GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Zuwendungsempfänger

Heinrich-Heine-Universität Düsseldorf

Thema der Förderung

Genomgraph-basierte Werkzeuge und Workflows für die Genomik und personalisierte Medizin

Projektleitung

Prof. Dr. Tobias Marschall

Förderkennzeichen

031L0184A

Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Schlussbericht „GraphGenomes“

Teilprojekt A, Laufzeit: 01. April 2019 – 31. Januar 2023
Förderkennzeichen: 031L0184A
Vorhabenbezeichnung: Genomgraph-basierte Werkzeuge und Workflows für die Genomik und personalisierte Medizin

Teil I - Kurzbericht

Stand der Technik und Aufgabenstellung

Die Genomsequenzierung (oder Exomsequenzierung) verändert derzeit die medizinische Forschung: Sie trägt dazu bei, die genetischen Grundlagen von Krankheiten zu erforschen und ist der Schlüssel zur Präzisionsmedizin. Die Sequenzierung des Genoms einer Patient/in ermöglicht die Bestimmung der individuellen genetischen Voraussetzungen und kann somatische genetische Varianten in einzelnen Zellen oder Geweben aufdecken. Diese Informationen können genutzt werden, um Genotyp und Phänotyp bei Krankheiten – von seltenen Mendelschen Krankheiten bis hin zu Autoimmunkrankheiten, Herz-Kreislauf-Erkrankungen und Krebs – in Verbindung zu bringen. Zu den Anwendungen gehören die Untersuchung der molekularen Grundlagen von Krankheiten, die Ableitung individueller Krankheitsrisiken, die Früherkennung von Krankheiten und die Auswahl von Therapien. Obwohl die Genomsequenzierung zu einem zentralen Ansatz in der Krankheitsforschung geworden ist, weisen die am häufigsten verwendeten Ansätze für das Read-Alignment und die Erkennung genetischer Varianten noch einige wesentliche Einschränkungen auf: Sie beruhen in der Regel auf dem Vergleich von DNA-Fragmenten mit einem Referenzgenom, ein Schritt, der unweigerlich fehlschlägt, wenn sich das Genom einer Proband/in lokal erheblich von dieser Referenz unterscheidet. Dieses Problem beeinträchtigt nicht nur die Analyse menschlicher genetischer Varianten, sondern erstreckt sich auch auf die Charakterisierung anderer "omics"-Daten, einschließlich Transkriptom- und Epigenomdaten und sogar Einzelzellsequenzierungsdaten, die alle in ähnlicher Weise auf die Zuordnung von Reads zu einer Position im Referenzgenom Datenverarbeitung angewiesen sind.

Zum Projektstart gab es bereits einige erste Ansätze (sowohl von den Projektpartnern als auch von anderen Arbeitsgruppen) zum Alignment von Reads und der anschließenden Bestimmung von genetischen Varianten durch den Vergleich einzelner Fragmente mit einem Genomgraphen, der auch als Pangenom bezeichnet wird, anstatt mit einem einzelnen Referenzgenom. Das übergeordnete Ziel unseres Projekts bestand darin, diese Ansätze weiterzuentwickeln und einer möglichst breiten Nutzerbasis zur Verfügung zu stellen um die Grundlage für die Anwendung von Pangenomen in der biomedizinischen Forschung und der personalisierten Medizin zu schaffen.

Teilprojekt A umfasste vier wesentliche Entwicklungsziele/Meilensteine:

- M2.1: Weiterentwicklung unseres Tools GraphAligner zum Alignieren von langen Reads an Genomgraphen.

- M2.2: Entwicklung von Werkzeugen um mit Hilfe von langen Reads Varianten zu genotypisieren, die in Genomgraphen kodiert sind.
- M2.3: Erweiterung des Genomgraphen mit Hilfe von langen Reads.
- M2.4: Haplotyperekonstruktion auf Graphenebene mit Hilfe von langen Reads.

Ablauf und wesentliche Ergebnisse

Die Arbeiten zu M2.1 waren bereits nach dem ersten Jahr weitestgehend abgeschlossen. Diese Arbeiten sind in einem Artikel dokumentiert (Rautiainen and Marschall, Genome Biology, 21:253, 2020). Das Software-Tool GraphAligner erlaubt es jetzt, lange Reads an Genomgraphen zu alignieren und erreicht dabei Geschwindigkeiten, die gängige Tools zum Sequenzalignment (Sequenz an Sequenz) übersteigt. Inzwischen werden weitere Ein- und Ausgabeformate unterstützt (VG und GFA zur Eingabe, JSON, GAM und GAF zur Ausgabe), die eine Interoperabilität ermöglichen, welche zur Konstruktion von komplexeren Workflows benötigt wird. Der Source Code von GraphAligner ist offen und steht auf GitHub unter der MIT-Lizenz zur Verfügung. Parallel dazu sind die Arbeiten zu M2.2 vorangeschritten. Eine vollständige Implementierung zum Prozessieren von Long Reads, die an Graphen aligniert sind, liegt vor. Das bedeutet, das nun GAF-Dateien, in denen solche Alignments gespeichert sind, eingelesen werden können und statt der sonst üblichen BAM-Dateien verwendet werden können um Genotypisierung (M2.2) und Haplotypisierung (M2.4) durchzuführen. Hierzu haben wir zusätzliche Verarbeitungsschritte implementiert um eine Ordnung auf Graphen (im GFA-Format) zu definieren und GAF-Dateien in Bezug auf diese Ordnung zu sortieren. Um die Ziele aus M2.3 zu erreichen, haben wir einen Workflow names SVarp entwickelt. Dieser ist in der Lage, Long Reads zu identifizieren, die Signaturen für strukturelle Varianten tragen, d.h. die Hinweise liefern, dass der Genomgraph an einer bestimmten Stelle erweitert werden muss. Diese werden dann zu Clustern zusammengefasst und mittels lokalem Genomassembly zusammengesetzt. Die resultierenden Sequenzen, die wir „SVtigs“ nennen, können dann als Eingabe für Algorithen zur Graphkonstruktion dienen. Somit steht ein kompletter Workflow für M2.3 nun zur Verfügung.

Zusammenarbeit mit anderen Forschungseinrichtungen

Wie oben dargestellt, haben wir im Zuge von M2.1 das Softwarewerkzeug GraphAligner (weiter)entwickelt. In Zusammenarbeit mit dem Telomere-to-Telomere Consortium (T2T) konnten wir GraphAligner anwenden um schwierige Regionen des menschlichen Genoms zu rekonstruieren, was zur ersten vollständigen Rekonstruktion eines Genoms beigetragen hat (Nurk et al., Science, 2022). Nach dem Start von GraphGenomes wurde vom NIH/NHGRI (USA) das Human Pangenome Reference Consortium (HPRC) gegründet. Dort werden in großem Maßstab Daten zur Konstruktion eines menschlichen Pangenoms produziert. Wir haben uns in das HPRC eingebracht und wenden dort Tools an, die wir im Rahmen von GraphGenomes entwickelt haben. Kürzlich ist das Manuskript des HPRC *„A draft human pangenome reference“* erschienen (Liao*, Asri*, Ebler*, et al., Nature, 2023), zu dem wir beigetragen haben. Dieser Meilenstein der Genomik hat weltweit ein sehr großes Medienecho hervorgerufen, mit Berichterstattung in vielen wichtigen Medien (z.B. Süddeutsche Zeitung, FAZ, CNN, Washington Post und viele andere, insgesamt 314 Medien laut https://nature.altmetric.com/details/148167380, abgerufen am 3.7.2023).

Schlussbericht „GraphGenomes“

Teilprojekt A, Laufzeit: 01. April 2019 – 31. Januar 2023
Förderkennzeichen: 031L0184A
Vorhabenbezeichnung: Genomgraph-basierte Werkzeuge und Workflows für die Genomik und personalisierte Medizin

Teil II: Eingehende Darstellung

1 Stand der Technik und Aufgabenstellung

Die Genomsequenzierung (oder Exomsequenzierung) verändert derzeit die medizinische Forschung: Sie trägt dazu bei, die genetischen Grundlagen von Krankheiten zu erforschen und ist der Schlüssel zur Präzisionsmedizin. Die Sequenzierung des Genoms einer Patient/in ermöglicht die Bestimmung der individuellen genetischen Voraussetzungen und kann somatische genetische Varianten in einzelnen Zellen oder Geweben aufdecken. Diese Informationen können genutzt werden, um Genotyp und Phänotyp bei Krankheiten – von seltenen Mendelschen Krankheiten bis hin zu Autoimmunkrankheiten, Herz-Kreislauf-Erkrankungen und Krebs – in Verbindung zu bringen. Zu den Anwendungen gehören die Untersuchung der molekularen Grundlagen von Krankheiten, die Ableitung individueller Krankheitsrisiken, die Früherkennung von Krankheiten und die Auswahl von Therapien. Obwohl die Genomsequenzierung zu einem zentralen Ansatz in der Krankheitsforschung geworden ist, weisen die am häufigsten verwendeten Ansätze für das Read-Alignment und die Erkennung genetischer Varianten noch einige wesentliche Einschränkungen auf: Sie beruhen in der Regel auf dem Vergleich von DNA-Fragmenten mit einem Referenzgenom, ein Schritt, der unweigerlich fehlschlägt, wenn sich das Genom einer Proband/in lokal erheblich von dieser Referenz unterscheidet. Dieses Problem beeinträchtigt nicht nur die Analyse menschlicher genetischer Varianten, sondern erstreckt sich auch auf die Charakterisierung anderer "omics"-Daten, einschließlich Transkriptom- und Epigenomdaten und sogar Einzelzellsequenzierungsdaten, die alle in ähnlicher Weise auf die Zuordnung von Reads zu einer Position im Referenzgenom Datenverarbeitung angewiesen sind.

Bereits zum Projektstart hatten uns populationsweite Sequenzierungsprojekte wie das 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015; Sudmant et al., 2015) und das Genome of the Netherlands Project (Genome of the Netherlands Consortium, 2014; Hehir-Kwa et al., 2016) große Kataloge häufig vorkommender genetischer Varianten in europäischen und anderen Populationen auf der ganzen Welt beschert. Aber diese Fülle an Informationen wurde nicht beim Read-Alignment berücksichtigt, woraus sich Verzerrungen bei der Bestimmung von genetischen Varianten ergeben. Die stärksten Verzerrungen treten auf, wenn die Sequenz einer Probe erheblich von der Referenz abweicht. Beispiele hierfür sind der Major Histocompatibility Complex (MHC) und andere immunogenetische Regionen im Genom. Diese Segmente tragen eine Fülle genomischer Strukturvariationen (SV), die bei einer Vielzahl häufiger und seltener Krankheiten eine Rolle spielen (1000 Genomes Project Consortium et al.,

2015; Sudmant et al., 2015). Um diese Einschränkungen zu überwinden, werden Datenstrukturen benötigt, die in der Lage sind, das heutige Wissen über die menschliche genetische Variation darzustellen. Auf Graphen basierende Genomdarstellungen sind vielversprechend, um dies zu erreichen (Computational Pan-Genomics Consortium, 2016; Paten et al., 2017). Prototypen die solche Konzept implementieren, hatten gezeigt, dass sie Verzerrungen beim Alignment von DNA-Fragmenten beseitigen (Rakocevic et al., 2017). Andernfalls können beispielsweise Insertionsallele inkorrekt zugeordnet werden (Garrison et al., 2017). Die Vorteile für eine verbesserte Genotypisierung bekannter Varianten sowie für eine verbesserte Entdeckung neuer Varianten in polymorphen Bereichen des Genoms waren bereits gut dokumentiert: Die Fehlerraten bei der HLA-Typisierung aus Ganzgenomsequenzierungsdaten sinken von 10 % bei Standardansätzen auf etwa 1 % bei Verwendung von Genomgraphen (A. Dilthey et al., 2015; A. T. Dilthey et al., 2016). Bei SNPs kann die Rate der verpassten Calls mit GraphTyper um den Faktor fünf reduziert werden (von 3,6 % auf 0,7 %) (Eggertsson et al., 2017). Ein noch drastischerer Effekt wurde für SVs berichtet, bei denen mehr als 8.800 zusätzliche Varianten >50bp mit graphbasierten Referenzen wiederhergestellt werden können und etwa die Hälfte der Deletionen >50bp und die Mehrheit (>70 %) der Insertionen >50bp mit Standard-Workflows verpasst werden (Maretty et al., 2017).

Trotz dieser konzeptionellen Fortschritte, die zum Teil von den Antragstellern vorangetrieben wurden (Dr. Dilthey: (A. Dilthey et al., 2015; A. T. Dilthey et al., 2016) und Dr. Marschall: (Computational Pan-Genomics Consortium, 2016; Garg et al., to appear 2018; Ghaffaari & Marschall, 2019; Rautiainen & Marschall, 2017)), wurden Genomgraphen in der Genomforschung noch nicht routinemäßig angewandt. Bislang fehlten interoperable Frameworks sowie gebrauchsfertige menschlichen Genomegraphen, die öffentlich zur Verfügung stehen. Darüber hinaus gab es insbesondere offene Herausforderungen bei der Anwendung von Genomgraphen auf lange Reads, die von Sequenzierinstrumenten der dritten Generation (z.B. Pacific Biosciences und Oxford Nanopore) erzeugt werden. Dieses Projekt zielte darauf, diese Hemnisse zu beseitigen, um die Verwendung von Genomgraphen zu ermöglichen. Teilprojekt A adressiert insbesondere die Verwendung von Graphengenomen in Verbindung mit Long-Read-Sequenzierungstechniken und umfasste vier wesentliche Entwicklungsziele/Meilensteine:

- M2.1: Weiterentwicklung unseres Tools GraphAligner zum Alignieren von langen Reads an Genomgraphen.
- M2.2: Entwicklung von Werkzeugen um mit Hilfe von langen Reads Varianten zu genotypisieren, die in Genomgraphen kodiert sind.
- M2.3: Erweiterung des Genomgraphen mit Hilfe von langen Reads.
- M2.4: Haplotyperekonstruktion auf Graphenebene mit Hilfe von langen Reads.

2 Ablauf und wesentliche Ergebnisse

Die Arbeiten zu M2.1 waren bereits im ersten Jahr weitestgehend abgeschlossen und der entsprechende Milestone ist erreicht worden. Diese Arbeiten sind in einem Zeitschriftenartitkel

dokumentiert (Rautiainen & Marschall, 2020). Das Software-Tool GraphAligner erlaubt es jetzt, lange Reads an Genomgraphen zu alignieren und erreicht dabei Geschwindigkeiten, die gängige Tools zum Sequenzalignment (Sequenz an Sequenz) übersteigt. Es werden verschiedene Ein- und Ausgabeformate unterstützt (VG und GFA zur Eingabe, JSON, GAM und GAF zur Ausgabe), die eine Interoperabilität ermöglichen, welche wir zur Konstruktion von komplexeren Workflows benötigen. GraphAligner findet in der wissenschaftlichen Community einen große Beachtung und der Artikel wurde bereits 101 mal zitiert (GoogleScholar, Stand 8/2023). GraphAligner ist ein universell einsetzbares Werkzeug, das sowohl unmittelbar für die Ziele in diesem Projekt Anwendung findet, aber auch darüber hinaus. So konnte GraphAligner erfreulicherweise einen wichtigen Beitrag zu einem historischen Erfolg beitragen – der ersten kompletten Rekonstruktion eines menschlichen Genoms (Nurk et al., 2022).

Im Verlauf des Projekts sind durch unsere Zusammenarbeit noch mehr methodische Querverbindungen zwischen M1.2 (Graphbasierte Genominferenz aus kurzen Reads) und M2.2 (Genotypisierung aus langen Reads) deutlich geworden. Daher haben die Projektpartner eng (untereinander als auch mit einem weiteren Drittmittelprojekt[1]) zusammengearbeitet um eine Softwareplattform (PanGenie) zu schaffen, die als Basis zum Erreichen beider Milestones genutzt werden kann. Zunächst haben wir den Anwendungsfall der kurzen Reads in den Mittelpunkt gestellt (M1.2) und entsprechende Evaluationen durchgeführt. An dieser Arbeit waren alle drei Projektpartner beteiligt und es ist im Fachjournal Nature Genetics veröffentlicht worden (Ebler et al., 2022). Die Ergebnisse zeigen, dass graphbasierte Genominferenz äußerst erfolgreich ist und die Zahl der für kurze Reads zugänglichen strukturellen Varianten verdoppeln kann.

Motiviert durch diese Erfolge haben wird PanGenie in Verschiedenen Kontexten angewendet. Im Rahmen des Human Genome Structural Variation Consortium (HGSVC), dem mit Tobias Marschall und Jan Korbel zwei Projekt-PIs angehören, konnten wir zeigen, dass PanGenie in der Tat den Weg ebnet für ein Einschluss von sehr viel mehr genetischen Varianten in genomweiten Assoziationsstudien. Die entsprechenden Resultate wurden in Science veröffentlicht (Ebert et al., 2021). Ferner war PanGenie in der Lage, potentielle Träger seltener struktureller genetischer Varianten (in dem Fall Inversionen), die im Zusammenhang mit Erbkrankheiten stehen, zu identifizieren. Die entsprechenden Ergebnisse sind ebenfalls inzwischen publiziert (Porubsky et al., 2022). Frau Ebler, die über GraphGenomes finanzierte Wissenschaftlerin, hat sich in das Human Pangenome Reference Consortium (HPRC) eingebracht und Pangenie dort angewendet. Wir konnten zeigen, dass durch diese Kombination im Schnitte mehr als 18.000 strukturelle Varianten pro Genom erkannt werden können, während mit Standardmethoden, die keinen Genomgraphen/Pangenom verwenden, weniger als 10.000 erkannt werden. Das entsprechende Manuskript des HPRC „A draft human pangenome reference“ (Liao et al., 2023) wurde 2023 in Nature veöffentlichen. Dieser Meilenstein der Genomik hat weltweit ein sehr großes Medienecho hervorgerufen, mit Berichterstattung in vielen wichtigen Medien (z.B. Süddeutsche Zeitung, FAZ, CNN, Washington Post und viele

1NIH/NHGRI Grant No. U01 HG010973: “Representing structural haplotypes and complex genetic variation in pan-genome graphs”, Prof. Marschall is a co-PI.

andere, insgesamt 315 Medien laut Altmetric[2]).

Parallel dazu sind die Arbeiten zu M2.2 vorangeschritten. Frau Ebler (die durch das Projekt finanzierte Wissenschaftlerin) hat gemeinsam mit Samarendra Pani (zunächste Masterstudent, dann durch Haushaltsmittel finanzierter Wissenschaftlicher Mitarbeiter) einen Prototypen zur graphbasierten Genotypisierung mit Hilfe von langen Reads entwickelt. Eine vollständige Implementierung zum Prozessieren von Long Reads, die an Graphen aligniert sind, liegt nun vor. Das bedeutet, das nun GAF-Dateien, in denen solche Alignments gespeichert sind, eingelesen werden können und statt der sonst üblichen BAM-Dateien verwendet werden können um Genotypisierung (M2.2) und Haplotypisierung (M2.4) durchzuführen. Hierbei sind zunächst zusätzliche Schwierigkeiten aufgetreten, da ein initialer Prototyp zunächst zu langsam und speicherintensiv war. Durch die Zusammenarbeit mit Herrn Pani konnte die zusätzliche Arbeit aber bewältigt werden. Konkret waren zusätzliche Verarbeitungsschritte nötig um eine Ordnung auf Graphen (im GFA-Format) zu definieren und GAF-Dateien in Bezug auf diese Ordnung zu sortieren.

Um die Ziele aus M2.3 zu erreichen, haben wir einen Workflow names SVarp entwickelt. Dieser ist in der Lage, Long Reads zu identifisieren, die Signaturen für strukturelle Varianten tragen, d.h. die Hinweise liefern, dass der Genomgraph an einer bestimmten Stelle erweitert werden muss. Diese werden dann zu Clustern zusammengefasst und mittels lokalem Genomassembly zusammengesetzt. Die resultierenden Sequenzen, die wir „SVtigs“ nennen, können dann als Eingabe für Algorithen zur Graphkonstruktion dienen. Somit steht ein kompletter Workflow für M2.3 nun zur Verfügung.

3 Positionen des zahlenmäßigen Nachweises

Neben den Personalkosten hatten wir noch Mittel für einen leistungsfähigen Laptop beantragt und verwendet. Dieser Laptop war erforderlich, ein Entwickeln und Testen der Software aus diesem Projekt auf den weniger leistungsstarken Geräten, die als Grundausstattung zur Verfügung stehen, nicht möglich war.

4 Nutzen und Verwertbarkeit der Ergebnisse

Der in der Vorhabenbeschreibung dargestellte Verwertungsplan ist weiterhin voll anwendbar. Das GraphGenomes-Projekt hat Analysemethoden basierend auf Genomgraphen entwickelt, getestet und als Open Source Software der wissenschaftlichen Gemeinschaft bereitgestellt. Frei nutzbare Genomgraphen sind nach enger Zusammenarbeit durch das Human Pangenome Reference Consortium (HPRC) verfügbar gemacht worden. Die wissenschaftliche Beschreibung und Evaluationen unserer Software-Komponenten, sowie deren Anwendungen, sind (oder werden) in wissenschaftlichen Fachzeitschriften publiziert.

2https://nature.altmetric.com/details/148167380, abgerufen am 22.8.2023

5 Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Die rapiden Fortschritte bei der Produktion von langen Reads, insbesondere beim „circular consensus sequencing“ (Wenger et al., 2019), und bei Algorithmen zum Genomassembly, haben die Datenbasis zur Konstruktion von Pangenomen während der Projektlaufzeit dramatisch verändert. Dies hat es erfordert, dass wir uns mit der Frage beschäftigt haben, wie mit diesen neuen Datentypen entsprechende Genomassemblies erstellt werden können. Interessanterweise gibt es methodischen Berührungspunkte zu M2.3. Entsprechende Ergebnisse zum Genomassembly haben wir beschrieben (Porubsky et al., 2021).

Wie oben beschrieben, hat sich eine enge Zusammenarbeit zwischen dem vom NIH finanzierten HPRC und unserem Projekt GraphGenomes ergeben. Zum einen hat zwischen GraphGenomes (031L0184A) und dem Projekt “Representing structural haplotypes and complex genetic variation in pan-genome graphs” (NIH/NHGRI Grant No. U01 HG010973) Synergie entwickelt (die Arbeitspakete sind jedoch disjunkt) und es ist die Software Pangenie aus dieser Zusammenarbeit entstanden.

Das vom NHGRI/NIH mit erheblichen Mitteln betriebene Human Pangenome Reference Consortium (HPRC) hat in großem Maße Daten für die Erstellung eines Pangenoms erzeugt. Das ist sehr erfreulich, denn GraphGenomes hat ja keine Datenproduktion vorgesehen. Dadurch haben sich die beiden Projekte sehr gut ergänzt. Während die von uns entwickelten Methoden im HPRC eingesetzt wurden, so profitierte unsere Methodenentwicklung unmittelbar von den Daten, die in HPRC produziert wurden.

7 Erfolgte oder geplanten Veröffentlichungen

Folgende Veröffentlichungen sind im Kontext von GraphGenomes erschienen:

Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, Buonaiuto S, Chang XH, Cheng H, Chu J, Colonna V, Eizenga JM, Feng X, Fischer C, Fulton RS, Garg S, Groza C, Guarracino A, Harvey WT, Heumos S, Howe K, Jain M, Lu T-Y, Markello C, Martin FJ, Mitchell MW, Munson KM, Mwaniki MN, Novak AM, Olsen HE, Pesout T, Porubsky D, Prins P, Sibbesen JA, Sirén J, Tomlinson C, Villani F, Vollger MR, Antonacci-Fulton LL, Baid G, Baker CA, Belyaeva A, Billis K, Carroll A, Chang P-C, Cody S, Cook DE, Cook-Deegan RM, Cornejo OE, Diekhans M, Ebert P, Fairley S, Fedrigo O, Felsenfeld AL, Formenti G, Frankish A, Gao Y, Garrison NA, Giron CG, Green RE, Haggerty L, Hoekzema K, Hourlier T, Ji HP, Kenny EE, Koenig BA, Kolesnikov A, Korbel JO, Kordosky J, Koren S, Lee H, Lewis AP, Magalhães H, Marco-Sola S, Marijon P, McCartney A, McDaniel J, Mountcastle J, Nattestad M, Nurk S, Olson ND, Popejoy AB, Puiu D, Rautiainen M, Regier AA, Rhie A, Sacco S, Sanders AD, Schneider VA, Schultz BI, Shafin K, Smith MW, Sofia HJ, Abou Tayoun AN, et al. A draft human pangenome reference. ***Nature***. 2023;617:312–324.

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, Yilmaz F, Zhao X, Hsieh P, Lee J, Kumar S, Lin J, Rausch T, Chen Y, Ren J, Santamarina M, Höps W, Ashraf H, Chuang NT, Yang X, Munson KM, Lewis AP, Fairley

S, Tallon LJ, Clarke WE, Basile AO, Byrska-Bishop M, Corvelo A, Evani US, Lu T-Y, Chaisson MJP, Chen J, Li C, Brand H, Wenger AM, Ghareghani M, Harvey WT, Raeder B, Hasenfeld P, Regier AA, Abel HJ, Hall IM, Flicek P, Stegle O, Gerstein MB, Tubio JMC, Mu Z, Li YI, Shi X, Hastie AR, Ye K, Chong Z, Sanders AD, Zody MC, Talkowski ME, Mills RE, Devine SE, Lee C, Korbel JO, Marschall T, Eichler EE. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. ***Science*** [Internet]. 2021;372. Available from: http://dx.doi.org/10.1126/science.abf7117

Porubsky D, Höps W, Ashraf H, Hsieh P, Rodriguez-Martin B, Yilmaz F, Ebler J, Hallast P, Maria Maggiolini FA, Harvey WT, Henning B, Audano PA, Gordon DS, Ebert P, Hasenfeld P, Benito E, Zhu Q, Human Genome Structural Variation Consortium (HGSVC), Lee C, Antonacci F, Steinrücken M, Beck CR, Sanders AD, Marschall T, Eichler EE, Korbel JO. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. ***Cell***. 2022;185:1986–2005.e26.

Porubsky D, Ebert P, Audano PA, Vollger MR, Harvey WT, Marijon P, Ebler J, Munson KM, Sorensen M, Sulovari A, Haukness M, Ghareghani M, Human Genome Structural Variation Consortium, Lansdorp PM, Paten B, Devine SE, Sanders AD, Lee C, Chaisson MJP, Korbel JO, Eichler EE, Marschall T. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. ***Nat Biotechnol*.** 2021;39:302–308.

Dabbaghie F, Ebler J, Marschall T. BubbleGun: enumerating bubbles and superbubbles in genome graphs. ***Bioinformatics*.** 2022;38:4217–4219.

Ebler J, Ebert P, Clarke WE, Rausch T, Audano PA, Houwaart T, Mao Y, Korbel JO, Eichler EE, Zody MC, Dilthey AT, Marschall T. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. ***Nat Genet*.** 2022;54:518–525.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, Aganezov S, Hoyt SJ, Diekhans M, Logsdon GA, Alonge M, Antonarakis SE, Borchers M, Bouffard GG, Brooks SY, Caldas GV, Chen N-C, Cheng H, Chin C-S, Chow W, de Lima LG, Dishuck PC, Durbin R, Dvorkina T, Fiddes IT, Formenti G, Fulton RS, Fungtammasan A, Garrison E, Grady PGS, Graves-Lindsay TA, Hall IM, Hansen NF, Hartley GA, Haukness M, Howe K, Hunkapiller MW, Jain C, Jain M, Jarvis ED, Kerpedjiev P, Kirsche M, Kolmogorov M, Korlach J, Kremitzki M, Li H, Maduro VV, Marschall T, McCartney AM, McDaniel J, Miller DE, Mullikin JC, Myers EW, Olson ND, Paten B, Peluso P, Pevzner PA, Porubsky D, Potapova T, Rogaev EI, Rosenfeld JA, Salzberg SL, Schneider VA, Sedlazeck FJ, Shafin K, Shew CJ, Shumate A, Sims Y, Smit AFA, Soto DC, Sović I, Storer JM, Streets A, Sullivan BA, Thibaud-Nissen F, Torrance J, Wagner J, Walenz BP, Wenger A, Wood JMD, Xiao C, Yan SM, Young AC, Zarate S, Surti U, McCoy RC, Dennis MY, Alexandrov IA, Gerton JL, O'Neill RJ, Timp W, Zook JM, Schatz MC, Eichler EE, et al. The complete sequence of a human genome. ***Science*.** 2022;376:44–53.

Rautiainen M, Marschall T. GraphAligner: rapid and versatile sequence-to-graph alignment.

***Genome Biol*.** 2020;21:253.

Darüber hinaus planen wir noch weitere Publikationen, die die Ergebnisse von M2.2 und M2.3 näher beschreiben.

8 Literatur

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74.

Computational Pan-Genomics Consortium. (2016). Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbw089

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

Dilthey, A. T., Gourraud, P.-A., Mentzer, A. J., Cereb, N., Iqbal, Z., & McVean, G. (2016). High-Accuracy HLA Type Inference from Whole-Genome Sequencing Data Using Population Reference Graphs. *PLoS Computational Biology*, *12*(10), e1005151.

Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., Yilmaz, F., Zhao, X., Hsieh, P., Lee, J., Kumar, S., Lin, J., Rausch, T., Chen, Y., Ren, J., … Eichler, E. E. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, *372*(6537). https://doi.org/10.1126/science.abf7117

Ebler, J., Ebert, P., Clarke, W. E., Rausch, T., Audano, P. A., Houwaart, T., Mao, Y., Korbel, J. O., Eichler, E. E., Zody, M. C., Dilthey, A. T., & Marschall, T. (2022). Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nature Genetics*, *54*(4), 518–525.

Eggertsson, H. P., Jonsson, H., Kristmundsdottir, S., Hjartarson, E., Kehr, B., Masson, G., Zink, F., Hjorleifsson, K. E., Jonasdottir, A., Jonasdottir, A., Jonsdottir, I., Gudbjartsson, D. F., Melsted, P., Stefansson, K., & Halldorsson, B. V. (2017). Graphtyper enables population-scale genotyping using pangenome graphs. *Nature Genetics*, *49*(11), 1654–1660.

Garg, S., Rautiainen, M., Novak, A. M., Garrison, E., Durbin, R., & Marschall, T. (to appear 2018). A graph-based approach to diploid genome assembly. *Bioinformatics* .

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Lin, M. F., Paten, B., & Durbin, R. (2017). Sequence variation aware genome references and read mapping with the variation graph toolkit. In *bioRxiv* (p. 234856). https://doi.org/10.1101/234856

Genome of the Netherlands Consortium. (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, *46*(8), 818–825.

Ghaffaari, A., & Marschall, T. (2019). Fully-sensitive seed finding in sequence graphs using a hybrid index. *Bioinformatics* , *35*(14), i81–i89.

Hehir-Kwa, J. Y., Marschall, T., Kloosterman, W. P., Francioli, L. C., Baaijens, J. A., Dijkstra, L. J., Abdellaoui, A., Koval, V., Thung, D. T., Wardenaar, R., Renkens, I., Coe, B. P., Deelen, P., de Ligt, J., Lameijer, E.-W., van Dijk, F., Hormozdiari, F., Genome of the Netherlands

Consortium, Uitterlinden, A. G., … Guryev, V. (2016). A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nature Communications*, *7*, 12989.

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., … Paten, B. (2023). A draft human pangenome reference. *Nature*, *617*(7960), 312–324.

Maretty, L., Jensen, J. M., Petersen, B., Sibbesen, J. A., Liu, S., Villesen, P., Skov, L., Belling, K., Theil Have, C., Izarzugaza, J. M. G., Grosjean, M., Bork-Jensen, J., Grove, J., Als, T. D., Huang, S., Chang, Y., Xu, R., Ye, W., Rao, J., … Schierup, M. H. (2017). Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature*, *548*(7665), 87–91.

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., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53.

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

Porubsky, D., Ebert, P., Audano, P. A., Vollger, M. R., Harvey, W. T., Marijon, P., Ebler, J., Munson, K. M., Sorensen, M., Sulovari, A., Haukness, M., Ghareghani, M., Human Genome Structural Variation Consortium, Lansdorp, P. M., Paten, B., Devine, S. E., Sanders, A. D., Lee, C., Chaisson, M. J. P., … Marschall, T. (2021). Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nature Biotechnology*, *39*(3), 302–308.

Porubsky, D., Höps, W., Ashraf, H., Hsieh, P., Rodriguez-Martin, B., Yilmaz, F., Ebler, J., Hallast, P., Maria Maggiolini, F. A., Harvey, W. T., Henning, B., Audano, P. A., Gordon, D. S., Ebert, P., Hasenfeld, P., Benito, E., Zhu, Q., Human Genome Structural Variation Consortium (HGSVC), Lee, C., … Korbel, J. O. (2022). Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell*, *185*(11), 1986–2005.e26.

Rakocevic, G., Semenyuk, V., Spencer, J., Browning, J., Johnson, I., Arsenijevic, V., Nadj, J., Ghose, K., Suciu, M. C., Ji, S.-G., Demir, G., Li, L., Toptas, B. C., Dolgoborodov, A., Pollex, B., Glotova, I., Komar, P., Li, Y., Popovic, M., … Kural, D. (2017). Fast and Accurate Genomic Analyses using Genome Graphs. In *bioRxiv* (p. 194530). https://doi.org/10.1101/194530

Rausch, T., Jones, D. T. W., Zapatka, M., Stütz, A. M., Zichner, T., Weischenfeldt, J., Jäger, N., Remke, M., Shih, D., Northcott, P. A., Pfaff, E., Tica, J., Wang, Q., Massimi, L., Witt, H., Bender, S., Pleier, S., Cin, H., Hawkins, C., … Korbel, J. O. (2012). Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell*, *148*(1-2), 59–71.

Rautiainen, M., & Marschall, T. (2017). Aligning sequences to general graphs in O(V + mE) time. In *bioRxiv* (p. 216127). https://doi.org/10.1101/216127

Rautiainen, M., & Marschall, T. (2020). GraphAligner: rapid and versatile sequence-to-graph

alignment. *Genome Biology*, *21*(1), 253.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81.

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*. https://doi.org/10.1038/s41587-019-0217-9

Yates, L. R., & Campbell, P. J. (2012). Evolution of the cancer genome. *Nature Reviews. Genetics*, *13*(11), 795–806.