerated pangenomics. Quantum computing could revolutionize the analysis of pangenomes [165] by significantly speeding up the comparison and mapping of individual genomes to complex sequence graphs, a process that is currently computationally intensive. It could also enhance the efficiency of analyzing large-scale genomic datasets, enabling more rapid insights into genetic diversity and personalized medicine.

Despite these improvements in read mapping, there are other fields where pangenomic methods were already successfully applied: GWAS [176, 81], improving crop breeding [80, 98, 162], genetic diversity [99], disease resistance in plants [51], pathogen evolution [171] and personalized medicine [136]. For these use cases the ODGI toolkit presented in this thesis is well prepared. For example, its capability to calculate a dissimilarity index between each pair of genomes enables scientists to find structural variants that are strongly associated with binary phenotypes in cows [109]. When researching the rice pangenome, it was discovered that there is a significant difference in the number of unique genes of Asian (10,101) and African (1259) rice [128]. Further highlighting the advantages of reference-unbiased methods like PGGB, the pangenome variation graph of 82 *A. thaliana* genomes contains up to 30% more nucleotides compared to reference-based (graph) methods. On average, reference-based graphs excludes 1.8% of the sequence in *H. sapiens* chromosome 6 to 22.1% in *E. coli* in comparison to a PGGB graph [61]. Recent studies on the transmissible cancer affecting Tasmanian devils, as discussed by Dr. Rodrigo Hamede [148], highlight the importance of comprehensive pangenome approaches. The use of pangenomics could be key in understanding and managing diseases like the Tasmanian devil facial tumor disease (DFTD), where traditional reference-based methods might miss critical genetic variations involved in disease transmission and evolution [72, 138].

But why is the community holding back? One reasonable worry is the potential incompatibility of traditional single linear reference tools and resources when compared to the novel pangenomic ones. For example, genomic positions on a single reference genome might be represented differently in a pangenomic context. This would affect annotation and other downstream analysis. Large data portals like EMBL/EBI are aware of this and transition their resources, like ENSEMBL [70], to better align with the pangenomics world, with a backwards compatibility to the single linear reference world. So

the (pan)genomics world is in flux. I believe that at some point we will have the traditional resource and algorithms beneath the pangenomic ones. The biological questions asked will determine which ones to use. In practice, the ultimate aim is that a user is not aware which technology is currently being employed to answer hypotheses.

Despite the advances in pangenome graph algorithms like our PG-SGD, there remain significant gaps in interactive visualization tools. Waragraph [49] and BandageNG face challenges with interactivity and scalability. As was shown, BandageNG struggles with large-scale graphs, making it difficult to visualize extensive pangenomic data effectively. Furthermore, existing tools often lack advanced interactive features such as dynamic zooming, filtering, and querying, which are crucial for in-depth exploration of complex datasets. A proper pangenomic browser with a bridge to traditional genomics does not exist. What one would wish for in the pangenome variation graph community is a tool that is created following a professional design process employed by e.g., the developers of PanVA [156]: *Discover*, *Design*, *Implement*, and *Deploy* [126]. Before the implementation, there was extensive discussion with long-term collaborators, visualization experts, and genome scientist from academia and the industry interested in (pan)genomics visualization. As is the *de facto* standard in the visualization community, a subsequent scientific user evaluation of the implemented software as well as use case evaluations was conducted. Future tools need to incorporate several key features: Enhanced interactivity is essential, enabling users to zoom, pan, and filter large pangenomic datasets effectively. Solutions that improve scalability will be crucial for handling the growing size and complexity of pangenome graphs. Moreover, integrating pangenomic data with traditional genomic resources will facilitate a more comprehensive analysis. By focusing on these areas, we can develop more robust and user-friendly visualization tools, significantly advancing our ability to analyze and interpret pangenomic data.

Another question that arises when working with large numbers of genomes: How many genomes are enough? Do we have to sequence and assemble every single individual? Is a pangenome growth curve helping here? In my opinion, the answer depends on several factors: (i) The genomic complexity of a species: This refers to the structural and functional intricacies of an organism's genome. For species with complex genomes—characterized by large numbers of genes, extensive repetitive regions, or extensive structural variation—thousands to millions of genomes might be needed to fully capture their genomic diversity. (ii) The genetic diversity of a species: Species with high genetic diversity may require tens of thousands of genomes to capture rare variants and understand population-specific differences. (iii) The research goals: For example identifying the core pangenome might require fewer genomes. (iv) The technological advances: Advances in sequencing technology and decreases in cost make it more feasible to sequence large numbers of genomes. One key technology here would be single–cell whole genome sequencing. This would allow us to e.g. build pangenome graphs from cells of a tumor tissue for a very fine granular GP research as never seen before. Improved bioinformatics tools for assembling and analyzing pangenomes can help maximize the information obtained from each genome, potentially reducing the total number

required. Ultimately, a universal solution does not exist. A pangenome growth curve can provide valuable guidance, but the primary determinant should be the research questions being addressed.

Assuming we sequenced every human being, we would be faced with several philosophical implications and questions. First, the concept of genetic uniqueness would be put to test. We would need to consider whether every individual is truly genetically unique or if there are more fundamental commonalities that define humanity as a whole. This raises questions about the nature of human identity: Is it primarily shaped by our genome, or are environmental factors and personal experiences more influential? Although having comprehensive genomic data would promote an in-depth understanding, its practical relevance and utility in everyday applications would remain uncertain. To address this, privacy protections must be reinforced, and genetic counseling should help individuals interpret their data within a broader context. One key ethical concern would be balancing the management such of a vast and detailed data pool with the autonomy of individuals' genetic information. If your genetic material shapes your social status, how would this turn out? For example, there are already ongoing debates about whole-genome sequencing of newborns [143]. The conclusive argument is that for us as a species, we could live a bodily healthier life while for an individual, there are lot's of ethical, moral, technical, economic, and unknown issues involved. Furthermore, the availability of complete genomic data introduces the risk of misuse, such as the creation of targeted biological threats. The ability to manipulate or potentially eradicate individuals based on specific genetic traits poses serious ethical dilemmas and social explosives. Frameworks like the Genetic Information Nondiscrimination Act (GINA) could help safeguard against misuse of genetic information. However, to have a holistic understanding of human health and behavior would require integrating this genomic data with knowledge of environmental or lifestyle factors. Another hope is that such a data source would bring us as humans closer: It can highlight commonalities despite some individual differences. To summarize, such an endeavor would invite ongoing reflection on the broader implications of human genetics and the role of technology in shaping our future. A question for the reader here is: Would you like to get your genome sequenced and make it publicly available?

As has already been mentioned, pangenome graph tools are not only capable to work with genomic data, but also with transcriptome [133], genetic [96], or proteomic [34] data presenting a paradigm shift for multiomics applications. Consequently, pangenomics does not only improve the characterization of genotypes on their own, but will play a key role in future multiomics analysis for the enhanced GP discovery.

I used this integrated discussion to share my educated opinion about the benefits that multiomics and pangenomic models offer for integrative research, especially in interdisciplinary projects as well as sketching some future directions of these fields.

5 Conclusion

In my thesis, I explored integrative analysis methods aimed at mitigating analysis bias and enhancing genotype-phenotype discovery. Studying biological complexity requires the use of high-dimensional and multi-modal data derived from various omics technologies. Solely relying on a single data source comes with significant constraints, as it fails to adequately model intricate molecular mechanisms. Multifaceted data analysis holds transformative potential by integrating diverse biological data sources, fostering a thorough understanding of complex molecular mechanisms, and driving advancements in research, healthcare, and targeted therapies. Integrating biological data from different origins can bridge the gap between experts from varied research backgrounds, facilitating interdisciplinary research and yielding more precise and informed conclusions.

Progress in pangenomics has historically been prevented by selection bias in genomic analyses, with genome inference technologies like genotype arrays or NGS limited to easily detectable variants. This led to a skewed understanding of genomic diversity. Recent advances in long-read sequencing technology allow the exploration of *the dark parts of a genome*. The T2T Consortium's first complete human genome assembly, exemplifies how scientists can tackle reference bias both *in vitro* and *in silico* for a more advanced genotype representation. This shift towards comprehensive genomic understanding is further exemplified by the HPRC's release of the first draft human pangenome reference, highlighting the potential of integrating diverse genomic data to mitigate reference bias.

The tools and methods I developed contribute significantly to the second phase of pangenomics, which leverages recent technological advances to achieve a more comprehensive understanding of genomic diversity. The integration of biological data from various sources and the development of novel pangenomic tools underscore the transformative potential of pangenomics and multiomics in advancing genotype-phenotype research and bridging interdisciplinary gaps in the scientific community. So ...
Is it time to change the reference genome? [6] - I think it is!

Bibliography

[1] Al-Amrani, S., Al-Jabri, Z., Al-Zaabi, A., Alshekaili, J., and Al-Khabori, M. (2021). Proteomics: Concepts and applications in human medicine. *World Journal of Biological Chemistry*, **12**(5), 57–69.

[2] Amat, J. A., Garrido, A., Rendón-Martos, M., Portavia, F., and Rendón, M. A. (2022). Plumage coloration in greater flamingos phoenicopterus roseus is affected by interactions between foraging site, body condition and sex. *Ardeola*, **69**(2).

[3] Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J., and Stegle, O. (2020). Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, **21**(1), 111.

[4] Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I. T., Novak, A. M., Deran, A., Fang, Q., Xie, D., Feng, S., Stiller, J., Genereux, D., Johnson, J., Marinescu, V. D., Alföldi, J., Harris, R. S., Lindblad-Toh, K., Haussler, D., Karlsson, E., Jarvis, E. D., Zhang, G., and Paten, B. (2020). Progressive cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, **587**(7833), 246–251.

[5] Ball, M. P. (2007). Punnett square mendel flowers.

[6] Ballouz, S., Dobin, A., and Gillis, J. A. (2019). Is it time to change the reference genome? *Genome Biology*, **20**(1).

[7] Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., and Edwards, D. (2020). Plant pan-genomes are the new reference. *Nature Plants*, **6**(8), 914–920.

[8] Bayer, P. E., Petereit, J., Durant, E., Monat, C., Rouard, M., Hu, H., Chapman, B., Li, C., Cheng, S., Batley, J., and Edwards, D. (2022). Wheat panache: A pangenome graph database representing presence–absence variation across sixteen bread wheat genomes. *The Plant Genome*, **15**(3).

[9] Behera, S., Catreux, S., Rossi, M., Truong, S., Huang, Z., Ruehle, M., Visvanath, A., Parnaby, G., Roddey, C., Onuchic, V., Cameron, D. L., English, A., Mehtalia, S., Han, J., Mehio, R., and Sedlazeck, F. J. (2024). Comprehensive and accurate genome analysis at scale using dragen accelerated algorithms. *bioRxiv*.

[10] Benkirane, H., Pradat, Y., Michiels, S., and Cournède, P.-H. (2023). Customics: A versatile deep-learning based strategy for multi-omics integration. *PLOS Computational Biology*, **19**(3), e1010921.

[11] Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., and Milanesi, L. (2016). Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics*, **17**(S2).

[12] Beyer, W., Novak, A. M., Hickey, G., Chan, J., Tan, V., Paten, B., and Zerbino, D. R. (2019). Sequence tube maps: making graph genomes intuitive to commuters. *Bioinformatics*, **35**(24), 5318–5320.

[13] Bonner, W. A., Hulett, H. R., Sweet, R. G., and Herzenberg, L. A. (1972). Fluorescence activated cell sorting. *Review of Scientific Instruments*, **43**(3), 404–409.

[14] Bovine Pan-Genome Consortium (2024). Bovine Pan-Genome Consortium. `https://njdbickhart.github.io/` (last accessed August 2024).

[15] Browning, S. R. and Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, **12**(10), 703–714.

[16] Chandra, G., Gibney, D., and Jain, C. (2024). Haplotype-aware sequence alignment to pangenome graphs. *Genome Research*, page gr.279143.124.

[17] Chang, K., Creighton, C. J., Davis, C., Donehower, L., Drummond, J., Wheeler, D., Ally, A., Balasundaram, M., Birol, I., Butterfield, Y. S. N., Chu, A., Chuah, E., Chun, H.-J. E., Dhalla, N., Guin, R., Hirst, M., Hirst, C., Holt, R. A., Jones, S. J. M., Lee, D., Li, H. I., Marra, M. A., Mayo, M., Moore, R. A., Mungall, A. J., Robertson, A. G., Schein, J. E., Sipahimalani, P., Tam, A., Thiessen, N., Varhol, R. J., Beroukhim, R., Bhatt, A. S., Brooks, A. N., Cherniack, A. D., Freeman, S. S., Gabriel, S. B., Helman, E., Jung, J., Meyerson, M., Ojesina, A. I., Pedamallu, C. S., Saksena, G., Schumacher, S. E., Tabak, B., Zack, T., Lander, E. S., Bristow, C. A., Hadjipanayis, A., Haseley, P., Kucherlapati, R., Lee, S., Lee, E., Luquette, L. J., Mahadeshwar, H. S., Pantazi, A., Parfenov, M., Park, P. J., Protopopov, A., Ren, X., Santoso, N., Seidman, J., Seth, S., Song, X., Tang, J., Xi, R., Xu, A. W., Yang, L., Zeng, D., Auman, J. T., Balu, S., Buda, E., Fan, C., Hoadley, K. A., Jones, C. D., Meng, S., Mieczkowski, P. A., Parker, J. S., Perou, C. M., Roach, J., Shi, Y., Silva, G. O., Tan, D., Veluvolu, U., Waring, S., Wilkerson, M. D., Wu, J., Zhao, W., Bodenheimer, T., Hayes, D. N., Hoyle, A. P., Jeffreys, S. R., Mose, L. E., Simons, J. V., Soloway, M. G., Baylin, S. B., Berman, B. P., Bootwalla, M. S., Danilova, L., Herman, J. G., Hinoue, T., Laird, P. W., Rhie, S. K., Shen, H., Triche, T., Weisenberger, D. J., Carter, S. L., Cibulskis, K., Chin, L., Zhang, J., Getz, G., Sougnez, C., Wang, M., Dinh, H., Doddapaneni, H. V., Gibbs, R., Gunaratne, P., Han, Y., Kalra, D., Kovar, C., Lewis, L., Morgan, M., Morton, D., Muzny, D., Reid, J., Xi, L., Cho, J., DiCara, D., Frazer,

S., Gehlenborg, N., Heiman, D. I., Kim, J., Lawrence, M. S., Lin, P., Liu, Y., Noble, M. S., Stojanov, P., Voet, D., Zhang, H., Zou, L., Stewart, C., Bernard, B., Bressler, R., Eakin, A., Iype, L., Knijnenburg, T., Kramer, R., Kreisberg, R., Leinonen, K., Lin, J., Liu, Y., Miller, M., Reynolds, S. M., Rovira, H., Shmulevich, I., Thorsson, V., Yang, D., Zhang, W., Amin, S., Wu, C.-J., Wu, C.-C., Akbani, R., Aldape, K., Baggerly, K. A., Broom, B., Casasent, T. D., Cleland, J., Creighton, C., Dodda, D., Edgerton, M., Han, L., Herbrich, S. M., Ju, Z., Kim, H., Lerner, S., Li, J., Liang, H., Liu, W., Lorenzi, P. L., Lu, Y., Melott, J., Mills, G. B., Nguyen, L., Su, X., Verhaak, R., Wang, W., Weinstein, J. N., Wong, A., Yang, Y., Yao, J., Yao, R., Yoshihara, K., Yuan, Y., Yung, A. K., Zhang, N., Zheng, S., Ryan, M., Kane, D. W., Aksoy, B. A., Ciriello, G., Dresdner, G., Gao, J., Gross, B., Jacobsen, A., Kahles, A., Ladanyi, M., Lee, W., Lehmann, K.-V., Miller, M. L., Ramirez, R., Rätsch, G., Reva, B., Sander, C., Schultz, N., Senbabaoglu, Y., Shen, R., Sinha, R., Sumer, S. O., Sun, Y., Taylor, B. S., Weinhold, N., Fei, S., Spellman, P., Benz, C., Carlin, D., Cline, M., Craft, B., Ellrott, K., Goldman, M., Haussler, D., Ma, S., Ng, S., Paull, E., Radenbaugh, A., Salama, S., Sokolov, A., Stuart, J. M., Swatloski, T., Uzunangelov, V., Waltman, P., Yau, C., Zhu, J., Hamilton, S. R., Abbott, S., Abbott, R., Dees, N. D., Delehaunty, K., Ding, L., Dooling, D. J., Eldred, J. M., Fronick, C. C., Fulton, R., Fulton, L. L., Kalicki-Veizer, J., Kanchi, K.-L., Kandoth, C., Koboldt, D. C., Larson, D. E., Ley, T. J., Lin, L., Lu, C., Magrini, V. J., Mardis, E. R., McLellan, M. D., McMichael, J. F., Miller, C. A., O'Laughlin, M., Pohl, C., Schmidt, H., Smith, S. M., Walker, J., Wallis, J. W., Wendl, M. C., Wilson, R. K., Wylie, T., Zhang, Q., Burton, R., Jensen, M. A., Kahn, A., Pihl, T., Pot, D., Wan, Y., Levine, D. A., Black, A. D., Bowen, J., Network, T. C. G. A. R., Center, G. C., Center, G. D. A., Center, S., Center, D. C., Site, T. S., and Center, B. C. R. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, **45**(10), 1113–1120.

[18] Chen, C., Wang, J., Pan, D., Wang, X., Xu, Y., Yan, J., Wang, L., Yang, X., Yang, M., and Liu, G. (2023). Applications of multi-omics analysis in human diseases. *MedComm*, **4**(4).

[19] Cheong, S.-H. and Si, Y.-W. (2022). Force-directed algorithms for schematic drawings and placement: A survey.

[20] Chessel, D. and Hanafie, M. (1996). *Analysis of the co-inertia of K tables*, volume 44. Revue de statistique appliquée.

[21] Cheung, S. K. C., Chuang, P.-K., Huang, H.-W., Hwang-Verslues, W. W., Cho, C. H.-H., Yang, W.-B., Shen, C.-N., Hsiao, M., Hsu, T.-L., Chang, C.-F., and Wong, C.-H. (2015). Stage-specific embryonic antigen-3 (SSEA-3) and β3GalT5 are cancer specific and significant markers for breast cancer stem cells. *Proceedings of the National Academy of Sciences*, **113**(4), 960–965.

[22] Chierici, M., Bussola, N., Marcolini, A., Francescatto, M., Zandonà, A., Trastulla, L., Agostinelli, C., Jurman, G., and Furlanello, C. (2020). Integrative network fusion: A multi-omics approach in molecular profiling. *Frontiers in Oncology*, **10**.

[23] Chin, C.-S., Behera, S., Khalak, A., Sedlazeck, F. J., Sudmant, P. H., Wagner, J., and Zook, J. M. (2023). Multiscale analysis of pangenomes enables improved representation of genomic diversity for repetitive and clinically relevant genes. *Nature Methods*, **20**(8), 1213–1221.

[24] Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.-C., Agarwala, R., McLaren, W. M., Ritchie, G. R., Albracht, D., Kremitzki, M., Rock, S., Kotkiewicz, H., Kremitzki, C., Wollam, A., Trani, L., Fulton, L., Fulton, R., Matthews, L., Whitehead, S., Chow, W., Torrance, J., Dunn, M., Harden, G., Threadgold, G., Wood, J., Collins, J., Heath, P., Griffiths, G., Pelan, S., Grafham, D., E. Eichler, E., Weinstock, G., Mardis, E. R., Wilson, R. K., Howe, K., Flicek, P., and Hubbard, T. (2011). Modernizing reference genome assemblies. *PLoS Biology*, **9**(7), e1001091.

[25] Churchill, F. B. (1974). William johannsen and the genotype concept. *Journal of the History of Biology*, **7**(1), 5–30.

[26] Cicherski, A. and Dojer, N. (2023). *From de Bruijn Graphs to Variation Graphs – Relationships Between Pangenome Models*, page 114–128. Springer Nature Switzerland.

[27] Coarfa, C., Grimm, S. L., Rajapakshe, K., Perera, D., Lu, H.-Y., Wang, X., Christensen, K. R., Mo, Q., Edwards, D. P., and Huang, S. (2021). Reverse-phase protein array: Technology, application, data processing, and integration. *Journal of Biomolecular Techniques*, **32**(1), 15–29.

[28] Cochetel, N., Minio, A., Guarracino, A., Garcia, J. F., Figueroa-Balderas, R., Massonnet, M., Kasuga, T., Londo, J. P., Garrison, E., Gaut, B. S., and Cantu, D. (2023). A super-pangenome of the north american wild grape species. *Genome Biology*, **24**(1).

[29] Collado-Torres, L., Nellore, A., Kammers, K., Ellis, S. E., Taub, M. A., Hansen, K. D., Jaffe, A. E., Langmead, B., and Leek, J. T. (2017). Reproducible RNA-seq analysis using recount2. *Nature Biotechnology*, **35**(4), 319–321.

[30] Computational Pan-Genomics Consortium (2018). Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*, **19**(1), 118–135.

[31] Conesa, A. and Beck, S. (2019). Making multi-omics data accessible to researchers. *Scientific Data*, **6**(1).

[32] Cornell University (2021). $5M grant will tackle pangenomics computing challenge. `https://news.cornell.edu/stories/2021/11/5m-grant-will-tackle-pangenomics-computing-challenge` (last accessed August 2024).

[33] da Veiga Leprevost, F., Grüning, B. A., Alves Aflitos, S., Röst, H. L., Uszkoreit, J., Barsnes, H., Vaudel, M., Moreno, P., Gatto, L., Weber, J., Bai, M., Jimenez, R. C., Sachsenberg, T., Pfeuffer, J., Vera Alvarez, R., Griss, J., Nesvizhskii, A. I., and Perez-Riverol, Y. (2017). Biocontainers: an open-source and community-driven framework for software standardization. *Bioinformatics*, **33**(16), 2580–2582.

[34] Dabbaghie, F., Srikakulam, S. K., Marschall, T., and Kalinina, O. V. (2023). PanPA: generation and alignment of panproteome graphs. *Bioinformatics Advances*, **3**(1).

[35] Demartini, D. R., Pasquali, G., and Carlini, C. R. (2013). An overview of proteomics approaches applied to biopharmaceuticals and cyclotides research. *Journal of Proteomics*, **93**, 224–233.

[36] Di Tommaso, P. *et al.* (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, **35**(4), 316–319.

[37] Dilthey, A., Cox, C., Iqbal, Z., Nelson, M. R., and McVean, G. (2015). Improved genome inference in the mhc using a population reference graph. *Nature Genetics*, **47**(6), 682–688.

[38] Ding, W., Baumdicker, F., and Neher, R. A. (2017). panX: pan-genome analysis and exploration. *Nucleic Acids Research*, **46**(1), e5–e5.

[39] DOLÉDEC, S. and CHESSEL, D. (1994). Co-inertia analysis: an alternative method for studying species–environment relationships. *Freshwater Biology*, **31**(3), 277–294.

[40] Durant, E., Sabot, F., Conte, M., and Rouard, M. (2021). Panache: a web browser-based viewer for linearized pangenomes. *Bioinformatics*.

[41] Durbin, R. M. *et al.* (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319), 1061–1073.

[42] Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D.,

Zhao, P., Zhong, F., Korlach, J., and Turner, S. (2009). Real-time dna sequencing from single polymerase molecules. *Science*, **323**(5910), 133–138.

[43] Eizenga, J. M., Novak, A. M., Sibbesen, J. A., Heumos, S., Ghaffaari, A., Hickey, G., Chang, X., Seaman, J. D., Rounthwaite, R., Ebler, J., Rautiainen, M., Garg, S., Paten, B., Marschall, T., Sirén, J., and Garrison, E. (2020). Pangenome graphs. *Annual Review of Genomics and Human Genetics*, **21**(1), 139–162.

[44] Ewels, P. *et al.* (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**(19), 3047–3048.

[45] Ewels, P. *et al.* (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, **38**(3), 276–278.

[46] Feinberg, A. P. and Fallin, M. D. (2015). Epigenetics at the crossroads of genes and the environment. *JAMA*, **314**(11), 1129.

[47] Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989). Electrospray ionization for mass spectrometry of large biomolecules. *Science*, **246**(4926), 64–71.

[48] Fischer, C. *et al.* (2022). gfaestus. `https://github.com/chfi/gfaestus` (last accessed August 2024.

[49] Fischer, C. *et al.* (2024). waragraph. `https://github.com/chfi/waragraph` (last accessed August 2024).

[50] Frolova, A. and Wilczyński, B. (2018). Distributed bayesian networks reconstruction on the whole genome scale. *PeerJ*, **6**, e5692.

[51] Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., Burzynski-Chang, E. A., Fish, T. L., Stromberg, K. A., Sacks, G. L., Thannhauser, T. W., Foolad, M. R., Diez, M. J., Blanca, J., Canizares, J., Xu, Y., van der Knaap, E., Huang, S., Klee, H. J., Giovannoni, J. J., and Fei, Z. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature Genetics*, **51**(6), 1044–1051.

[52] Gao, Y., Liu, Y., Ma, Y., Liu, B., Wang, Y., and Xing, Y. (2020). abpoa: an simd-based c library for fast partial order alignment using adaptive band. *Bioinformatics*, **37**(15), 2209–2211.

[53] Gao, Y., Yang, X., Chen, H., Tan, X., Yang, Z., Deng, L., Wang, B., Kong, S., Li, S., Cui, Y., Lei, C., Wang, Y., Pan, Y., Ma, S., Sun, H., Zhao, X., Shi, Y., Yang, Z., Wu, D., Wu, S., Zhao, X., Shi, B., Jin, L., Hu, Z., Mao, C., Fan, S., Gao, Q., Dai, J., Bu, F., He, G., Wu, Y., Yuan, H., Li, J., Chen, C., Yang, J., Wei, C., Jin, X., Shen, X., Lu, Y., Chu, J., Ye, K., and Xu, S. (2023). A pangenome reference of 36 chinese populations. *Nature*, **619**(7968), 112–121.

[54] Garrison, E. (2019a). Graphical pangenomics.

[55] Garrison, E. (2019b). Untangling graphical pangenomics. `https://ekg.github.io/2019/07/09/Untangling-graphical-pangenomics` (last accessed August 2024).

[56] Garrison, E. (2024a). Gfalace. `https://github.com/pangenome/gfalace` (last accessed August 2024.

[57] Garrison, E. (2024b). IMPG. `https://github.com/pangenome/impg` (last accessed August 2024.

[58] Garrison, E. and Guarracino, A. (2022). Unbiased pangenome graphs. *Bioinformatics*, **39**(1).

[59] Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., Paten, B., and Durbin, R. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, **36**(9), 875–879.

[60] Garrison, E., Guarracino, A., Heumos, S., Villani, F., Bao, Z., Tattini, L., Hagmann, J., Vorbrugg, S., Marco-Sola, S., Kubica, C., Ashbrook, D. G., Thorell, K., Rusholme-Pilcher, R. L., Liti, G., Rudbeck, E., Nahnsen, S., Yang, Z., Moses, M. N., Nobrega, F. L., Wu, Y., Chen, H., de Ligt, J., Sudmant, P. H., Soranzo, N., Colonna, V., Williams, R. W., and Prins, P. (2023). Building pangenome graphs.

[61] Garrison, E., Guarracino, A., Heumos, S., Villani, F., Bao, Z., Tattini, L., Hagmann, J., Vorbrugg, S., Marco-Sola, S., Kubica, C., Ashbrook, D. G., Thorell, K., Rusholme-Pilcher, R. L., Liti, G., Rudbeck, E., Golicz, A. A., Nahnsen, S., Yang, Z., Moses, M. N., Nobrega, F. L., Wu, Y., Chen, H., de Ligt, J., Sudmant, P. H., Huang, S., Weigel, D., Soranzo, N., Colonna, V., Williams, R. W., and Prins, P. (2024). Building pangenome graphs. *Accepted at Nature Methods*.

[62] Gautreau, G., Bazin, A., Gachet, M., Planel, R., Burlot, L., Dubois, M., Perrin, A., Médigue, C., Calteau, A., Cruveiller, S., Matias, C., Ambroise, C., Rocha, E. P. C., and Vallenet, D. (2020). PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLOS Computational Biology*, **16**(3), e1007732. Publisher: Public Library of Science.

[63] Gholami, A. M., Hahne, H., Wu, Z., Auer, F. J., Meng, C., Wilhelm, M., and Kuster, B. (2013). Global proteome analysis of the nci-60 cell line panel. *Cell Reports*, **4**(3), 609–620.

[64] Gonnella, G., Niehus, N., and Kurtz, S. (2018). GfaViz: flexible and interactive visualization of GFA sequence graphs. *Bioinformatics*, **35**(16), 2853–2855.

[65] Guarracino, A., Heumos, S., Nahnsen, S., Prins, P., and Garrison, E. (2022). ODGI: understanding pangenome graphs. *Bioinformatics*, **38**(13), 3319–3326.

[66] Guarracino, A., Buonaiuto, S., de Lima, L. G., Potapova, T., Rhie, A., Koren, S., Rubinstein, B., Fischer, C., Abel, H. J., Antonacci-Fulton, L. L., Asri, M., Baid, G., Baker, C. A., Belyaeva, A., Billis, K., Bourque, G., Carroll, A., Chaisson, M. J. P., Chang, P.-C., Chang, X. H., Cheng, H., Chu, J., Cody, S., Cook, D. E., Cook-Deegan, R. M., Cornejo, O. E., Diekhans, M., Doerr, D., Ebert, P., Ebler, J., Eichler, E. E., Eizenga, J. M., Fairley, S., Fedrigo, O., Felsenfeld, A. L., Feng, X., Flicek, P., Formenti, G., Frankish, A., Fulton, R. S., Gao, Y., Garg, S., Garrison, N. A., Giron, C. G., Green, R. E., Groza, C., Haggerty, L., Hall, I., Harvey, W. T., Haukness, M., Haussler, D., Heumos, S., Hickey, G., Hoekzema, K., Hourlier, T., Howe, K., Jain, M., Jarvis, E. D., Ji, H. P., Kenny, E. E., Koenig, B. A., Kolesnikov, A., Korbel, J. O., Kordosky, J., Lee, H., Lewis, A. P., Li, H., Liao, W.-W., Lu, S., Lu, T.-Y., Lucas, J. K., Magalhães, H., Marco-Sola, S., Marijon, P., Markello, C., Marschall, T., Martin, F. J., McCartney, A., McDaniel, J., Miga, K. H., Mitchell, M. W., Monlong, J., Mountcastle, J., Munson, K. M., Mwaniki, M. N., Nattestad, M., Novak, A. M., Nurk, S., Olsen, H. E., Olson, N. D., Paten, B., Pesout, T., Popejoy, A. B., Porubsky, D., Prins, P., Puiu, D., Rautiainen, M., Regier, A. A., Sacco, S., Sanders, A. D., Schneider, V. A., Schultz, B. I., Shafin, K., Sibbesen, J. A., Sirén, J., Smith, M. W., Sofia, H. J., Tayoun, A. N. A., Thibaud-Nissen, F., Tomlinson, C., Tricomi, F. F., Villani, F., Vollger, M. R., Wagner, J., Walenz, B., Wang, T., Wood, J. M. D., Zimin, A. V., Zook, J. M., Gerton, J. L., Phillippy, A. M., Colonna, V., and Garrison, E. (2023). Recombination between heterologous human acrocentric chromosomes. *Nature*, **617**(7960), 335–343.

[67] Guarracino, A., Mwaniki, N., Marco-Sola, S., and Garrison, E. (2024). wfmash: whole-chromosome pairwise alignment using the hierarchical wavefront algorithm. `https://github.com/waveygang/wfmash` (last accessed August 2024).

[68] Guo, L., Wang, X., Ayhan, D. H., Rhaman, M. S., Yan, M., Jiang, J., Wang, D., Zheng, W., Mei, J., Ji, W., Jiao, J., Chen, S., Sun, J., Yi, S., Meng, D., Wang, J., Bhuiyan, M. N., Qin, G., Guo, L., Yang, Q., Zhang, X., Sun, H., Liu, C., and Ye, W. (2024). Super pangenome of grapevines empowers improvement of the oldest domesticated fruit. *bioRxiv*.

[69] Hachul, S. and Jünger, M. (2005). Large-graph layout with the fast multipole multilevel method. Working paper, Universität zu Köln.

[70] Harrison, P. W., Amode, M. R., Austine-Orimoloye, O., Azov, A. G., Barba, M., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S. K., Boddu, S., Branco Lins, P. R., Brooks, L., Ramaraju, S. B., Campbell, L. I., Martinez, M. C., Charkhchi, M., Chougule, K., Cockburn, A., Davidson, C., De Silva, N. H., Dodiya, K., Donaldson, S., El Houdaigui, B., Naboulsi, T. E., Fatima, R., Giron, C. G., Genez,

T., Grigoriadis, D., Ghattaoraya, G. S., Martinez, J. G., Gurbich, T. A., Hardy, M., Hollis, Z., Hourlier, T., Hunt, T., Kay, M., Kaykala, V., Le, T., Lemos, D., Lodha, D., Marques-Coelho, D., Maslen, G., Merino, G. A., Mirabueno, L. P., Mushtaq, A., Hossain, S. N., Ogeh, D. N., Sakthivel, M. P., Parker, A., Perry, M., Piližota, I., Poppleton, D., Prosovetskaia, I., Raj, S., Pérez-Silva, J. G., Salam, A. I. A., Saraf, S., Saraiva-Agostinho, N., Sheppard, D., Sinha, S., Sipos, B., Sitnik, V., Stark, W., Steed, E., Suner, M.-M., Surapaneni, L., Sutinen, K., Tricomi, F. F., Urbina-Gómez, D., Veidenberg, A., Walsh, T. A., Ware, D., Wass, E., Willhoft, N. L., Allen, J., Alvarez-Jarreta, J., Chakiachvili, M., Flint, B., Giorgetti, S., Haggerty, L., Ilsley, G. R., Keatley, J., Loveland, J. E., Moore, B., Mudge, J. M., Naamati, G., Tate, J., Trevanion, S. J., Winterbottom, A., Frankish, A., Hunt, S. E., Cunningham, F., Dyer, S., Finn, R. D., Martin, F. J., and Yates, A. D. (2023). Ensembl 2024. *Nucleic Acids Research*, **52**(D1), D891–D899.

[71] Hartl, D. L. and Clark, A. G. (2007). *Principles of population genetics*. Oxford University Press, New York, NY, 4 edition.

[72] Hawkins, C., Baars, C., Hesterman, H., Hocking, G., Jones, M., Lazenby, B., Mann, D., Mooney, N., Pemberton, D., Pyecroft, S., Restani, M., and Wiersma, J. (2006). Emerging disease and population decline of an island endemic, the tasmanian devil sarcophilus harrisii. *Biological Conservation*, **131**(2), 307–324.

[73] Hein, J. (1989). A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given. *Molecular Biology and Evolution*.

[74] Heumos, S., Dehn, S., Bräutigam, K., Codrea, M. C., Schürch, C. M., Lauer, U. M., Nahnsen, S., and Schindler, M. (2022). Multiomics surface receptor profiling of the NCI-60 tumor cell panel uncovers novel theranostics for cancer immunotherapy. *Cancer Cell International*, **22**(1), 311.

[75] Heumos, S., Heuer, M. F., Hanssen, F., Heumos, L., Guarracino, A., Heringer, P., Ehmele, P., Prins, P., Garrison, E., and Nahnsen, S. (2024a). Cluster efficient pangenome graph construction with nf-core/pangenome. *bioRxiv*.

[76] Heumos, S., Guarracino, A., Schmelzle, J.-N. M., Li, J., Zhang, Z., Hagmann, J., Nahnsen, S., Prins, P., and Garrison, E. (2024b). Pangenome graph layout by path-guided stochastic gradient descent. *Bioinformatics*, **40**(7).

[77] Hickey, G., Monlong, J., Ebler, J., Novak, A. M., Eizenga, J. M., Gao, Y., Abel, H. J., Antonacci-Fulton, L. L., Asri, M., Baid, G., Baker, C. A., Belyaeva, A., Billis, K., Bourque, G., Buonaiuto, S., Carroll, A., Chaisson, M. J. P., Chang, P.-C., Chang, X. H., Cheng, H., Chu, J., Cody, S., Colonna, V., Cook, D. E., Cook-Deegan, R. M., Cornejo, O. E., Diekhans, M., Doerr, D., Ebert, P., Ebler, J., Eichler, E. E., Fairley, S.,

Fedrigo, O., Felsenfeld, A. L., Feng, X., Fischer, C., Flicek, P., Formenti, G., Frankish, A., Fulton, R. S., Garg, S., Garrison, E., Garrison, N. A., Giron, C. G., Green, R. E., Groza, C., Guarracino, A., Haggerty, L., Hall, I. M., Harvey, W. T., Haukness, M., Haussler, D., Heumos, S., Hoekzema, K., Hourlier, T., Howe, K., Jain, M., Jarvis, E. D., Ji, H. P., Kenny, E. E., Koenig, B. A., Kolesnikov, A., Korbel, J. O., Kordosky, J., Koren, S., Lee, H., Lewis, A. P., Liao, W.-W., Lu, S., Lu, T.-Y., Lucas, J. K., Magalhães, H., Marco-Sola, S., Marijon, P., Markello, C., Marschall, T., Martin, F. J., McCartney, A., McDaniel, J., Miga, K. H., Mitchell, M. W., Mountcastle, J., Munson, K. M., Mwaniki, M. N., Nattestad, M., Nurk, S., Olsen, H. E., Olson, N. D., Pesout, T., Phillippy, A. M., Popejoy, A. B., Porubsky, D., Prins, P., Puiu, D., Rautiainen, M., Regier, A. A., Rhie, A., Sacco, S., Sanders, A. D., Schneider, V. A., Schultz, B. I., Shafin, K., Sibbesen, J. A., Sirén, J., Smith, M. W., Sofia, H. J., Tayoun, A. N. A., Thibaud-Nissen, F., Tomlinson, C., Tricomi, F. F., Villani, F., Vollger, M. R., Wagner, J., Walenz, B., Wang, T., Wood, J. M. D., Zimin, A. V., Zook, J. M., Marschall, T., Li, H., and Paten, B. (2023). Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nature Biotechnology*.

[78] Hu, H., Li, R., Zhao, J., Batley, J., and Edwards, D. (2024). Technological development and advances for constructing and analyzing plant pangenomes. *Genome Biology and Evolution*, **16**(4).

[79] International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, **431**(7011), 931–945.

[80] Jayakodi, M., Padmarasu, S., Haberer, G., Bonthala, V. S., Gundlach, H., Monat, C., Lux, T., Kamal, N., Lang, D., Himmelbach, A., Ens, J., Zhang, X.-Q., Angessa, T. T., Zhou, G., Tan, C., Hill, C., Wang, P., Schreiber, M., Boston, L. B., Plott, C., Jenkins, J., Guo, Y., Fiebig, A., Budak, H., Xu, D., Zhang, J., Wang, C., Grimwood, J., Schmutz, J., Guo, G., Zhang, G., Mochida, K., Hirayama, T., Sato, K., Chalmers, K. J., Langridge, P., Waugh, R., Pozniak, C. J., Scholz, U., Mayer, K. F. X., Spannagl, M., Li, C., Mascher, M., and Stein, N. (2020). The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*, **588**(7837), 284–289.

[81] Jin, S., Han, Z., Hu, Y., Si, Z., Dai, F., He, L., Cheng, Y., Li, Y., Zhao, T., Fang, L., and Zhang, T. (2023). Structural variation (sv)-based pan-genome and gwas reveal the impacts of svs on the speciation and diversification of allotetraploid cottons. *Molecular Plant*, **16**(4), 678–693.

[82] Johannsen, W. (2014). The genotype conception of heredity. *International Journal of Epidemiology*, **43**(4), 989–1000.

[83] Jonkheer, E. M., van Workum, D.-J. M., Sheikhizadeh Anari, S., Brankovics, B., de Haan, J. R., Berke, L., van der Lee, T. A. J., de Ridder, D., and Smit, S. (2022).

Pantools v3: functional annotation, classification and phylogenomics. *Bioinformatics*, **38**(18), 4403–4405.

[84] Jose, A., Kulkarni, P., Thilakan, J., Munisamy, M., Malhotra, A. G., Singh, J., Kumar, A., Rangnekar, V. M., Arya, N., and Rao, M. (2024). Integration of panomics technologies and three-dimensional in vitro tumor models: an approach toward drug discovery and precision medicine. *Molecular Cancer*, **23**(1).

[85] Karasikov, M., Mustafa, H., Danciu, D., Zimmermann, M., Barber, C., Rätsch, G., and Kahles, A. (2020). Indexing all life's known biological sequences. *bioRxiv*.

[86] Kennedy, L. B. and Salama, A. K. S. (2020). A review of cancer immunotherapy toxicity. *CA: A Cancer Journal for Clinicians*, **70**(2), 86–104.

[87] Köster, J. and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, **28**(19), 2520–2522.

[88] Krakau, S. (2024). nf-co2footprint. `https://github.com/nextflow-io/nf-co2footprint` (last accessed August 2024).

[89] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.-F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin,

S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H.-C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G. R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F. A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S.-P., Yeh, R.-F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Patrinos, A., and Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

[90] Lee, C., Grasso, C., and Sharlow, M. F. (2002). Multiple sequence alignment using partial order graphs. *Bioinformatics*, **18**(3), 452–464.

[91] Leonard, A. S., Crysnanto, D., Fang, Z.-H., Heaton, M. P., Vander Ley, B. L., Herrera, C., Bollwein, H., Bickhart, D. M., Kuhn, K. L., Smith, T. P. L., Rosen, B. D., and Pausch, H. (2022). Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. *Nature Communications*, **13**(1).

[92] Li, H. (2019a). On a reference pan-genome model. `https://lh3.github.io/2019/07/08/on-a-reference-pan-genome-model` (last accessed August 2024).

[93] Li, H. (2019b). On a reference pan-genome model (Part II). `https://lh3.github.io/2019/07/12/on-a-reference-pan-genome-model-part-ii` (last accessed August 2024).

[94] Li, H., Feng, X., and Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, **21**(1).

[95] Li, H., Wang, S., Chai, S., Yang, Z., Zhang, Q., Xin, H., Xu, Y., Lin, S., Chen, X., Yao, Z., Yang, Q., Fei, Z., Huang, S., and Zhang, Z. (2022). Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. *Nature Communications*, **13**(1), 682.

[96] Li, H., Marin, M., and Farhat, M. R. (2024a). Exploring gene content with pangene graphs. *Bioinformatics*, **40**(7).

[97] Li, J., Schmelzle, J.-N., Du, Y., Heumos, S., Guarracino, A., Guidi, G., Prins, P., Garrison, E., and Zhiru, Z. (2024b). Rapid GPU-Based Pangenome Graph Layout. *International Conference for High Performance Computing, Networking, Storange, and Analysis (SC)*.

[98] Li, N., He, Q., Wang, J., Wang, B., Zhao, J., Huang, S., Yang, T., Tang, Y., Yang, S., Aisimutuola, P., Xu, R., Hu, J., Jia, C., Ma, K., Li, Z., Jiang, F., Gao, J., Lan, H., Zhou, Y., Zhang, X., Huang, S., Fei, Z., Wang, H., Li, H., and Yu, Q. (2023). Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. *Nature Genetics*, **55**(5), 852–860.

[99] Lian, Q., Huettel, B., Walkemeier, B., Mayjonade, B., Lopez-Roques, C., Gil, L., Roux, F., Schneeberger, K., and Mercier, R. (2024). A pan-genome of 69 arabidopsis thaliana accessions reveals a conserved genome structure throughout the global species range. *Nature Genetics*, **56**(5), 982–991.

[100] Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., Garg, S., Groza, C., Guarracino, A., Harvey, W. T., Heumos, S., Howe, K., Jain, M., Lu, T.-Y., Markello, C., Martin, F. J., Mitchell, M. W., Munson, K. M., Mwaniki, M. N., Novak, A. M., Olsen, H. E., Pesout, T., Porubsky, D., Prins, P., Sibbesen, J. A., Sirén, J., Tomlinson, C., Villani, F., Vollger, M. R., Antonacci-Fulton, L. L., Baid, G., Baker, C. A., Belyaeva, A., Billis, K., Carroll, A., Chang, P.-C., Cody, S., Cook, D. E., Cook-Deegan, R. M., Cornejo, O. E., Diekhans, M., Ebert, P., Fairley, S., Fedrigo, O., Felsenfeld, A. L., Formenti, G., Frankish, A., Gao, Y., Garrison, N. A., Giron, C. G., Green, R. E., Haggerty, L., Hoekzema, K., Hourlier, T., Ji, H. P., Kenny, E. E., Koenig, B. A., Kolesnikov, A., Korbel, J. O., Kordosky, J., Koren, S., Lee, H., Lewis, A. P., Magalhães, H., Marco-Sola, S., Marijon, P., McCartney, A., McDaniel, J., Mountcastle, J., Nattestad, M., Nurk, S., Olson, N. D., Popejoy, A. B., Puiu, D., Rautiainen, M., Regier, A. A., Rhie, A., Sacco, S., Sanders, A. D., Schneider, V. A., Schultz, B. I., Shafin, K., Smith, M. W., Sofia, H. J., Abou Tayoun, A. N., Thibaud-Nissen, F., Tricomi, F. F., Wagner, J., Walenz, B., Wood, J. M. D., Zimin, A. V., Bourque, G., Chaisson, M. J. P., Flicek, P., Phillippy, A. M., Zook, J. M., Eichler, E. E., Haussler, D., Wang, T., Jarvis, E. D., Miga, K. H., Garrison, E., Marschall, T., Hall, I. M., Li, H., and Paten, B. (2023). A draft human pangenome reference. *Nature*, **617**(7960), 312–324.

[101] Liao, X., Makris, M., and Luo, X. M. (2016). Fluorescence-activated cell sorting for purification of plasmacytoid dendritic cells from the mouse bone marrow. *Journal of Visualized Experiments*, (117).

[102] Liu, Y., Beyer, A., and Aebersold, R. (2016). On the dependency of cellular protein levels on mrna abundance. *Cell*, **165**(3), 535–550.

[103] Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.-A., Zhang, H., Liu, Z., Shi, M., Huang, X., Li, Y., Zhang, M., Wang, Z., Zhu, B., Han, B., Liang, C., and Tian, Z. (2020). Pan-genome of wild and cultivated soybeans. *Cell*, **182**(1), 162–176.e13.

[104] Logsdon, G. A., Rozanski, A. N., Ryabov, F., Potapova, T., Shepelev, V. A., Catacchio, C. R., Porubsky, D., Mao, Y., Yoo, D., Rautiainen, M., Koren, S., Nurk, S., Lucas, J. K., Hoekzema, K., Munson, K. M., Gerton, J. L., Phillippy, A. M., Ventura, M., Alexandrov, I. A., and Eichler, E. E. (2024). The variation and evolution of complete human centromeres. *Nature*, **629**(8010), 136–145.

[105] Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., Fleming, J., Siminoff, L., Traino, H., Mosavel, M., Barker, L., Jewell, S., Rohrer, D., Maxim, D., Filkins, D., Harbach, P., Cortadillo, E., Berghuis, B., Turner, L., Hudson, E., Feenstra, K., Sobin, L., Robb, J., Branton, P., Korzeniewski, G., Shive, C., Tabor, D., Qi, L., Groch, K., Nampally, S., Buia, S., Zimmerman, A., Smith, A., Burges, R., Robinson, K., Valentino, K., Bradbury, D., Cosentino, M., Diaz-Mayoral, N., Kennedy, M., Engel, T., Williams, P., Erickson, K., Ardlie, K., Winckler, W., Getz, G., DeLuca, D., MacArthur, D., Kellis, M., Thomson, A., Young, T., Gelfand, E., Donovan, M., Meng, Y., Grant, G., Mash, D., Marcus, Y., Basile, M., Liu, J., Zhu, J., Tu, Z., Cox, N. J., Nicolae, D. L., Gamazon, E. R., Im, H. K., Konkashbaev, A., Pritchard, J., Stevens, M., Flutre, T., Wen, X., Dermitzakis, E. T., Lappalainen, T., Guigo, R., Monlong, J., Sammeth, M., Koller, D., Battle, A., Mostafavi, S., McCarthy, M., Rivas, M., Maller, J., Rusyn, I., Nobel, A., Wright, F., Shabalin, A., Feolo, M., Sharopova, N., Sturcke, A., Paschal, J., Anderson, J. M., Wilder, E. L., Derr, L. K., Green, E. D., Struewing, J. P., Temple, G., Volpi, S., Boyer, J. T., Thomson, E. J., Guyer, M. S., Ng, C., Abdallah, A., Colantuoni, D., Insel, T. R., Koester, S. E., Little, A. R., Bender, P. K., Lehner, T., Yao, Y., Compton, C. C., Vaught, J. B., Sawyer, S., Lockhart, N. C., Demchok, J., and Moore, H. F. (2013). The genotype-tissue expression (gtex) project. *Nature Genetics*, **45**(6), 580–585.

[106] Mendel, G. (1866). Versuche über pflanzen-hybriden. *Verhandlungen des Naturforschenden Vereins zu Brünn*, **4**, 3–47. Reprinted in various editions; translated as "Experiments on Plant Hybridization" in Proceedings of the Natural History Society of Brünn.

[107] Meng, C., Kuster, B., Culhane, A. C., and Gholami, A. M. (2014). A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*, **15**(1).

[108] Mikheyev, A. S. and Tin, M. M. Y. (2014). A first look at the oxford nanopore minion sequencer. *Molecular Ecology Resources*, **14**(6), 1097–1102.

[109] Milia, S., Leonard, A. S., Mapel, X. M., Bernal Ulloa, S. M., Drögemüller, C., and Pausch, H. (2024). Taurine pangenome uncovers a segmental duplication upstream of kit associated with depigmentation in white-headed cattle. *bioRxiv*.

[110] Minkin, I., Pham, S., and Medvedev, P. (2016). Twopaco: an efficient algorithm to build the compacted de bruijn graph from many complete genomes. *Bioinformatics*, **33**(24), 4024–4032.

[111] Noll, N., Molari, M., Shaw, L. P., and Neher, R. A. (2022). Pangraph: scalable bacterial pan-genome graph construction. *bioRxiv*.

[112] Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., Caldas, G. V., Chen, N.-C., Cheng, H., Chin, C.-S., Chow, W., de Lima, L. G., Dishuck, P. C., Durbin, R., Dvorkina, T., Fiddes, I. T., Formenti, G., Fulton, R. S., Fungtammasan, A., Garrison, E., Grady, P. G. S., Graves-Lindsay, T. A., Hall, I. M., Hansen, N. F., Hartley, G. A., Haukness, M., Howe, K., Hunkapiller, M. W., Jain, C., Jain, M., Jarvis, E. D., Kerpedjiev, P., Kirsche, M., Kolmogorov, M., Korlach, J., Kremitzki, M., Li, H., Maduro, V. V., Marschall, T., McCartney, A. M., McDaniel, J., Miller, D. E., Mullikin, J. C., Myers, E. W., Olson, N. D., Paten, B., Peluso, P., Pevzner, P. A., Porubsky, D., Potapova, T., Rogaev, E. I., Rosenfeld, J. A., Salzberg, S. L., Schneider, V. A., Sedlazeck, F. J., Shafin, K., Shew, C. J., Shumate, A., Sims, Y., Smit, A. F. A., Soto, D. C., Sović, I., Storer, J. M., Streets, A., Sullivan, B. A., Thibaud-Nissen, F., Torrance, J., Wagner, J., Walenz, B. P., Wenger, A., Wood, J. M. D., Xiao, C., Yan, S. M., Young, A. C., Zarate, S., Surti, U., McCoy, R. C., Dennis, M. Y., Alexandrov, I. A., Gerton, J. L., O'Neill, R. J., Timp, W., Zook, J. M., Schatz, M. C., Eichler, E. E., Miga, K. H., and Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, **376**(6588), 44–53.

[113] Perez-Riverol, Y., Bai, M., da Veiga Leprevost, F., Squizzato, S., Park, Y. M., Haug, K., Carroll, A. J., Spalding, D., Paschall, J., Wang, M., del Toro, N., Ternent, T., Zhang, P., Buso, N., Bandeira, N., Deutsch, E. W., Campbell, D. S., Beavis, R. C., Salek, R. M., Sarkans, U., Petryszak, R., Keays, M., Fahy, E., Sud, M., Subramaniam, S., Barbera, A., Jiménez, R. C., Nesvizhskii, A. I., Sansone, S.-A., Steinbeck, C., Lopez, R., Vizcaíno, J. A., Ping, P., and Hermjakob, H. (2017). Discovering and linking public omics data sets using the omics discovery index. *Nature Biotechnology*, **35**(5), 406–409.

[114] Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., He, Q., Ou, S., Zhang, H., Li, X., Li, X., Li, Y., Liao, Y., Gao, Q., Tu, B., Yuan, H., Ma, B., Wang, Y., Qian, Y., Fan, S., Li, W., Wang, J., He, M., Yin, J., Li, T., Jiang, N., Chen, X., Liang, C., and Li, S. (2021). Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell*, **184**(13), 3542–3558.e16.

[115] Rautiainen, M. and Marschall, T. (2020). GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biology*, **21**(1).

[116] Recht, B., Re, C., Wright, S., and Niu, F. (2011). Hogwild!: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

[117] Rhie, A., Nurk, S., Cechova, M., Hoyt, S. J., Taylor, D. J., Altemose, N., Hook, P. W., Koren, S., Rautiainen, M., Alexandrov, I. A., Allen, J., Asri, M., Bzikadze, A. V., Chen, N.-C., Chin, C.-S., Diekhans, M., Flicek, P., Formenti, G., Fungtammasan, A., Garcia Giron, C., Garrison, E., Gershman, A., Gerton, J. L., Grady, P. G. S., Guarracino, A., Haggerty, L., Halabian, R., Hansen, N. F., Harris, R., Hartley, G. A., Harvey, W. T., Haukness, M., Heinz, J., Hourlier, T., Hubley, R. M., Hunt, S. E., Hwang, S., Jain, M., Kesharwani, R. K., Lewis, A. P., Li, H., Logsdon, G. A., Lucas, J. K., Makalowski, W., Markovic, C., Martin, F. J., Mc Cartney, A. M., McCoy, R. C., McDaniel, J., McNulty, B. M., Medvedev, P., Mikheenko, A., Munson, K. M., Murphy, T. D., Olsen, H. E., Olson, N. D., Paulin, L. F., Porubsky, D., Potapova, T., Ryabov, F., Salzberg, S. L., Sauria, M. E. G., Sedlazeck, F. J., Shafin, K., Shepelev, V. A., Shumate, A., Storer, J. M., Surapaneni, L., Taravella Oill, A. M., Thibaud-Nissen, F., Timp, W., Tomaszkiewicz, M., Vollger, M. R., Walenz, B. P., Watwood, A. C., Weissensteiner, M. H., Wenger, A. M., Wilson, M. A., Zarate, S., Zhu, Y., Zook, J. M., Eichler, E. E., O'Neill, R. J., Schatz, M. C., Miga, K. H., Makova, K. D., and Phillippy, A. M. (2023). The complete sequence of a human Y chromosome. *Nature*, **621**(7978), 344–354.

[118] Ribas, A. and Wolchok, J. D. (2018). Cancer immunotherapy using checkpoint blockade. *Science*, **359**(6382), 1350–1355.

[119] Rice, E. S., Alberdi, A., Alfieri, J., Athrey, G., Balacco, J. R., Bardou, P., Blackmon, H., Charles, M., Cheng, H. H., Fedrigo, O., Fiddaman, S. R., Formenti, G., Frantz, L. A. F., Gilbert, M. T. P., Hearn, C. J., Jarvis, E. D., Klopp, C., Marcos, S., Mason, A. S., Velez-Irizarry, D., Xu, L., and Warren, W. C. (2023). A pangenome graph reference of 30 chicken genomes allows genotyping of large and complex structural variants. *BMC Biology*, **21**(1).

[120] Sandhu, C., Qureshi, A., and Emili, A. (2018). Panomics for precision medicine. *Trends in Molecular Medicine*, **24**(1), 85–101.

[121] SanDiegOmics (2024). 2023 Sequencing Market Share – The Tide is Turning... Very Slowly. `https://sandiegomics.com/2023-sequencing-market-share-the-tide-is-turning-very-slowly/` (last accessed August 2024).

[122] Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, **74**(12), 5463–5467.

[123] Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R. P., Banday, S., Mishra, A. K., Das, G., and Malonia, S. K. (2023). Next-generation sequencing technology: Current trends and advancements. *Biology*, **12**(7), 997.

[124] Schneeberger, K., Hagmann, J., Ossowski, S., Warthmann, N., Gesing, S., Kohlbacher, O., and Weigel, D. (2009). Simultaneous alignment of short reads against multiple genomes. *Genome Biology*, **10**(9), R98.

[125] Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., Fulton, R. S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K. M., Auger, K., Chow, W., Collins, J., Harden, G., Hubbard, T., Pelan, S., Simpson, J. T., Threadgold, G., Torrance, J., Wood, J. M., Clarke, L., Koren, S., Boitano, M., Peluso, P., Li, H., Chin, C.-S., Phillippy, A. M., Durbin, R., Wilson, R. K., Flicek, P., Eichler, E. E., and Church, D. M. (2017). Evaluation of grch38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, **27**(5), 849–864.

[126] Sedlmair, M., Meyer, M., and Munzner, T. (2012). Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics*, **18**(12), 2431–2440.

[127] Selves, J., Long-Mira, E., Mathieu, M.-C., Rochaix, P., and Ilié, M. (2018). Immunohistochemistry for diagnosis of metastatic carcinomas of unknown primary site. *Cancers*, **10**(4), 108.

[128] Shang, L., Li, X., He, H., Yuan, Q., Song, Y., Wei, Z., Lin, H., Hu, M., Zhao, F., Zhang, C., Li, Y., Gao, H., Wang, T., Liu, X., Zhang, H., Zhang, Y., Cao, S., Yu, X., Zhang, B., Zhang, Y., Tan, Y., Qin, M., Ai, C., Yang, Y., Zhang, B., Hu, Z., Wang, H., Lv, Y., Wang, Y., Ma, J., Wang, Q., Lu, H., Wu, Z., Liu, S., Sun, Z., Zhang, H., Guo, L., Li, Z., Zhou, Y., Li, J., Zhu, Z., Xiong, G., Ruan, J., and Qian, Q. (2022). A super pan-genomic landscape of rice. *Cell Research*, **32**(10), 878–896.

[129] Sheikhizadeh, S., Schranz, M. E., Akdel, M., de Ridder, D., and Smit, S. (2016). Pantools: representation, storage and exploration of pan-genomic data. *Bioinformatics*, **32**(17), i487–i493.

[130] Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, **25**(22), 2906–2912.

[131] Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., Boorgula, M. P., Chavan, S., Vergara, C., Ortega, V. E., Levin, A. M., Eng, C., Yazdanbakhsh, M., Wilson, J. G., Marrugo, J., Lange, L. A., Williams, L. K., Watson, H., Ware, L. B., Olopade, C. O., Olopade, O., Oliveira, R. R., Ober, C., Nicolae, D. L., Meyers, D. A., Mayorga, A., Knight-Madden, J., Hartert, T., Hansel, N. N., Foreman, M. G., Ford, J. G., Faruque, M. U., Dunston, G. M., Caraballo, L., Burchard, E. G., Bleecker, E. R., Araujo, M. I., Herrera-Paz, E. F., Campbell, M., Foster, C., Taub, M. A., Beaty, T. H., Ruczinski, I., Mathias, R. A., Barnes, K. C., and Salzberg, S. L. (2018). Assembly of a pan-genome from deep sequencing of 910 humans of african descent. *Nature Genetics*, **51**(1), 30–35.

[132] Shoemaker, R. H. (2006). The nci60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer*, **6**(10), 813–823.

[133] Sibbesen, J. A., Eizenga, J. M., Novak, A. M., Sirén, J., Chang, X., Garrison, E., and Paten, B. (2023). Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. *Nature Methods*, **20**(2), 239–247.

[134] Singh, V., Pandey, S., and Bhardwaj, A. (2022). From the reference human genome to human pangenome: Premise, promise and challenge. *Frontiers in Genetics*, **13**.

[135] Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., Sibbesen, J. A., Hickey, G., Chang, P.-C., Carroll, A., Gupta, N., Gabriel, S., Blackwell, T. W., Ratan, A., Taylor, K. D., Rich, S. S., Rotter, J. I., Haussler, D., Garrison, E., and Paten, B. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, **374**(6574).

[136] Sirén, J., Eskandar, P., Ungaro, M. T., Hickey, G., Eizenga, J. M., Novak, A. M., Chang, X., Chang, P.-C., Kolmogorov, M., Carroll, A., Monlong, J., and Paten, B. (2024). Personalized pangenome references. *Nature Methods*.

[137] Slatko, B. E., Gardner, A. F., and Ausubel, F. M. (2018). Overview of next-generation sequencing technologies. *Current Protocols in Molecular Biology*, **122**(1).

[138] Stammnitz, M. R., Gori, K., Kwon, Y. M., Harry, E., Martin, F. J., Billis, K., Cheng, Y., Baez-Ortega, A., Chow, W., Comte, S., Eggertsson, H., Fox, S., Hamede, R., Jones, M., Lazenby, B., Peck, S., Pye, R., Quail, M. A., Swift, K., Wang, J., Wood, J., Howe, K., Stratton, M. R., Ning, Z., and Murchison, E. P. (2023). The evolution of two transmissible cancers in tasmanian devils. *Science*, **380**(6642), 283–293.

[139] Stanislaus, R., Carey, M., Deus, H. F., Coombes, K., Hennessy, B. T., Mills, G. B., and Almeida, J. S. (2008). Rppaml/rims: A metadata format and an information management system for reverse phase protein arrays. *BMC Bioinformatics*, **9**(1).

[140] Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, **14**(9), 865–868.

[141] Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. *Bioinformatics and Biology Insights*, **14**, 117793221989905.

[142] Swain, S. M., Shastry, M., and Hamilton, E. (2022). Targeting her2-positive breast cancer: advances and future directions. *Nature Reviews Drug Discovery*, **22**(2), 101–126.

[143] Szalai, C. (2023). Arguments for and against the whole-genome sequencing of newborns. *Am. J. Transl. Res.*, **15**(10), 6255–6263.

[144] Talenti, A., Powell, J., Hemmink, J. D., Cook, E. A. J., Wragg, D., Jayaraman, S., Paxton, E., Ezeasor, C., Obishakin, E. T., Agusi, E. R., Tijjani, A., Marshall, K., Fisch, A., Ferreira, B. R., Qasim, A., Chaudhry, U., Wiener, P., Toye, P., Morrison, L. J., Connelley, T., and Prendergast, J. G. D. (2022). A cattle graph genome incorporating global breed diversity. *Nature Communications*, **13**(1), 910.

[145] Tarazona, S., Balzano-Nogueira, L., and Conesa, A. (2018). *Multiomics Data Integration in Time Series Experiments*, page 505–532. Elsevier.

[146] Tarazona, S., Arzalluz-Luque, A., and Conesa, A. (2021). Undisclosed, unmet and neglected challenges in multi-omics studies. *Nature Computational Science*, **1**(6), 395–402.

[147] TAUB, FLOYD, E., DeLEO, J. M., and THOMPSON, E. B. (1983). Sequential comparative hybridizations analyzed by computerized image processing can identify and quantitate regulated rnas. *DNA*, **2**(4), 309–327.

[148] THIS PODCAST WILL KILL YOU (2024). Episode 147 Tasmanian Devil Facial Tumor Disease: Sympathy for the Devil. `https://thispodcastwillkillyou.com/2024/07/30/episode-147-tasmanian-devil-facial-tumor-disease-sympathy-for-the-devil/` (last accessed August 2024).

[149] Thul, P. J. and Lindskog, C. (2017). The human protein atlas: A spatial map of the human proteome. *Protein Science*, **27**(1), 233–244.

[150] Treangen, T. J. and Salzberg, S. L. (2011). Repetitive dna and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, **13**(1), 36–46.

[151] Treangen Lab (2022). Some of our favorite papers from 2021. `https://www.treangenlab.com/post/2021-papers/#pangenomes` (last accessed August 2024).

[152] Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., Wernerus, H., Björling, L., and Ponten, F. (2010). Towards a knowledge-based human protein atlas. *Nature Biotechnology*, **28**(12), 1248–1250.

[153] Uhlen, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., and Pontén, F. (2015). Tissue-based map of the human proteome. *Science*, **347**(6220).

[154] Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., and Pontén, F. (2015). Tissue-based map of the human proteome. *Science*, **347**(6220).

[155] Vakili, D., Radenkovic, D., Chawla, S., and Bhatt, D. L. (2021). Panomics: New databases for advancing cardiology. *Frontiers in Cardiovascular Medicine*, **8**.

[156] van den Brandt, A., Jonkheer, E. M., van Workum, D.-J. M., van de Wetering, H., Smit, S., and Vilanova, A. (2024). Panva: Pangenomic variant analysis. *IEEE Transactions on Visualization and Computer Graphics*, **30**(8), 4895–4909.

[157] van Dijk, L. R., Manson, A. L., Earl, A. M., Garimella, K. V., and Abeel, T. (2024). Fast and exact gap-affine partial order alignment with POASTA. *bioRxiv*.

[158] Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, **27**(5), 737–746.

[159] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R.,

Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S. C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science*, **291**(5507), 1304–1351.

[160] Villani, F., Guarracino, A., Ward, R. R., Green, T., Emms, M., Pravenec, M., Prins, P., Garrison, E., Williams, R. W., Chen, H., and Colonna, V. (2024). Pangenome reconstruction in rats enhances genotype-phenotype mapping and novel variant discovery. *bioRxiv*.

[161] Vivian, J., Rao, A. A., Nothaft, F. A., Ketchum, C., Armstrong, J., Novak, A.,

Pfeil, J., Narkizian, J., Deran, A. D., Musselman-Brown, A., Schmidt, H., Amstutz, P., Craft, B., Goldman, M., Rosenbloom, K., Cline, M., O'Connor, B., Hanna, M., Birger, C., Kent, W. J., Patterson, D. A., Joseph, A. D., Zhu, J., Zaranek, S., Getz, G., Haussler, D., and Paten, B. (2017). Toil enables reproducible, open source, big biomedical data analyses. *Nature Biotechnology*, **35**(4), 314–316.

[162] Wang, Y., Wang, Z., Chen, Y., Lan, T., Wang, X., Liu, G., Xin, M., Hu, Z., Yao, Y., Ni, Z., Sun, Q., Guo, W., and Peng, H. (2024). Genomic insights into the origin and evolution of spelt (triticum spelta l.) as a valuable gene pool for modern wheat breeding. *Plant Communications*, **5**(5), 100883.

[163] Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**(1), 57–63.

[164] Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, **45**(10), 1113–1120.

[165] Welcome Sanger Institute (2024). Researchers aim to analyse pangenomes using quantum computing. `https://www.sanger.ac.uk/news_item/researchers-aim-to-analyse-pangenomes-using-quantum-computing/` (last accessed August 2024).

[166] Wick, R. R., Schultz, M. B., Zobel, J., and Holt, K. E. (2015). Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, **31**(20), 3350–3352.

[167] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific Data*, **3**(1).

[168] Wratten, L., Wilm, A., and Göke, J. (2021). Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature Methods*, **18**(10), 1161–1168.

[169] Xu, X., Zhang, Q., Li, M., Lin, S., Liang, S., Cai, L., Zhu, H., Su, R., and Yang, C. (2022). Microfluidic single-cell multiomics analysis. *VIEW*, **4**(1).

[170] Yang, X., Lee, W.-P., Ye, K., and Lee, C. (2019). One reference genome is not enough. *Genome Biology*, **20**(1).

[171] Yang, Z., Guarracino, A., Biggs, P. J., Black, M. A., Ismail, N., Wold, J. R., Merriman, T. R., Prins, P., Garrison, E., and de Ligt, J. (2023). Pangenome graphs in infectious disease: a comprehensive genetic variation analysis of neisseria meningitidis leveraging oxford nanopore long reads. *Frontiers in Genetics*, **14**.

[172] Yokoyama, T. T., Sakamoto, Y., Seki, M., Suzuki, Y., and Kasahara, M. (2019). MoMI-G: modular multi-scale integrated genome graph browser. *BMC Bioinformatics*, **20**(1), 548.

[173] Zhang, Z. H., Jhaveri, D. J., Marshall, V. M., Bauer, D. C., Edson, J., Narayanan, R. K., Robinson, G. J., Lundberg, A. E., Bartlett, P. F., Wray, N. R., and Zhao, Q.-Y. (2014). A comparative study of techniques for differential expression analysis on rna-seq data. *PLoS ONE*, **9**(8), e103207.

[174] Zheng, J. X., Pawar, S., and Goodman, D. F. (2019). Graph drawing by stochastic gradient descent. *IEEE Transactions on Visualization and Computer Graphics*, **25**(9), 2738–2748.

[175] Zhong, C., Chen, C., Wang, L., and Ning, K. (2021). Integrating pan-genome with metagenome for microbial community profiling. *Computational and Structural Biotechnology Journal*, **19**, 1458–1466.

[176] Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., Wu, Y., Cheng, L., Fang, Y., Wu, K., Zhang, J., Lyu, H., Lin, T., Gao, Q., Saha, S., Mueller, L., Fei, Z., Städler, T., Xu, S., Zhang, Z., Speed, D., and Huang, S. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. *Nature*, **606**(7914), 527–534.

[177] Zipf, G. K. (1932). *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press, Cambridge, MA and London, England.

Appendix A

Appendix

A.1 Awards

1. IGGSy 2024 in Ascona, Switzerland: Student travel award 700 CHF.
2. ISMB Bio-Ontologies 2020, virtual: Best poster prize for *Semantic Variation Graphs: Ontologies for Pangenome Graphs*.

A.2 Invitations

1. Japan NBDC/DBCLS Biohackathon 2019 in Fukuoka, Japan: Invited long talk symposium speaker *VG Browser: Interactive Visualization of Genome Variation Graphs*.
2. Institute for Medical Biometry and Bioinformatics October 2021 in Düsseldorf, Germany: Invited Talk *Exploring pangenome graphs and possible applications*.
3. MemPanG23 in Memphis, TN, USA: Invited organizer, instructor, and chair of pangenomics course and symposium *MemPanG23*.
4. Nextflow Summit 2023 in Barcelona, Spain: Invited talk *Cluster scalable pangenome graph construction with nf-core/pangenome*.
5. HPRC HUGO24 Workshop in Rome, Italy: Invited instructor of pangenomics course for the HPRC.
6. MemPanG24 in Memphis, TN, USA: Invited organizer, instructor, and chair of pangenomics course and symposium *MemPanG24*.

A.3 Grants

1. Research grant *Pantograph* by the Ministry of Economics and Energy (BMWi): 190,000€ to research pangenome graph visualization. 2019.

A.4 Teaching

A.4.1 QBiC

1. Data Management for Quantitative Biology Summer 2020: Tutor.
2. Grundlagen der Bioinformatik Summer 2022: Tutor.
3. Biomedical Data Management Summer 2023: Tutor.
4. M3 Workshop 2024: Speaker.

A.4.2 External

1. Utrecht Bioinformatics Center 2022, virtual: Teaching assistant virtual course *Advanced Bioinformatics: data mining and data integration for life sciences*.
2. MemPanG23 in Memphis, TN, USA: Invited organizer, instructor, and chair of pangenomics course and symposium *MemPanG23*.
3. HPRC HUGO24 Workshop in Rome, Italy: Invited instructor of pangenomics course for the HPRC.
4. MemPanG24 in Memphis, TN, USA: Invited organizer, instructor, and chair of pangenomics course and symposium *MemPanG24*.

A.5 Mentoring

1. BSc supervision *Die Konstruktion eines Lodderomyces elongisporus Pangenomgraphen*. May23-Aug23. Tübingen, Germany.
2. MSc supervision *Joining medical data and pangenome graphs using the semantic web*. Oct23-Mar24. Tübingen, Germany.

A.6 Hackathons

1. NBDC/DBCLS Biohackathon 2019 in Fukuoka, Japan: Co-Project leader *Pantograph*, playing around with SequenceTubeMap and SPARQL.
2. Computomics Hackathon November 2019 in Tübingen, Germany: Progressing *Pantograph*, playing around with pangenome graphs and SPARQL.
3. COVID-19 Biohackathon 2020, virtual: Co-Project leader *Pangenome Browser* and *Pangenome Ontology*.
4. Crusco Biohackathon August 2020 in Lavello, Italy: Progressing *PG-SGD* with Andrea Guarracino and Erik Garrison in Lavello.
5. ELIXIR Europe BioHackathon 2020, virtual: Project leader *Federated Interoperable Annotated Variation Graphs*.
6. nf-core/hackathon March 2021, virtual: Starting nf-core/pangenome.
7. ELIXIR Europe BioHackathon 2021 in Barcelona, Spain: progressing nf-core/pangenome.

8. Pangenomics Bio Hacking 2021, virtual: Pangenome expert and participant.
9. nf-core/hackathon March 2022, virtual: Progressing nf-core/pangenome.
10. nf-core/hackathon March 2023, virtual: Progressing nf-core/pangenome.
11. nf-core/hackathon October 2023 in Barcelona, Spain: Progressing nf-core/pangenome.
12. nf-core/hackathon March 2024, virtual and Tübingen, Germany: Co-team leader group *pipelines*. Finalizing nf-core/pangenome.

A.7 Miscellaneous

1. Associate member of the Human Pangenome Reference Consortium.
2. Reviewer for Oxford Bioinformatics.
3. Reviewer for United Kingdom Research and Innovation (UKRI).

A.8 Printouts Of Core Publications

Publications contributing to this doctoral thesis, as listed in Chapter 1.1, are included in the following appendix. The corresponding citations can be found in Chapter 1.1. All publications are printed as published.

A.8.1 License Information

Heumos *et al. Cancer Cell International* (2022) 22:311
https://doi.org/10.1186/s12935-022-02710-y

Cancer Cell International

RESEARCH Open Access

Multiomics surface receptor profiling of the NCI-60 tumor cell panel uncovers novel theranostics for cancer immunotherapy

Simon Heumos[1,2†], Sandra Dehn[3†], Konstantin Bräutigam[4], Marius C. Codrea[1], Christian M. Schürch[5], Ulrich M. Lauer[6,7], Sven Nahnsen[1,2] and Michael Schindler[3*]

Abstract

Background: Immunotherapy with immune checkpoint inhibitors (ICI) has revolutionized cancer therapy. However, therapeutic targeting of inhibitory T cell receptors such as PD-1 not only initiates a broad immune response against tumors, but also causes severe adverse effects. An ideal future stratified immunotherapy would interfere with cancer-specific cell surface receptors only.

Methods: To identify such candidates, we profiled the surface receptors of the NCI-60 tumor cell panel via flow cytometry. The resulting surface receptor expression data were integrated into proteomic and transcriptomic NCI-60 datasets applying a sophisticated multiomics multiple co-inertia analysis (MCIA). This allowed us to identify surface profiles for skin, brain, colon, kidney, and bone marrow derived cell lines and cancer entity-specific cell surface receptor biomarkers for colon and renal cancer.

Results: For colon cancer, identified biomarkers are CD15, CD104, CD324, CD326, CD49f, and for renal cancer, CD24, CD26, CD106 (VCAM1), EGFR, SSEA-3 (B3GALT5), SSEA-4 (TMCC1), TIM1 (HAVCR1), and TRA-1-60R (PODXL). Further data mining revealed that CD106 (VCAM1) in particular is a promising novel immunotherapeutic target for the treatment of renal cancer.

Conclusion: Altogether, our innovative multiomics analysis of the NCI-60 panel represents a highly valuable resource for uncovering surface receptors that could be further exploited for diagnostic and therapeutic purposes in the context of cancer immunotherapy.

Keywords: Immunotherapy, cancer, Multiomics, Theranostics, Flow cytometry, FACS, NCI-60, Receptorome

Background

Implementation of cancer immunotherapy by immune checkpoint inhibitors (ICIs) is one of the most recent transforming developments in oncology that strongly helped to improve overall survival of patients suffering from various cancers [1]. Its principle is based on the antibody-mediated blockage of inhibitory immune signaling exerted by tumor cells to unleash the immune system, with the overall goal to achieve a sustainable tumor elimination by the host's intrinsic immune response. The first established ICIs target the PD1-PDL1 inhibitory axis on T cells to activate the cytotoxic T lymphocyte (CTL) response [1] and block signaling of the T cell inhibitory receptor CTLA-4 [2]. In this context, alternative strategies are to activate not only T cells but also NK-cells [3], or to target general immune-inhibitory signaling axes,

†Simon Heumos and Sandra Dehn contributed equally to this work

*Correspondence: michael.schindler@med.uni-tuebingen.de

[3] Institute for Medical Virology and Epidemiology of Viral Diseases, University Hospital Tübingen, Tübingen, Germany
Full list of author information is available at the end of the article

i.e., the CD47-SIRPα pathway [4]. Conceivably, due to the nature of this approach, cancer immunotherapy via ICIs is unspecific, can have severe adverse effects and has a certain risk of complete therapy failure or could even result in an aggravation of the addressed malignancies [5]. Hence, there are ongoing efforts to improve immunotherapies, especially in terms of specificity, so that only or at least mainly tumor cells are killed [6].

Such approaches are based, for example, on bispecific antibodies that bind to a tumor antigen and are designed to crosslink T cells to the tumor cells via secondary binding to CD3 or another T cell-specific antigen [7]. In this way, cytotoxic T cells should be specifically recruited to and kill malignant tumor cells. This concept is further elaborated to recruit NK cells, and also has been developed in the context of chimeric antigen receptor (CAR) T cells [7, 8]. CAR T cells are engineered to express an artificial T cell receptor tumor-specific antigen. Altogether, efforts are undertaken to improve efficacy of immunotherapy and reduce side-effects by specific targeting of malignant tumor cells.

The key to improving and precisely targeting cancer immunotherapy is the knowledge of tumor-specific biomarkers accessible at the cell surface. Therefore, we hypothesized that a novel systematic screening of cancer cell lines using a distinct comprehensive flow cytometric approach should enable the identification of hitherto unidentified cancer entity-specific cell surface receptors. For this, we took advantage of the NCI-60 tumor cell panel, a collection of 60 different human cancer cell lines that were established to facilitate systematic screening of antitumor drugs (https://dtp.cancer.gov/discovery_development/nci-60/cell_list.htm) [9–12].

We systematically characterized the expression of 332 receptors on the surface of the NCI-60 tumor cell collection using an array of flow cytometry-applicable antibodies. The NCI-60 panel has already been comprehensively characterized via transcriptomics and proteomics [13, 14]. While these latter approaches facilitate the identification of tumor biomarkers, they do not give any information on differential cell surface expression which is a prerequisite to exploit them as immunotherapeutic targets. Therefore, building on these high-quality public data sources, we analyzed the receptorome using flow cytometry and present an integrated three-layer multiomics approach. As a result of our cell surface receptor profiling using the NCI-60 tumor cell panel, we identified tumor biomarkers and immunotherapeutic targets that are readily accessible on the surface of human cancer cells via well-characterized antibodies. We anticipate new avenues for the development of highly specific and targeted immunotherapeutic approaches using the presented data as a resource.

Methods

Cell culture

The NCI-60 tumor cell panel from the US National Cancer Institute was purchased from Charles River Laboratories (Charles River Laboratories Inc., New York, NY, USA). All cell lines were cultivated in RPMI-1640 medium supplemented with 10% fetal calf serum (FCS), 2 mM L-glutamine and 100 µg ml^{-1} penicillin–streptomycin. Cells were cultured at 37 °C in an atmosphere of 5% CO_2.

Flow cytometric cell surface receptor screening

Before NCI-60 cells were used for flow cytometric analyses they were cultured from nitrogen stocks and allowed to grow for at least 2 weeks (four to five passages at maximum). Cells were detached by Accutase treatment and stained with the LegendScreen Human PE kit (BioLegend) using 332 PE-conjugated antibodies essentially as described before [15, 16]. Cell surface expression of the 332 receptors was measured via flow cytometry using the MACSQuant VYB Analyzer (Miltenyi Biotec). Flow cytometry data was analyzed with the FlowLogic (Miltenyi-Inivai) software to obtain the mean fluorescence intensity (MFI) values of each analyzed receptor.

General data curation and quality control

For data quality control, to compare the receptor MFIs across 2 weeks, and for the MCIA analysis, R version 3.3.2 was used. For all other analysis the R version was 4.1.3 [17]. All analysis scripts including input and output data can be found at https://github.com/qbicsoftware/QMSFC. The repository comes with a detailed README and Anaconda environments to ensure reproducibility of the results [18]. For all data sets, as BR.MDAMB468 is not present in the microarray data, we removed it from the data set. We also removed tumor cell line ME-LOX-IMVI as it is lacking any melanin production and therefore it is most likely not a melanoma cell line [19]. We harmonized the cell names and annotated the cells to the respective tissue type [20, 21].

Flow cytometry data curation

For quality control, isotype control samples were analyzed separately from the data related to specific cell surface staining. The 10 isotype controls were removed from the full FACS data and considered individually as visible in Additional file 1: Fig. S1, Additional file 2: Fig. S2, Additional file 3: Fig. S3, Additional file 4: Fig. S4, Additional file 5: Fig. S5, Additional file 6: Fig. S6. Furthermore, for a set of cell lines, a second independent legend screen was conducted 1 week after the first sampling to check for reproducibility of the procedure, this data is available in Additional file 7: Dataset S1. The 25 cell lines of which

a 2nd week measurement was performed were independently analyzed from the original FACS data (MFI values) set. For each of the cell lines, a between paired samples correlation test using the R function correlation test with method Spearman was executed. Before testing, a Shapiro-Wilk test of normality ensured that none of the measurements follow a normal distribution. For each receptor of each cell line pair, the log2 fold change was calculated. Furthermore, a between paired samples correlation test of all week 1 MFI value versus all week 2 MFI values was conducted. Before testing, an Anderson-Darling test for normality was performed. For further data analyses week 1 measurements only were used to have consistent data for all cell lines. FACS data (MFI values) of the receptor expression of the tumor NCI-60 cell line panel was log transformed (base 10). For all other downstream analyses, the raw expression values were used that are summarized in the Additional file 8: Dataset S2.

Microarray data curation

Robust multi-array average (RMA) [22] normalized microarray data of the NCI-60 cell lines was fetched from the Gene Expression Omnibus using accession number GSE32474 [12, 23–27]. The HGNC symbols for the Affymetrix U133 Plus 2.0 chip were downloaded from the Ensemble BioMart Portal [28]. The microarray probes were annotated with the HGNC symbols. All unannotated probes were discarded. We observed 37,989 distinct HGNC-Affymetrix identifiers and 19,473 unique HGNC symbols. We merged the expression values with the same HGNC symbol using their mean.

Proteome data curation

MaxQuant [29] processed proteome data of the NCI-60 cell line panel was downloaded from PRIDE [30] using project number PXD005946 [14]. Label-free quantification (LFQ) values were log transformed (base 10), and previously identified contaminant proteins [14], labeled in the dataset with "CON" or "REV", were removed. Both cell lines HOP92 and SR were two times in the proteomics data set. For each duplicate, we removed the cell line with more missing values. As the subsequent MCIA analysis can't be performed with missing values, we only kept proteins with 60 measurements, 514 in total.

Hierarchical clustering

Hierarchical clustering was performed using Spearman correlation [31] for the distance metric with ward.D2 for the agglomeration.

MCIA

The omicade4 R package [19] was applied as an exploratory analysis to the transcriptomic study, the proteomic study and the FACS study of 58 cell lines. A customized version of the plotting function of omicade4 adds colors from the RColorBrewer package (https://CRAN.R-project.org/package=RColorBrewer). Using the information given in the sample and feature space of the MCIA, cell tissue type hits of FACS (LE, ME, CO, RE, CNS) were manually selected with selectVar. The resulting FACS receptors were annotated with gene identifiers for downstream comparison with the RNA-Seq recount2 data analysis.

RNAseq data analysis

TCGA recount2 data of CNS, RE, CO, ME, LE was downloaded from the recount2 portal. For each of these, initial data processing was performed with the recount Bioconductor package [32]. Only samples with non-empty metadata were kept. Samples were filtered for "Primary Tumor" and "Solid Tissue Normal". Only for the RE and CO tissue types there were "Solid Tissue Normal" samples. For CO, we obtained 500 tumor and 41 normal samples, and for RE we obtained 899 tumor and 129 normal samples, respectively. For both tissue type data, DESeq2 was applied in order to identify differentially expressed genes [33]. The experimental design formula was:

$$\sim gdc\_cases.demographic.gender + gdc\_cases.demographic.race + gdc\_cases.samples.sample\_type.$$

Genes were classified as differentially expressed with a p-adjusted value < 0.05.

Human Protein Atlas data analysis

Available data for tissue protein expression in the Human Protein Atlas (HPA, [34]) was used for a systematic investigation of the eight differentially expressed molecules as identified by multi-omics. The expression of those eight molecules (CD24, CD26, CD106 [VCAM1], EGFR, SSEA4 [TMCC1], TIM1 [HAVCR1], SSEA3 [B3GALT5], TRA-1-60R [PODXL]) at the protein level was compared in normal renal tubules (in tissue cores of maximum 11 patients) vs. renal cell carcinoma, clear cell and non-clear cell type (in tissue cores of maximum 40 patients). For some of these proteins, immunohistochemistry data from multiple different antibody clones were available which is shown in Additional file 9: Table S1, all of which were included in the analysis. For each patient/kidney sample, up to two different tissue cores were available. Normal and tumor kidney samples were not matched. The cohort ($n = 23$) of renal cell carcinomas comprised 21 clear cell renal cell carcinomas (91.3%) and two non-clear cell renal cell carcinomas (8.7%) for CD106 (VCAM1). For EGFR ($n = 40$), 31 clear cell renal cell carcinomas (77.5%) and nine non-clear cell renal cell carcinomas (22.5%) were included.

Immunohistochemistry staining results were visually reviewed and jointly scored for each tissue core by two pathologists (K.B. and C.M.S.). The scoring was based on staining intensity and the amount of positive cells in a three-tiered manner, as follows: Intensity: 0: no expression; 1: weak expression; 2: moderate expression; 3: strong expression. Amount of positive cells: 0: none; 1: < 25%; 2: 25–75%; and 3: > 75% positive cells. Finally, the two values were multiplied, resulting in a modified immunoreactivity score (IRS, [35]) of values ranging between 0 and 9. In cases with two available tissue cores, the mean IRS was used for further analysis.

Data were visualized using GraphPad PRISM v.9.0.0. IRS scores of normal vs. tumor tissue were compared using the unpaired Mann-Whitney test. Two-tailed exact p-values < 0.05 were considered statistically significant. Limitations of the analyses: Firstly, the tissue cores of normal kidney tissue did not match tissue cores of renal adenocarcinoma. In addition, the sample size in the Human Protein Atlas was partially very limited and possibly affected statistical analysis. Moreover, some antibodies showed paradox and contradictory staining, an important technical limitation.

The Cancer Proteome Atlas analysis

The Cancer Protein Atlas portal was accessed (https://tcpaportal.org/tcpa/differential_analysis.html) in order to explore the functional proteomics landscape of all renal cell cancer subtypes in regard to EGFR expression. On the web interface "By tumor type" and Pan-Can 32 were selected. For "Select tumor A" always "Kidney renal clear cell carcinoma (KIRC) (445 samples)" was chosen. For "Select tumor B" either "Kidney Chromophobe (KICH) (63 samples)" or "Kidney renal papillary cell carcinoma (KIRP) (208 samples)" was selected. The results were filtered by Protein Marker ID = EGFR and Gene(s) = EGFR.

Results

Cell surface receptor expression of the NCI-60 tumor cell panel

The NCI-60 tumor cell panel is a collection of 60 different human cancer cell lines representing 9 different tumor entities: leukemia (LE), lung cancer (LC), colon cancer (CO), cancer of the central nervous system (CNS), melanoma (ME), ovarian cancer (OV), renal cancer (RE), prostate cancer (PR) and breast cancer (BR).

This panel is frequently used in cancer research, and various studies have performed detailed analyses of the transcriptome and proteome of these cell lines to identify cancer-specific biomarkers or potential therapeutic targets. However, surface residing biomarkers might be missed by the latter approaches, since when employing standard-proteomics membranes are largely excluded in the non-soluble fraction and dynamic internalization processes of surface receptors, shedding and other mechanism that alter the receptorome are not detected by transcriptomics [15]. Furthermore, it remains unclear if biomarkers are also differentially expressed at the cell surface and may therefore be suitable targets for antibody-mediated immunotherapy. To close this important gap, we performed a medium-throughput flow cytometric profiling of the NCI-60 tumor cell panel with an arrayed set of 342 PE-labeled antibodies (Fig. 1 and Additional file 8). To probe for measurement reproducibility, 25 out of the 60 tumor cell lines were randomly selected for a subsequent measurement 1 week later, showing no significant differences (Spearman correlation coefficient $r = 0.91$, $p < 2.2e{-}16$) in the absolute mean fluorescence intensity (MFI)-values with same laser settings (Additional file 7). Importantly, data was matched to the corresponding isotype controls (IC) considering their background signal when calculating the true intensities. Detailed IC data is presented in Additional files 1, 2, 3, 4, 5 and 6,

Multiomics analyses integrates flow cytometric data, proteomics and transcriptomics

It is important to note that differentially expressed receptors might in a first instance represent general tissue markers, as well as cancer specific markers. Hence, our subsequent workflow aims to screen for common markers first, eliminate those and then perform subfiltering for tissue and cancer specific markers.

To identify the most robustly expressed cell surface receptors on the various tumor cell lines, we integrated our FACS data with previously generated transcriptomic (Tx) and proteomic (Px) profiles of the NCI-60 tumor cell panel [13, 14, 19]. To this end, we used multiple co-inertia analysis (MCIA) to exploit the potential of the various omics data sets [19].

During data curation, two cell lines had to be removed (see "Methods"), eventually ending up with 58 tumor cell lines, each yielding 332 features based on flow cytometry, 514 features from Px and 19,437 from Tx. Spearman distance was used to create a dendrogram of the various cell lines, visualizing their relationship and revealing potential differences in tumor cell line annotation upon utilization of different omics data (Fig. 2).

Of note, flow cytometry-based characterization, i.e., relationships between tumor cell lines based on expression of cell surface receptors, was comparable to the annotation based on Tx data, and superior to Px. Flow cytometric analysis revealed specific clusters of the respective cell lines, in detail of central nervous system, renal, melanoma, colon and leukemia tumor cell lines, indicating that these might harbor unique identifiers.

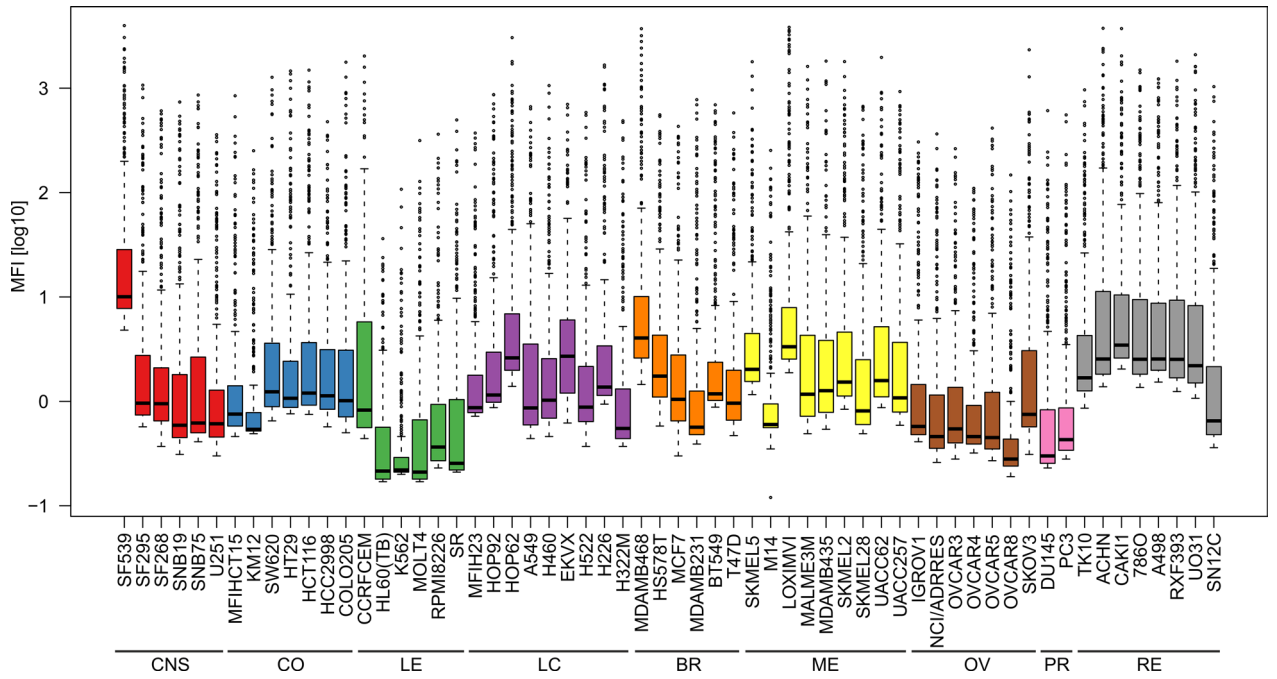

Fig. 1 Flow cytometric profiling of the NCI-60 tumor cell panel. The NCI-60 tumor cell lines were analyzed for expression of 332 cell surface receptors by flow cytometry (see "Methods"). Shown is the distribution of the cell surface expression of all measured receptors as $\log_{10}$ mean fluorescence intensity (MFI). *CNS* cancer of the central nervous system, *CO* colon cancer, *LE* leukemia, *LC* lung cancer, *BR* breast cancer, *ME* melanoma, *OV* ovarian cancer, *PR* prostate cancer, *RE* renal cancer

For Tx, only lung cancer-derived tumor cell lines showed a more comprehensive clustering when compared to FACS, while ovarian, prostate and breast cancer cell lines were dispersed in all analyses (Fig. 2).

Comprehensive MCIA integrating all three omics approaches (Fig. 3) revealed distinct clusters for the individual cancer types (Fig. 3a). In fact, we observed a significant contribution of the flow cytometric surface proteome data to the molecular profile of the individual tumor cell lines. Furthermore, different tumor types showed similar molecular patterns, as they formed clusters within the sample space. In agreement to the relationship analyses based on the various single omics techniques (Fig. 2), the MCIA revealed unique signatures predominantly for melanoma (ME; yellow) and leukemia (LE; green), but also identifiable clusters for cancer of the central nervous system (CNS; red), renal cancer (RE; grey) and colon cancer (CO; blue) (Fig. 3a).

Figure 3b shows the variables (transcripts, proteins and surface receptors) and their contribution to the clustering. Moreover, scree plot analysis helped us to determine the number of factors that should be considered (Fig. 3c), hence the number of principal components (PCs) that contribute to variability and help to discriminate tumor entities. The first 10 PCs already account for 74.7% of the variance and therefore were included into MCIA. The pseudo-eigenvalues of the whole NCI-60 data set (Fig. 3d), including Tx, Px and flow-cytometry, demonstrates that the integration of the receptorome data crucially contributes to the total variance of the MCIA and is hence an essential denominator to identify biomarkers. Table 1 lists cell surface receptors that were identified as tissue specific biomarkers based on our MCIA. As anticipated, from the lack of clustering (Figs. 2 and 3a), no MCIA hits could be retrieved for lung-, breast, ovarian- and prostate cancer, which might also reflect the large heterogeneity of these cancer entities.

Identification of tumor biomarkers by integration of MCIA into recount2 RNA-seq data

As discussed, the cell surface receptors identified by the MCIA do not necessarily represent tumor biomarkers, as we are lacking healthy control tissue to compare with the tumor cell lines. As an example, leukocyte specific marker CD2 is a T cell antigen. Therefore, while it is not surprising that MCIA identified CD2 as a leukemia specific receptor in comparison to all the other cancer entities, it is conceivable that CD2 is not a tumor marker. Having now identified differentially expressed surface markers by FACS and MCIA, we would now ideally

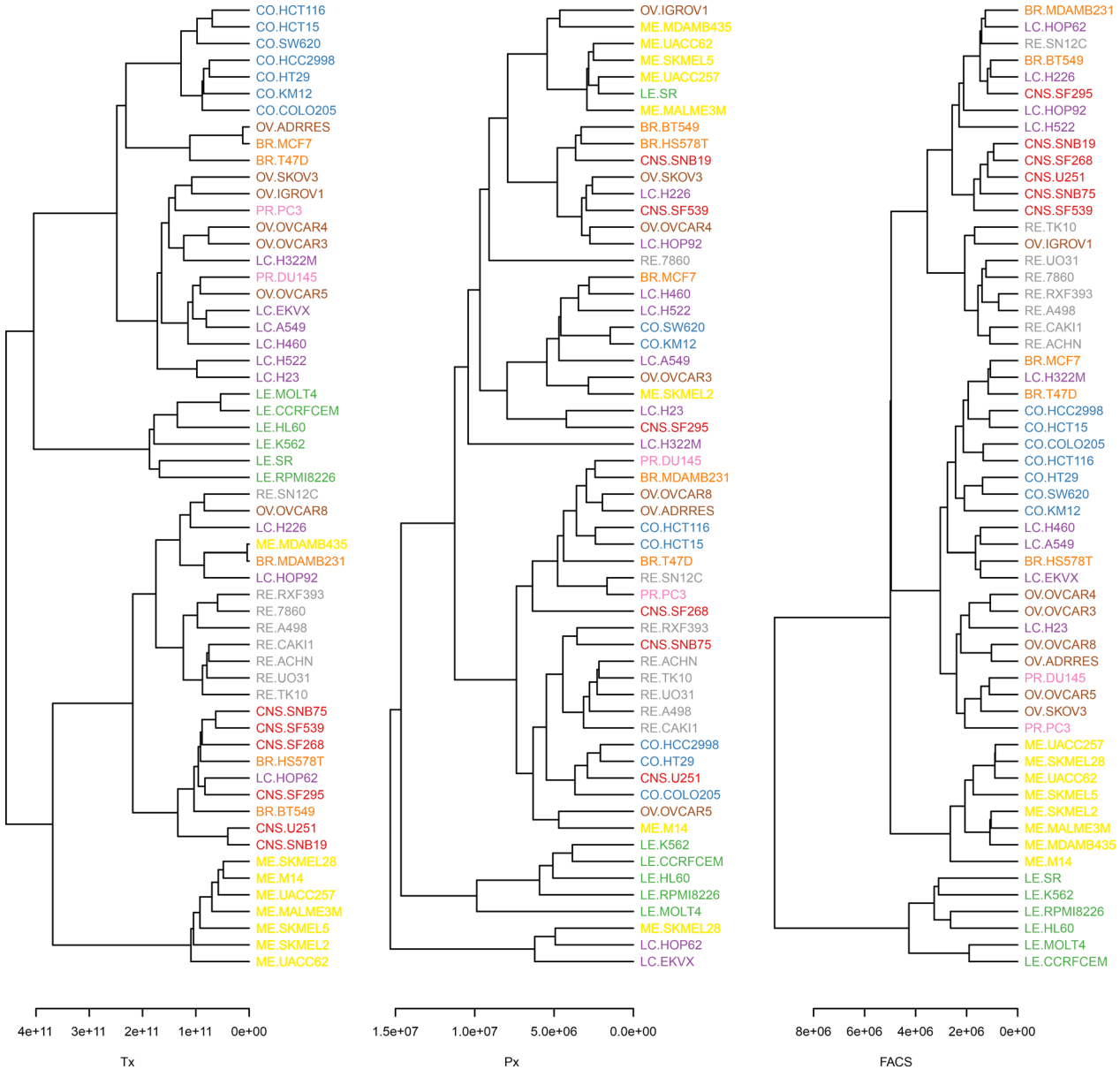

Fig. 2 Tumor cell line relationship using different omics techniques. A dendrogram of Tx (left), Px (middle) and flow cytometric (right) data obtained from the 58 tumor cell lines from the NCI-60 tumor cell panel was build using Spearman distance calculation. Tumor cell lines were color coded according to the cancer entity they represent. Blue, colon cancer (CO); brown, ovarian cancer (OV); orange, breast cancer (BR); pink, prostate cancer (PR); purple; lung cancer (LC); green, leukemia (LE); grey, renal cancer (RE); red, cancer of the central nervous system (CNS); yellow, melanoma (ME)

match the data with receptoromes of healthy control tissue. Due to the lack of such data, we validated our MCIA hits with The Cancer Genome Atlas (TCGA; www.cancer.gov/tcga) RNAseq count data, utilizing the recount2 resource results (https://jhubiostatistics.shinyapps.io/recount/ [32]). This allows to compare data from malignant with those from healthy tissues and hence to discriminate tissue markers from tumor markers. To our surprise, we could only retrieve for colon and renal cancer comprehensive recount2 RNAseq data from tumor as well as healthy tissues. For our other cancer types that distinctly clustered by MCIA, we could retrieve no TCGA RNAseq count data comparing tumor to normal tissue, precluding this type of analyses. The differentially expressed genes for colon and renal cancer are listed in Additional file 10: Dataset S3 and Additional file 11:

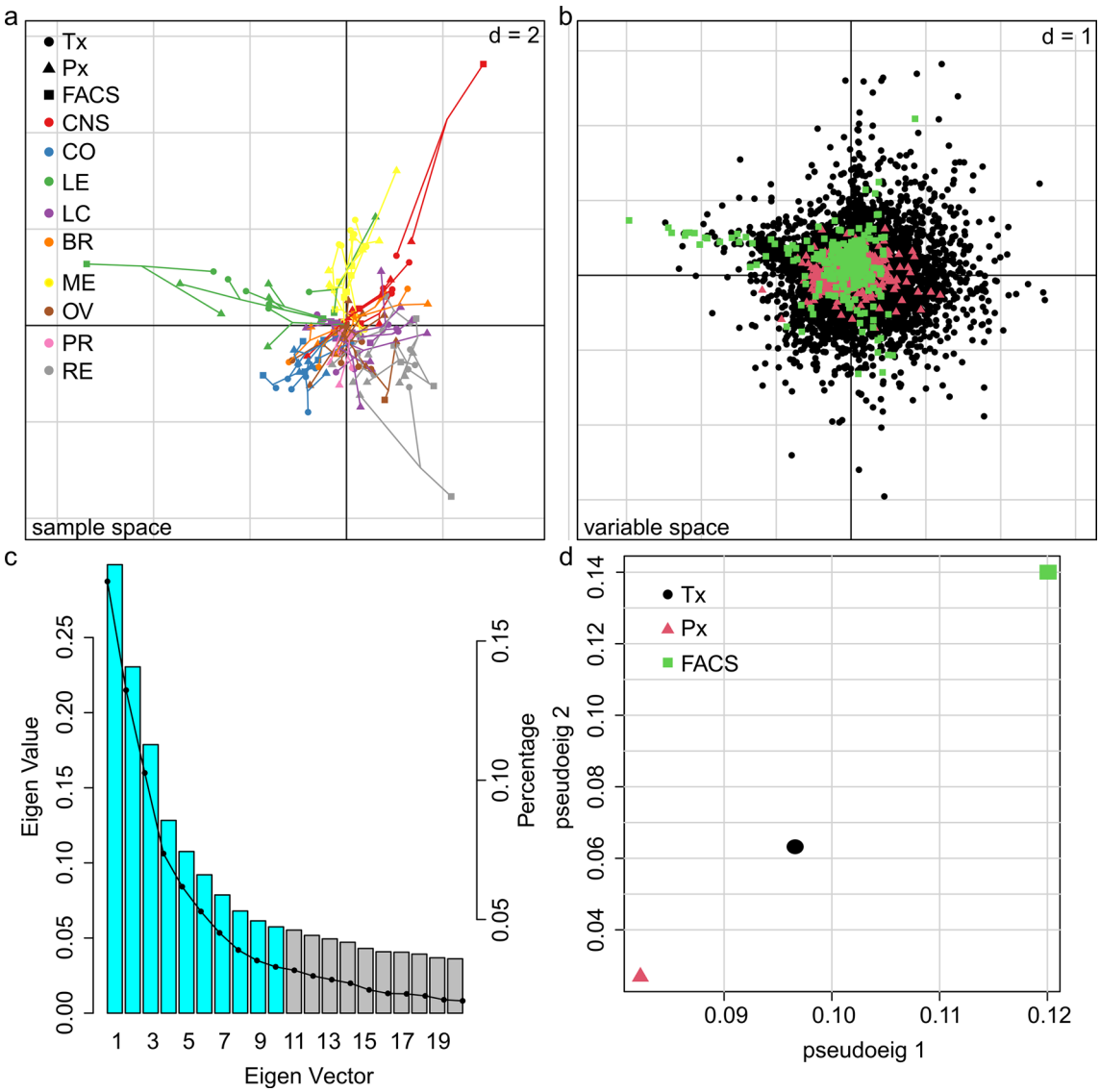

Fig. 3 Multiple co-inertia analysis (MCIA) of the NCI-60 panel data. **a** (top left) The first two principal components of the MCIA plot show similar trends in microarray (Tx, circle), proteomics (Px, triangle), and flow cytometry (FACS, square) profiles, suggesting that the most variant sources of biological information are similar. The type of shape indicates the respective omics platform. Shapes are connected by lines joining a common point representing the maximized covariance reference structure derived from the MCIA analysis. The length of a line models the divergence between the data from the same tumor cell line. Colors represent the nine NCI-60 different tissues covered by the tumor cell lines. Central nervous system (CNS) and leukemia (LE) cell lines are separated along the first axis (PC1, horizontal). Melanoma (ME) was projected on the positive side of the second axis (PC2, vertical). **b** (top right) The variable space with data from the different omics techniques is colored coded (Tx, black; Px, red; FACS, green). A tissue specific feature will be projected in the direction of this tissue. The larger the distance from the origin, the more potentially significant a feature is. **c** (bottom left) A scree plot showing the eigenvalues on the y-axis and the number of PCs on the x-axis. Used to rationalize the number of PCs included in the analysis. **d** (bottom right) The pseudo-eigenvalue space of the NCI-60 data sets summarizes the consensus between the platforms, highlighting which omics technique contributes more to the total variance (Tx, black; Px, red; flow cytometry, green)

Dataset S4. For both cancer entities, the five (colon) and eight (renal) identified MCIA receptors were differentially expressed between healthy and malignant tissues and hence represent potential surface accessible tumor biomarkers (Table 2).

We next plotted the relative MFIs of the tumor biomarkers and directly compared receptor expression on colon and renal tumor cell lines in the NCI-60 tumor cell panel to the other cell lines (Fig. 4). All receptors showed increased expression on colon and renal cancer

Table 1 Tumor cell line specific cell surface markers based on MCIA analysis

Colon cancer (CO)	Renal cancer (RE)	Melanoma (ME)	Central nervous system (CNS)	Leukemia (LE)	
CD104	CD106	CD1a	CD105	CD100	CD28
CD15	CD24	CD213a2	CD273B7DC	CD102	CD38
CD324	CD26	CD317	CD275B7H2	CD11a	CD4
CD326	EGFR	CD39	CD49e	CD18	CD45
CD49f	SSEA-3	CD49d	CD80	CD184	CD48
	SSEA-4	Integrin(a9b1)	MSCA1MSC	CD1b	CD5
	TIM1			CD1c	CD50
	TRA-1-60-R			CD1d	CD7
				CD2	CD84
				CD27	CD8a

Table 2 CO and RE cancer cell surface biomarkers based on integrated MCIA and recount2

log2FoldChange	pvalue	padj	receptor ID	Gene ID
Colon cancer: recount2 analysis filtered by MCIA hits				
0.659	6.844e−10	3.064e−9	CD49f	ITGA6
0.503	1.923e−06	6.064e−06	CD15	FUT4
0.328	0.016	0.029	CD104	ITGB4
−0.431	3.118e−06	9.609e−06	CD326	EPCAM
−0.489	2.994e−8	1.144e−7	CD324	CDH1
Renal cancer: recount2 analysis filtered by MCIA hits				
2.304	1.021e−36	1.237e−35	TIM1	HAVCR1
1.726	4.503e−36	5.309e−35	CD106	VCAM1
1.464	7.690e−63	2.500e−61	SSEA-4	TMCC1
1.371	6.499e−11	2.119e−10	SSEA-3	B3GALT5
1.200	1.296e−32	1.323e−31	EGFR	EGFR
0.681	9.772e−7	2.383e−06	CD26	DPP4
0.464	1.575e−8	4.375−8	CD24	CD24
−1.856	7.622e−41	1.1002e−39	TRA-1-60-R	PODXL

cell lines as compared to the remaining tumor cell lines from the NCI-60 panel. Expression of CD15, CD104, CD324, CD326 and CD49f was specifically enriched on the surface of colon cancer-derived cell lines (Fig. 4a, left and primary FACS data in in Fig. 4b). Renal cancer cell lines in comparison to all other cancer entities expressed significantly higher cell surface levels of CD24, CD26, CD106 (VCAM1), TIM1, SSEA-3 (B3GALT5), SSEA-4 (TMCC1), TRA-1-60R (PODXL) and EGFR (Fig. 4a right and primary FACS data in Fig. 4b).

Immunohistochemical analysis and validation of kidney tumor biomarkers using the human protein atlas

We next addressed the expression of kidney tumor biomarkers at the protein level by using data from the Human Protein Atlas (HPA, www.proteinatlas.org, [34]). CD106 (VCAM1) protein expression tended to be lower in healthy renal tubules compared to renal adenocarcinoma (Fig. 5a–c). For EGFR, we observed the same trend with higher expression in the tumors (Fig. 5d, e). whereas one EGFR antibody clone had the opposite phenotype (Fig. 5e). Pooled analysis without that clone showed a significantly higher protein expression of EGFR in tumors compared to healthy tubules ($p=0.0030$; exact, two-tailed; sum of ranks 199.5, 1397; Mann–Whitney $U=121.5$; Fig. 5f). There were no significant differences in expression between clear cell and non-clear cell renal cell carcinomas for CD106 ($n=23$; $p=0.787$; $df=5$, value = 2.428, chi-squared test) and EGFR ($n=44$; without CAB068186 antibody; $p=0.432$; $df=6$; value = 5.919, chi-squared test).

For CD24, we also observed moderately higher protein levels in tumors, which is depicted in Additional file 12: Fig. S7. For TRA-1-60R (PODXL), there were no significant differences detected (Additional file 12, panel d, e). Interestingly, and in contrast to the expected outcomes from our mRNA analysis, we observed significantly lower expression levels of TIM1 (HAVCR1), SSEA4 (TMCC1), CD26 and SSEA3 (B3GALT5) in tumors compared to healthy renal tubules (Additional file 12, panel f–o). The whole data are summarized in Additional file 8: Table S1.

Hence, while previous studies including data summarized in the HPA provided a list of promising candidates potentially suitable to use as cancer biomarkers, our data now critically expands this knowledge to markers that are surface accessible and are therefore candidates for tumor specific immunotherapy. Especially CD106 (VCAM1) and EGFR seem promising immunotherapeutic targets that show higher surface expression levels in the context of renal cancer.

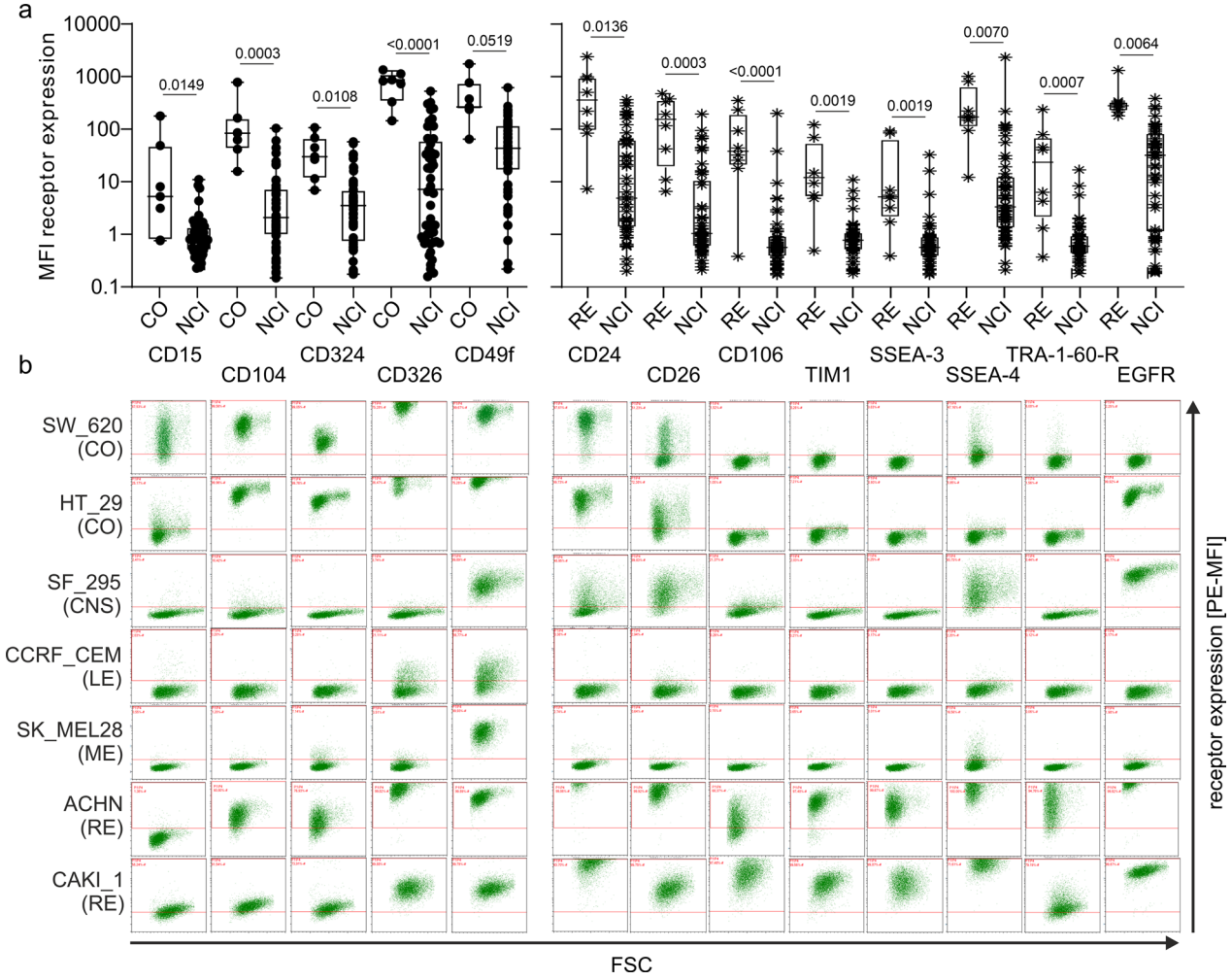

Fig. 4 Differential cell surface receptor expression on CO and RE cancer cell lines as compared to other tumor cell lines. **a** Shown is the relative MFI of each of the depicted receptors of each CO (n = 7, left) or RE (n = 8, right) cell line in comparison to all other tumor cell lines in the NCI-60 panel (NCI). Each symbol represents the specific cell surface receptor expression MFI of one cell line. Statistical significance of overall differences in receptor expression (MFI) of CO or RE versus all other cell lines was calculated with a one-way ANOVA with multiple comparisons and a Kruskal–Wallis test assuming non-gaussian distribution. **b** Primary FACS plots from representative cell lines of the various tumor entities. Plotted is the size of the cells (forward scatter, FSC) vs. the relative fluorescence expression of the respective cell surface marker (PE, mean fluorescence intensity, MFI). Note that no representative FACS plots are shown for ovarian cancer (OV), breast cancer (BR), prostate cancer (PR) and lung cancer (LC) because these cancer entities did not form clusters based on flow cytometric receptor expression (compare Fig. 2)

Receptor expression of kidney cancer subtypes by functional proteomics

Reverse Phase Protein Array (RPPA) data of the "The Cancer Proteome Atlas" (TCPA, https://tcpaportal.org/tcpa/, [36]) was assessed in order to explore the EGFR functional proteomics landscape of all renal cell cancer subtypes evaluated in this portal: These are "Kidney Chromophobe" (KICH, n = 63), "Kidney renal clear cell carcinoma" (KIRC, n = 445), and "Kidney renal papillary cell carcinoma" (KIRP, n = 208). Since CD106 (VACM1) and EGFR seemed to be promising targets based on our previous analysis we checked the dataset for these two markers. No data for VCAM1 was found. For EGFR, KICH and KIRP were compared against KIRC, since KIRC is the most frequent renal cancer subtype. Both comparisons revealed a significant difference when comparing the expression levels of EGFR (Table 3).

This result is in slight contrast to HPA (Fig. 5), by which we did not find kidney cancer subtype specific differences in EGFR expression (Fig. 5). This is most likely due to the much smaller sample size in HPA with only nine non-clear cell renal cell carcinoma. More importantly, the TCPA analysis provides an independent confirmation of the high expression of EGFR in

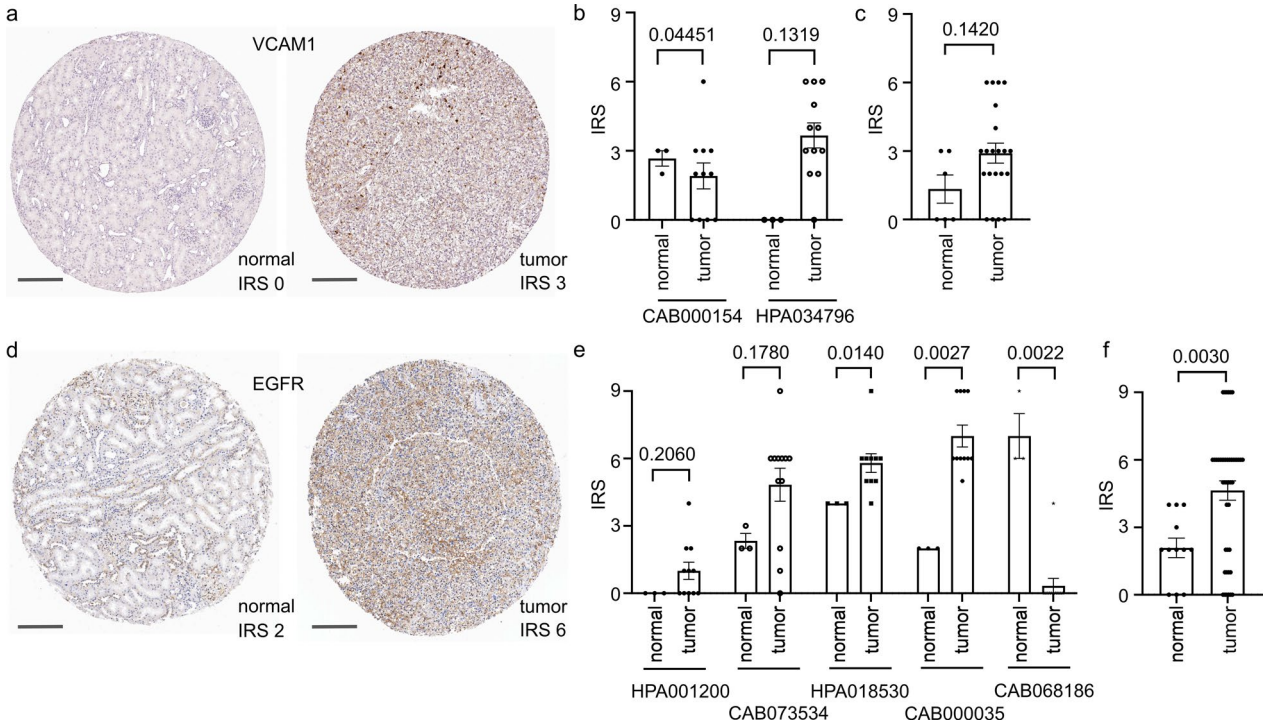

Fig. 5 Expression of VCAM1 and EGFR in healthy vs. tumor tissue based on HPA data. **a** Representative tissue cores from the HPA showing immunohistochemistry (IHC) staining for VCAM1. Left core, normal kidney; right core, renal cancer. **b**, **c** Summarized data for immunoreactive scores (IRS) for normal (pooled n = 6) and tumor kidney (pooled n = 23) from two antibodies shown separately (**b**) and pooled (**c**). **d** Representative tissue cores from the HPA showing IHC staining for EGFR. Left core, normal kidney; right core, renal cancer. **e**, **f** Summarized data for IRS for normal (pooled n = 12) and tumor kidney (pooled n = 44) from five antibodies shown separately (**e**) and pooled (**f**). In **f**, antibody CAB068186 which showed opposite results was not included. Scale bars, 200 µm

renal cell cancer and further indicates that EGFR could be a subtype specific biomarker for kidney cancer.

Discussion

By employing a comprehensive flow cytometric screen and combined multiomics analyses (integrating previously defined Tx and Px datasets), we identified novel specific cell surface expression patterns in five of the nine cancer entities represented by the NCI-60 tumor cell panel.

In the first instance, these receptors do not represent tumor biomarkers, as their identification is based on comparisons within the NCI-60 tumor cell panel. However, by subsequent cross-analysis with the TCGA recount2 RNAseq data, which also comprises healthy tissue for comparison [32], we could deconvolute our data to identify tumor specific biomarkers for two tumor entities, i.e., colon and renal cancer.

Table 3 Kidney cancer (RE) subtype specific expression analysis of EGFR via functional proteomics (TCPA)

Comparison	Expression A	Expression B	pvalue
KIRC vs. KICH	0.74176	0.32398	1.2851e−10
KIRC vs. KIRP	0.74176	0.21612	1.8623e−72

In theory, MCIA of the NCI-60 tumor cell panel with Tx and Px data alone, followed by annotation of biomarkers to their specific localization, also could have enabled the identification of cell surface specific receptor expression. However, implementation of the flow cytometric screening data provides several advantages. First, it is clear from the MCIA that integration of the flow cytometric data critically expands the sample space and strongly contributes to the overall variation (Fig. 3). Second, by employing flow cytometry, which is an antibody-based detection method, we already pre-screen for surface markers that can be detected and targeted by antibodies, anticipating a subsequent exploitation of these receptors for both diagnostic and immunotherapeutic (i.e., theranostic) applications. On the other hand, combining flow cytometry with the already available Tx and Px data by MCIA strongly enhanced confidence in our hits.

The initial MCIA revealed specific cell surface markers allowing to discriminate colon cancer (CO), melanoma

(ME), renal cancer (RE), cancer of the central nervous system (CNS) and leukemia (LE) cancer cell lines from all other cell lines within the NCI-60 tumor cell panel (Table 1). On the other hand, for the other four cancer entities, i.e., lung cancer (LC), breast cancer (BR), ovarian cancer (OV) and prostate cancer (PR) we failed to annotate a specific cell surface marker expression pattern, which might be due to the large heterogeneity of these tumor types.

We were surprised about our difficulties in finding accessible and reliable data to compare our MCIA derived potential tumor biomarkers with healthy tissue. Only for healthy colon and renal tissue we succeeded to extract healthy tissue RNAseq count data from the recount2 data base, allowing to cross-validate our cancer cell line-derived markers for differential expression between malignant and healthy tissue. Direct investigation of the TCGA portal revealed that healthy tissue RNAseq count data for skin, bone marrow, and lymph nodes is missing. Hence, further work to obtain omics data from healthy tissues in case of melanoma (ME), cancer of the central nervous system (CNS) and leukemia (LE) is warranted to obtain tumor markers for the latter tissues, too.

Ultimately, based on our MCIA, we identified CD15, CD104 (Integrin-β4), CD324 (E-cadherin), CD326 (EpCAM), and CD49f as biomarkers for colon cancer and CD24, CD26 (DPP4), CD106 (VCAM1), TIM-1, SSEA-3 (B3GALT5), SSEA-4 (TMCC1), TRA-1-60-R (PODXL) and EGFR for renal cancer. Of note, the colon cancer biomarkers identified by our approach have been proposed as potential tumor markers in colorectal cancer before [37–43]. These findings raise confidence in our data and independently confirm the stringency in our experimental screening and MCIA on the one hand, but also suggest following up on these receptors as structures for tumor-targeted immunotherapy. Similarly, for renal cancer, EGFR and TIM-1 are established tumor biomarkers for which immunotherapy has been proposed [44–46] and a phase I clinical trial with the goal to treat renal cell carcinoma with a TIM-1 targeting antibody indicated efficacy with manageable adverse effects [47]. Beyond that, CD24 was also proposed as a renal cancer biomarker [48, 49] and is discussed as a "hot candidate" for targeted immunotherapy, as it emerged as a novel "don't eat me"-signal that is expressed on various tumors and prevents their phagocytosis by macrophages [50, 51]. The role of CD26 is less explored in renal cancer even though it is suggested as a target for immunotherapy in the context of other cancer entities including the generation of CD26 directed CAR-T cells [52–56]. There is less experimental evidence establishing the other receptors identified by us as renal cancer-specific biomarkers. SSEA-3 and SSEA-4 as well as TRA-1-60-R are described as stem cell markers with some associations to renal cancer, that might be linked with aggressive tumor progression in vivo [57–61].

Likewise, CD106 (VCAM1) might be an interesting and potential novel cell surface biomarker as well as an attractive target for immunotherapy in the context of renal cancer. Currently, the role of VCAM1 in renal cancer is less explored, with data pointing towards protective roles in this cancer entity [62], as well as a potential involvement of VCAM1 in tumor immune evasion [63]. Furthermore, VCAM1 plays an important role in anti-tumor T cell responses and T cell infiltration into tumors [64, 65]. For VCAM1, EGFR and CD24 our findings were validated with healthy and tumor tissue data obtained from the Human Protein Atlas. This set of data provides further independent indications that VCAM1, EGFR and CD24 are highly expressed in renal cancers and that their high expression is a poor prognostic marker for survival [49, 66]. Given that EGFR and CD24 are already proposed and in the process of being therapeutically exploited, we now propose VCAM1 as a novel biomarker and target for cancer-specific stratified immunotherapy.

Furthermore, "The Cancer Proteome Atlas" resource enabled us to investigate kidney cancer subtype specific expression of EGFR, indicating that this receptor is a specific biomarker for clear cell renal cell carcinoma.

Conclusion

Altogether, our work adds a comprehensive panel of potential surface accessible cancer biomarkers which need to be further characterized by mining of databases and being evaluated in primary patient tumor tissue and healthy control tissues. Thus, our distinct approach demonstrates the power of open data integration and open science for fundamental and translational research. Furthermore, efforts to develop immunotherapeutic approaches utilizing these biomarkers, as for instance CAR T cells and bispecific antibodies for specific and direct elimination of these tumors are highly important and warranted.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12935-022-02710-y.

Additional file 1: Figure S1. Boxplot showing MFIs of all Isotype antibody controls.

Additional file 2: Figure S2. Boxplot showing MFIs of all mouse Isotype antibody controls.

Additional file 3: Figure S3. Boxplot showing MFIs of all rat Isotype antibody controls.

Additional file 4: Figure S4. Boxplot showing MFIs of AHIgGITCL Isotype antibody controls.

Additional file 5: Figure S5. Boxplot showing MFIs of all Mouse antibodies without Mouse IgG3 antibodies.

Additional file 6: Figure S6. Boxplot showing MFIs of Mouse IgG3 antibodies only.

Additional file 7: Dataset S1. Correlation antibody staining, dataset showing the correlation in antibody staining from two different biological replicates in 2 subsequent weeks.

Additional file 8: Dataset S2. Dataset including all mean fluorescence intensity values (MFIs] from all FACS screens done with the NCI-60 panel.

Additional file 9: Table S1. Table summarizing the results of the HPA analysis related to the biomarkers identified for renal cancer.

Additional file 10: Dataset S3. Dataset summarizing the analysis of the Recount2 RNAseq data analysis for differential RNA levels in healthy versus tumor colon tissue.

Additional file 11: Dataset S4. Dataset summarizing the analysis of the Recount2 RNAseq data analysis for differential RNA levels in healthy versus tumor renal tissue.

Additional file 12: Figure S7. This figure shows the results of the analysis of biomarker expression in normal and tumor kidney by immunohistochemistry based on HPA data.

Acknowledgements
We thank Stefan Czemmel from QBiC for helpful discussions regarding the data curation and MCIA analysis and Daniel Sauter (Institute for Medical Virology, University Hospital Tübingen) for critical reading of the manuscript, feedback and valuable comments.

Author contributions
SH did data curation and quality control, performed the MCIA with support from MCC and SN and did the RNAseq analysis and TCPA data exploration; SD cultured the NCI-60 panel and analyzed all cell lines by flow cytometry, and well assisted in data analyses; KB and CMS analyzed the HPA data; UML provided resources and assisted in data interpretation; SN provided infrastructure and assisted in data interpretation and analyses; MS planned, designed and supervised the overall study, provided resources, analyzed data and wrote the manuscript. All authors edited the manuscript draft together to its final form. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. This work was in part funded by basic research support given from the University Hospital Tübingen, Medical Faculty. SH acknowledges funding from the Central Innovation Programm (ZIM) for SMEs of the Federal Ministry for Economic Affairs and Energy of Germany. SN acknowledges Germany's Excellence Strategy (iFIT) - EXC 2180–390900677.

Availability of data and materials
All data generated and analyzed during this study are included in this published manuscript. The analysis scripts are available at https://github.com/qbicsoftware/QMSFC.

Declarations

Ethics approval and consent to participate
For this study no ethics approval was necessary as the experiments are based upon the NCI-60 tumor cell panel. No primary human patient material was used and no animal experiments were conducted.

Consent for publication
All authors gave their consent to publish, All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Quantitative Biology Center (QBiC), University of Tübingen, 72076 Tübingen, Germany. [2]Biomedical Data Science, Dept. of Computer Science, University of Tübingen, 72076 Tübingen, Germany. [3]Institute for Medical Virology and Epidemiology of Viral Diseases, University Hospital Tübingen, Tübingen, Germany. [4]Institute of Pathology, University of Bern, 3008 Bern, Switzerland. [5]Department of Pathology and Neuropathology, University Hospital and Comprehensive Cancer Center Tübingen, Tübingen, Germany. [6]Department of Internal Medicine VIII, Medical Oncology and Pneumology, Virotherapy Center Tübingen (VCT), Medical University Hospital Tübingen, 72076 Tübingen, Germany. [7]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Partner Site Tübingen, 72076 Tübingen, Germany.

Received: 25 July 2022 Accepted: 30 August 2022
Published online: 11 October 2022

References

1. Ribas A, Wolchok JD. Cancer immunotherapy using checkpoint blockade. Science. 2018;359(6382):1350–5.
2. Marin-Acevedo JA, Kimbrough EO, Lou Y. Next generation of immune checkpoint inhibitors and beyond. J Hematol Oncol. 2021;14(1):45.
3. Shimasaki N, Jain A, Campana D. NK cells for cancer immunotherapy. Nat Rev Drug Discov. 2020;19(3):200–18.
4. Jiang Z, Sun H, Yu J, Tian W, Song Y. Targeting CD47 for cancer immunotherapy. J Hematol Oncol. 2021;14(1):180.
5. Kennedy LB, Salama AKS. A review of cancer immunotherapy toxicity. CA Cancer J Clin. 2020;70(2):86–104.
6. Leko V, Rosenberg SA. Identifying and targeting human tumor antigens for T cell-based immunotherapy of solid tumors. Cancer Cell. 2020;38(4):454–72.
7. Sun Q, Melino G, Amelio I, Jiang J, Wang Y, Shi Y. Recent advances in cancer immunotherapy. Discov Oncol. 2021;12(1):27.
8. Yeo D, Giardina C, Saxena P, Rasko JEJ. The next wave of cellular immunotherapies in pancreatic cancer. Mol Ther Oncolytics. 2022;24:561–76.
9. Chabner BA. NCI-60 cell line screening: a radical departure in its time. J Natl Cancer Inst. 2016. https://doi.org/10.1093/jnci/djv388.
10. Abaan OD, Polley EC, Davis SR, Zhu YJ, Bilke S, Walker RL, Pineda M, Gindin Y, Jiang Y, Reinhold WC, et al. The exomes of the NCI-60 panel: a genomic resource for cancer biology and systems pharmacology. Cancer Res. 2013;73(14):4372–82.
11. Monks A, Zhao Y, Hose C, Hamed H, Krushkal J, Fang J, Sonkin D, Palmisano A, Polley EC, Fogli LK, et al. The NCI transcriptional pharmacodynamics workbench: a tool to examine dynamic expression profiling of therapeutic response in the NCI-60 cell line panel. Cancer Res. 2018;78(24):6807–17.
12. Reinhold WC, Sunshine M, Varma S, Doroshow JH, Pommier Y. Using CellMiner 1.6 for systems pharmacology and genomic analysis of the NCI-60. Clin Cancer Res. 2015;21(17):3841–52.
13. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, et al. A gene expression database for the molecular pharmacology of cancer. Nat Genet. 2000;24(3):236–44.
14. Gholami AM, Hahne H, Wu Z, Auer FJ, Meng C, Wilhelm M, Kuster B. Global proteome analysis of the NCI-60 cell line panel. Cell Rep. 2013;4(3):609–20.
15. Graessel A, Hauck SM, von Toerne C, Kloppmann E, Goldberg T, Koppensteiner H, Schindler M, Knapp B, Krause L, Dietz K, et al. A Combined omics approach to generate the surface atlas of human Naive CD4+ T cells during early T-cell receptor activation. Mol Cell Proteom. 2015;14(8):2085–102.
16. Businger R, Kivimaki S, Simeonov S, Vavouras Syrigos G, Pohlmann J, Bolz M, Muller P, Codrea MC, Templin C, Messerle M, et al. Comprehensive analysis of human cytomegalovirus- and HIV-mediated plasma membrane remodeling in macrophages. mBio. 2021;12(4):e0177021.
17. Team RC. R: a language and environment for statistical computing. R Foundation for Statistical Computing; 2021. https://www.R-project.org/.
18. Documentation A. Anaconda software distribution. In: 2-2.4.0 edn: Anaconda Inc.; 2020.
19. Meng C, Kuster B, Culhane AC, Gholami AM. A multivariate approach to the integration of multi-omics datasets. BMC Bioinform. 2014;15:162.

20. Prasad VV, Gopalan RO. Continued use of MDA-MB-435, a melanoma cell line, as a model for human breast cancer, even in year, 2014. NPJ Breast Cancer. 2015;1:15002.
21. Ke W, Yu P, Wang J, Wang R, Guo C, Zhou L, Li C, Li K. MCF-7/ADR cells (re-designated NCI/ADR-RES) are not derived from MCF-7 breast cancer cells: a loss for breast cancer multidrug-resistant research. Med Oncol. 2011;28(Suppl 1):135–41.
22. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249–64.
23. Pfister TD, Reinhold WC, Agama K, Gupta S, Khin SA, Kinders RJ, Parchment RE, Tomaszewski JE, Doroshow JH, Pommier Y. Topoisomerase I levels in the NCI-60 cancer cell line panel determined by validated ELISA and microarray analysis and correlation with indenoisoquinoline sensitivity. Mol Cancer Ther. 2009;8(7):1878–84.
24. Giovinazzi S, Sirleto P, Aksenova V, Morozov VM, Zori R, Reinhold WC, Ishov AM. Usp7 protects genomic stability by regulating Bub3. Oncotarget. 2014;5(11):3728–42.
25. Kohn KW, Zeeberg BM, Reinhold WC, Pommier Y. Gene expression correlations in human cancer cell lines define molecular interaction networks for epithelial phenotype. PLoS ONE. 2014;9(6):e99269.
26. Reinhold WC, Varma S, Sunshine M, Rajapakse V, Luna A, Kohn KW, Stevenson H, Wang Y, Heyn H, Nogales V, et al. The NCI-60 methylome and its integration into CellMiner. Cancer Res. 2017;77(3):601–12.
27. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, et al. NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res. 2009;37(Database issue):D885–90.
28. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A. BioMart—biological queries made easy. BMC Genom. 2009;10:22.
29. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008;26(12):1367–72.
30. Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, Inuganti A, Griss J, Mayer G, Eisenacher M, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res. 2019;47(D1):D442–50.
31. Dodge Y. Spearman rank correlation coefficient. In: The concise encyclopedia of statistics. New York: Springer; 2008. p. 502–5.
32. Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT. Reproducible RNA-seq analysis using recount. Nat Biotechnol. 2017;35(4):319–21.
33. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
34. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, et al. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol. 2010;28(12):1248–50.
35. Fedchenko N, Reifenrath J. Different approaches for interpretation and reporting of immunohistochemistry analysis results in the bone tissue—a review. Diagn Pathol. 2014;9:221.
36. Li J, Lu Y, Akbani R, Ju Z, Roebuck PL, Liu W, Yang JY, Broom BM, Verhaak RG, Kane DW, et al. TCPA: a resource for cancer functional proteomics data. Nat Methods. 2013;10(11):1046–7.
37. Giordano G, Febbraro A, Tomaselli E, Sarnicola ML, Parcesepe P, Parente D, Forte N, Fabozzi A, Remo A, Bonetti A, et al. Cancer-related CD15/FUT4 overexpression decreases benefit to agents targeting EGFR or VEGF acting as a novel RAF-MEK-ERK kinase downstream regulator in metastatic colorectal cancer. J Exp Clin Cancer Res. 2015;34:108.
38. Carlsen L, Huntington KE, El-Deiry WS. Immunotherapy for colorectal cancer: mechanisms and predictive biomarkers. Cancers (Basel). 2022;14(4):1028.
39. Jang TJ, Park JB, Lee JI. The expression of CD10 and CD15 Is Progressively increased during colorectal cancer development. Korean J Pathol. 2013;47(4):340–7.
40. Li M, Jiang X, Wang G, Zhai C, Liu Y, Li H, Zhang Y, Yu W, Zhao Z. ITGB4 is a novel prognostic factor in colon cancer. J Cancer. 2019;10(21):5223–33.
41. Chen GT, Waterman ML. Cancer: leaping the E-cadherin hurdle. EMBO J. 2015;34(18):2307–9.
42. Guo L, Fu J, Sun S, Zhu M, Zhang L, Niu H, Chen Z, Zhang Y, Guo L, Wang S. MicroRNA-143-3p inhibits colorectal cancer metastases by targeting ITGA6 and ASAP3. Cancer Sci. 2019;110(2):805–16.
43. Haraguchi N, Ishii H, Mimori K, Ohta K, Uemura M, Nishimura J, Hata T, Takemasa I, Mizushima T, Yamamoto H, et al. CD49f-positive cell population efficiently enriches colon cancer-initiating cells. Int J Oncol. 2013;43(2):425–30.
44. Ciardiello F, Caputo R, Bianco R, Damiano V, Pomatico G, Pepe S, Bianco AR, Agrawal S, Mendelsohn J, Tortora G. Cooperative inhibition of renal cancer growth by anti-epidermal growth factor receptor antibody and protein kinase A antisense oligonucleotide. J Natl Cancer Inst. 1998;90(14):1087–94.
45. Zhang Q, Tian K, Xu J, Zhang H, Li L, Fu Q, Chai D, Li H, Zheng J. Synergistic effects of cabozantinib and EGFR-specific CAR-NK-92 cells in renal cell carcinoma. J Immunol Res. 2017;2017:6915912.
46. Thomas LJ, Vitale L, O'Neill T, Dolnick RY, Wallace PK, Minderman H, Gergel LE, Forsberg EM, Boyer JM, Storey JR, et al. Development of a novel antibody-drug conjugate for the potential treatment of ovarian, lung, and renal cell carcinoma expressing TIM-1. Mol Cancer Ther. 2016;15(12):2946–54.
47. McGregor BA, Gordon M, Flippot R, Agarwal N, George S, Quinn DI, Rogalski M, Hawthorne T, Keler T, Choueiri TK. Safety and efficacy of CDX-014, an antibody-drug conjugate directed against T cell immunoglobulin mucin-1 in advanced renal cell carcinoma. Invest New Drugs. 2020;38(6):1807–14.
48. Lee HJ, Kim DI, Kwak C, Ku JH, Moon KC. Expression of CD24 in clear cell renal cell carcinoma and its prognostic significance. Urology. 2008;72(3):603–7.
49. Arik D, Can C, Dundar E, Kabukcuoglu S, Pasaoglu O. Prognostic significance of CD24 in clear cell renal cell carcinoma. Pathol Oncol Res. 2017;23(2):409–16.
50. Altevogt P, Sammar M, Huser L, Kristiansen G. Novel insights into the function of CD24: a driving force in cancer. Int J Cancer. 2021;148(3):546–59.
51. Barkal AA, Brewer RE, Markovic M, Kowarsky M, Barkal SA, Zaro BW, Krishnan V, Hatakeyama J, Dorigo O, Barkal LJ, et al. CD24 signalling through macrophage Siglec-10 is a target for cancer immunotherapy. Nature. 2019;572(7769):392–6.
52. Enz N, Vliegen G, De Meester I, Jungraithmayr W. CD26/DPP4—a potential biomarker and target for cancer therapy. Pharmacol Ther. 2019;198:135–59.
53. Inamoto T, Yamochi T, Ohnuma K, Iwata S, Kina S, Inamoto S, Tachibana M, Katsuoka Y, Dang NH, Morimoto C. Anti-CD26 monoclonal antibody-mediated G1-S arrest of human renal clear cell carcinoma Caki-2 is associated with retinoblastoma substrate dephosphorylation, cyclin-dependent kinase 2 reduction, p27(kip1) enhancement, and disruption of binding to the extracellular matrix. Clin Cancer Res. 2006;12(11 Pt 1):3470–7.
54. Nishida H, Hayashi M, Morimoto C, Sakamoto M, Yamada T. CD26 is a potential therapeutic target by humanized monoclonal antibody for the treatment of multiple myeloma. Blood Cancer J. 2018;8(11):99.
55. Varona A, Blanco L, Perez I, Gil J, Irazusta J, Lopez JI, Candenas ML, Pinto FM, Larrinaga G. Expression and activity profiles of DPP IV/CD26 and NEP/CD10 glycoproteins in the human renal cancer are tumor-type dependent. BMC Cancer. 2010;10:193.
56. Zhou S, Li W, Xiao Y, Zhu X, Zhong Z, Li Q, Cheng F, Zou P, You Y, Zhu X. A novel chimeric antigen receptor redirecting T-cell specificity towards CD26(+) cancer cells. Leukemia. 2021;35(1):119–29.
57. Lou YW, Wang PY, Yeh SC, Chuang PK, Li ST, Wu CY, Khoo KH, Hsiao M, Hsu TL, Wong CH. Stage-specific embryonic antigen-4 as a potential therapeutic target in glioblastoma multiforme and other cancers. Proc Natl Acad Sci USA. 2014;111(7):2482–7.
58. Saito S, Aoki H, Ito A, Ueno S, Wada T, Mitsuzuka K, Satoh M, Arai Y, Miyagi T. Human alpha2,3-sialyltransferase (ST3Gal II) is a stage-specific embryonic antigen-4 synthase. J Biol Chem. 2003;278(29):26474–9.
59. Maruyama R, Saito S, Bilim V, Hara N, Itoi T, Yamana K, Nishiyama T, Arai Y, Takahashi K, Tomita Y. High incidence of GalNAc disialosyl lactotetraosylceramide in metastatic renal cell carcinoma. Anticancer Res. 2007;27(6C):4345–50.
60. Yoon JY, Gedye C, Paterson J, Ailles L. Stem/progenitor cell marker expression in clear cell renal cell carcinoma: a potential relationship with the immune microenvironment to be explored. BMC Cancer. 2020;20(1):272.
61. Fiedorowicz M, Khan MI, Strzemecki D, Orzel J, Welniak-Kaminska M, Sobiborowicz A, Wieteska M, Rogulski Z, Cheda L,

Wargocka-Matuszewska W, et al. Renal carcinoma CD105—/CD44— cells display stem-like properties in vitro and form aggressive tumors in vivo. Sci Rep. 2020;10(1):5379.
62. Shioi K, Komiya A, Hattori K, Huang Y, Sano F, Murakami T, Nakaigawa N, Kishida T, Kubota Y, Nagashima Y, et al. Vascular cell adhesion molecule 1 predicts cancer-free survival in clear cell renal carcinoma patients. Clin Cancer Res. 2006;12(24):7339–46.
63. Wu TC. The role of vascular cell adhesion molecule-1 in tumor immune evasion. Cancer Res. 2007;67(13):6003–6.
64. Riegler J, Gill H, Ogasawara A, Hedehus M, Javinal V, Oeh J, Ferl GZ, Marik J, Williams S, Sampath D, et al. VCAM-1 density and tumor perfusion predict T-cell infiltration and treatment response in preclinical models. Neoplasia. 2019;21(10):1036–50.
65. Nakajima K, Ino Y, Yamazaki-Itoh R, Naito C, Shimasaki M, Takahashi M, Esaki M, Nara S, Kishi Y, Shimada K, et al. IAP inhibitor, embelin increases VCAM-1 levels on the endothelium, producing lymphocytic infiltration and antitumor immunity. Oncoimmunology. 2020;9(1):1838812.
66. Wang S, Yu ZH, Chai KQ. Identification of EGFR as a novel key gene in clear cell renal cell carcinoma (ccRCC) through bioinformatics analysis and meta-analysis. Biomed Res Int. 2019;2019:6480865.

Publisher's Note

Bioinformatics, 2024, **40(7)**, btae363
https://doi.org/10.1093/bioinformatics/btae363
Advance Access Publication Date: 3 July 2024
Original Paper

OXFORD

Genome analysis

Pangenome graph layout by Path-Guided Stochastic Gradient Descent

Simon Heumos [1,2,3,4,†], Andrea Guarracino [5,6,†], Jan-Niklas M. Schmelzle [7,8], Jiajie Li [8], Zhiru Zhang [8], Jörg Hagmann [9], Sven Nahnsen [1,2,3,4], Pjotr Prins [5], Erik Garrison [5,*]

[1]Quantitative Biology Center (QBiC), University of Tübingen, 72076 Tübingen, Germany
[2]Biomedical Data Science, Department of Computer Science, University of Tübingen, 72076 Tübingen, Germany
[3]M3 Research Center, University Hospital Tübingen, 72076 Tübingen, Germany
[4]Institute for Bioinformatics and Medical Informatics (IBMI), University of Tübingen, 72076 Tübingen, Germany
[5]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN 38163, United States
[6]Genomics Research Centre, Human Technopole, 20157 Milan, Italy
[7]Department of Computer Engineering, School of Computation, Information and Technology (CIT), Technical University of Munich, 80333 Munich, Germany
[8]School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853, United States
[9]Computomics GmbH, 72072 Tübingen, Germany

*Corresponding author. Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Translational Research Building, 71 South Manassas, Memphis, TN 38163, United States. E-mail: egarris5@uthsc.edu

[†]Equal contribution.

Associate Editor: Peter Robinson

Abstract

Motivation: The increasing availability of complete genomes demands for models to study genomic variability within entire populations. Pangenome graphs capture the full genomic similarity and diversity between multiple genomes. In order to understand them, we need to see them. For visualization, we need a human-readable graph layout: a graph embedding in low (e.g. two) dimensional depictions. Due to a pangenome graph's potential excessive size, this is a significant challenge.

Results: In response, we introduce a novel graph layout algorithm: the Path-Guided Stochastic Gradient Descent (PG-SGD). PG-SGD uses the genomes, represented in the pangenome graph as paths, as an embedded positional system to sample genomic distances between pairs of nodes. This avoids the quadratic cost seen in previous versions of graph drawing by SGD. We show that our implementation efficiently computes the low-dimensional layouts of gigabase-scale pangenome graphs, unveiling their biological features.

Availability and implementation: We integrated PG-SGD in *ODGI* which is released as free software under the MIT open source license. Source code is available at https://github.com/pangenome/odgi.

1 Introduction

Reference genomes are widely used in genomics, serving as a foundation for a variety of analyses, including gene annotation, read mapping, and variant detection (Singh *et al.* 2022). However, this linear model is becoming obsolete given the accessibility to hundreds or even thousands of high-quality genomes. A single genome cannot fully represent the genetic diversity of any species, resulting in reference bias (Ballouz *et al.* 2019). In contrast, a pangenome models the entire set of genomic elements of a given population (Tettelin *et al.* 2008, Computational Pan-Genomics Consortium 2018, Eizenga *et al.* 2020, Sherman and Salzberg 2020). Pangenomes can be represented as a sequence graph incorporating sequences as nodes and their relationships as edges (Hein 1989). In the variation graph model (Garrison *et al.* 2018), genomes are encoded as paths traversing the nodes in the graph.

A graph layout is the arrangement of nodes and edges in an N-dimensional space. Graph layout algorithms aim to find optimal node coordinates in order to minimize overlapping nodes or edges, reduce edge crossings, and promote an intuitive understanding of the graph. One popular approach is force-directed graph drawing (Cheong and Si 2022) which uses physical simulation to produce esthetic layouts. The classical approach combines repulsive forces on all vertices and attractive forces on adjacent vertices. This is prone to get stuck in local minima, but multi-layer strategies such as the Fast Multipole Multilevel Method (FM^3) (Hachul and Jünger 2005) or Stochastic Gradient Descent (SGD) implementations alleviate such a problem (Zheng *et al.* 2019). SGD uses the gradient of its individual terms to approximate the gradient of a sum of functions.

A *pangenome* graph layout can provide a human-readable visualization of genetic variation between multiple genomes. However, Zheng *et al.* (2019)'s algorithm has a quadratic up front cost in the number of nodes to find pairwise distances to guide the layout, making it impossible to apply to

Received: 10 November 2023; **Revised:** 20 February 2024; **Editorial Decision:** 24 March 2024

pangenome graphs with millions of nodes. Also, existing generic graph layout approaches ignore the biological information inherent in pangenome graphs. One such bioinformatics tool is *BandageNG*, the current state of the art for genome graph visualization. It uses FM^3 which only considers the nodes and edges of a graph.

In practice, MultiDimensional Scaling (MDS) is applied to minimize the difference between the visual distance and theoretical graph distance. This can be accomplished by using pairwise node distances to minimize an energy function. Since pangenome graphs represent genomes as paths in the graph, a reasonable distance metric would be the nucleotide distance between a pair of nodes traversed by the same path. Such path sampling would overcome the quadratic costs of previous versions of graph drawing by SGD.

Typically, force-directed layouts are hard to compute (Wang *et al.* 2014). Although, *BandageNG* applies FM^3 for layout generation, its parallelism is bound by the number of connected graph components. Alternatively, the lock-free HOGWILD! method offers a highly parallelizable and thus scalable SGD approach that can be applied when the optimization problem is sparse (Recht *et al.* 2011).

Here, we present a new pangenome graph layout algorithm which applies a Path-Guided SGD (PG-SGD) to use the paths as an embedded positional system to find distances between nodes, moving pairs of nodes in parallel with a modified HOGWILD! strategy. The algorithm computes the pangenome graph layout that best reflects the nucleotide sequences in the graph. To our knowledge, no generic graph layout algorithm takes into account such path encoded biological information when computing a graph's layout.

PG-SGD can be extended in any number of dimensions. In the ODGI toolkit (Guarracino *et al.* 2022), we provide implementations for 1D and 2D layouts. These algorithms have already been successfully applied to construct and visualize large-scale pangenome graphs of the Human Pangenome Reference Consortium (HPRC) (Guarracino *et al.* 2023, Liao *et al.* 2023). In addition, we show that PG-SGD is almost an order of magnitude faster than *BandageNG*.

2 Algorithm

While PG-SGD is inspired by Zheng *et al.* (2019), we designed the algorithm to work on the variation graph model (Definition 2.1).

Definition 2.1. Variation graphs are a mathematical formalism to represent pangenome graphs (Garrison 2019). In the variation graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{P})$, nodes (or vertices) $\mathcal{V} = v_1 \dots v_{|\mathcal{V}|}$ contain nucleotide sequences. Each node v_i has a unique identifier i and an implicit reverse complement $\bar{v}_i$. The node strand o represents the node orientation. Edges $\mathcal{E} = e_1 \dots e_{|\mathcal{E}|}$ connect ordered pairs of node strands ($e_i = (o_a, o_b)$), defining the graph topology. Paths $\mathcal{P} = p_1 \dots p_{|\mathcal{P}|}$ are series of connected steps s_i that refer to node strands in the graph ($p_i = s_1 \dots s_{|p_i|}$); the paths represent the genomes embedded in the graph.

We report PG-SGD's pseudocode in Algorithm 1 and its schematic in Fig. 1. In brief, the algorithm moves one pair of nodes (v_i, v_j) at a time, minimizing the difference between the

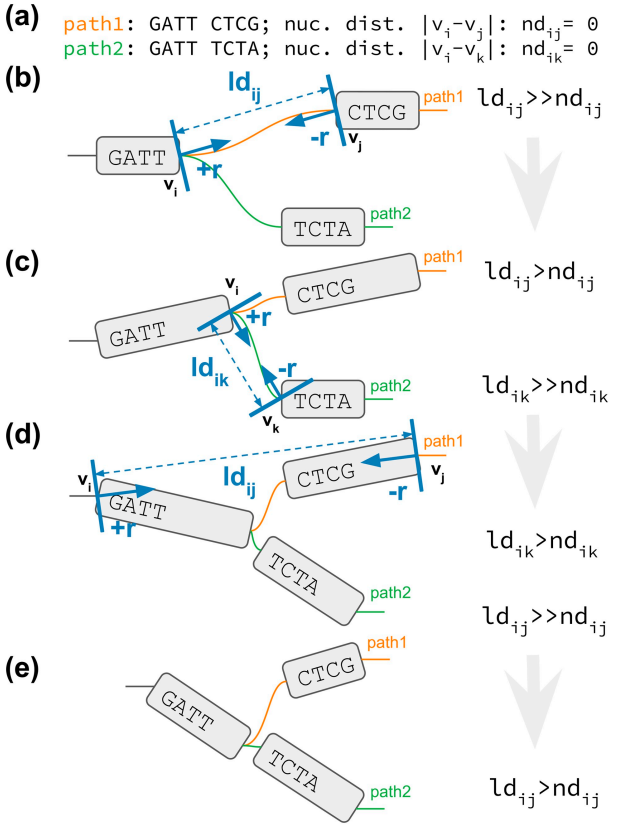

Figure 1. 2D PG-SGD update operation sketches. (a) The path information of the graph. *path1* and *path2* both visit the same first node. Then their sequence diverges and they visit distinct nodes. (b–e) v_i/v_j or v_i/v_k is the current pair of nodes to update. ld_{ij}/ld_{ik} is the current layout distance. r, $-r$ is the current size of the update. (b) Initial graph layout highlighting the future update of the two nodes of *path1*. (c) The graph layout after the first update. The nodes appear longer now, because we updated at the end of the nodes. Highlighted is the future update of the two nodes of *path2*. (d) The graph layout after the second update. Highlighted is the future update of the two nodes of *path1*. (e) Final graph layout after three updates using the 2D PG-SGD.

layout distance ld_{ij} of the two nodes and the nucleotide distance nd_{ij} of the same nodes as calculated along a path that traverses them. In the 2D layouts, nodes have two ends. When moving a pair of nodes, we actually move one end of each node. For clarification, an example is given in Fig. 1. v_i is the node associated with the step s_i sampled uniformly from all the steps in $\mathcal{P}$. v_j is the node associated with the step s_j sampled from the same path of s_i by drawing a uniform or a Zipfian distribution (Zipf 1932). The difference between nd_{ij} and ld_{ij} guides the update of the node coordinates in the layout. The magnitude r of the update depends on the learning rate μ. The number of iterations steers the annealing step size η which determines the learning rate μ. A large η in the first iterations leads to a globally linear (in 1D) or planar (in 2D) layout. By decreasing η, the layout adjustments become more localized, ensuring that the nodes are positioned to best reflect the nucleotide distances in the paths (i.e. in the genomes).

Originating from empirical inspection of word frequency tables, Zipf's law states that a word with rank n occurs $1/n$ times as the most frequent one. This law is modeled by the Zipf distribution. Sampling s_j from a Zipf distribution fixed

Algorithm 1: Pseudocode of PG-SGD in 1D.

PG-SGD $(\mathcal{G})$**:**
- **input:** variation graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{P})$
- **output:** N-dimensional layout $\mathcal{L}$ with $|\mathcal{V}|$ nodes
- $\mathcal{XP} \leftarrow$ `PathIndex`$(\mathcal{G})$ `// for path position lookup`
- $\mathcal{L} \leftarrow$ `LayoutInitialization`$(\mathcal{V}, N)$
- $\mathcal{Z} \leftarrow$ `InitZipf`$(\mathcal{G}, \mathcal{XP})$ `// Zipfian distribution`
- **for** η *in annealing schedule*:
  - **for each** *planned term update*:
    - $s_i \leftarrow$ `Unif`$(\mathcal{XP})$ `// uniform sampling of a step from` $\mathcal{P}$
    - $p \leftarrow$ `Path`$(s_i, \mathcal{XP})$ `// path of` s_i
    - **if** $(cooling \mid\mid flip)$ **then**
      - $s_j \leftarrow$ `Unif(StepCount`$(p, \mathcal{XP}))$ `// uniform sampling of a step from` p
    - **else**
      - $s_j \leftarrow$ `Zipf`(p) `// Zipfian sampling of a step from` p
    - **end**
    - $p_i \leftarrow$ `StepPos`(s_i) `// nuc. position`
    - $p_j \leftarrow$ `StepPos`(s_j) `// nuc. position`
    - $nd_{ij} \leftarrow ||p_i - p_j||$ `// nuc. distance`
    - $ld_{ij} \leftarrow ||l_i - l_j||$ `// layout distance`
    - $w_{ij} \leftarrow \frac{1.0}{nd_{ij}}$ `// term weight`
    - $\mu \leftarrow w_{ij}\eta$ `// learning rate`
    - **if** $\mu > 1$**:**
      - $\mu \leftarrow 1$
    - **end**
    - $\delta \leftarrow \mu \cdot \frac{ld_{ij} - nd_{ij}}{2}$ `// the actual delta`
    - **if** $abs(\delta) <= 0$ **then**
      - $STOP$ `// we can't optimize more`
    - $r \leftarrow \frac{\delta}{ld_{ij}}$ `// size of the update`
    - $r_x \leftarrow r \cdot (l_i - l_j)$ `// update size normalized by layout distance`
    - $l_i \leftarrow l_i + r_x$ `// update` v_i `coordinates`
    - $l_j \leftarrow l_j - r_x$ `// update` v_j `coordinates`
  - **end**
- **end**
- **end**

in the s_i's path position space increases the possibility to draw a nucleotide position close to s_i. So there is a high chance to use small nucleotide distances nd_{ij} to refine the layout of nodes comprising a few base pairs. The Zipf distribution is also long-tailed, with many occurrences of low frequency events. However, extremely long-range correlations might not be captured sufficiently, resulting in collapsed layouts for structures that are otherwise linear. To provide balance between global and local layout updates, in half of the updates (*flip* flag in Algorithm 1), the s_j is sampled uniformly instead from a Zipf distribution, with uniform sampling being more favorable for global updates. Furthermore, to enhance local linearity (in 1D) or planarity (in 2D) of the graph layout, a *cooling* phase skews the Zipfian distribution after half of iterations have been completed. This increases the likelihood of sampling smaller nucleotide distances for the layout updates.

3 Implementation

We implemented PG-SGD in ODGI (Guarracino *et al.* 2022): the 1D version can be found in *odgi sort* and the 2D version in *odgi layout*. To efficiently retrieve path nucleotide positions, we implemented a path index. This index is a strict subset of the XG index (Garrison *et al.* 2018) where we avoid to use succinct SDSL data structures (Gog *et al.* 2014). Instead, we rely on bit-compressed integer vectors, enabling efficient retrieval of path nucleotide positions to quickly compute nucleotide distances without having to store all pairwise distances between nodes in memory. This approach ensures to scale on large pangenome graphs representing thousands of whole genomes.

Graph layout initialization can significantly influence the quality of the final layout. In the 1D implementation, by default, nodes are placed in the same order as they appear in the input graph, although we also provide support for random layout initialization. In 2D, we offer several layout initialization techniques. One approach places nodes in the first layout dimension according to their order in the input graph, adding either uniform or Gaussian noise in the second dimension. Another strategy arranges nodes along a Hilbert curve, an approach that often favors the creation of planar final layouts. We also support fixing node positions to keep nodes in the same order as they are in a selected path, such as a reference genome. This feature allows us to build reference-focused graph layouts (Supplementary Fig. S1d).

Our implementation is multithreaded and uses shared memory for storing the layout in a vector, according to the HOGWILD! strategy (Recht *et al.* 2011). Threads perform layout updates without any locking for additional speed up. This approach is feasible since pangenome graphs are typically sparse (Guarracino *et al.* 2022), with low average node degree. As a result, the updates only modify small parts of the entire layout. While the HOGWILD! SGD algorithm writes the layout updates to a shared non-atomic double vector, PG-SGD stores node coordinates in a vector of atomic doubles. This vector prevents any potential memory overwrites. Our tests revealed basically no performance loss with respect to the non-atomic counterpart.

4 Results

4.1 Performance

We apply the 2D PG-SGD to the human pangenome (Liao *et al.* 2023) from the HPRC to show the scalability of the algorithm. Experiments were conducted on a cluster with 24 `Regular` nodes (32 cores/64 threads with two AMD EPYC 7343 processors with 512 GB RAM) and 4 `HighMem` nodes (64 cores/128 threads with two AMD EPYC 7513 processors with 2048 GB RAM). We downloaded pangenome graphs for each autosome (24 in total) and for the mitochondrial DNA. Each graph represents 90 whole human haplotypes: 44 diploid individuals plus the GRCh38 (Schneider *et al.* 2017) and CHM13 (Nurk *et al.* 2021) haploid human references (see Supplementary Table S1 for graph statistics). When applied to these pangenome graphs using one `Regular` node for each calculation, *odgi layout*'s 2D PG-SGD implementation obtains the graph layouts in 50 min on average, with the highest run time observed being chromosome 16 (Supplementary Table S1). This is expected since chromosome 16 has one of the highest levels of segmentally duplicated sequence among the human autosomes

(Martin *et al.* 2004). Repetitive sequences lead to graph nodes with a very high number of path steps, which are computationally expensive to work with (Guarracino *et al.* 2022). Memory consumption is 29.66 GB of RAM on average, with the memory peak again occurring with chromosome 16, due to the path index building phase. Given its scalability, we applied 2D PG-SGD to the full graph with all chromosomes together using a `HighMem` node (Supplementary Table S1). To compare, *BandageNG* (https://github.com/asl/BandageNG, last accessed July 2023), the current state of the art for graph visualization, was used to calculate a 2D layout of each of the HPRC pangenome graphs. For a fair comparison, we did not rely on *BandageNG*'s interactive GUI application, but we executed *BandageNG layout*, which directly emits a 2D graph layout similar to *odgi layout*. *BandageNG* was not able to produce a layout for the full graph within 7 days, hitting the wall clock time limit of the cluster. On average, PG-SGD is ∼8× faster than *BandageNG* while using ∼2× less memory.

4.2 Pangenome graph layouts reveal biological features

Graph visualization is essential for understanding pangenome graphs and the genome variation they represent. We show how 2D PG-SGD allows us gaining insight into biological data by looking at the graph layout structure. In Fig. 2a, the chromosomes of the HPRC graph show the large-scale structural variations in the centromeres. Focusing on the major histocompatibility complex (MHC) of chromosome 6 (Fig. 2b), the 2D layout reveals the positions and diversity of all MHC genes (Fig. 2c). In Fig. 2d, the C4A and C4B genes are highlighted. Complementary, we provide various 1D visualizations in Supplementary Fig. S1.

5 Discussion

We presented PG-SGD, the first layout algorithm for pangenome graphs that leverages the biological information available within the genomes represented in the graph. Other generic graph layout algorithms, such as the one offered by *BandageNG*, ignore this additional information. Our implementation efficiently computes the layout of pangenome graphs representing thousands of whole genomes.

Graph visualization is key for understanding genome variations and the layouts produced by PG-SGD offer an unprecedented high-level perspective on pangenome variation. We implemented PG-SGD to generate layouts in 1D and 2D. These graph projections have already been employed in constructing and analyzing the first draft human pangenome reference (Liao *et al.* 2023), as well as in the discovery of heterologous recombination of human acrocentric chromosomes (Guarracino *et al.* 2023). Furthermore, they are applied in the creation and analysis of pangenome graphs for any species (Guarracino *et al.* 2022, Garrison *et al.* 2023). Of note, there still remains a gap in interactive and scalable solutions that merge layouts of large pangenome graphs with annotation. Our algorithm will underpin new pangenome graph browsers for studying graph layouts and the genome variation they represent (https://github.com/chfi/waragraph, last accessed July 2023).

The performance analysis shows that our 2D implementation outperforms *BandageNG* when handling large, complex pangenome graphs. While *BandageNG* was not able to deliver a layout of the whole HPRC graph within 1 week, our 2D PG-SGD calculated one within one day. There are some possible optimization approaches for future work to further improve the performance of PG-SGD, making it possible for interactive use. The data structure could be optimized to improve cache performance. Moreover, the high-degree of parallelism could be further exploited by using a GPU. In *BandageNG*, one cannot select the number of threads for the calculations. They are automatically chosen by the number of connected components of the graph to draw. This limits its parallelism and leads to an unbalanced workload. Since *BandageNG* was primarily designed for assembly graphs, one may have to adjust its parameters dependent on the input graph, in order to boost the layout generation or to adjust the highlighting of desired graph features.

The classical force model of state of the art generic graph algorithms, such as FM^3-based ones, places nodes according to their attractive and repulsive forces. This force can be seen as equivalent to how our 2D PG-SGD moves the nodes' ends in 2D: If the nucleotide distance of the randomly chosen path steps is smaller than the layout distance of the nodes' ends, we move them closer together ("attractive force"), else we move them further away ("repulsive force"). However, the key difference here is that this approach is path-guided: paths represent biological sequences in pangenome graphs, so it is as if PG-SDG considers a "biological force" for placing the graph nodes. Theoretically, it would be possible to combine our approach with a force-directed one. Combining both methods, we might get the best of both worlds: multi-threadable PG-SGD iteratively applied to different graph layout-levels. We can imagine that such an approach can lead to a further speedup when calculating the layout. However, for generic graphs, this would only work if path information for each node could be added: we would replace the classical physical simulation approach with our path-guided method. If such information is not available, one could randomly cover the graph with paths. This function is already provided in *odgi cover*. However, this is an NP-hard problem and our preliminary solutions proved ineffective.

With assembly graphs we face the same problem: they usually do not carry path information during each assembly step. One could map the initial assembly reads back against the assembly graph in order to build paths through the graph. This would allow us to obtain a layout using PG-SGD.

PG-SGD can be extended to any number of dimensions. It can be seen as a graph embedding algorithm that converts high-dimensional, sparse pangenome graphs into low-dimensional, dense, and continuous vector spaces, while preserving its biologically relevant information. This enables the application of machine learning algorithms that use the graph layout for variant detection and classification. Our future research involves leveraging these graph projections to detect structural variants and to identify and correct assembly errors. Moreover, we are considering extending the algorithm to RNA and protein sequences to support pantranscriptome graphs (Sibbesen *et al.* 2023) and panproteome graphs (Dabbaghie *et al.* 2023), respectively.

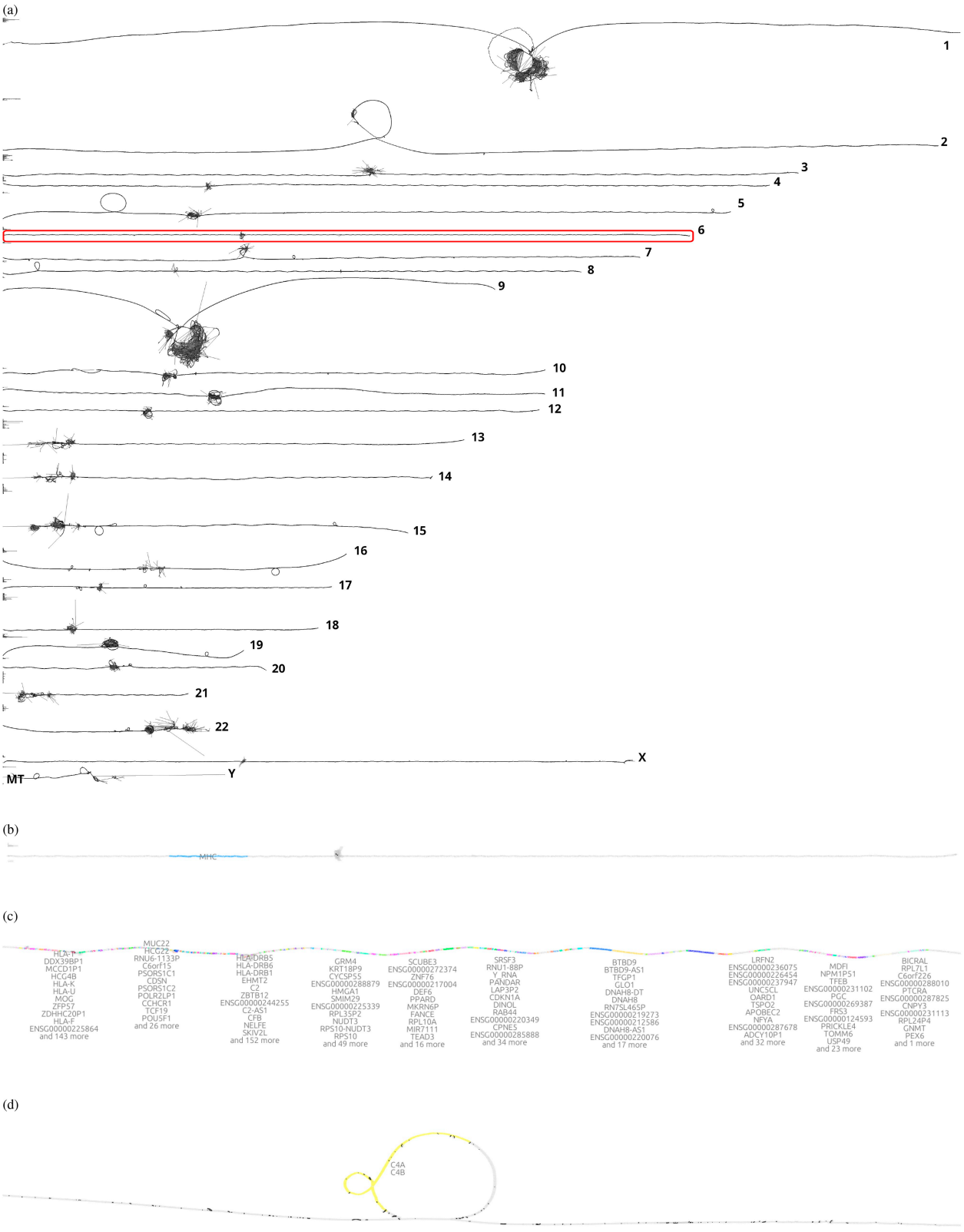

Figure 2. 2D visualizations of all chromosomes of the Human Pangenome Reference Consortium (HPRC) 90 haplotypes pangenome graph, chromosome 6, the major histocompatibility complex (MHC), and the complement component 4 (C4). (a) *odgi draw* layout of the HPRC pangenome graph 90 haplotypes. Displayed are all 24 autosomes and the mitochondrial chromosome. A red rectangle highlights chromosome 6 which is shown in the subfigure below. (b) *gfaestus* screenshot of the chromosome 6 layout. Colored in blue is the MHC. The hairball in the middle is the centromere. The black structures in the centromere are edges. (c) *gfaestus* screenshot of the MHC. All MHC genes are color annotated and the names of the genes appear as a text overlay. (d) *gfaestus* screenshot of the region around C4, specifically color highlighting genes C4A and C4B. The black lines are the edges of the graph.

Acknowledgements

We thank Matthias Seybold from the Quantitative Biology Center for maintaining the Core Facility Cluster.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

J.H is employed by Computomics GmbH.

Funding

This work was supported by the BMBF-funded de. NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) [031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, and 031A538A]. S.H. acknowledges funding from the Central Innovation Programme (ZIM) for SMEs of the Federal Ministry for Economic Affairs and Energy of Germany. S.N. acknowledges Germany's Excellence Strategy (CMFI), EXC-2124 and (iFIT)–EXC 2180–390900677. A.G. acknowledges efforts by Nicole Soranzo to establish a pangenome research unit at the Human Technopole in Milan, Italy. J.-N.M.S., J.L., Z.Z., P.P., and E.G. acknowledge funding from the NSF PPoSS Award #2118709. The authors gratefully acknowledge support from National Institutes of Health/NIDA U01DA047638 (E.G.), National Institutes of Health/NIGMS R01GM123489 (E.G. and P.P.).

Data availability

Software versions, code, and links to data used to prepare this manuscript can be found at https://github.com/pangenome/sorting-paper. Animations of the algorithm are deposited at https://doi.org/10.5281/zenodo.8288999.

References

Ballouz S, Dobin A, Gillis JA *et al.* Is it time to change the reference genome? *Genome Biol* 2019;**20**:159.

Cheong S-H, Si Y-W. Force-directed algorithms for schematic drawings and placement: a survey. *Inf Vis* 2019;**9**:65–91.

Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Brief Bioinform* 2018; **19**:118–35.

Dabbaghie F, Srikakulam SK, Marschall T *et al.* PanPA: generation and alignment of panproteome graphs. *Bioinformatics* 2023;**3**:vbad167.

Eizenga JM, Novak AM, Sibbesen JA *et al.* Pangenome graphs. *Annu Rev Genomics Hum Genet* 2020;**21**:139–62.

Garrison E. *Graphical pangenomics*. Apollo – University of Cambridge Repository 2019.

Garrison E, Sirén J, Novak AM *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol* 2018;**36**:875–9.

Garrison E, Guarracino A, Heumos S *et al.* Building pangenome graphs. bioRxiv 2023.

Gog S, Beller T, Moffat A *et al.* From theory to practice: plug and play with succinct data structures. In: *13th International Symposium on Experimental Algorithms, (SEA 2014). Springer International Publishing* 2014, 326–37.

Guarracino A, Heumos S, Nahnsen S *et al.* ODGI: understanding pangenome graphs. *Bioinformatics* 2022;**38**:3319–26.

Guarracino A, Buonaiuto S, de Lima LG *et al.* Recombination between heterologous human acrocentric chromosomes. *Nature* 2023; **617**:335–43.

Hachul S, Jünger M. Large-graph layout with the fast multipole multilevel method. Working Paper, Universität zu Köln, 2005.

Hein J. A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given. *Mol Biol Evol* 1989;**6**:649–68.

Liao W-W, Asri M, Ebler J *et al.* A draft human pangenome reference. *Nature* 2023;**617**:312–24.

Martin J, Han C, Gordon LA *et al.* The sequence and analysis of duplication-rich human chromosome 16. *Nature* 2004;**432**:988–94.

Nurk S, Koren S, Rhie A *et al.* The complete sequence of a human genome. *Science* 2022;**376**:44–53.

Recht B, Re C, Wright S *et al.* Hogwild!: a lock-free approach to parallelizing stochastic gradient descent. *Advances in Neural Information Processing Systems*; 24. Curran Associates, Inc., 2011.

Schneider VA, Graves-Lindsay T, Howe K *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* 2017;**27**:849–64.

Sherman RM, Salzberg SL. Pan-genomics in the human genome era. *Nat Rev Genet* 2020;**21**:243–54.

Sibbesen JA, Eizenga JM, Novak AM *et al.* Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. *Nat Methods* 2023;**20**:239–47.

Singh V, Pandey S, Bhardwaj A *et al.* From the reference human genome to human pangenome: premise, promise and challenge. *Front Genet* 2022;**13**:1042550.

Tettelin H, Riley D, Cattuto C *et al.* Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol* 2008;**11**:472–7.

Wang L, Wang X, Wang Q *et al.* Research on force-directed algorithm optimization methods. In: *Proceedings of the 2014 International Conference on e-Education, e-Business and Information Management (ICEEIM 2014), Shanghai, China*. Atlantis Press 2014.

Zheng JX, Pawar S, Goodman DFM *et al.* Graph drawing by stochastic gradient descent. *IEEE Trans Vis Comput Graph* 2019;**25**:2738–48.

Zipf GK. *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge, MA/London, England: Harvard University Press, 1932.

Bioinformatics, 2024, 40, 1–6
https://doi.org/10.1093/bioinformatics/btae363
Original Paper

Journal
doi.10.1093/xxxxxx/yyyyyy
Advance Access Publication Date: Day Month Year
Genome Analysis

Genome Analysis

Cluster efficient pangenome graph construction with nf-core/pangenome

Simon Heumos [1,2,3,4,*], Michael F. Heuer [], Friederike Hanssen [1,2,3,4], Lukas Heumos [5,6,7], Andrea Guarracino [8,9], Peter Heringer [1,2,3,4], Philipp Ehmele [5], Pjotr Prins [8], Erik Garrison [8], Sven Nahnsen [1,2,3,4,*]

[1]Quantitative Biology Center (QBiC) Tübingen, University of Tübingen, Tübingen, Germany
[2]Biomedical Data Science, Dept. of Computer Science, University of Tübingen, Tübingen, Germany
[3]M3 Research Center, University Hospital Tübingen, Tübingen, Germany
[4]Institute for Bioinformatics and Medical Informatics (IBMI), Eberhard-Karls University of Tübingen, Tübingen, Germany
[5]Institute of Computational Biology, Department of Computational Health, Helmholtz Munich, Germany
[6]Comprehensive Pneumology Center with the CPC-M bioArchive, Helmholtz Zentrum Munich, Member of the German Center for Lung Research (DZL), Munich, Germany
[7]TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany
[8]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, 71 S Manassas St, Memphis, 38163, Tennessee, USA
[9]Human Technopole, Viale Rita Levi-Montalcini 1, 20157, Milan, Italy

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Pangenome graphs offer a comprehensive way of capturing genomic variability across multiple genomes. However, current construction methods often introduce biases, excluding complex sequences or relying on references. The PanGenome Graph Builder (PGGB) addresses these issues. To date, though, there is no state-of-the-art pipeline allowing for easy deployment, efficient and dynamic use of available resources, and scalable usage at the same time.
Results: To overcome these limitations, we present *nf-core/pangenome*, a reference-unbiased approach implemented in Nextflow following nf-core's best practices. Leveraging biocontainers ensures portability and seamless deployment in HPC environments. Unlike PGGB, nf-core/pangenome distributes alignments across cluster nodes, enabling scalability. Demonstrating its efficiency, we constructed pangenome graphs for 1000 human chromosome 19 haplotypes and 2146 *E. coli* sequences, achieving a two to threefold speedup compared to PGGB without increasing greenhouse gas emissions.
Availability: nf-core/pangenome is released under the MIT open-source license, available on GitHub and Zenodo, with documentation accessible at https://nf-co.re/pangenome/1.1.2/docs/usage.
Contact: simon.heumos@qbic.uni-tuebingen.de, sven.nahnsen@qbic.uni-tuebingen.de

1 Introduction

The availability of high-quality collections of population-wide whole-genome assemblies (Liao *et al.*, 2023; Kang *et al.*, 2023; Weller *et al.*, 2023; Zhou *et al.*, 2022; Liu *et al.*, 2020; Leonard *et al.*, 2022) offers new opportunities to study sequence evolution and variation within and between genomic populations. A challenge is simultaneously representing and analyzing hundreds to thousands of genomes at a gigabase scale. One solution here is a pangenome. It models a population's entire set of genomic sequences (Ballouz *et al.*, 2019). In contrast to reference-based genomic approaches, which relate sequences to a linear genome, pangenomics relates each new sequence to all the others represented in the pangenome (The Computational Pan-Genomics Consortium, 2016; Eizenga *et al.*, 2020; Sherman and Salzberg, 2020) minimizing reference-bias. Pangenomes can be described as sequence graphs which store DNA sequences in nodes with edges connecting the nodes as they occur in the individual sequences (Hein, 1989). Genomes are encoded as paths traversing the nodes (Garrison *et al.*, 2018).

Current pangenome graph construction methods exclude complex sequences or are reference-biased (Chin *et al.*, 2023; Minkin *et al.*, 2016). One recent approach that overcomes such limitations is the PanGenome Graph Builder (PGGB) pipeline (Garrison *et al.*, 2023). PGGB iteratively refines an all-to-all whole-genome alignment graph that lets us explore sequence conservation and variation, infer phylogeny, and identify recombination events. PGGB was already extensively evaluated (Garrison *et al.*, 2023; Andreace *et al.*, 2023) and applied to build the first draft human pangenome reference (Liao *et al.*, 2023). However, PGGB is implemented in bash: This (a) makes it difficult to deploy on HPC systems, (b) does not allow for a fine granular tuning of computing resources for different steps of the pipeline (Sztuka *et al.*, 2024), and (c) limits its cluster scalability because PGGB can only use the resources of one node. These limitations greatly hinder the broad application of large-scale pangenomes.

To compensate for that, we wrote *nf-core/pangenome*, a reference-unbiased approach to construct pangenome graphs. Mirroring PGGB, nf-core/pangenome is implemented in Nextflow (Di Tommaso *et al.*, 2017). In contrast to PGGB, nf-core/pangenome can distribute the quadratic all-to-all base-level alignments across nodes of a cluster by splitting the approximate alignments into problems of equal size. We benchmarked the time spent on base-pair level alignments and show that it is reduced linearly with an increase in alignment problem chunks. We showcase the workflow's scalability by applying it to 1000 chromosome 19 human haplotypes, and to 2146 *E. coli* sequences, which were built in less than half the time PGGB required while not increasing the CO2 equivalent (CO2e) emissions.

2 Material and Methods

2.1 Pipeline overview

The pipeline's (Fig. 1a) input is a FASTA file compressed with *bgzip* (Li *et al.*, 2009) containing the sequences to create the graph. Sequence names should follow the Pangenome Sequence Naming specification (PanSN-spec) (Garrison, 2021). The primary output is a pangenome variation graph (Garrison *et al.*, 2018) in the Graphical Fragment Assembly (GFA) format version 1 (GFA Working Group, 2016).

2.1.1 Core workflow

The core workflow of nf-core/pangenome is an exact mirror of PGGB (Fig. 1a). The pipeline comes with additional enhancements: (a) All concurrent processes can be run in parallel. (b) Each process can be given individual computing resources.

The first step in the nf-core/pangenome pipeline is the all-to-all alignment of the input sequences with the whole-chromosome pairwise sequence aligner WFMASH (Guarracino *et al.*, 2024). This avoids reference, order, or orientation bias, and allows each sequence in the pangenome to serve as a reference when exploring related variation. In the pangenome graph induction step SEQWISH (Garrison and Guarracino, 2022), an alignment to variation graph inducer, converts the sequence alignments into a variation graph. We then normalize the graph with the variation graph simplification algorithm SMOOTHXG (Garrison *et al.*, 2023): A 1-dimensional (1D) graph embedding (Heumos *et al.*, 2023) orders the graphs' nodes to best-match the nucleotide distances of the genomic paths of the graph. Next, the graph is split into partially overlapping segments. The sequences of each segment are realigned with a local Multiple Sequence Alignment (MSA) kernel, partial order alignment (POA) (Lee *et al.*, 2002). Afterwards, the segments are laced back together into a variation graph. By default, the SMOOTHXG process is applied 3 times in order to smoothen the edge effects at the boundaries of the segments. Finally, we employ GFAFFIX (Liao *et al.*, 2023) to systematically condense redundant nodes within the graph.

Basic graph build quality is evaluated with ODGI: Optimized Dynamic Genome/Graph Implementation (Guarracino *et al.*, 2022) for understanding pangenome graphs. ODGI reports basic graph statistics and diagnostic 1D and 2D visualizations. Optionally, nf-core/pangenome calls variants against any (reference) path(s) in the graph using *vg deconstruct* (Garrison *et al.*, 2018). Finally, graph statistics and visualizations are summarized in a MultiQC (Ewels *et al.*, 2016) report. Pipeline implementation details are given in Suppl. 5.1.

3 Results

3.1 Alignment jobs distribution evaluation

Generating all-vs-all alignments is a computationally quadratic problem. To evaluate nf-core/pangenome's alignment jobs scalability (detailed in Suppl. 5.5), we applied it to 1024 *E. coli* genomes with varying numbers of chunks. nf-core/pangenome's alignment jobs distribution linearly reduces the time spent on base-pair level alignments with increased chunks. The CO2 consumption is not influenced by the number of chunks (Suppl. Fig. 5.5).

3.2 Building a 1000 haplotypes chr19 pangenome graph

The Human Pangenome Resource Consortium (HPRC) recently built a draft human pangenome reference of 90 haplotypes (Liao *et al.*, 2023). However, haplotype data for thousands of individuals already exists generated by the 1000 Genomes Project (1KGP) (Durbin *et al.*, 2010). As a use case study, we used nf-core/pangenome to build a pangenome graph of 1000 chromosome 19 haplotypes (Kuhnle *et al.*, 2020) within 3 days. The CO2e was 22.52 kg. PGGB built the same graph within 7 days. In Fig. 1b the pangenome growth curve generated with PANACUS (Liao *et al.*, 2023) shows a growth of the number of nucleotides with an increasing number of haplotypes. The size of the softcore pangenome does not change with increasing numbers of haplotypes.

3.3 Building a 2146 sequences *E. coli* pangenome graph

To evaluate the pipeline's scalability, we built a pangenome graph of 2146 *E. coli* sequences. nf-core/pangenome built the pangenome graph in 10 days, emitting 175.18 kg of CO2e. Due to wall clock time restrictions on our cluster, PGGB was not able to finish the graph construction within 30 days. To build a reasonable pangenome growth curve (Fig. 1c) we dropped all paths containing "plasmid" (130 in total) in their name. The softcore pangenome of the graph does not change with an increasing number of haplotypes (stable at 3Mb of sequence), but the general growth curve is steep.

4 Discussion

We implemented nf-core/pangenome, an easy-to-install, portable, and cluster-scalable pipeline for the unbiased construction of pangenome variation graphs. It is the first pangenomic nf-core pipeline enabling the comparative analysis of gigabase-scale pangenome datasets. The pipeline's core workflow steps were already successfully applied to *Neisseria meningitidis* (Yang *et al.*, 2023), wild grapes (Cochetel *et al.*, 2023), humans (Guarracino *et al.*, 2023; Liao *et al.*, 2023), grapevines (Guo *et al.*, 2024), taurines (Milia *et al.*, 2024), and rats (Villani *et al.*, 2024) underpinning the community effort to focus on a best-practice workflow to create reference-unbiased and sequence complete pangenome graphs. The modular domain-specific language (DSL) 2 pipeline structure eases the exchange of key processes with alternative tools, the extent of the pipeline with new tools, and the integration of parts of the pipeline with other (sub-)workflows.

We have shown that we are able to perform all-vs-all base pair level alignments of thousands of sequences. When executed on an HPC, nf-core/pangenome's parallel workflow accelerates graph construction compared to PGGB. PGGB's inability to assign individual computational

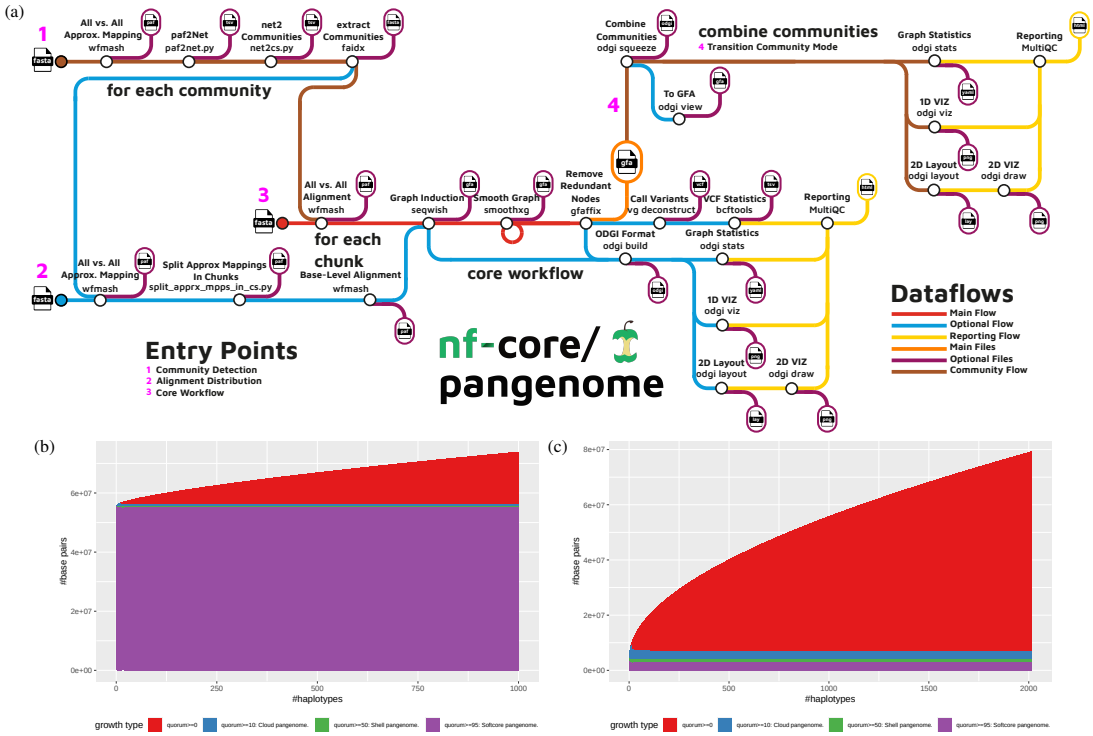

Fig. 1. (a) Schematic representation of the nf-core/pangenome workflow processes and detailed analysis steps. The input consists of one FASTA file containing all sequences. The pipeline comes with 3 major entry points: Community detection (1), alignment distribution (2), and core workflow (3). Optional community detection (1) is performed on the input sequences. If selected, the heavy all-to-all baise-pair level alignments (2) can be split into problems of equal size. nf-core/pangenome's core workflow (3) is a direct mirror of PGGB. If running in community mode, all communal graphs are combined into one (4) and the subsequent quality control subworkflow is executed. The output is a pangenome graph in GFA format. (b) + (c) Pangenome growth curves of the built pangenome graphs. Growth type is defined as the minimum fraction of haplotypes that must share a graph feature after each time a haplotype is added to the growth histograph. $quorum >= 0$: All sequences without any filtering are considered. $quorum >= 10$: Sequences traversed by at least 10% of the haplotypes. $quorum >= 50$: Sequences traversed by at least 50% of haplotypes. $quorum >= 95$: Sequences traversed by 95% of haplotypes. (b) Pangenome growth curve of the chromosome 19 pangenome graph of 1000 haplotypes. (c) Pangenome growth curve of the E. coli pangenome graph of 2013 haplotypes.

resources to each pipeline step leads to the allocation of one whole node of an HPC, despite the fact that some processes can only make use of one thread. This blocks valuable CPU cycles for other users working on the same HPC and ultimately can lead to additional costs. In contrast, nf-core/pangenome does not have such limitations: Nextflow's process management enables the optimal workload of given compute resources which can be especially important when running a pipeline in commercial clouds.

Competing pipelines don't use any workflow management system to connect their processes (Chin *et al.*, 2023), or their workflow language of choice is e.g. Toil (Vivian *et al.*, 2017; Hickey *et al.*, 2023) which makes them less user-friendly, less cluster efficient, and less portable (Wratten *et al.*, 2021). nf-core/pangenome is currently the only pangenomics pipeline that is optionally monitoring its CO2 footprint. The measurements have shown that constructing extensive pangenome graphs, such as the 2146 *E. coli* graph, requires a considerable amount of energy. Therefore, before executing environmentally questionable experiments, we would recommend thoroughly assessing both the rationale and the methodology.

Although, we expect our pipeline to scale well for future pangenome graph construction challenges, such as for the next HPRC phase which targets 300 individuals, there still is potential for further optimization: IMplicit Pangenome Graph (IMPG) (https://github.com/ekg/impg), a tool that extracts homologous loci from all genomes mapped to a specific target region. This would allow us to break the whole genome multiple alignments into smaller pieces, construct a pangenome graph for each piece, and lace these together into a full graph with https://github.com/pangenome/gfalace.

We anticipate the pipeline, or its parts, will enhance current single linear reference analysis methods to explore whole population variation instead of focusing on one reference only. Looking ahead, pangenome construction pipelines like nf-core/pangenome will play a pivotal role in studying entire populations, single-cell whole genome sequencing analysis, and constructing personalized (medical) pangenome references (Sirén *et al.*, 2023).

Software and data availability

Code and links to data resources used to build this manuscript and its figures, can be found in the paper's public repository: https://github.com/subwaystation/pangenome-paper.

Acknowledgments

We thank Matthias Seybold from QBiC for maintaining the Core Facility Cluster. We thank Sabrina Krakau from QBiC for giving feedback to the nf-co2footprint plugin section. We are grateful to the nf-core community

for their support during the implementation of the pipeline. From the nf-core community, we want to thank Matthias Hörtenhuber, Maxime Garcia, Susanne Jodoin, Julia Mir Petrol, Adam Talbot, and Gisela Gabernet.

Funding

S.H. acknowledges funding from the Central Innovation Programme (ZIM) for SMEs of the Federal Ministry for Economic Affairs and Energy of Germany. This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A). A.G. acknowledges support from the Human Technopole. S.N. acknowledges support from iFIT funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2180—390900677 and CMFI under EXC 2124–390838134.

Competing interests

Author L.H. is employed by LaminLabs.

References

Andreace, F. *et al.* (2023). Comparing methods for constructing and representing human pangenome graphs. *Genome Biology*, **24**(1).

Ballouz, S. *et al.* (2019). Is it time to change the reference genome? *Genome Biology*, **20**(1), 159.

Breitwieser, F. P. *et al.* (2019). Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Research*, **29**(6), 954–960.

Chin, C.-S. *et al.* (2023). Multiscale analysis of pangenomes enables improved representation of genomic diversity for repetitive and clinically relevant genes. *Nature Methods*, **20**(8), 1213–1221.

Cochetel, N. *et al.* (2023). A super-pangenome of the north american wild grape species. *Genome Biology*, **24**(1).

Di Tommaso, P. *et al.* (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, **35**(4), 316–319.

Durbin, R. M. *et al.* (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319), 1061–1073.

Eizenga, J. M. *et al.* (2020). Pangenome graphs. *Annual Review of Genomics and Human Genetics*, **21**(1), 139–162.

Ewels, P. *et al.* (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**(19), 3047–3048.

Garrison, E. (2021). Pansn-spec: Pangenome sequence naming. https://github.com/pangenome/PanSN-spec.

Garrison, E. and Guarracino, A. (2022). Unbiased pangenome graphs. *Bioinformatics*, **39**(1).

Garrison, E. *et al.* (2018). Variation Graph Toolkit Improves Read Mapping by Representing Genetic Variation in the Reference. *Nature Biotechnology*, **36**(9), 875–879.

Garrison, E. *et al.* (2023). Building pangenome graphs. *bioRxiv*.

GFA Working Group (2016). Graphical fragment assembly (gfa) format specification. https://github.com/GFA-spec/GFA-spec.

Guarracino, A. *et al.* (2022). ODGI: understanding pangenome graphs. *Bioinformatics*, **38**(13), 3319–3326.

Guarracino, A. *et al.* (2023). Recombination between heterologous human acrocentric chromosomes. *Nature*, **617**(7960), 335–343.

Guarracino, A. *et al.* (2024). wfmash: whole-chromosome pairwise alignment using the hierarchical wavefront algorithm. https://github.com/waveygang/wfmash.

Guo, L. *et al.* (2024). Super pangenome of grapevines empowers improvement of the oldest domesticated fruit.

Hein, J. (1989). A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given. *Molecular Biology and Evolution*.

Heumos, S. *et al.* (2023). Pangenome graph layout by path-guided stochastic gradient descent.

Hickey, G. *et al.* (2023). Pangenome graph construction from genome alignments with minigraph-cactus. *Nature Biotechnology*, **42**(4), 663–673.

Kang, M. *et al.* (2023). The pan-genome and local adaptation of arabidopsis thaliana. *Nature Communications*, **14**(1).

Kuhnle, A. *et al.* (2020). Efficient construction of a complete index for pan-genomics read alignment. *Journal of Computational Biology*, **27**(4), 500–513.

Lannelongue, L. *et al.* (2021). Green algorithms: Quantifying the carbon footprint of computation. *Advanced Science*, **8**(12).

Lee, C. *et al.* (2002). Multiple sequence alignment using partial order graphs. *Bioinformatics*, **18**(3), 452–464.

Leonard, A. S. *et al.* (2022). Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. *Nature Communications*, **13**(1).

Li, H. *et al.* (2009). The sequence alignment/map format and samtools. *Bioinformatics*, **25**(16), 2078–2079. 19505943[pmid].

Liao, W.-W. *et al.* (2023). A draft human pangenome reference. *Nature*, **617**(7960), 312–324.

Liu, Y. *et al.* (2020). Pan-genome of wild and cultivated soybeans. *Cell*, **182**(1), 162–176.e13.

Milia, S. *et al.* (2024). Taurine pangenome uncovers a segmental duplication upstream ofkitassociated with depigmentation in white-headed cattle.

Minkin, I. *et al.* (2016). Twopaco: an efficient algorithm to build the compacted de bruijn graph from many complete genomes. *Bioinformatics*, **33**(24), 4024–4032.

Sayers, E. W. *et al.* (2021). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, **50**(D1), D20–D26.

Sherman, R. M. and Salzberg, S. L. (2020). Pan-genomics in the human genome era. *Nature Reviews Genetics*, **21**(4), 243–254.

Sirén, J. *et al.* (2023). Personalized pangenome references.

Sztuka, M. *et al.* (2024). Nextflow vs. plain bash: different approaches to the parallelization of SNP calling from the whole genome sequence data. *NAR Genomics and Bioinformatics*, **6**(2), lqae040.

The Computational Pan-Genomics Consortium (2016). Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*, page bbw089.

Traag, V. A. *et al.* (2019). From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, **9**(1).

Villani, F. *et al.* (2024). Pangenome reconstruction in rats enhances genotype-phenotype mapping and novel variant discovery.

Vivian, J. *et al.* (2017). Toil enables reproducible, open source, big biomedical data analyses. *Nature Biotechnology*, **35**(4), 314–316.

Weller, C. A. *et al.* (2023). Highly complete long-read genomes reveal pangenomic variation underlying yeast phenotypic diversity. *Genome Research*, **33**(5), 729–740.

Wratten, L. *et al.* (2021). Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature Methods*, **18**(10), 1161–1168.

Yang, Z. *et al.* (2023). Pangenome graphs in infectious disease: a comprehensive genetic variation analysis of neisseria meningitidis leveraging oxford nanopore long reads. *Frontiers in Genetics*, **14**.

Zhou, Y. *et al.* (2022). Assembly of a pangenome for global cattle reveals missing sequences and novel structural variations, providing new insights into their diversity and evolutionary history. *Genome Research*, **32**(8), 1585–1601.

5 Supplement

5.1 Implementation

nf-core/pangenome is written in Nextflow using its latest domain-specific language (DSL) 2 syntax which facilitates a modular pipeline structure. Each software tool is an individual process that is implemented in its own module (https://nf-co.re/docs/contributing/modules). Processes are concatenated into subworkflows (https://nf-co.re/docs/contributing/subworkflows). Developed with the nf-core framework, the pipeline follows a set of best-practice guidelines ensuring high-quality development, maintenance, and testing standards. Specifically, we provide community support via a dedicated Slack channel (https://nfcore.slack.com/channels/pangenome), GitHub issues, and detailed documentation (https://nf-co.re/pangenome/1.1.2/docs/usage). Versioning and portability are enabled through (a) semantic versioning (https://semver.org/) of the pipeline via tagged releases on GitHub, (b) packaging software dependencies in archivable containers so that the software compute environment is the same across different systems, and (c) summarizing software versions and parameters in the MultiQC report of the pipeline. nf-core/ pangenome uses biocontainers to facilitate portability across different computing resources like HPC clusters, cloud platforms, or local machines. Code changes are evaluated with GitHub Actions' continuous integration (CI) using a pipeline-specific small test data set. For each new pipeline release, a full-size test is run on Amazon Web Services (AWS) validating the code integrity and cloud compatibility of real-world data sets. Specifically, a pangenome graph is created from the 8 Saccharomyces cerevisiae strains of the Yeast Population Reference Panel (YPRP) (Yue and Liti 2018). The results of such a run are available on the nf-core webpage (https://nf-co.re/pangenome/1.1.2/results/pangenome/results-0e8a38734ea3c0397f94416a0146a2972fe2db8b). Because we implemented our processes using DSL2 nf-core/modules (https://github.com/nf-core/modules), they can be distributed easily to other users to share commonly used processes or subworkflows across pipelines. This boosts the reuse of existing work done by the community to be integrated into future pipelines.

5.2 Chromosome community detection

Eukaryotic genomes are usually organized into chromosomes. Taking this into account during graph construction, the chromosome groupings from the input sequence are examined. Specifically, the homologies detected in the all-to-all WFMASH mapping step are put into the Leiden (Traag *et al.*, 2019) clustering algorithm. The edge weight is $mapped\_length * mapped\_identity$. For each of the resulting communities, the nf-core/pangenome core workflow is executed in parallel. The communal graphs are joined into one and a final round of quality control is applied (see Fig 1a, brown tubes). In practice, this works well for large input sequences with a large mapping length (>1Mb) filter, which was demonstrated by Guarracino *et al.* (2023) when exploring the recombination between heterologous human acrocentric chromosomes.

5.3 Compute environment

We applied the nf-core/pangenome pipeline v1.1.2 to various inputs evaluating both the scalability of the all-vs-all alignment step as well as the pipeline as a whole. We used Nextflow version 23.10.1.5891 and Singularity version 3.8.7-1.el8 for each pipeline run. Experiments were conducted on our core facility cluster (CFC) with 24 Regular nodes (32 cores / 64 threads with two AMD EPYC 7343 processors with 512 GB RAM and 2 TB scratch space) and 4 HighMem nodes (64 cores / 128 threads with two AMD EPYC 7513 processors with 2048 GB RAM and 4TB scratch space). Each Nextflow process was given at most 64 threads. This ensures a fair run time comparison with PGGB v0.5.4 which was always executed on one Regular node via Slurm.

5.4 Estimation of the carbon footprint of pipeline runs

We also estimated the carbon dioxide equivalent (CO2e) emissions of each nf-core/pangenome pipeline run using the nf-co2footprint Nextflow plugin (https://github.com/nextflow-io/nf-co2footprint) v1.0.0-beta. Using the Nextflow resource usage metrics and information about the power consumption of the compute system, the plugin first estimates the energy consumption for each pipeline task. It then uses the consumed energy's location-specific carbon intensity to estimate the respective CO2e emission. The calculations are based on the carbon footprint computation method developed in the Green Algorithms project (www.green-algorithms.org) (Lannelongue *et al.*, 2021).

5.5 Alignment jobs distribution

The computationally heavy all versus all base-pair level alignments can be distributed across nodes of a cluster: First, WFMASH is run in mapping mode (WFMASH MAP), finding all sequence homologies using approximate alignments. The resulting Pairwise mApping Format (PAF) file is split into chunks of equal problem size. The number of chunks is manually selected. The value can be guided by the number and size of the input sequences, and by the available hardware. Assuming the number of chunks equals the number of nodes on a cluster, then potentially each base-pair level alignment (WFMASH ALIGN) can be run in parallel on each node (Fig. 1a, cyan tubes). All resulting PAFs are then forwarded to the pipeline's core workflow which is continued at the SEQWISH process.

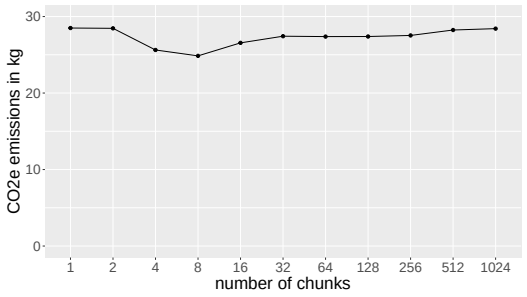

Fig. S1. Base-pair level alignment evaluation. CO2e emissions are stable across varying numbers of chunks.

5.6 1KGP chromosome 19 data set

The FASTA of the chromosome 19 data set was downloaded in December 2023 from http://dolomit.cs.tu-dortmund.de/chr19.1000.fa.xz. The data set is described in (Kuhnle *et al.*, 2020). Statistics of the built pangenome graph can be seen in Supplementary Fig. 5.6. The initial graph contains over 97% of Ns. We applied *odgi crush* (odgi version 0.8.6), which crushes consecutive Ns of all nodes containing Ns into just one N per node, to the 1KGP chromosome 19 pangenome graph. This brings down the number of Ns from over 3B to exactly 6000 (Suppl. Fig. 5.6). The 2D visualization (Suppl. Fig. 5.6) is perfectly linear without any large SVs present hinting that only short-read data was used to create the haplotype sequences. In contrast, the 2D layout of the HPRC PGGB chromosome 19 pangenome graph (Heumos *et al.*, 2023) clearly presents SVs, especially in the centromere's location. Investigating these complex regions of a human

Sample Name	Length	Nodes	Edges	Paths	Components	A	C	T	G	N
chr19.1000	3 393 915 450	2 594 408	3 498 791	1 000	1	19 633 450	16 968 263	19 854 852	17 458 885	3 320 000 000
chr19.1000.crush	73 921 450	2 594 408	3 498 791	1 000	1	19 633 450	16 968 263	19 854 852	17 458 885	6 000

Fig. S2. Screenshot of the output of ODGI's MultiQC module displaying the vital graph statistics calculated by odgi stats of the 1000 Genomes Project 1000 haplotypes chromosome 19 pangenome graphs. In the crushed graph consecutive Ns of all nodes containing Ns were merged into just one N per node. A: Number of adenine bases in the graph. C: Number of cytosine bases in the graph. T: Number of thymine bases in the graph. G: Number of guanine bases in the graph. N: Number of bases with unknown base identity.

Fig. S3. odgi draw 2D layout displaying the graph topology of the crushed 1KGP pangenome graph. Structural variation would appear as bubbles.

chromosome is only possible when using long-read assemblies for graph construction.

5.7 *E.coli* data set

The 2146 full length E. coli sequences originate from Genbank (Sayers *et al.*, 2021) and were downloaded 18 months ago. The initial pangenome consisted of 2 graphical components (Suppl. Fig. 5.7). This means that no strong homologies were found in some sequences. There can be many reasons for additional graph components: (a) The chosen sequence identity during the WFMASH mapping was not low enough. Although we went for a low 90% sequence identity (as was done by Garrison *et al.* (2023)), we still observe this additional graph component, so its sequence must be quite dissimilar to all other sequences. (b) There is human contamination in the bacterial sequences (Breitwieser *et al.*, 2019). (c) Some sequences from GenBank may be of a low quality or were misassembled. We then used odgi explode to extract the largest graphical component, applied *odgi crush* and dropped all paths containing "plasmid" in their path name with *odgi paths*. This left us with one component and 2013 paths. In the 2D visualization, we observe a highly connected graph (Suppl. Fig. 5.7). All the reasons mentioned above, but especially horizontal gene transfer could explain this phenomenon. Therefore, there are a lot of edge crossings in the pangenome graph. The long stretches is dangling sequence. We speculate that here the 88 thousand Ns could play role.

Sample Name	Length	Nodes	Edges	Paths	Components	A	C	T	G	N
ecoli_2146	140 465 404	6 118 071	8 972 252	2 146	2	21 671 240	19 111 201	21 699 33	19 131 791	58 851 837
ecoli2146.pan.explode	140 321 750	6 113 871	8 966 644	2 143	1	21 631 63	19 077 54	21 663 73	19 096 998	58 851 837
ecoli2146.pan.explode.crush	81 557 898	6 113 871	8 966 644	2 143	1	21 631 63	19 077 54	21 663 73	19 096 998	87 985
ecoli2146.pan.explode.crush.no_plasmids	81 557 898	6 113 871	8 966 644	2 013	1	21 631 63	19 077 54	21 663 73	19 096 998	87 985

Fig. S4. Excerpt of the 2146 sequences E. coli pangenome graph's MultiQC report. Displayed are vital graph statistics by MultiQC's ODGI module. A: Number of adenine bases in the graph. C: Number of cytosine bases in the graph. T: Number of thymine bases in the graph. G: Number of guanine bases in the graph. N: Number of bases with unknown base identity.

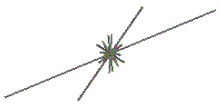

Fig. S5. odgi draw 2D layout visualization of the 2013 haplotypes E. coli pangenome graph.

Bioinformatics, 38(13), 2022, 3319–3326
https://doi.org/10.1093/bioinformatics/btac308
Advance Access Publication Date: 13 May 2022
Original Paper

OXFORD

Genome analysis

ODGI: understanding pangenome graphs

Andrea Guarracino [1,†], Simon Heumos [2,3,†], Sven Nahnsen [2,3], Pjotr Prins[4] and Erik Garrison [4,*]

[1]Genomics Research Centre, Human Technopole, Milan 20157, Italy, [2]Quantitative Biology Center (QBiC), University of Tübingen, Tübingen 72076, Germany, [3]Biomedical Data Science, Department of Computer Science, University of Tübingen, Tübingen 72076, Germany and [4]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN 38163, USA

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Peter Robinson

Received on November 9, 2021; revised on March 18, 2022; editorial decision on April 23, 2022

Abstract

Motivation: Pangenome graphs provide a complete representation of the mutual alignment of collections of genomes. These models offer the opportunity to study the entire genomic diversity of a population, including structurally complex regions. Nevertheless, analyzing hundreds of gigabase-scale genomes using pangenome graphs is difficult as it is not well-supported by existing tools. Hence, fast and versatile software is required to ask advanced questions to such data in an efficient way.

Results: We wrote Optimized Dynamic Genome/Graph Implementation (ODGI), a novel suite of tools that implements scalable algorithms and has an efficient in-memory representation of DNA pangenome graphs in the form of variation graphs. ODGI supports pre-built graphs in the Graphical Fragment Assembly format. ODGI includes tools for detecting complex regions, extracting pangenomic loci, removing artifacts, exploratory analysis, manipulation, validation and visualization. Its fast parallel execution facilitates routine pangenomic tasks, as well as pipelines that can quickly answer complex biological questions of gigabase-scale pangenome graphs.

Availability and implementation: ODGI is published as free software under the MIT open source license. Source code can be downloaded from https://github.com/pangenome/odgi and documentation is available at https://odgi.readthedocs.io. ODGI can be installed via Bioconda https://bioconda.github.io/recipes/odgi/README.html or GNU Guix https://github.com/pangenome/odgi/blob/master/guix.scm.

Contact: egarris5@uthsc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A pangenome models the full set of genomic elements in a given species or clade (Computational Pan-Genomics Consortium, 2018; Eizenga *et al.*, 2020b; Tettelin *et al.*, 2008). In contrast to reference-based approaches which relate samples to a single genome, these data structures encode the mutual relationships between all the genomes represented (Ballouz *et al.*, 2019). A class of methods to represent pangenomes involves sequence graphs (Hein, 1989; Paten *et al.*, 2017) where homologous regions between genomes are compressed into single representations of all alleles present in the pangenome. In sequence graphs, node labels are genomic sequences with edges connecting those nodes. A bidirected sequence graph can represent both strands of DNA. On this model, variation graphs add the concept of paths representing linear DNA sequences as traversals through the nodes of the graph (Garrison *et al.*, 2018). For example, a path can be a genome, haplotype, contig or read.

Pangenome graphs can be constructed by multiple sequence alignment (Grasso and Lee, 2004; Lee *et al.*, 2002) or by transitively reducing an alignment between sequences to an equivalent, labeled sequence graph (Garrison, 2019; Kehr *et al.*, 2014). Current methods to build these graphs are still under active development (Armstrong *et al.*, 2020; Garrison *et al.*, 2021; Li *et al.*, 2020), but they have largely settled on a common data model, represented in the Graphical Fragment Assembly (GFA) format (GFA Working Group, 2016). This standardization supports the development of a reference set of tools that operate on the pangenome graph model.

Pangenome graphs let us encode any kind of variation, allowing the generation of comprehensive data systems that builds the basis for the analyses of genome evolution. The Human Pangenome

Reference Consortium (HPRC) and Telomere-to-Telomere (T2T) consortium (Jarvis *et al.*, 2022; Logsdon *et al.*, 2021; Miga *et al.*, 2020; Nurk *et al.*, 2021) have recently demonstrated that high-quality haploid and diploid *de novo* assemblies can be routinely generated from third-generation long read sequencing data. We anticipate that *de novo* assemblies of similar quality will become common, leading to demand for methods to analyze pangenomes.

Although pangenome graphs are data structures of utility to researchers (Baaijens *et al.*, 2019; Computational Pan-Genomics Consortium, 2018; Garrison *et al.*, 2018; Hickey *et al.*, 2020; Sibbesen *et al.*, 2021), the scientific community still lacks a toolset capable of operating on gigabase-scale pangenome graphs constructed from whole-genome assemblies. Such an effort began with the VG toolkit (Garrison *et al.*, 2018), but its tools do not efficiently handle pangenome graphs presenting complex motifs that result from repetitive sequences. Here, we refocus the effort with the Optimized Dynamic Genome/Graph Implementation (ODGI) toolkit, a compatible, but independent pangenome graph interrogation and transformation system specifically implemented to handle the data scales encountered when working with pre-built constructed pangenomes comprising hundreds of haplotype-resolved genomes. ODGI offers a set of standard operations on the variation graph data model (Fig. 1), generalizing 'genome arithmetic' concepts, like those found in BEDTools (Quinlan and Hall, 2010), to work on pangenome graphs. Furthermore, it provides a variety of tools for graph visualization, sorting and liftover projections, all critical to understand and exploit pangenome graphs.

2 Model

A pangenome graph is a sequence model that encodes the mutual alignment of many genomes (Eizenga *et al.*, 2020b; Garrison, 2019). In the variation graph, $V = (N, E, P)$, nodes $N = n_1 \dots n_{|N|}$ contain genomic sequences. Each node n_i has an identifier i and an implicit reverse complement $\overline{n_i}$, and a node strand s corresponds to one of such orientations. Edges $E = e_1 \dots e_{|E|}$ represent ordered pairs of node strands: $e_i = (s_a, s_b)$. Paths $P = p_1 \dots p_{|P|}$ describe walks over node strands: $p_i = s_1 \dots s_{|p_i|}$. When used as a pangenome graph, V expresses sequences, haplotypes, contigs and annotations as paths. By containing both the sequences and information about their relative variations, the variation graph provides a complete and powerful foundation for many bioinformatic applications.

3 Implementation

The ODGI toolkit builds on existing approaches to efficiently store and manipulate pangenome graphs in the form of variation graphs (Garrison *et al.*, 2018). Similar to other efficient libraries presenting the HandleGraph model (Eizenga *et al.*, 2020a), the implementation of ODGI's tools rests on three key properties which hold for most pangenome graphs:

1. They are relatively sparse, with low average node degree.
2. They can be sorted so that most edges go between nodes that are close together in the sort order.
3. Their embedded paths are locally similar to each other.

These properties are used to build efficient dynamic variation graph data structures (Eizenga *et al.*, 2020a; Siren *et al.*, 2020). Sparsity (1) allows us to encode edges E using adjacency lists rather than matrices or hash tables. The local linear structure of the graph (2) lets us assign node identifiers that increase along the linear components of the graph, which supports a compact storage of edges and path steps as relativistic (usually small) differences rather than absolute (always large) integer identifiers. Path similarity (3) allows us to write local compressors that reduce the storage cost of collections of path steps.

ODGI improves on prior efforts, based on issues that arose during our work with high-quality *de novo* assemblies that cover almost all parts of the human genome (Logsdon *et al.*, 2021; Nurk *et al.*, 2021). In particular, we find that it is necessary to support graphs with regions of very high numbers of path traversals (high depth of path coverage of some nodes, the so-called node depth). Such motifs can occur in collapsed structures generated by ambiguous sequence homology relationships in repeats found in the centromeres and other segmental duplications. If we cannot process such regions, we cannot understand them, and our only option is to build graphs that do not include them. Our goal is to build tools that allow for a wide range of uses of pangenome graphs, including cases with potentially high path depth. To seamlessly represent such difficult regions, we followed an approach implemented in the dynamic version of the Graph BWT (GBWT) (Siren *et al.*, 2020) and built a node-centric, dynamic, compressed model of the paths. This design supports node-local modification and update of the graph, which lets us build and modify the graph and its paths in parallel.

We store the graph in a vector of node structures, each of which presents a node-local view of the graph sequence, topology and path layout (Algorithm 1). Expressed in terms of the variation graph V, ODGI's core *Node* structure includes a decoder that maps the neighbors of each node to a dense range of integers. For a given $Node_i$ and neighbor $Node_j$, the decoder itself does not store the *id* of $Node_j$, but rather a compact representation of the relative difference between the node ids: $\delta = Node_i.id - Node_j.id$. This keeps the size of the encoding small, per common pangenome graph property (2). We define the edges and path steps traversing the node in terms of this alphabet of $\boldsymbol{\delta}$'s.Each structure contains the sequence of the node ($Node_i.sequence$), its edges in both directions ($Node_i.edges$), and a vector of path steps that describes the previous and next steps in paths that walk across the node ($Node_i.path\_steps$). For efficiency, $Node_i.sequence$ is stored as a plain string, while the *edges* and *path_steps* are stored using a dynamic succinct integer vector that requires $O(2nw)$ bits for the edges and $O(5nw)$ bits for the path steps, where n is the number of steps on the node and w is $\approx log_2(n)$ (Prezza, 2017).

To allow edit operations in parallel, each node structure includes a byte-width mutex *lock*. All changes on the graph can involve at most two *Node* structs at a time (both edge and path step representations are doubly linked). To avoid deadlocks, we acquire the node

Algorithm 1: ODGI's relativistically packed *Node* structure and the *Step* structure used to represent the paths as doubly linked lists.

Struct *Node* **contains**
  id $\in \mathbb{N}$ // an identifier
  lock // atomic locking primitive
  sequence = [A|T|G|C|N]+
  // bit-packed vector of edges
  edges $= (x_i, x_j)* : (i, j) \in [1 \dots \Sigma]^2$
  // bit-packed vector of id deltas
  decoding $x_1 \dots x_\Sigma \in \mathbb{N}^\Sigma$
  // bit-packed vector of path steps
  path_steps $[Step_1 \dots Step_n]*$
end
Struct *Step* **contains**
  path_id $\in \mathbb{N}$ // the path's global id
  is_rev $\in (0, 1)$ // the step orientation
  is_start $\in (0, 1)$ // if first step in path
  is_end $\in (0, 1)$ // if last step in path
  prev_$\boldsymbol{\delta}$ $\in [1 \dots \Sigma]$ // δ-encoded previous node
  prev_rank $\in \mathbb{N}$ // step rank on previous node
  next_$\boldsymbol{\delta}$ $\in [1 \dots \Sigma]$ // δ-encoded previous node
  next_rank $\in \mathbb{N}$ // step rank on next node
end

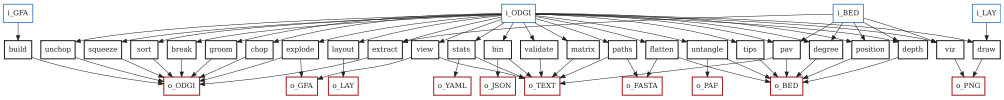

Fig. 1. Overview of the methods provided by ODGI (in black) and their supported input (in blue) and output (in red) data formats (A color version of this figure appears in the online version of this article.)

locks in ascending *Node.id* order and release them in descending order. In addition to node-local features of the graph, we must maintain some global information. Specifically, we record the start and end of paths, as well as a name to path id mapping in lock-free hash tables. The use of lock-free hash tables lets us avoid a global lock when looking up path or graph metadata, which would quickly become a bottleneck during parallel operations on the graph. By avoiding global locks, we implement many of the operations in ODGI using maximum parallelism available. This approach is key to enable our methods to scale to the largest pangenome graphs that we can currently build (with hundreds of vertebrate genomes).

4 Overview

ODGI provides a set of interrogative and manipulative operations on pangenome graphs. We have established these tools to support our exploration of graphs built from hundreds of large eukaryotic genomes. ODGI's tools are practical and able to work with high levels of graph complexity, even with regions where paths present very high depth nodes (10^5- to 10^6-fold depth). ODGI covers common operations that we have found to be essential when working with complex pangenome graphs:

- *odgi build* constructs the ODGI data model from GFA file (Section 4.1).
- *odgi view* converts the ODGI data model into GFA file (Section 4.1).
- *odgi viz* provides a linear visualization of the graph (Section 5.1).
- *odgi draw* renders a 2D image of the graph (Section 5.1).
- *odgi extract* excerpts subsets of the graph based on path ranges (Supplementary Section S.3).
- *odgi explode* breaks the graph into connected components (Supplementary Section S.3).
- *odgi squeeze* unifies disjoint graphs (Supplementary Section S.3).
- *odgi chop* breaks long nodes into shorter ones (Supplementary Section S.3).
- *odgi unchop* combines unitig nodes (Supplementary Section S.3).
- *odgi break* removes cycles in the graph (Supplementary Section S.3).
- *odgi prune* removes complex regions (Supplementary Section S.3).
- *odgi groom* resolves spurious inverting links (Supplementary Section S.3).
- *odgi position* lifts coordinates between path and graph positions (Section 5.2).
- *odgi untangle* deconvolutes paths relative to a reference (Section 5.2).
- *odgi tips* finds path end points relative to a reference (Supplementary Section S.2).
- *odgi sort* orders the graph nodes (Section 5.3).
- *odgi layout* establishes a 2D layout (Section 5.3).
- *odgi matrix* derives the pangenome matrix (Supplementary Section S.5).
- *odgi paths* lists and extracts paths in FASTA (Supplementary Section S.5).
- *odgi flatten* converts the graph to FASTA and BED (Supplementary Section S.5).
- *odgi pav* computes presence–absence variations (Supplementary Section S.5).
- *odgi stats* provides numerical properties of the graph (Section 5.4).
- *odgi bin* generates a summarized view of the graph (Supplementary Section S.5).
- *odgi depth* describes node depth over graph and path positions (Section 5.4).
- *odgi degree* describes node degree over graph and path positions (Section 5.4).

Each tool focuses on a small set of related operations. Most read or write the native ODGI format ('og' extension) (Fig. 1) and work with standard text-based data formats common to bioinformatics. This supports the implementation of flexible and composable graph processing pipelines based on graphs (GFA/ODGI) and standard bioinformatic data types representing positions, genomic ranges (BED) and pairwise mappings (PAF). We use variation graph paths to provide a universal coordinate system, representing annotations and pairwise sequence relationships using the paths as reference and query sequences. Thus, ODGI provides a set of interfaces that let us approach these graphs from the perspective of standard reference- and sequence-based data models. Indeed, by considering all paths in the graph as potential reference or query sequence, we make graphs invisible to downstream tools that operate on collections of sequences or rely on a reference sequence [e.g. SAMtools (Li *et al.*, 2009)], enabling interoperability. This approach benefits from the information in the graph without requiring that we build an entirely new set of bioinformatic methods to work in this difficult new pangenomic research context.

4.1 Building the ODGI model

ODGI maintains its own efficient binary format for storing graphs on disk. We begin by transforming the storage model of the standard GFAv1 (GFA Working Group, 2016) format (in which nodes, edges and paths are described independently) into the ODGI node-centric encoding with *odgi build*. This construction step can be a significant bottleneck, in particular as the size of the path set of the graph increases. The process itself is lossless. A graph in ODGI format represents everything that is in the input GFAv1 graph, without any loss of information. ODGI does not natively support GFAv2 or rGFA. GFAv2 is similar to GFAv1, but includes process-related annotations of assembly graphs not relevant for pangenome analyses. rGFA embeds a single coordinate hierarchy over the graph that links all sequences into a single base reference genome. This positional model depends on a particular graph induction algorithm Li *et al.* (2020). In contrast, ODGI implements coordinate translation dynamically (e.g. *odgi position* and *odgi untangle*), allowing use of any embedded genome as a reference. Its input graphs can represent any kind of alignment between the genomes. GFAv1 is fully capable of representing many reference genome coordinate systems simultaneously, which supports a reference-agnostic approach that uses the entire pangenome sequence space as a reference system. In doing so, our approach has the advantage of maintaining backward compatibility with existing tools based on genome sequences.

The ODGI data structure (Algorithm 1) allows algorithms that build and modify the graph to operate in parallel, without any global locks. In *odgi build*, we initially construct the node vector in a serial operation that scans across the input GFA file. Then, we serially add edges in the *Node.edges* vectors of pairs of nodes. Finally, we create paths in serial, and extend them in parallel by obtaining

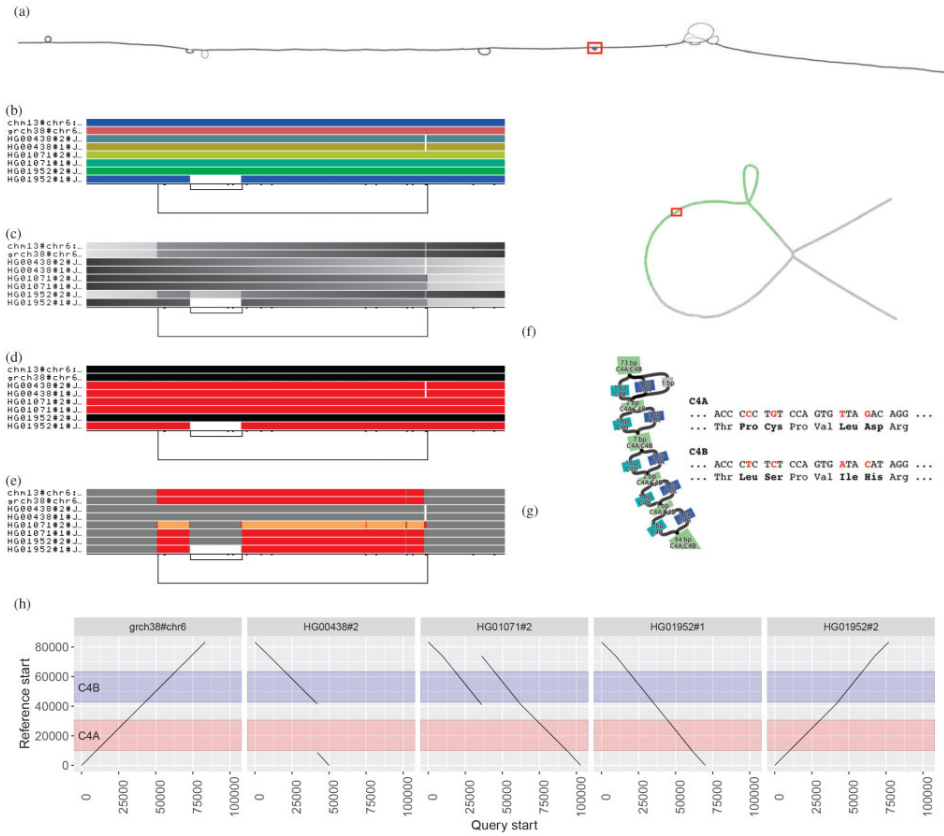

Fig. 2. Visualizing the major histocompatibility complex (MHC) and complement component 4 (C4) pangenome graphs. (**a**) *odgi draw* layout of the MHC pangenome graph extracted from a whole human pangenome graph of 90 haplotypes. The red rectangle highlights the C4 region. (**b–e**) *odgi viz* visualizations of the C4 pangenome graph, where eight paths are displayed: two reference genomes (CHM13 and GRCh38 on the top) and six haplotypes of three diploid individuals. (b) *odgi viz* default modality: the image shows a quite linear graph. The links at the bottom indicate the presence of a structural variant (long link) with another structural variant nested inside it (short link on the left). (c) Color by path position. The top two reference genomes and one haplotypes (HG01952#2) go from left to right, while five haplotypes go in the opposite direction, as indicated by the black color on their left. (d) *odgi viz* color by strandness: the red paths indicate the haplotypes that were assembled in reverse with respect to the two reference genomes. (e) *odgi viz* color by node depth: using the Spectra color palette with four levels of node depths, white indicates no depth, while gray, red and yellow indicate depth 1, 2 and greater than or equal to 3, respectively. Coloring by node depth, we can see that the two references present two different allele copies of the C4 genes, both of them including the HERV sequence. The entirely gray paths have one copy of these genes. HG01071#2 presents three copies of the *locus* (orange), of which one contains the HERV sequence (gray in the middle of the orange). In HG01952#1, the HERV sequence is absent. (**f**) Layout of the C4 pangenome graph made with the *Bandage* tool (Wick *et al.*, 2015) and annotated by using *odgi position*. Green nodes indicate the C4 genes (in red). The red rectangle highlights the regions where *C4A* and *C4B* genes differ. (**g**) Annotated *Bandage* layout of the C4 region where *C4A* and *C4B* genes differ due to single nucleotide variants leading to changes in the encoded protein sequences. Node labels were annoted by using *odgi position*. (**h**) Visualization of *odgi untangle* output in the C4 pangenome graph: the plots show the copy number status of the sequences in the C4 region with respect to the GRCh38 reference sequence, making clear, for example, that in HG00438#2, the *C4A* gene is missing (no black lines in the region annotated in red) (A color version of this figure appears in the online version of this article.)

the mutex *Node.lock* for pairs of nodes and by adding the path step in their *Node.path_steps* vectors. This parallelism speeds ODGI model construction by many-fold when testing against graphs made from assemblies produced by the HPRC (Section 5.5).

To support interchange with other pangenome tools or text-based processing, *odgi view* converts a graph in ODGI binary format to GFAv1. ODGI utilizes the PanSN (Garrison, 2021) specification to embed sample and haplotype information in the sequence name. This harmonizes the biosample information present in FASTA, GFA, PAF, VCF, BED, BEDPE, SAM/BAM and GFF/GTF formats related to the graph and its embedded genome sequences. By embedding all sequences into a single hierarchical namespace related to fundamental biological groupings in the input (e.g. biosample, individual, pooled group), PanSN allows us to utilize all assemblies in the pangenome as a combined reference coordinate model.

5 Results

Here, we apply our methods to a series of analyses, highlighting how ODGI can assist in exploring the biological features of pangenome graphs. We follow typical analyses that we have found critical to interpreting whole genome alignments represented in the variation graph model.

To simplify our exposition, we will extract small graph regions that are easy to interpret and describe. We focus on a handful of difficult loci from the human pangenome, extracting them from a prototype human pangenome graph built with the Pangenome Graph Builder pipeline (Garrison *et al.*, 2021). Pangenome graphs built from hundreds of haplotype-resolved *de novo* genome assemblies are very large, but it is often only necessary to work with only a small portion of the genomes represented, such as a specific locus (Fig. 2a) or a smaller region (Fig. 2b–g), or even a single gene (Fig. 3). This simplifies the downstream analyses and reduces the resources to work only with the extracted graphs. More on graph extraction and edit operations can be found at Supplementary Section S3.

5.1 Visualizing pangenome graphs

Visualization methods help us quickly gain insight into otherwise opaque biological data. We find visualization essential for understanding pangenome graphs. We pursue a novel approach to visualization with *odgi draw* and *odgi viz*, two tools that provide scalable

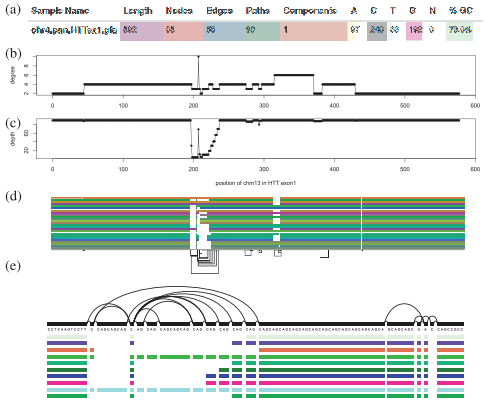

Fig. 3. Features of a 90-haplotype human pangenome graph of the exon 1 huntingtin gene (*HTTexon1*): (**a**) excerpt of vital statistics of the *HTTexon1* graph displayed by MultiQC's ODGI module. (**b**) Per nucleotide node degree distribution of CHM13 in the *HTTexon1* graph. Around position 200 there is a huge variation in node degree. (**c**) Per nucleotide node depth distribution of CHM13 in the *HTTexon1* graph. The alternating depth around position 200 indicates polymorphic variation complementing the above node degree analysis. (**d**) *odgi viz* visualization of the 23 largest gene alleles, CHM13 and GRCh38 of the *HTTexon1* graph. (**e**) *vg viz* nucleotide-level visualization of 10 gene alleles, CHM13, GRCH38 of the *HTTexon1* graph focusing on the CAG variable repeat region

ways of generating raster images showing the high-level structure of even large pangenome graphs (Fig. 2).

Using *odgi extract*, we extracted the major histocompatibility complex (MHC) locus from a 90-haplotype human chromosome 6 pangenome graph from the HPRC. Specifically, the graph contains the human references GRCh38, CHM13 and the contigs of 44 diploid individuals that encode all possible variations including those in telomeres and centromeres. The MHC genes are involved in antigen presentation, inflammation regulation, the complement system and the innate and adaptive immune responses (Shiina *et al.*, 2009). MHC genes are highly polymorphic, i.e. there are multiple different alleles across individuals in a population. Such variability becomes evident when we apply *odgi draw* to visualize the graph layout of a human MHC pangenome graph (Fig. 2a) (of note, *odgi layout* first generates the drawn projection, see Section 5.3). The visualization displays the graph topology in two dimensions (2D), with structural variation that appears as bubbles in the layout. A 2D rendering can be costly to compute, but we provide an implementation that scales linearly with pangenome sequence size, allowing us to apply it to large pangenome graphs.

The MHC locus includes the complement component 4 (C4) region, which encodes proteins involved in the complement system. In Figure 2a, C4 corresponds to the small bubble highlighted by the red rectangle. As an example use case, we took a closer look at the C4 region of the MHC by extracting it from the full MHC pangenome graph with *odgi extract*. Then, we visualized this subgraph by applying *odgi viz*, which produces binned, linearized renderings in 1 dimension (1D), where the graph is ordered in 1D across the horizontal axis, with each path represented by a row of the vertical axis (Fig. 2b–e). For each path, graph nodes are arranged from left to right, with the colored bars indicating the paths and the nodes they cross. White spaces indicate where paths do not traverse the nodes. Directly consecutive nodes are displayed with no white space between the two. The meaning of the colors depends on how *odgi viz* is executed. By default, path colors are derived from the path names (Fig. 2b), which are displayed on the left of the paths. The black lines on the bottom indicate the edges connecting the nodes and, therefore, represent the graph topology (see Supplementary Section S1 for a more detailed explanation). This visualization is computed in linear-time and offers a human-interpretable format suitable for understanding the topology and genome relationships in the pangenome graph. In humans, the C4 gene exists as two functionally distinct genes, *C4A* and *C4B*, which both vary in structure and copy number (Sekar *et al.*, 2016). In combination with the observed changes in path self-coverage, which represents copy number of a given path relative to the graph (Fig. 2e), the longer link at the bottom of Figure 2b–e indicates that the copy number status of these genes varies across the haplotypes represented in the pangenome. Moreover, the short nested variation on the left of the locus highlights that *C4A* and *C4B* genes segregate in both long and short genomic forms, distinguished by the presence or absence of a human endogenous retroviral (HERV) sequence.

Nevertheless, complex, non-linear graph structures are difficult to interpret in a low number of dimensions. To overcome this limitation, *odgi viz* supports multiple visualization modalities (Fig. 2c–e), making it easy to grasp the properties and shape of the graph. For example, we can color the paths by path position (Fig. 2c), with light gray indicating where paths begin and dark gray where they end. This visualization is suitable for understanding graph node order, as smooth color gradients indicate that the node order respects the linear paths' coordinate systems. Pangenome graphs can represent both strands of the genomic sequences of the DNA. We can display such information by coloring the paths by orientation, with paths colored where their sequence is reverse-complemented (red) or in direct orientation (black) with respect to the sequences of the graph nodes (Fig. 2d). Furthermore, we can use multiple color palettes to color the paths by how many times they traverse a node, which can be referred to as the path's depth or coverage of the node, the node depth. This highlights that in the C4 pangenome graph, the haplotypes present different number of copies of the C4 genes (Fig. 2e).

5.2 Untangling and navigating the pangenome

The key data in a pangenome graph is a representation of the alignment (i.e. the homology relationships) between genomic sequences. Navigating and understanding the graph requires coordinate systems to link other data to the sequences represented in the graph model. ODGI's tools use the embedded sequences to provide a universal coordinate space that is graph-independent, thereby remaining stable across different graphs built with the same sequences. Such a universal coordinate system allows us to support several kinds of 'lift-over' of coordinates between different sequences in the same or different graphs. As a demonstration, we took the C4 pangenome graph and added to its nodes gene annotation from GRCh38 (in GFF format file) using *odgi position* (Supplementary Section S2.1). The resulting TSV contains pairs of nodes and colors. Taking the graph and the TSV into Bandage (Wick *et al.*, 2015), the actual C4 genes are highlighted (Fig. 2f). Zooming to the nucleotide level, the annotation shows the single nucleotide differences of the *C4A* and the *C4B* genes (Fig. 2g).

odgi position can also translate graph and path positions between or within graphs, emitting the liftovers in BED format. For a precise translation process when conversing a query position to a reference position in a repeat region, we apply the *path jaccard* context mapping concept. It could be that the found reference node is visited several times by the reference. To ensure a precise translation, we select the reference position whose context (the multiset of *Node.id*s reached within a distance of e.g. 10 kbp) has the best jaccard metric when compared to the query context. For a more detailed explanation of the *path jaccard* concept see Supplementary Section S2.2.

To obtain a more precise overview of the locus in Figure 2b–e, we applied *odgi untangle* with GRCh38 as a reference. *odgi untangle* segments paths into linear segments by breaking these segments where the paths loop back on themselves. In this way, we obtain information on the position and copy number status of the sequences in the collapsed locus, in BEDPE or PAF format. In the representation in Figure 2h, the orientation of the line indicates if the copy number is in forward or in reverse orientation compared to GRCh38. *odgi untangle* is able to work with any sets of reference sequences, converting the graph to lift-over maps compatible with standard software for projecting annotations and alignments from one genome to another. An explanation of the untangling process is given in Supplementary Section S2.

5.3 Latent graph structure reveals underlying biology

Pangenome graphs can hide their underlying latent structures, introducing difficulties in the analysis and interpretation. Among the causes of this is the correct ordering of the graph nodes in a convenient number of dimensions. ODGI provides a variety of sorting algorithms to find the best graph node order in 1 or 2 dimensions, allowing us to understand the sparse structures typically found in pangenome graphs and the genetic variation they represent. *odgi sort* allows the chaining of these sorting algorithms. As many of the algorithms are affected by the initial node order, this allows us to generate sorting pipelines that progressively refine the graph ordering.

We applied several of *odgi sort*'s 1D algorithms to a 90-haplotype human MHC pangenome and a C4 subgraph (Supplementary Fig. S2). The randomly sorted MHC graph (Supplementary Fig. S2a) hides its linear graph structure, whereas our novel path-guided (PG) stochastic gradient descent (SGD) algorithm, PG-SGD, is able to produce a globally linear ordered graph revealing the C4 region (Supplementary Fig. S2b). This exploits path information to order the graph nodes. PG-SGD learns a 1D or 2D organization of the graph nodes that matches nucleotide distances in graph paths (i.e. the sequences embedded in the graph). To scale to large graphs, we learn this projection in parallel via a HOGWILD! approach (Niu *et al.*, 2011). PG-SGD can be seen as an adaptation of SGD-based drawing (Zheng *et al.*, 2019) to pangenome graphs. In parallel, each HOGWILD! thread updates the relative position of pairs of nodes so that their distance in the layout, or their order, better-matches their nucleotide distance in the paths running through the graph. Following standard SGD approaches, the learning rate is reduced as the algorithm progresses, and execution continues until the adjustments to the model fall below a target threshold ϵ.

A PG-SGD sorting of C4 compresses both sides of the variant bubble into one dimension, leading to an interrupted pattern of nodes across the copy-number variable region (Supplementary Fig. S2c). Subsequently applying a topological sort clarifies the graph's latent structure, simplifying interpretation (Supplementary Fig. S2d). To find the best order of graph nodes in 1D, *odgi sort*'s multiple sorting algorithms can be combined into a sorting pipeline to take advantage of the strength of each (results not shown). ODGI can project vector (in 1D) and matrix (2D) representations of the graph relative to these learned coordinate spaces. Based on this projection, we can trivially sort graph nodes in 1D. Moreover, we support the same concept in 2D in *odgi layout* by providing a 2D implementation of the PG-SGD algorithm (Fig. 2a). A detailed description of the node ordering process can be found at Supplementary Section S4. As we have shown above, the node order is crucial to understand the biological features of a pangenome graph.

5.4 Graph features highlight variation

Graphs statistics provide alternative ways to gain insight into pangenomes complexity revealing the overall structure, size and features of a graph and its sequences.

As a use case study (Fig. 3), we took a look at the metrics of a 90-haplotype human pangenome graph of the exon 1 huntingtin gene (*HTTexon1*). In particular, we obtained the number of nodes, edges, paths, components, bases, the graph length and the GC content with *odgi stats*. The output pangenome statistics in YAML textual file format was given to MultiQC's (Ewels *et al.*, 2016) newly added ODGI module. As can be seen in Figure 3a, we observe a very high GC content of 73.0% in the *HTTexon1* graph compared to the human genomic mean GC content of 40.9% (Piovesan *et al.*, 2019). This is in accordance with the literature (Neueder *et al.*, 2017). Despite this discovery, the MultiQC module provides an interactive way to comparatively explore statistics of an arbitrary number of graphs.

To investigate in detail which intricate regions in the *HTTexon1* graph are responsible for its genetic variation and high GC content, we took a look at the per nucleotide node degree (Fig. 3b) and node depth (Fig. 3c) distributions of CHM13 by using *odgi depth*'s and *odgi degree*'s BED output, respectively. The results indicate a highly

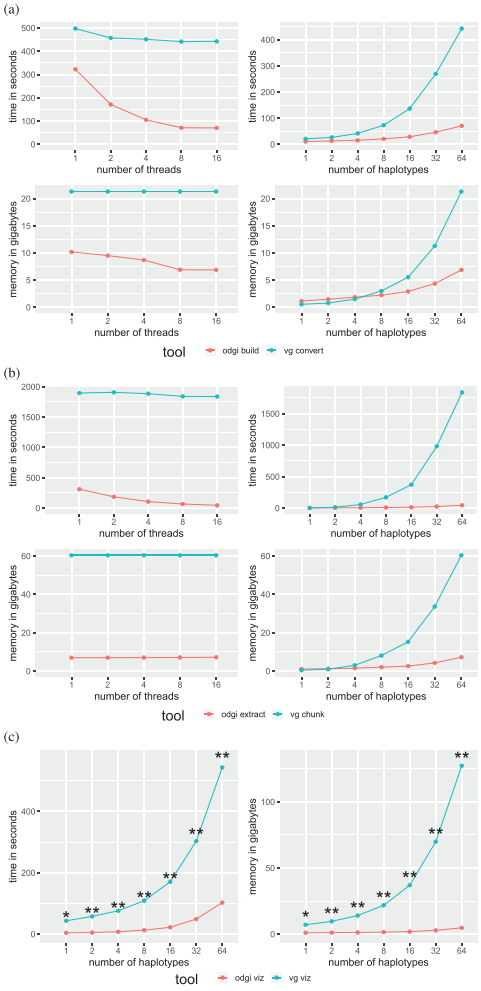

Fig. 4. Performance on a graph of human chromosome 6 from the HPRC. ODGI compares favorably to VG across all routine pangenomic tasks. Evaluations across threads were done using a 64 human haplotype graph. Evaluations across haplotypes were done using 16 threads. (**a**) Performance evaluation when translating a graph into the tools' respective native formats. (**b**) Performance evaluation when extracting the centromeric region from the HPRC graph. (**c**) Performance evaluation when visualizing a graph. Both tools were run with only one thread. *vg viz*: *A 816 MB SVG was produced which cannot be opened by any program. **All produced SVGs only contain an XML header, nothing else

polymorphic region around position 200 in the graph. Figure 3d supports this analysis. Zooming in on this region with *vg viz*, we can clearly identify the typical *HTTexon1* CAG variable repeat region (Fig. 3e). Figure 3b–d highlights the variant region around position 200 of CHM13, showing the variable number of glutamine residues of the different individuals as reported by Nance *et al.* (1999).

5.5 Performance evaluation

Although many of the operations that ODGI provides are unique, some are common with the existing VG toolkit. We compare with these to highlight the practical performance implications of our graph data structure design. Our results highlight the efficient parallel algorithm implementations enabled by this design.

We compared the efficiency of ODGI (v0.6.3-56-gebc493f 'Pulizia') and VG (v1.37.0 'Monchio') for routine pangenome tasks. In particular, we measured the execution time and memory usage (i) of transforming a GFAv1 file into a tool's native format, (ii) the extraction of a subgraph, (iii) the visualization of a pangenome graph and (iv) the finding of path positions in a pangenome graph. These graph operations are key when it comes to the understanding of

pangenome graphs. They are also a set of functions implemented in both toolkits. We ran these operations for a varying number of threads and haplotypes in the graph for a scaling analysis. We ran each evaluation configuration 10 times and report the mean of each run. All evaluations were performed on a VM in the German Network for Bioinformatics Infrastructure (deNBI) cloud with 28 cores and 256 GB of RAM. The presented results are from a 90-haplotype chromosome 6 human pangenome graph built with data from the HPRC. Specifically, the graph contains the human references GRCh38, CHM13 and the contigs of 44 diploid individuals that encode all possible variations including those in telomeres and centromeres. When transforming a GFAv1 file with VG, the static XG file format was used. The tools involved in the evaluation process require the XG format.

In general, ODGI makes comparatively better use of multithreading and requires much less memory (Fig. 4, Supplementary Table S4) across all operations. ODGI scales much better than VG when working with complex regions of the graph. For example, extracting a difficult centromeric subgraph (Fig. 4b), ODGI is up to 40 times faster and requires 8 times less memory than VG.

Both visualization tools can only make use of a single thread. For a 1 haplotype, graph *vg viz* produces a 816MB SVG which can't be opened by the standard programs to date. For larger graphs, *vg viz* runs through and produces SVGs with only the XML header. This makes it unusable for large graphs.

We also measured the disk space usage of GFAv1, ODGI's and VG's binary formats (Supplementary Table S5). While VG's XG occupies less disk space for smaller graphs, ODGI requires less space for graphs having 32 haplotypes or more. We hypothesize that this indicates the lower marginal cost for additional haplotypes when using ODGI's id delta encoding scheme.

6 Discussion

Pangenome graphs stand to become a ubiquitous model in genomics thanks to their capability to represent any genetic variant without being affected by reference bias (Eizenga *et al.*, 2020b). However, despite this great potential, their spread is impeded by the lack of tools capable of managing and analyzing pangenome graphs easily and efficiently.

By providing a set of standard analysis 'verbs' to interact with pangenome graphs, ODGI enables users to explore and discover important biological features captured in this flexible, inclusive model. It provides tools to easily transform, analyze, simplify, validate and visualize pangenome graphs at large scale. In particular, lifting over annotations and linearizing nested graph structures place the suite as the bridge between traditional linear reference genome analysis and pangenome graphs. With the increased adoption of long read sequencing we expect pangenomic tools to become increasingly common in the genomic studies at different taxonomic levels and in biomedical research. This progression is already afoot, particularly for targets that involve complex variation, such as cancer (The Computational Pan-Genomics Consortium, 2016), plant pangenomics (Bayer *et al.*, 2020, 2022; Li *et al.*, 2022; Liu *et al.*, 2020; Qin *et al.*, 2021) and metagenomics (Zhong *et al.*, 2021). Also, when studying animals like bovines (Bovine Pan-Genome Consortium, 2022; Leonard *et al.*, 2021; Talenti *et al.*, 2022).

Currently, bacterial pangenomes are best handled by specialized tools like PPanGGolin (Gautreau *et al.*, 2020), PanGraph (Noll *et al.*, 2022) or PanX (Ding *et al.*, 2018). The latter one doesn't build a graphical representation of a pangenome. But, it already has a very developed eco-system, which allows a detailed analysis of bacterial pangenomes using an interactive GUI. Unlike these approaches, which provide a monolithic, integrated solution to understanding pangenomes, ODGI is designed as a low-level toolkit that can work on a generic pangenome graph model frequently used by other existing methods. We hope that this design renders it useful to pangenome analysis pipeline authors. Other pangenome analysis platforms, like PanTools (Sheikhizadeh *et al.*, 2016) provide access to pangenome analyses at the scales we demonstrate with ODGI, but use specialized de Bruijn graph models to achieve this. In contrast ODGI supports the highly generic variation graph model, which has greater representational power than de Bruijn graphs.

ODGI will facilitate disentangling, describing and analyzing a much larger set of variation than previously was possible with tools that depend on short reads and reference genomes. Furthermore, users can even consider ODGI as a framework, taking advantage of its algorithms to develop new and more advanced tools that work on pangenome graphs, thus expanding the type of possible pangenomic analyses available to the scientific community.

The performance analysis shows that ODGI outperforms VG when handling large, complex pangenome graphs. Across the evaluation of key graph operations, ODGI's memory peak was 10GB. This makes it perfectly suited to be run interactively on a recent laptop. We expect that ODGI will be able to handle the next phase of the HPRC, a pangenome graph constructed from 300 individuals, without any problems.

While ODGI does not construct graphs from scratch nor is capable of extending them, it is already the backbone of the Pangenome Graph Builder pipeline (Garrison *et al.*, 2021). Its static, large-scale 1D and 2D visualizations of the pangenome graphs allow an unprecedented high-level perspective on variation in pangenomes, and have also been critical in the development of pangenome graph building methods. However, an interactive solution that combines the 1D and 2D layout of a graph with annotation and read mapping information across different zoom levels is still missing. Recent interactive pangenome graph browsers are reference-centric (Beyer *et al.*, 2019; Yokoyama *et al.*, 2019), have a limited predefined coordinate system (Durant *et al.*, 2021), or focus primarily on 2D representations (Gonnella *et al.*, 2019; Wick *et al.*, 2015). Our graph sorting and layout algorithms can provide the foundation for future tools of this type. We plan to focus on using these learned models to detect structural variation and assembly errors.

ODGI has allowed us to explore *context mapping* deconvolution of pangenome graph structures via the path jaccard metric. This resolves a major conceptual issue that has strongly guided existing algorithms to construct pangenome graphs. Previously, great efforts have been made to prevent the 'collapse' of non-orthologous sequences in the graph topology itself (Li *et al.*, 2020). This has been seen as essential to making these new bioinformatic models interpretable. While our presentation is primarily qualitative, our work demonstrates that we can mitigate this issue by exploiting the pangenome graph not as a static reference, but as a dynamic model of the mutual alignment of many genomic sequences. Because pangenome graphs can contain complete genomes, we are able to query them to polarize the information they contain in easily interpretable and reusable pairwise formats that are widely supported in bioinformatics. ODGI also projects variation graphs into vector and matrix representations that allow the direct application of machine learning and statistical models to the pangenome. We expect that ODGI will provide a reference interface between pangenomic and genomic approaches for understanding genome variation.

Acknowledgements

The authors thank members of the HPRC Pangenome Working Group for their insightful discussion and feedback, and members of the HPRC production teams for their development of resources used in our exposition.

Funding

The authors gratefully acknowledge support from National Institutes of Health/NIDA U01DA047638 (E.G.), National Institutes of Health/NIGMS R01GM123489 (E.G. and P.P.) and NSF PPoSS Award #2118709 (E.G. and P.P.). S.H. acknowledges funding from the Central Innovation Programme (ZIM) for SMEs of the Federal Ministry for Economic Affairs and Energy of Germany. S.N. acknowledges Germany's Excellence Strategy (CMFI), EXC-2124 and (iFIT)—EXC 2180–390900677. This work was supported by the BMBF-funded de. NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) [031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A and 031A532B].

Conflict of Interest: The authors have nothing to declare.

Data availability

Code and links to data resources used to build this manuscript and its figures, can be found in the paper's public repository: https://github.com/pangenome/odgi-paper.

References

Armstrong,J. *et al.* (2020) Progressive cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, **587**, 246–251.

Baaijens,J.A. *et al.* (2019) Full-length de novo viral quasispecies assembly through variation graph construction. *Bioinformatics*, **35**, 5086–5094.

Ballouz,S. *et al.* (2019) Is it time to change the reference genome? *Genome Biol.*, **20**, 159.

Bayer,P.E. *et al.* (2020) Plant pan-genomes are the new reference. *Nat. Plants*, **6**, 914–920.

Bayer,P.E. *et al.* (2022) Wheat panache – a pangenome graph database representing presence/absence variation across 16 bread wheat genomes. bioRxiv. https://doi.org/10.1101/2022.02.23.481560.

Beyer,W. *et al.* (2019) Sequence tube maps: making graph genomes intuitive to commuters. *Bioinformatics*, **35**, 5318–5320.

Bovine Pan-Genome Consortium. (2022) Bovine pan-genome consortium. https://njdbickhart.github.io/ (February 2022, date last accessed).

Computational Pan-Genomics Consortium. (2018) Computational pan-genomics: status, promises and challenges. *Brief. Bioinf.*, **19**, 118–135.

Ding,W. *et al.* (2018) panX: pan-genome analysis and exploration. *Nucleic Acids Res.*, **46**, e5.

Durant,E. *et al.* (2021) Panache: a web browser-based viewer for linearized pangenomes. *Bioinformatics*, **37**, 4556–4558.

Eizenga,J.M. *et al.* (2020a) Efficient dynamic variation graphs. *Bioinformatics*, **36**, 5139–5144.

Eizenga,J.M. *et al.* (2020b) Pangenome graphs. *Annu. Rev. Genomics Hum. Genet.*, **21**, 139–162.

Ewels,P. *et al.* (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048.

Garrison,E. (2019) Graphical Pangenomics (Doctoral thesis). https://doi.org/10.17863/CAM.41621.

Garrison,E. (2021) Pansn-spec*:* Pangenome Sequence Naming. https://github.com/pangenome/PanSN-spec (May 2022, date last accessed).

Garrison,E. *et al.* (2018) Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.*, **36**, 875–879.

Garrison,E. *et al.* (2021) The Pangenome Graph Builder. https://github.com/pangenome/pggb (May 2022, date last accessed).

Gautreau,G. *et al.* (2020) PPanGGOLiN: depicting microbial diversity via a partitioned pangenome graph. *PLoS Comput. Biol.*, **16**, e1007732.

GFA Working Group. (2016) Graphical Fragment Assembly *(*GFA*)* Format Specification. https://github.com/GFA-spec/GFA-spec (May 2022, date last accessed).

Gonnella,G. *et al.* (2019) GfaViz: flexible and interactive visualization of GFA sequence graphs. *Bioinformatics*, **35**, 2853–2855.

Grasso,C. and Lee,C. (2004) Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. *Bioinformatics*, **20**, 1546–1556.

Hein,J. (1989) A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given. *Mol. Biol. Evol.*, **6**, 649–68.

Hickey,G. *et al.* (2020) Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.*, **21**, 35.

Jarvis,E.D. *et al.* (2022) Automated assembly of high-quality diploid human reference genomes. bioRxiv. https://doi.org/10.1101/2022.03.06.483034.

Kehr,B. *et al.* (2014) Genome alignment with graph data structures: a comparison. *BMC Bioinformatics.*, **15**, 99.

Lee,C. *et al.* (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics*, **18**, 452–464.

Leonard,A.S. *et al.* (2021) Bovine pangenome reveals trait-associated structural variation from diverse assembly inputs. bioRxiv. https://doi.org/10.1101/2021.11.02.4466900.

Li,H. *et al.*; 1000 Genome Project Data Processing Subgroup. (2009) The sequence alignment/map format and samtools. *Bioinformatics*, **25**, 2078–2079.

Li,H. *et al.* (2020) The design and construction of reference pangenome graphs with minigraph. *Genome Biol.*, **21**, 265.

Li,H. *et al.* (2022) Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. *Nat. Commun.*, **13**, 682.

Liu,Y. *et al.* (2020) Pan-genome of wild and cultivated soybeans. *Cell*, **182**, 162–176.e13.

Logsdon,G.A. *et al.* (2021) The structure, function and evolution of a complete human chromosome 8. *Nature*, **593**, 101–107.

Miga,K.H. *et al.* (2020) Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, **585**, 79–84.

Nance,M.A. *et al.* (1999) Analysis of a very large trinucleotide repeat in a patient with juvenile Huntington's disease. *Neurology*, **52**, 392–394.

Neueder,A. *et al.* (2017) The pathogenic exon 1 HTT protein is produced by incomplete splicing in Huntington's disease patients. *Sci. Rep.*, **7**, 1307.

Niu,F. *et al.* (2011) Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. *NIPS'11: Proceedings of the 24th International Conference on Neural Information Processing Systems,* 693–701.

Noll,N. *et al.* (2022). Pangraph: scalable bacterial pan-genome graph construction. bioRxiv.

Nurk,S. *et al.* (2022) The complete sequence of a human genome. *Science*, **376**, 44–53.

Paten,B. *et al.* (2017) Genome graphs and the evolution of genome inference. *Genome Res.*, **27**, 665–676.

Piovesan,A. *et al.* (2019) On the length, weight and GC content of the human genome. *BMC Res. Notes*, **12**, 106.

Prezza,N. (2017) A framework of dynamic data structures for string processing. *Leibniz Internatnal Proceedings in Informatics*, 75.

Qin,P. *et al.* (2021) Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell*, **184**, 3542–3558.

Quinlan,A.R. and Hall,I.M. (2010) Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Sekar,A. *et al.*; Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2016) Schizophrenia risk from complex variation of complement component 4. *Nature*, **530**, 177–183.

Sheikhizadeh,S. *et al.* (2016) PanTools: representation, storage and exploration of pan-genomic data. *Bioinformatics*, **32**, 487–493.

Shiina,T. *et al.* (2009) The HLA genomic loci map: expression, interaction, diversity and disease. *J. Hum. Genet.*, **54**, 15–39.

Sibbesen,J.A. *et al.* (2021) Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. bioRxiv. https://doi.org/10.1101/2021.03.26.437240.

Siren,J. *et al.* (2020) Haplotype-aware graph indexes. *Bioinformatics*, **36**, 400–407.

Talenti,A. *et al.* (2022) A cattle graph genome incorporating global breed diversity. *Nat. Commun.*, **13**, 910.

Tettelin,H. *et al.* (2008) Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.*, **11**, 472–477.

The Computational Pan-Genomics Consortium. (2016) Computational pan-genomics: status, promises and challenges. *Brief. Bioinformatics*, **19**, bbw089.

Wick,R.R. *et al.* (2015) Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, **31**, 3350–3352.

Yokoyama,T.T. *et al.* (2019) MoMI-G: modular multi-scale integrated genome graph browser. *BMC Bioinformatics*, **20**, 548.

Zheng,J.X. *et al.* (2019) Graph drawing by stochastic gradient descent. *IEEE Trans. Vis. Comput. Graph.*, **25**, 2738–2748.

Zhong,C. *et al.* (2021) Integrating pan-genome with metagenome for microbial community profiling. *Comput. Struct. Biotechnol. J.*, **19**, 1458–1466.