10/07/23

Art, Nach- und Reichhaltigkeit genetischer Daten

Peter Krawitz

Institut für Genomische Statistik und Bioinformatik

Uni Bonn

Auszug aus dem Papier zur Dateninfrastruktur

(AG3, Sprecher: Berlage, Haack)

„[…] Genom-Rechenzentren als IT-Dienstleister für die Speicherung und bioinformatische Verarbeitung genomischer Rohdaten unter fachlicher Steuerung der Leistungserbringer bzw. ihrer Bioinformatik. Die geplante Ausweitung der Gesamtgenom-Sequenzierung in der Krankenversorgung wird zu einem steigenden Bedarf an Speicher- und Verarbeitungskapazitäten für die dabei entstehenden Daten führen. Hinzu kommen labortechnische Weiterentwicklungen (bessere genomische Abdeckung, long-read Verfahren), die diesen Bedarf zusätzlich erhöhen. Sowohl aus ökonomischer Sicht als auch zur Sicherstellung der für eine Versorgungsnutzung erforderlichen fachlichen Expertise ist die Zentralisierung von Teilen der Hard- und Software in genomDE daher mittel- bis langfristig unumgänglich. Die Genom-Rechenzentren sind verteilt, aber technisch und logisch stark integriert, sodass sie eine einheitliche Speicherung und Verarbeitung der Daten unterstützen können. […]“

Wieviel Speicherplatz und Rechenleistung brauchen wir?

Die (zu) einfache Antwort:

1. Das Genom ist circa 3 Gigabasen,
2. Unterschiede betreffen ca. 1%,
3. jede Zelle hat das gleiche Genom,
4. eine Base braucht ca. ein Byte und
5. mit Dragen

Science
THE HUMAN GENOME
nature
the human genome

=> benötigt man weniger als 1h für die Prozessierung und 1GB Speicher.

06:23
09:38

Genome Analyzer
IIx
illumina

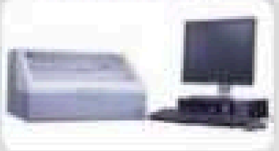
AUTOMATED WORKFLOW

1. SAMPLE PREPARATION

~6 hours (~3 hours hands-on)

- Sample collection, genomic DNA sheared
- DNA end-repair
- Adapter ligation

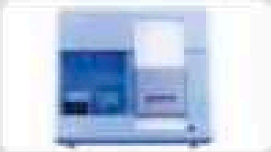
2. CLUSTER GENERATION



~5 hours (~10 min. preparation)

- Flow cell and pre-filled reagents placed into Cluster Station
- Samples applied to flow cell
- Complete walk-away automation

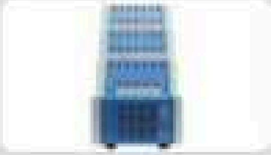
3. SEQUENCING BY SYNTHESIS



~5 days for 2 × 35 bp (~10 min. preparation)

- Flow cell and pre-filled reagents placed on Genome Analyzer
- Complete walk-away automation, including support for longer reads

4. PAIRED-END MODULE

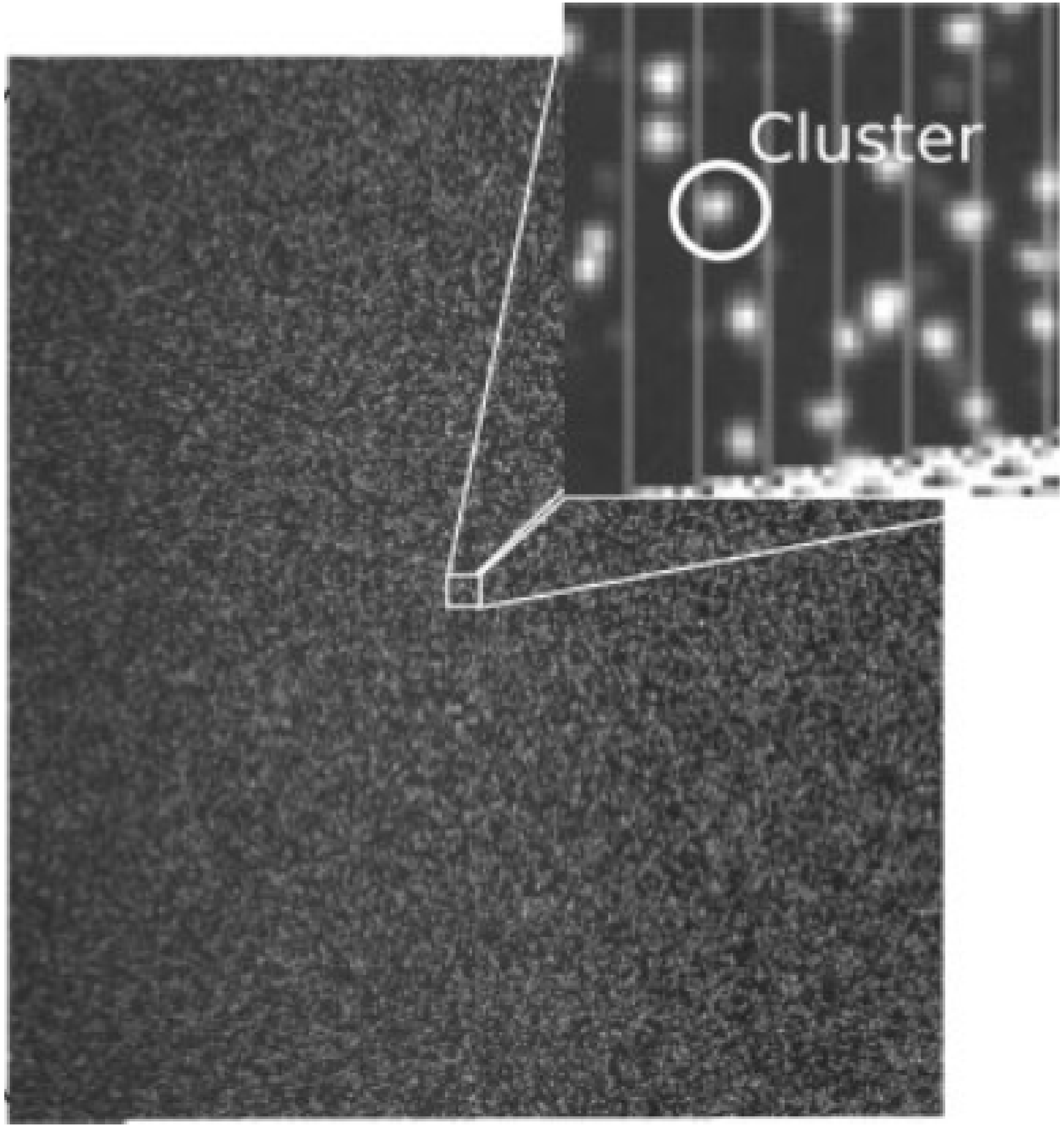


- Add-on module for automated reagent delivery
- Second read prepared and sequenced while flow cell remains on Genome Analyzer

5. DATA ANALYSIS

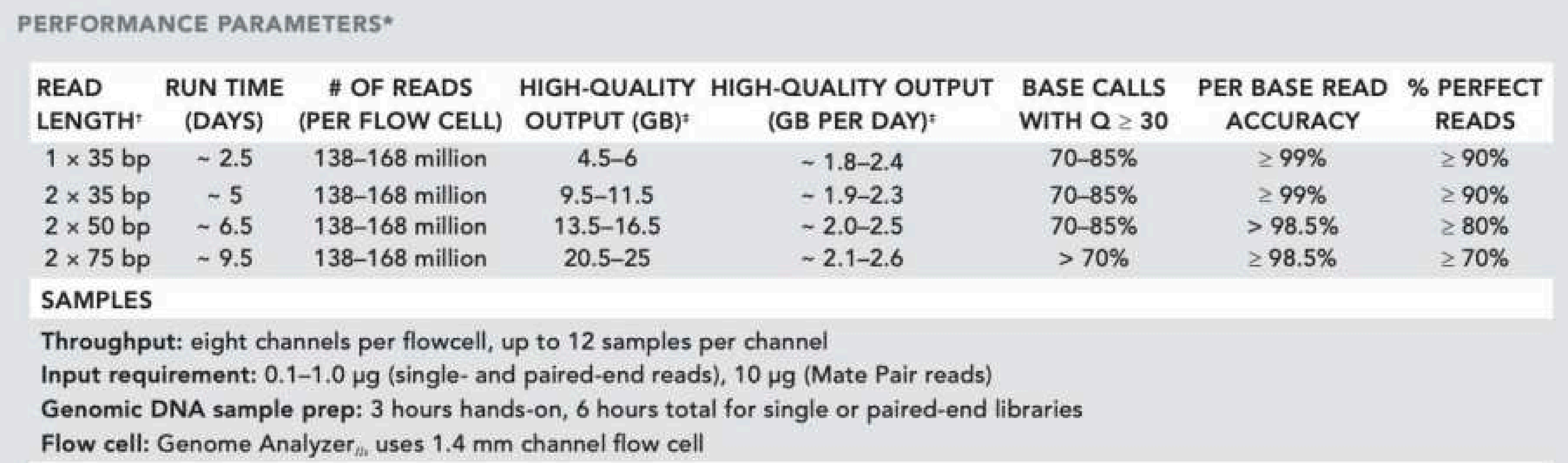
~1–2 days

- Real-time image analysis
- Data transfer to automated analysis pipeline
- Automated base calling
- Paired-end alignment
- Polymorphism detection



How much storage did a GS from Genome Analyzer need?

In each cycle, the flowcell is imaged in a series of non-overlapping regions. The number of regions (known in Illumina terminology as ‘tiles’) depends on the device configuration. The default configuration for a Genome Analyzer 2 is 100 tiles per lane, where there are eight lanes per flowcell. Tiles are separated by a margin to prevent clusters being imaged multiple times. The flowcell is moved under the camera in order to image each tile in each cycle. Therefore, during each cycle 800 tiles are imaged in four channels. Runs are typically around 37 cycles producing a total of 118 400 images. Each GA2 image is a 2048 × 1794 pixel 16 bit grey-scale TIFF (though only 12 bits contain data). Each pixel covers ~ 0.14 μ2 of the flowcell and ~ 9 pixels comprise one cluster object.



PERFORMANCE PARAMETERS*

READ LENGTH[†]	RUN TIME (DAYS)	# OF READS (PER FLOW CELL)	HIGH-QUALITY OUTPUT (GB)[‡]	HIGH-QUALITY OUTPUT (GB PER DAY)[‡]	BASE CALLS WITH Q ≥ 30	PER BASE READ ACCURACY	% PERFECT READS
1 × 35 bp	~ 2.5	138–168 million	4.5–6	~ 1.8–2.4	70–85%	≥ 99%	≥ 90%
2 × 35 bp	~ 5	138–168 million	9.5–11.5	~ 1.9–2.3	70–85%	≥ 99%	≥ 90%
2 × 50 bp	~ 6.5	138–168 million	13.5–16.5	~ 2.0–2.5	70–85%	> 98.5%	≥ 80%
2 × 75 bp	~ 9.5	138–168 million	20.5–25	~ 2.1–2.6	> 70%	≥ 98.5%	≥ 70%

SAMPLES

Throughput: eight channels per flowcell, up to 12 samples per channel

Input requirement: 0.1–1.0 µg (single- and paired-end reads), 10 µg (Mate Pair reads)

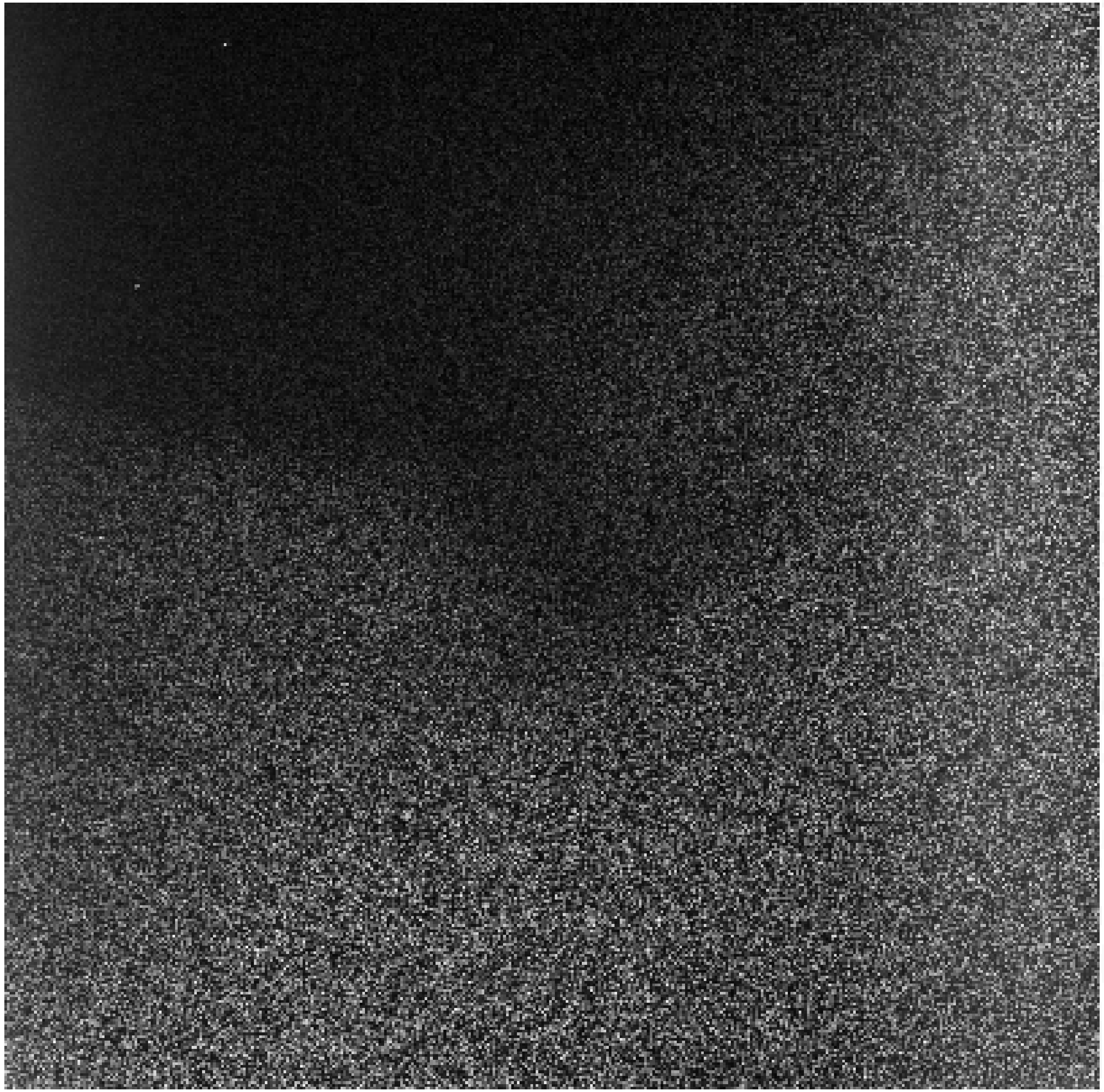
Genomic DNA sample prep: 3 hours hands-on, 6 hours total for single or paired-end libraries

Flow cell: Genome Analyzer$_{IIx}$ uses 1.4 mm channel flow cell

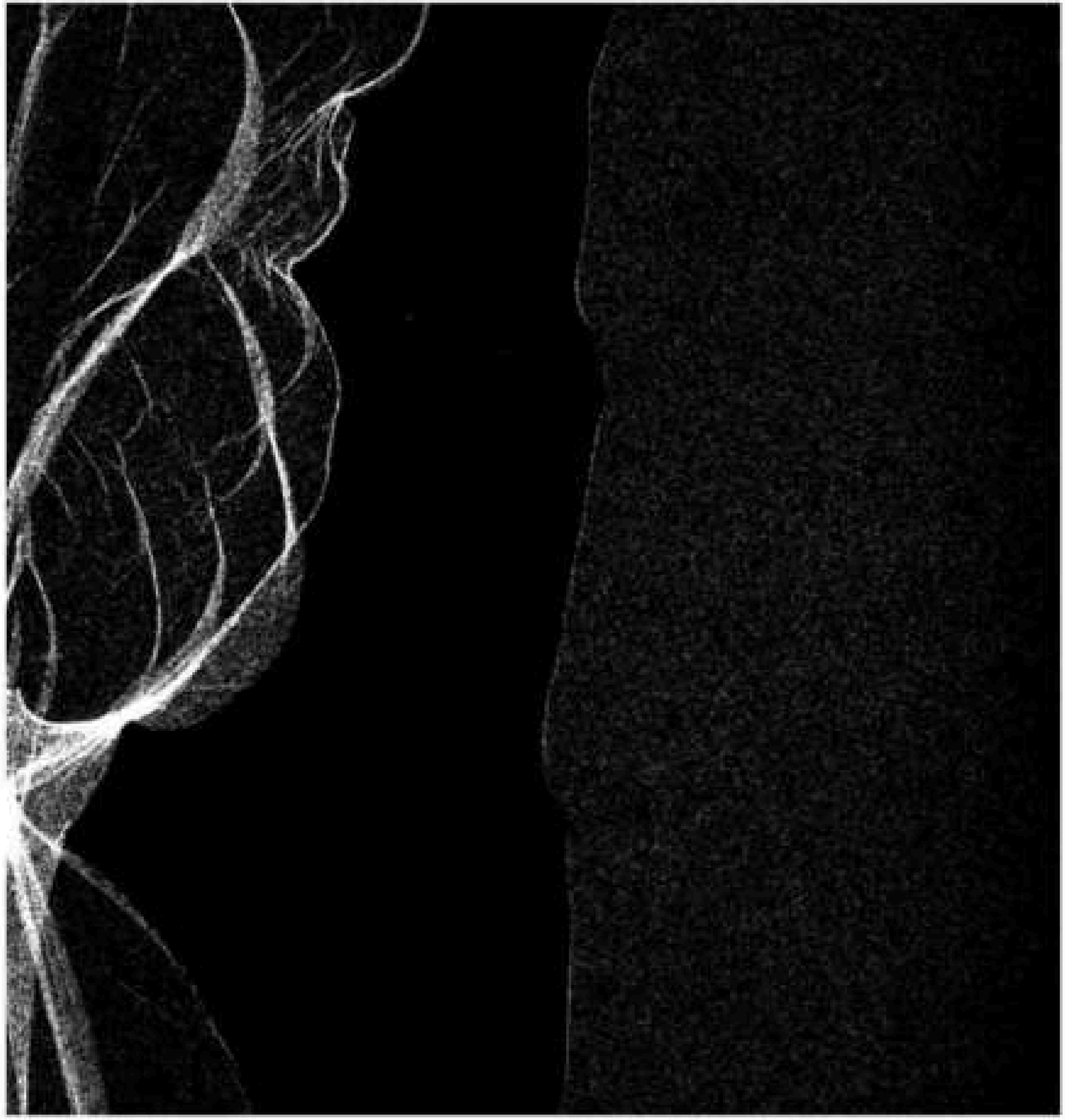
Die ehrliche Antwort vor 10 Jahren mit dem Genome Analyzer brauchte ein 30x Genom:

>2 Wochen

>10 TeraByte

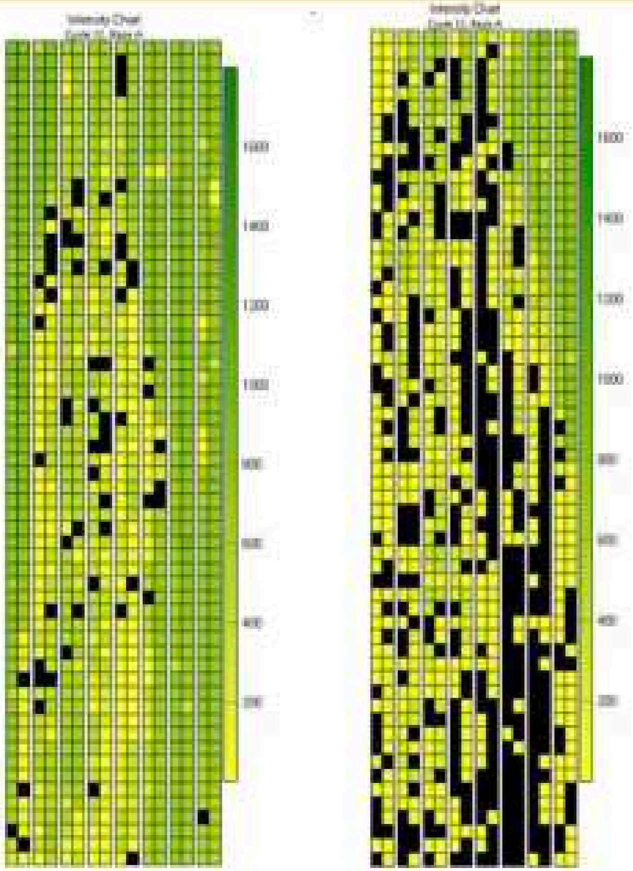


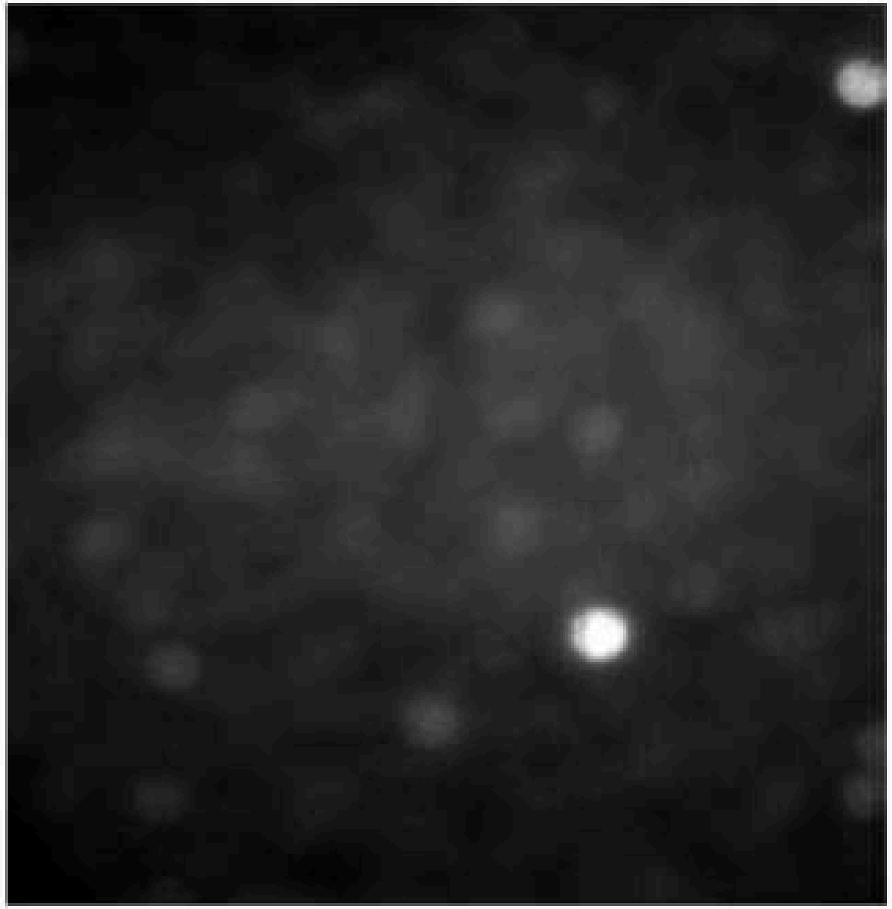
Uneven primer lawn,



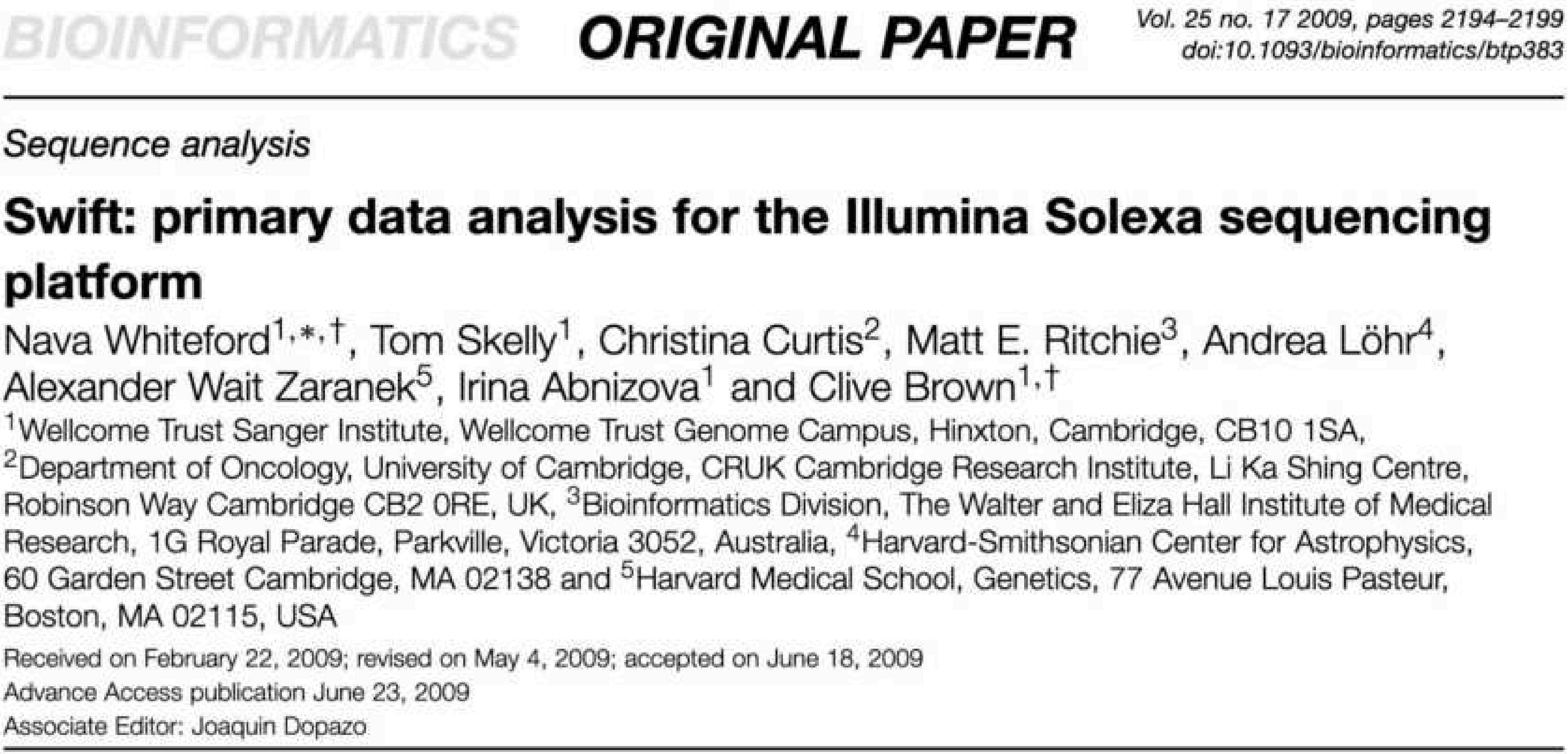
substrate peeling (when the Peltier block failed)

Fail Mode: Tile Out of Focus





- A distorted focus dot usually precedes a tile that is out of focus
- Possible cause: Oil or reagent (leaking from manifold) on surface of flow cell
- Possible cause: Problem with focus laser
- Possible cause: RTA cannot process the tile due to extreme density



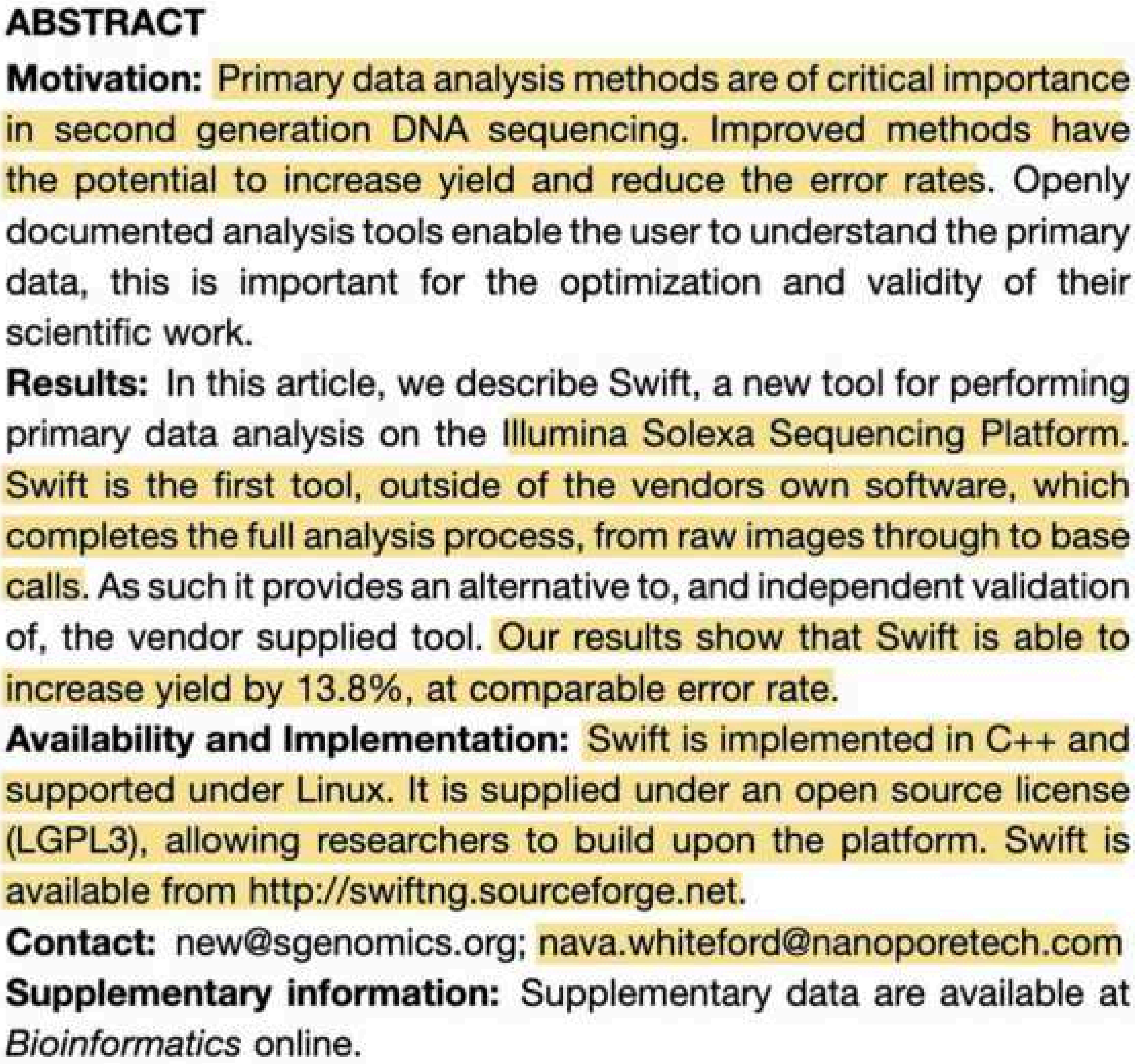
BIOINFORMATICS ORIGINAL PAPER
Vol. 25 no. 17 2009, pages 2194–2199
doi:10.1093/bioinformatics/btp383

Sequence analysis

Swift: primary data analysis for the Illumina Solexa sequencing platform

Nava Whiteford[1,*,†], Tom Skelly[1], Christina Curtis[2], Matt E. Ritchie[3], Andrea Löhr[4], Alexander Wait Zaranek[5], Irina Abnizova[1] and Clive Brown[1,†]
[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, [2]Department of Oncology, University of Cambridge, CRUK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way Cambridge CB2 0RE, UK, [3]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia, [4]Harvard-Smithsonian Center for Astrophysics, 60 Garden Street Cambridge, MA 02138 and [5]Harvard Medical School, Genetics, 77 Avenue Louis Pasteur, Boston, MA 02115, USA

Received on February 22, 2009; revised on May 4, 2009; accepted on June 18, 2009
Advance Access publication June 23, 2009
Associate Editor: Joaquin Dopazo



ABSTRACT

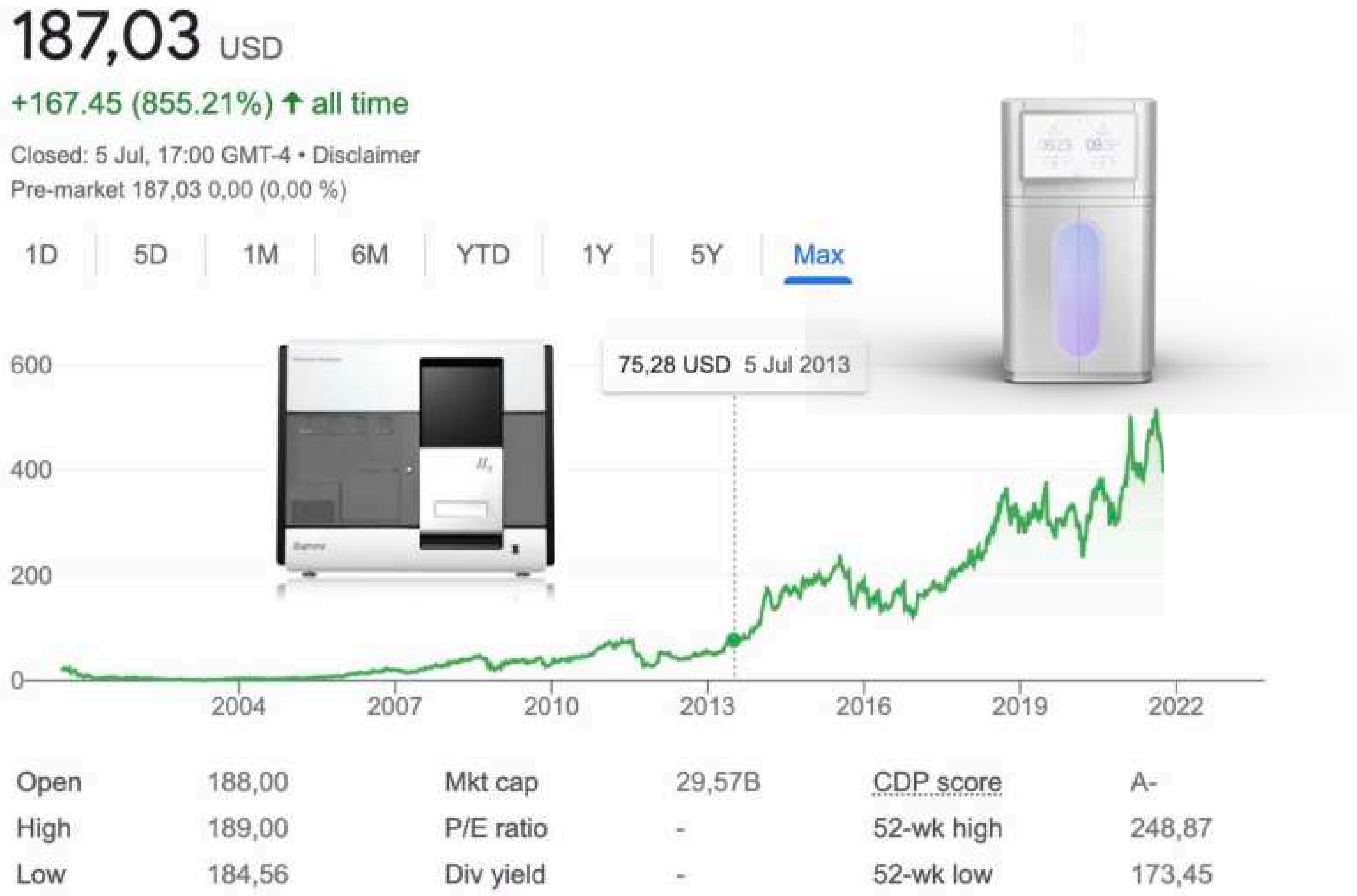
Motivation: Primary data analysis methods are of critical importance in second generation DNA sequencing. Improved methods have the potential to increase yield and reduce the error rates. Openly documented analysis tools enable the user to understand the primary data, this is important for the optimization and validity of their scientific work.

Results: In this article, we describe Swift, a new tool for performing primary data analysis on the Illumina Solexa Sequencing Platform. Swift is the first tool, outside of the vendors own software, which completes the full analysis process, from raw images through to base calls. As such it provides an alternative to, and independent validation of, the vendor supplied tool. Our results show that Swift is able to increase yield by 13.8%, at comparable error rate.

Availability and Implementation: Swift is implemented in C++ and supported under Linux. It is supplied under an open source license (LGPL3), allowing researchers to build upon the platform. Swift is available from http://swiftng.sourceforge.net.

Contact: new@sgenomics.org; nava.whiteford@nanoporetech.com

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



187,03 USD

+167.45 (855.21%) ↑ all time

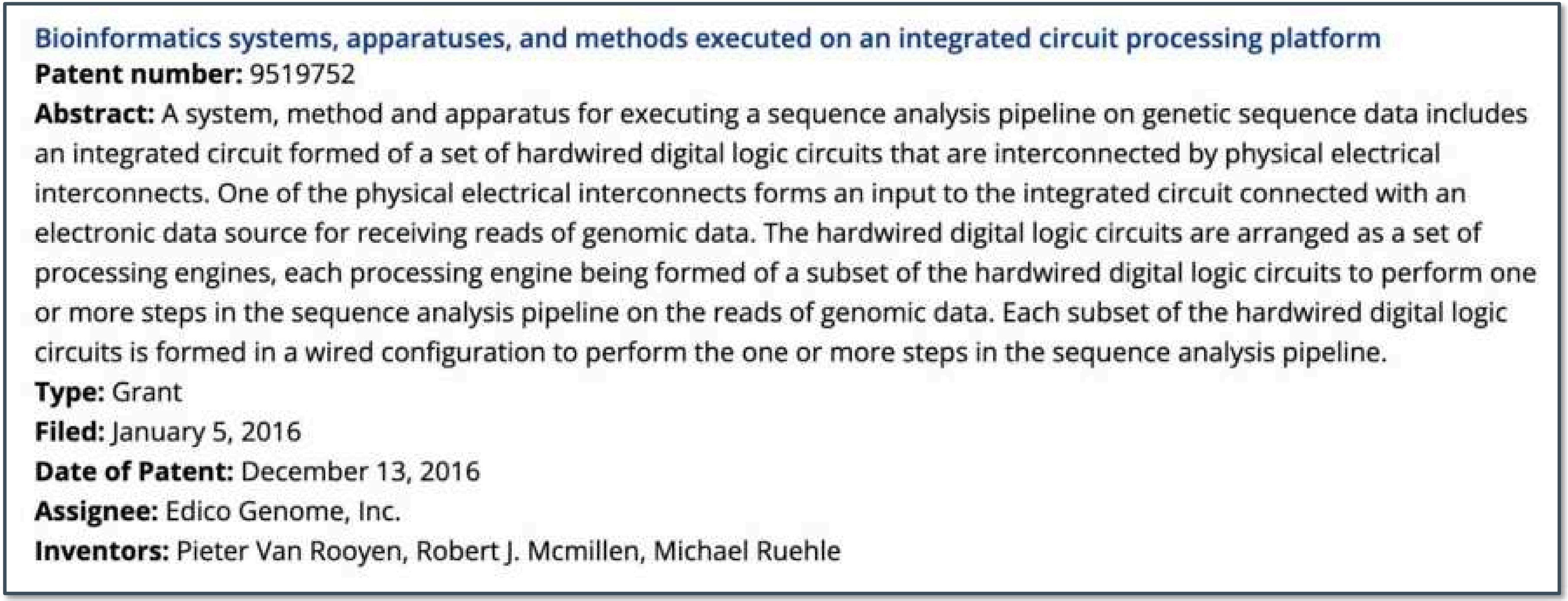
Closed: 5 Jul, 17:00 GMT-4 • Disclaimer

Pre-market 187,03 0,00 (0,00 %)

1D
5D
1M
6M
YTD
1Y
5Y
Max
75,28 USD 5 Jul 2013
600
400
200
0
2004
2007
2010
2013
2016
2019
2022

Open	188,00	Mkt cap	29,57B	CDP score	A-
High	189,00	P/E ratio	-	52-wk high	248,87
Low	184,56	Div yield	-	52-wk low	173,45

Innovation is in our DNA



Bioinformatics systems, apparatuses, and methods executed on an integrated circuit processing platform

Patent number: 9519752

Abstract: A system, method and apparatus for executing a sequence analysis pipeline on genetic sequence data includes an integrated circuit formed of a set of hardwired digital logic circuits that are interconnected by physical electrical interconnects. One of the physical electrical interconnects forms an input to the integrated circuit connected with an electronic data source for receiving reads of genomic data. The hardwired digital logic circuits are arranged as a set of processing engines, each processing engine being formed of a subset of the hardwired digital logic circuits to perform one or more steps in the sequence analysis pipeline on the reads of genomic data. Each subset of the hardwired digital logic circuits is formed in a wired configuration to perform the one or more steps in the sequence analysis pipeline.

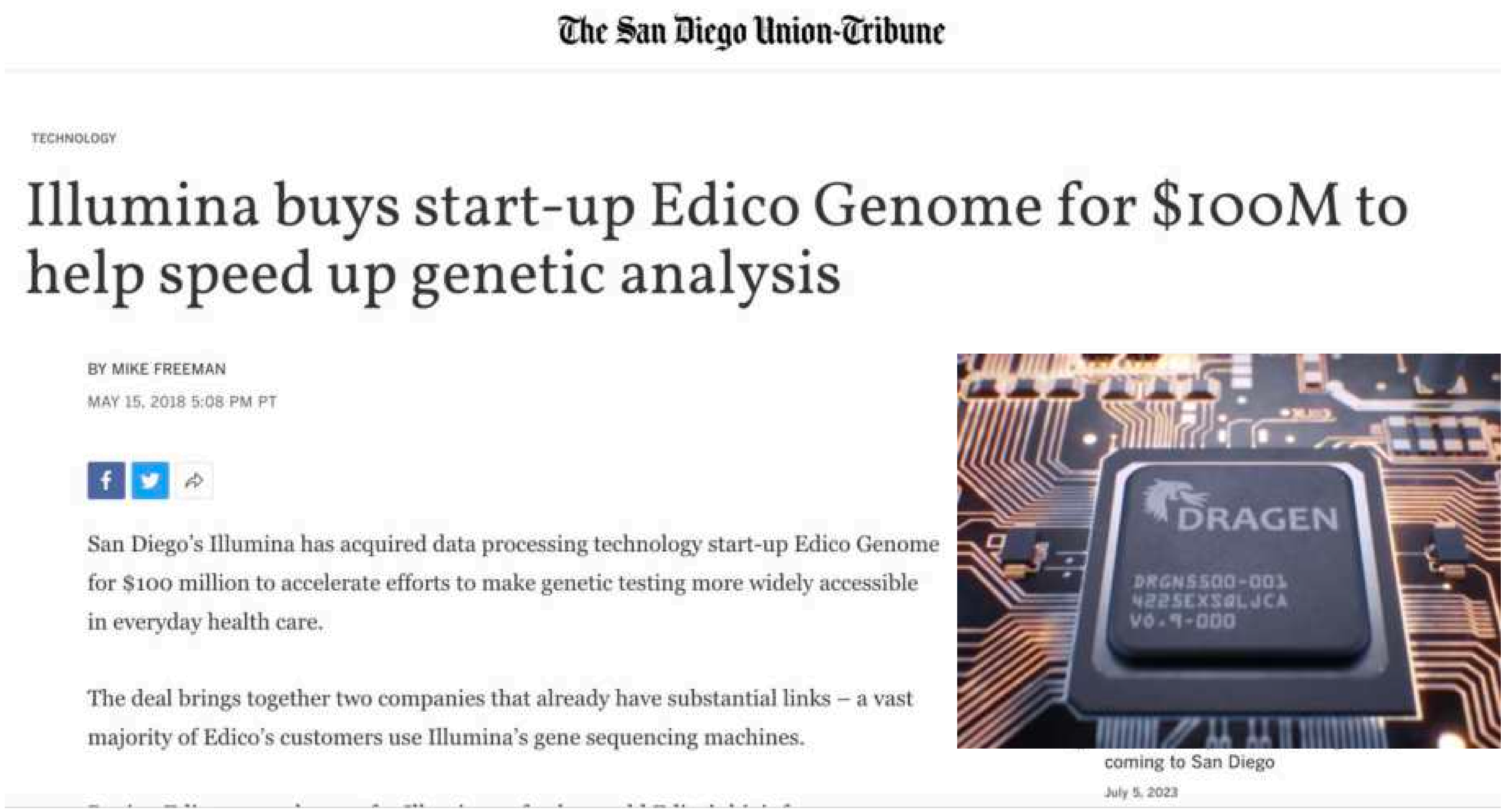
Type: Grant

Filed: January 5, 2016

Date of Patent: December 13, 2016

Assignee: Edico Genome, Inc.

Inventors: Pieter Van Rooyen, Robert J. Mcmillen, Michael Ruehle



The San Diego Union-Tribune

TECHNOLOGY

Illumina buys start-up Edico Genome for $100M to help speed up genetic analysis

BY MIKE FREEMAN

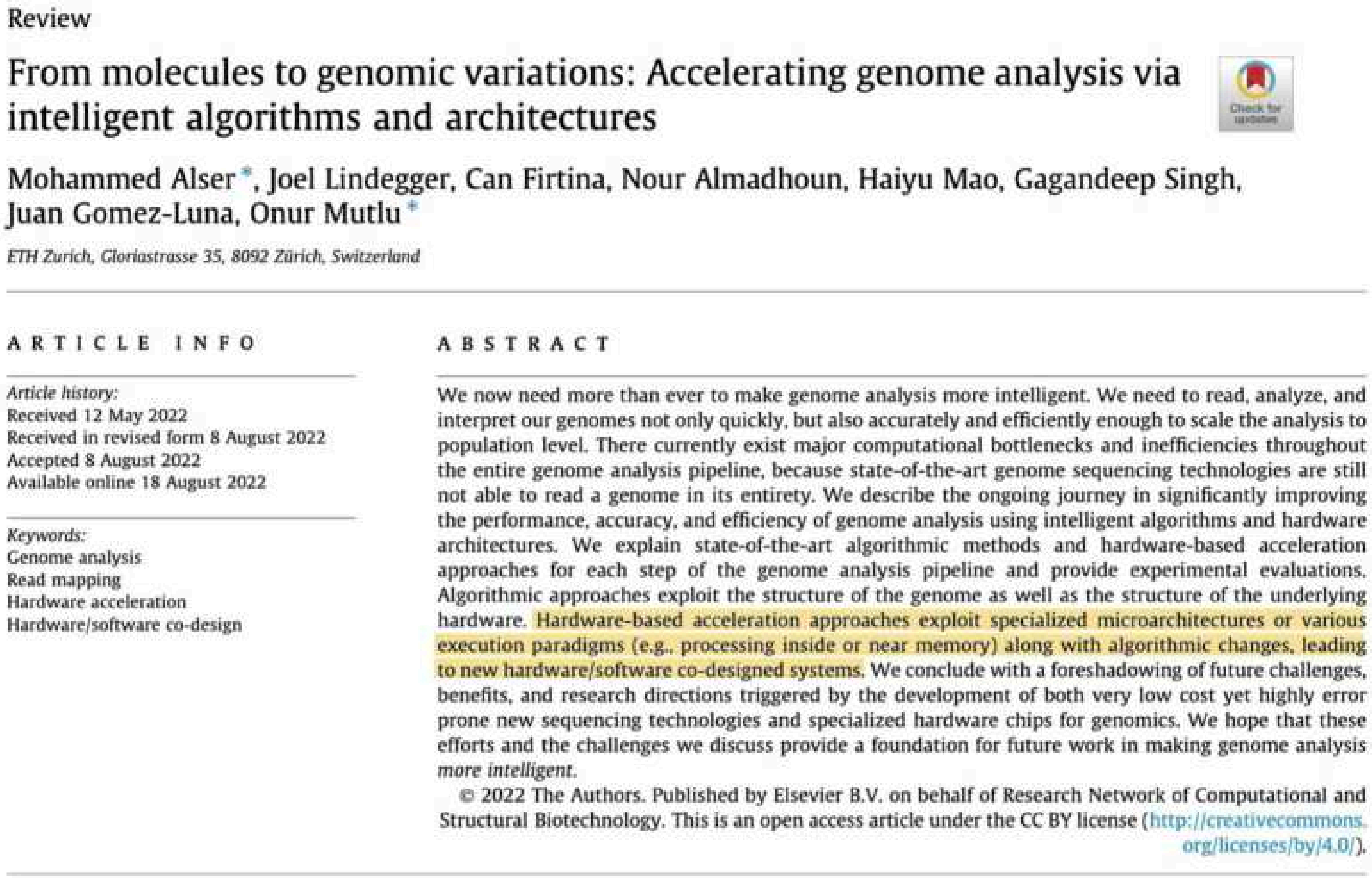
MAY 15, 2018 5:08 PM PT

San Diego's Illumina has acquired data processing technology start-up Edico Genome for $100 million to accelerate efforts to make genetic testing more widely accessible in everyday health care.

The deal brings together two companies that already have substantial links – a vast majority of Edico's customers use Illumina's gene sequencing machines.

coming to San Diego

July 5, 2023



Review

From molecules to genomic variations: Accelerating genome analysis via intelligent algorithms and architectures

Mohammed Alser *, Joel Lindegger, Can Firtina, Nour Almadhoun, Haiyu Mao, Gagandeep Singh, Juan Gomez-Luna, Onur Mutlu *

ETH Zurich, Gloriastrasse 35, 8092 Zürich, Switzerland

ARTICLE INFO

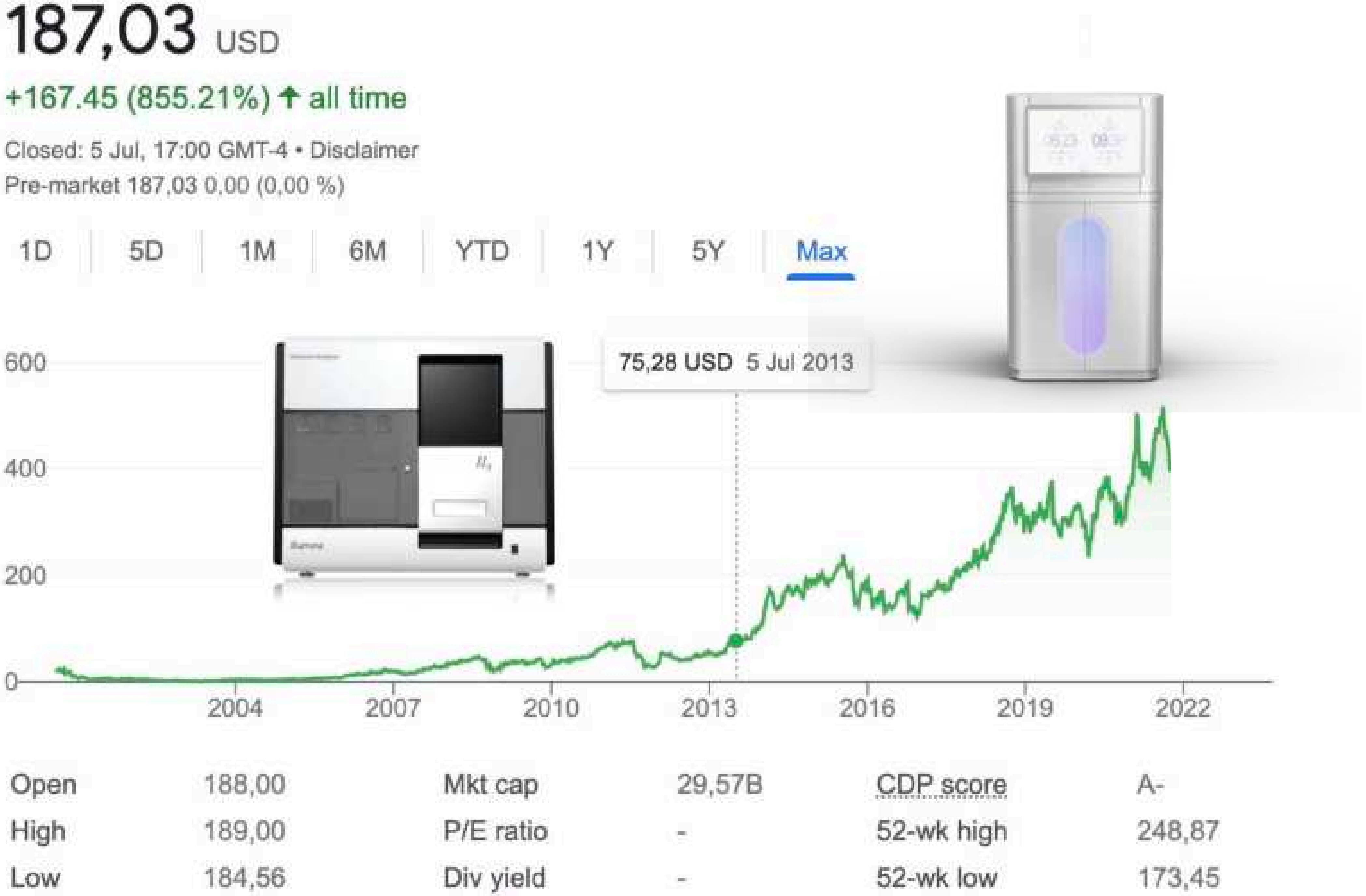
Article history:
Received 12 May 2022
Received in revised form 8 August 2022
Accepted 8 August 2022
Available online 18 August 2022

Keywords:
Genome analysis
Read mapping
Hardware acceleration
Hardware/software co-design

ABSTRACT

We now need more than ever to make genome analysis more intelligent. We need to read, analyze, and interpret our genomes not only quickly, but also accurately and efficiently enough to scale the analysis to population level. There currently exist major computational bottlenecks and inefficiencies throughout the entire genome analysis pipeline, because state-of-the-art genome sequencing technologies are still not able to read a genome in its entirety. We describe the ongoing journey in significantly improving the performance, accuracy, and efficiency of genome analysis using intelligent algorithms and hardware architectures. We explain state-of-the-art algorithmic methods and hardware-based acceleration approaches for each step of the genome analysis pipeline and provide experimental evaluations. Algorithmic approaches exploit the structure of the genome as well as the structure of the underlying hardware. Hardware-based acceleration approaches exploit specialized microarchitectures or various execution paradigms (e.g., processing inside or near memory) along with algorithmic changes, leading to new hardware/software co-designed systems. We conclude with a foreshadowing of future challenges, benefits, and research directions triggered by the development of both very low cost yet highly error prone new sequencing technologies and specialized hardware chips for genomics. We hope that these efforts and the challenges we discuss provide a foundation for future work in making genome analysis *more intelligent*.

© 2022 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).



187,03 USD

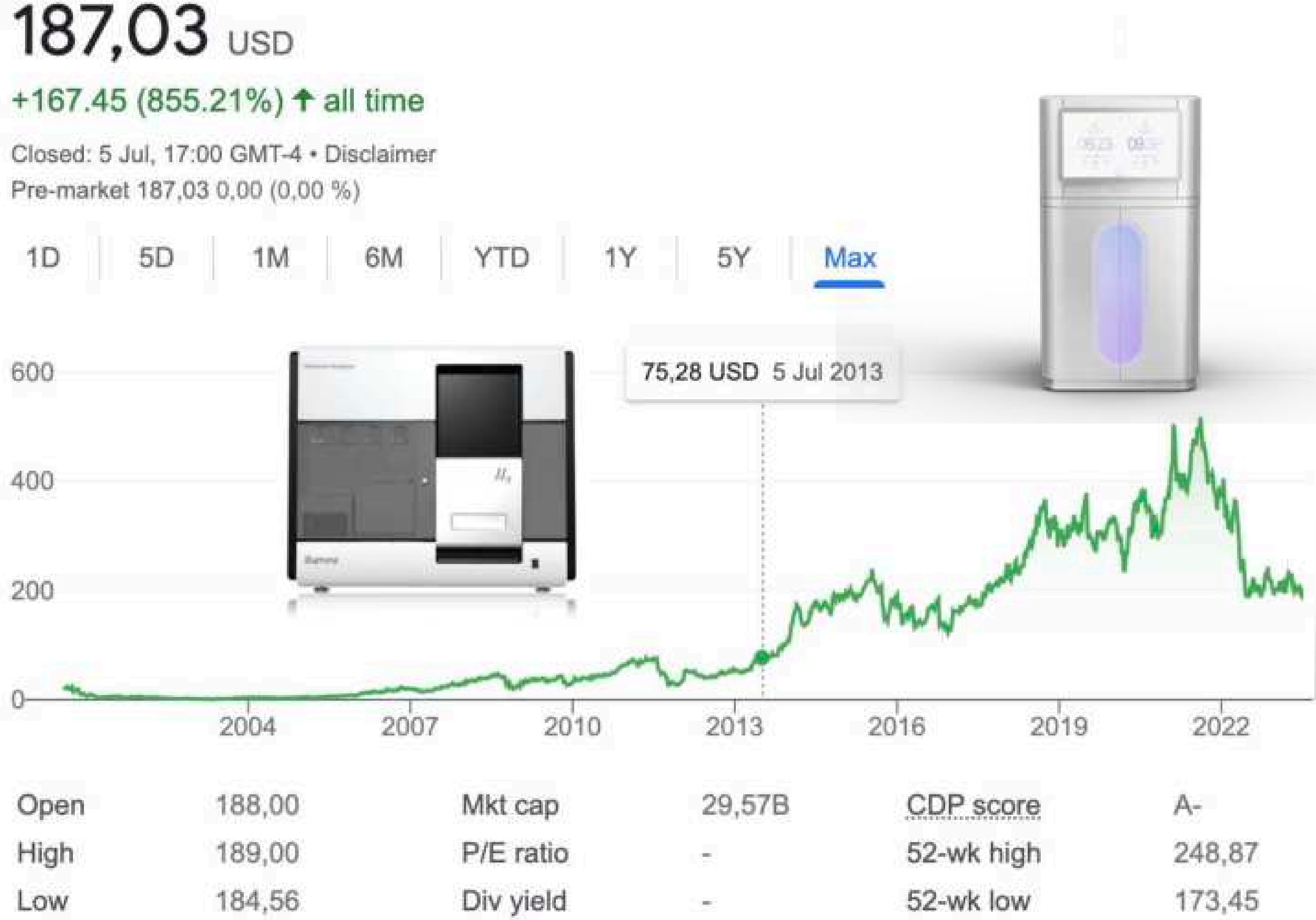
+167.45 (855.21%) ↑ all time

Closed: 5 Jul, 17:00 GMT-4 • Disclaimer

Pre-market 187,03 0,00 (0,00 %)

1D
5D
1M
6M
YTD
1Y
5Y
Max
75,28 USD 5 Jul 2013
600
400
200
0
2004
2007
2010
2013
2016
2019
2022

Open	188,00	Mkt cap	29,57B	CDP score	A-
High	189,00	P/E ratio	-	52-wk high	248,87
Low	184,56	Div yield	-	52-wk low	173,45



187,03 USD

+167.45 (855.21%) ↑ all time

Closed: 5 Jul, 17:00 GMT-4 • Disclaimer

Pre-market 187,03 0,00 (0,00 %)

1D
5D
1M
6M
YTD
1Y
5Y
Max
75,28 USD 5 Jul 2013
600
400
200
0
2004
2007
2010
2013
2016
2019
2022

Open	188,00	Mkt cap	29,57B	CDP score	A-
High	189,00	P/E ratio	-	52-wk high	248,87
Low	184,56	Div yield	-	52-wk low	173,45

Wieviel Speicherplatz und Rechenleistung brauchen wir?

Die (zu) einfache Antwort:

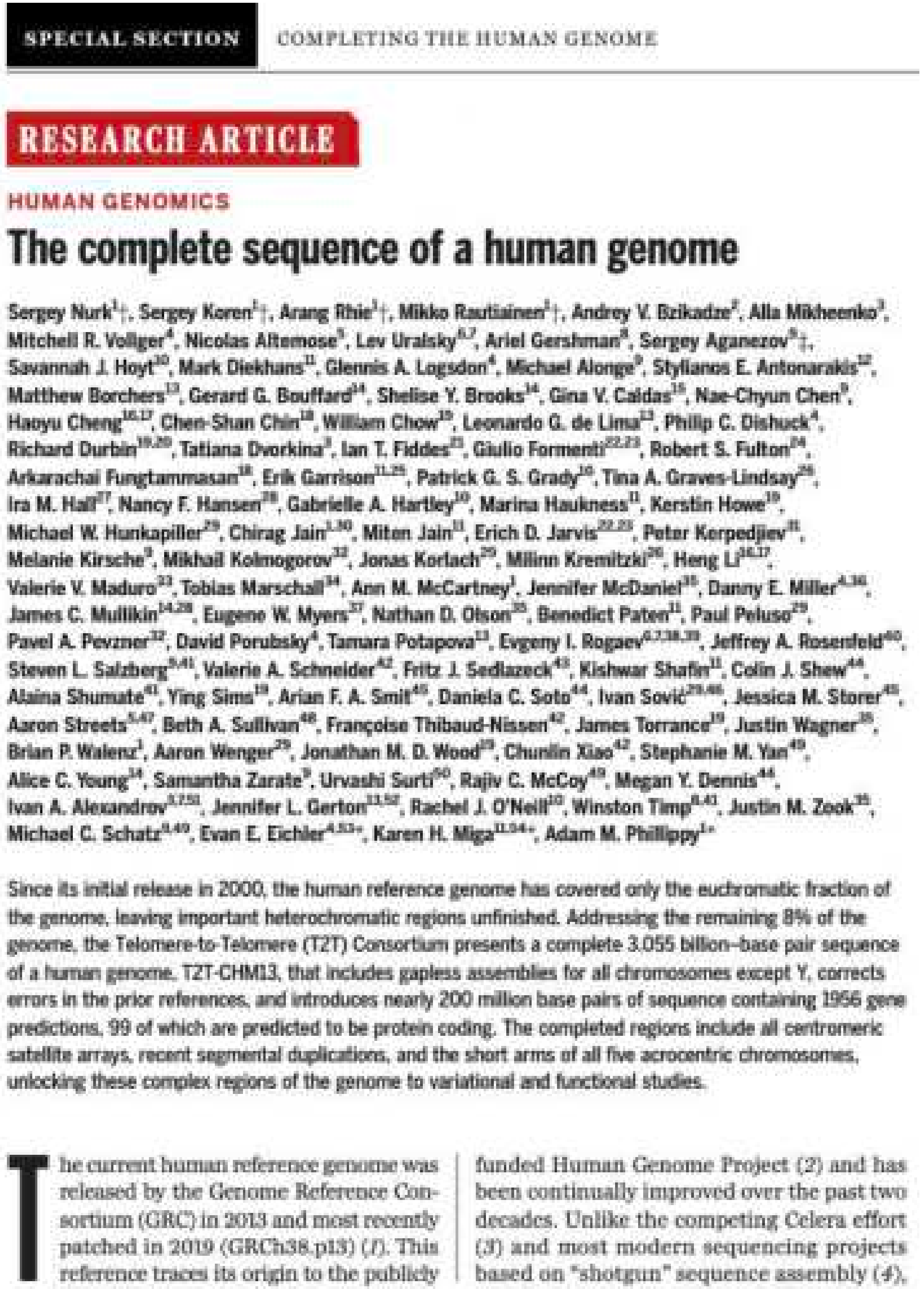
1. Das Genom ist circa 3 Gigabasen,
2. Unterschiede betreffen ca. 10%,

~~3. jede Zelle hat das gleiche Genom~~,

~~4. eine Base braucht ca. ein Byte und~~

~~5. mit Dragen~~

Science
THE HUMAN GENOME
nature
the human genome



SPECIAL SECTION COMPLETING THE HUMAN GENOME

RESEARCH ARTICLE

HUMAN GENOMICS

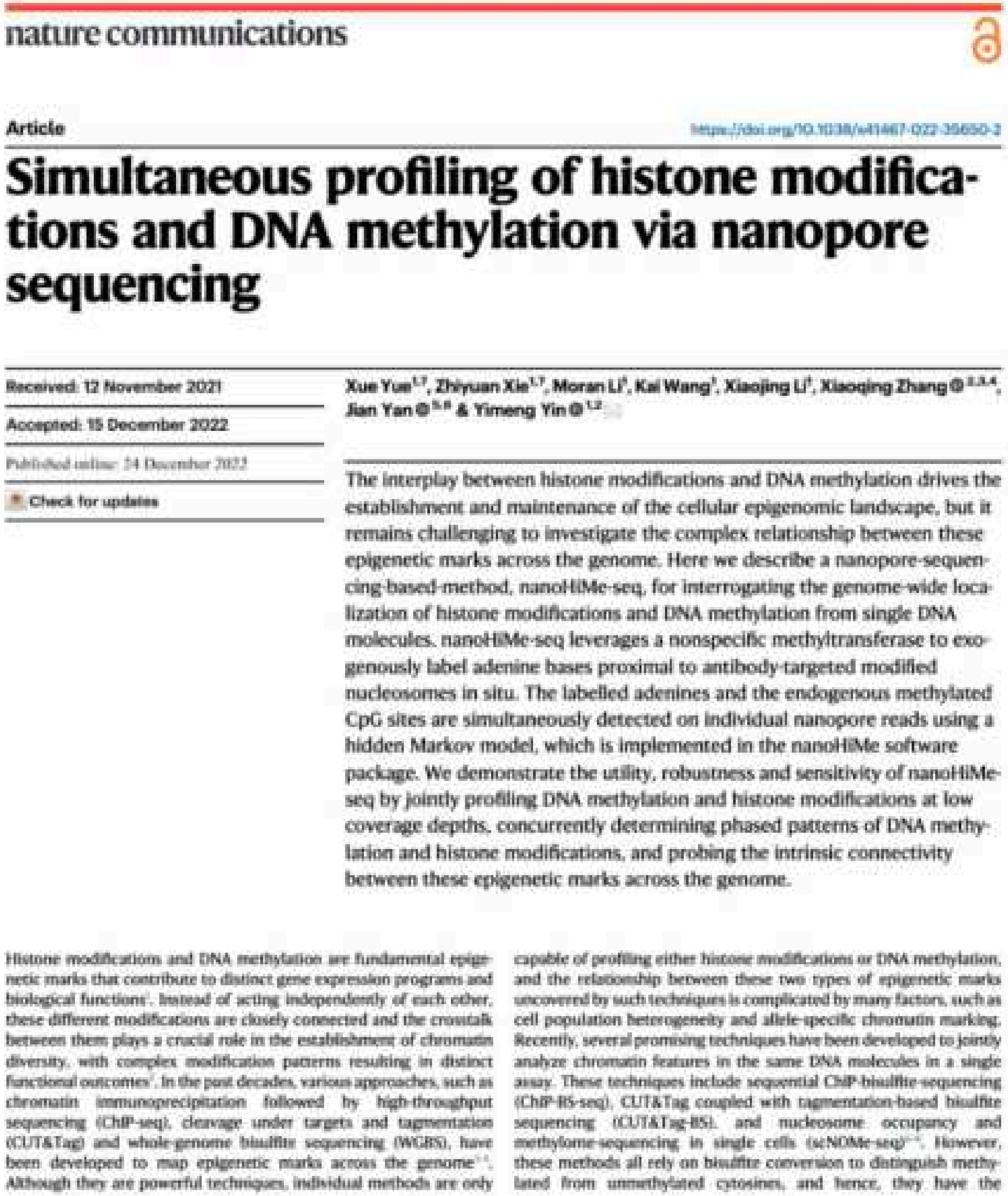
The complete sequence of a human genome

Sergey Nurk[1]†, Sergey Koren[1]†, Arang Rhie[1]†, Mikko Rautiainen[1]†, Andrey V. Bzikadze[2], Alla Mikheenko[3], Mitchell R. Vollger[4], Nicolas Altemose[5], Lev Uralsky[6,7], Ariel Gershman[8], Sergey Aganezov[9]‡, Savannah J. Hoyt[10], Mark Diekhans[11], Glennis A. Logsdon[4], Michael Alonge[9], Stylianos E. Antonarakis[12], Matthew Borchers[13], Gerard G. Bouffard[14], Shelise Y. Brooks[14], Gina V. Caldas[15], Nae-Chyun Chen[9], Haoyu Cheng[16,17], Chen-Shan Chin[18], William Chow[19], Leonardo G. de Lima[13], Philip C. Dishuck[4], Richard Durbin[19,20], Tatiana Dvorkina[3], Ian T. Fiddes[21], Giulio Formenti[22,23], Robert S. Fulton[24], Arkarachai Fungtammasan[18], Erik Garrison[11,25], Patrick G. S. Grady[10], Tina A. Graves-Lindsay[26], Ira M. Hall[27], Nancy F. Hansen[28], Gabrielle A. Hartley[10], Marina Haukness[11], Kerstin Howe[19], Michael W. Hunkapiller[29], Chirag Jain[1,30], Miten Jain[11], Erich D. Jarvis[22,23], Peter Kerpedjiev[31], Melanie Kirsche[9], Mikhail Kolmogorov[32], Jonas Korlach[29], Milinn Kremitzki[26], Heng Li[16,17], Valerie V. Maduro[33], Tobias Marschall[34], Ann M. McCartney[1], Jennifer McDaniel[35], Danny E. Miller[4,36], James C. Mullikin[14,28], Eugene W. Myers[37], Nathan D. Olson[35], Benedict Paten[11], Paul Peluso[29], Pavel A. Pevzner[32], David Porubsky[4], Tamara Potapova[13], Evgeny I. Rogaev[6,7,38,39], Jeffrey A. Rosenfeld[40], Steven L. Salzberg[9,41], Valerie A. Schneider[42], Fritz J. Sedlazeck[43], Kishwar Shafin[11], Colin J. Shew[44], Alaina Shumate[41], Ying Sims[19], Arian F. A. Smit[45], Daniela C. Soto[44], Ivan Sović[29,46], Jessica M. Storer[45], Aaron Streets[5,47], Beth A. Sullivan[48], Françoise Thibaud-Nissen[42], James Torrance[19], Justin Wagner[35], Brian P. Walenz[1], Aaron Wenger[29], Jonathan M. D. Wood[19], Chunlin Xiao[42], Stephanie M. Yan[49], Alice C. Young[14], Samantha Zarate[9], Urvashi Surti[50], Rajiv C. McCoy[49], Megan Y. Dennis[44], Ivan A. Alexandrov[3,7,51], Jennifer L. Gerton[13,52], Rachel J. O'Neill[10], Winston Timp[8,41], Justin M. Zook[35], Michael C. Schatz[9,49], Evan E. Eichler[4,53]*, Karen H. Miga[11,54]*, Adam M. Phillippy[1]*

Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished. Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion–base pair sequence of a human genome, T2T-CHM13, that includes gapless assemblies for all chromosomes except Y, corrects errors in the prior references, and introduces nearly 200 million base pairs of sequence containing 1956 gene predictions, 99 of which are predicted to be protein coding. The completed regions include all centromeric satellite arrays, recent segmental duplications, and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies.

The current human reference genome was released by the Genome Reference Consortium (GRC) in 2013 and most recently patched in 2019 (GRCh38.p13) (*1*). This reference traces its origin to the publicly funded Human Genome Project (*2*) and has been continually improved over the past two decades. Unlike the competing Celera effort (*3*) and most modern sequencing projects based on "shotgun" sequence assembly (*4*),

---



nature communications

Article

https://doi.org/10.1038/s41467-022-35650-2

Simultaneous profiling of histone modifications and DNA methylation via nanopore sequencing

Received: 12 November 2021

Accepted: 15 December 2022

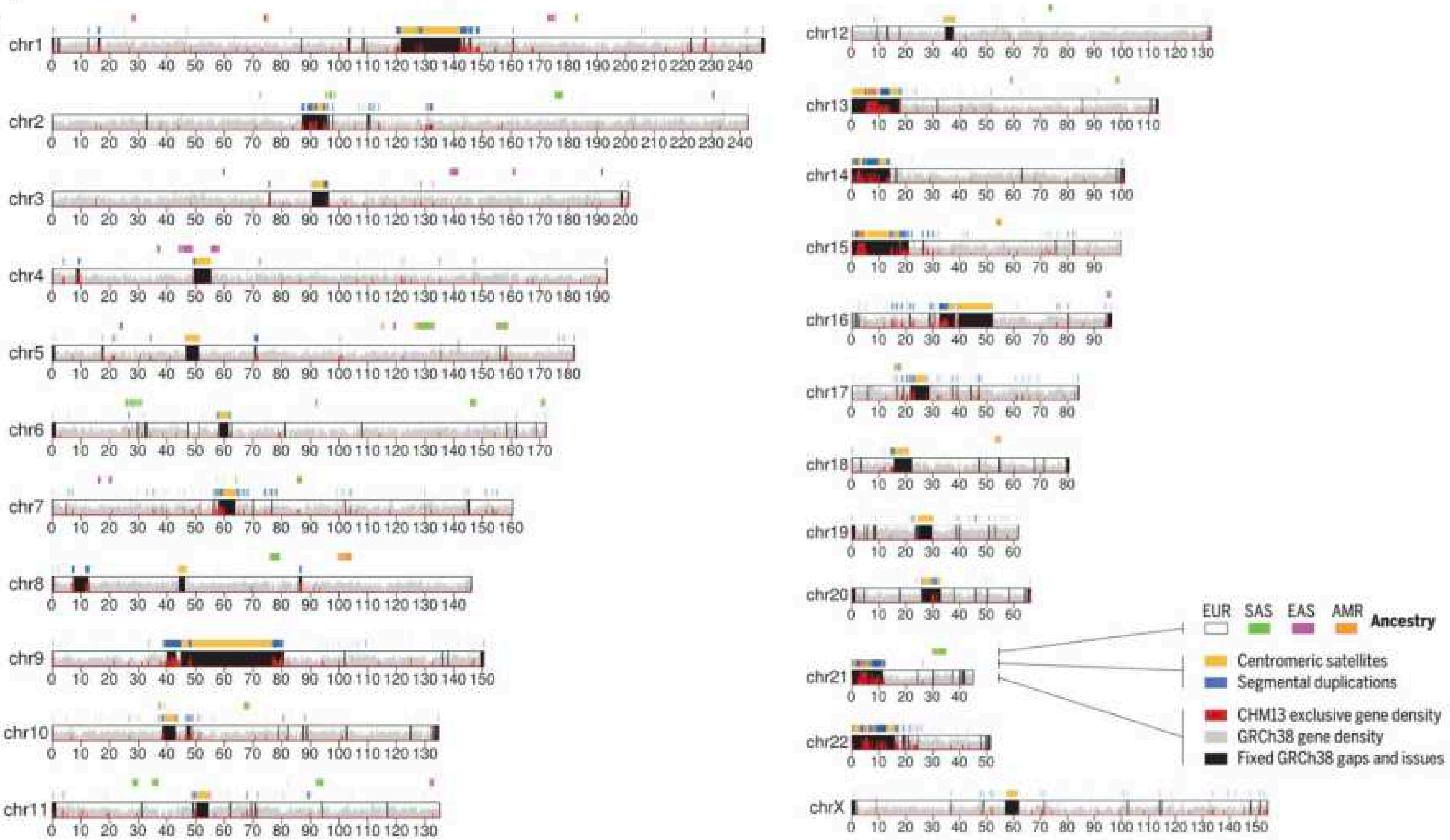
Published online: 24 December 2022

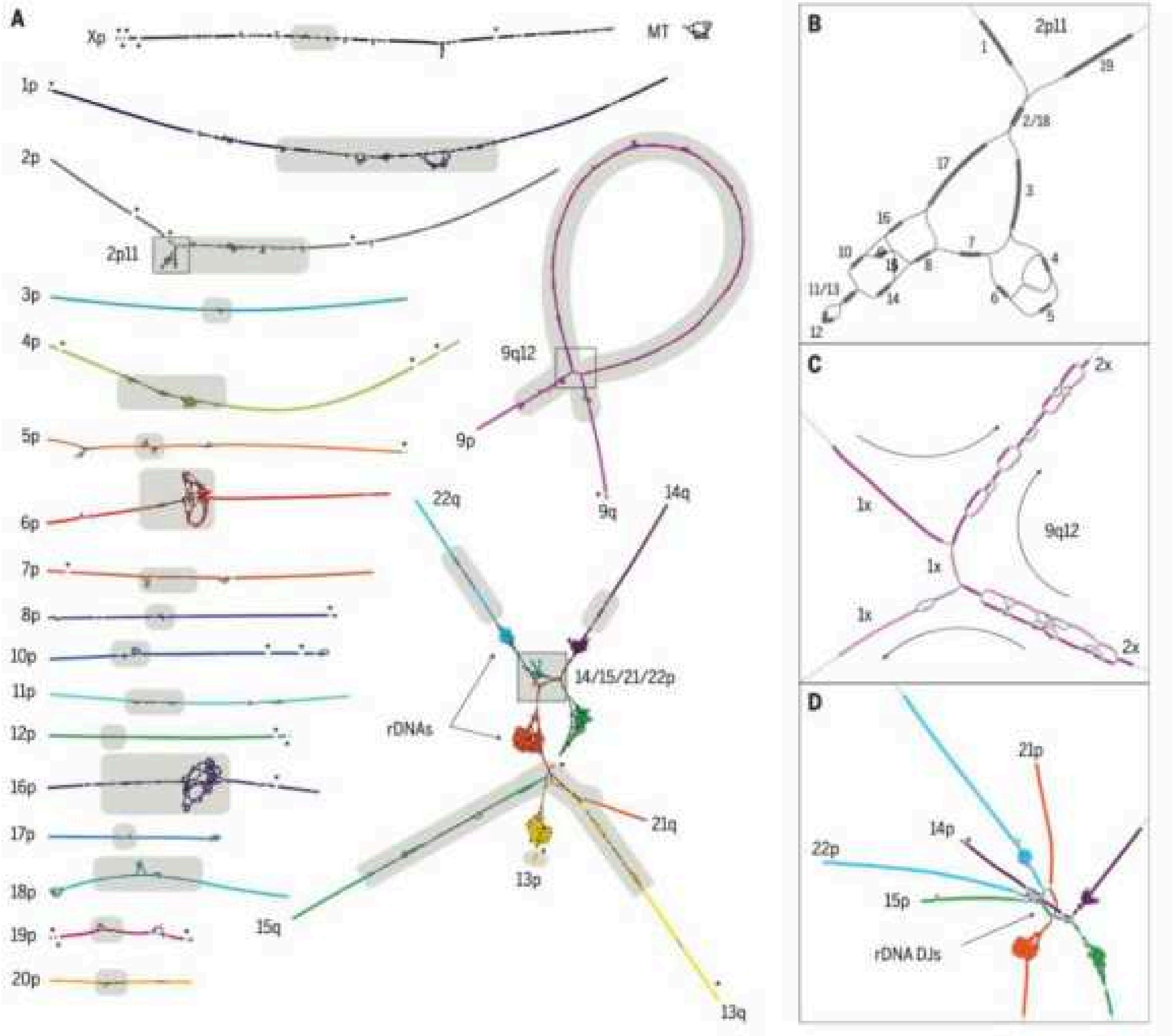
Check for updates

Xue Yue[1,7], Zhiyuan Xie[1,7], Moran Li[1], Kai Wang[1], Xiaojing Li[1], Xiaoqing Zhang[2,3,4], Jian Yan[5,6] & Yimeng Yin[1,2]

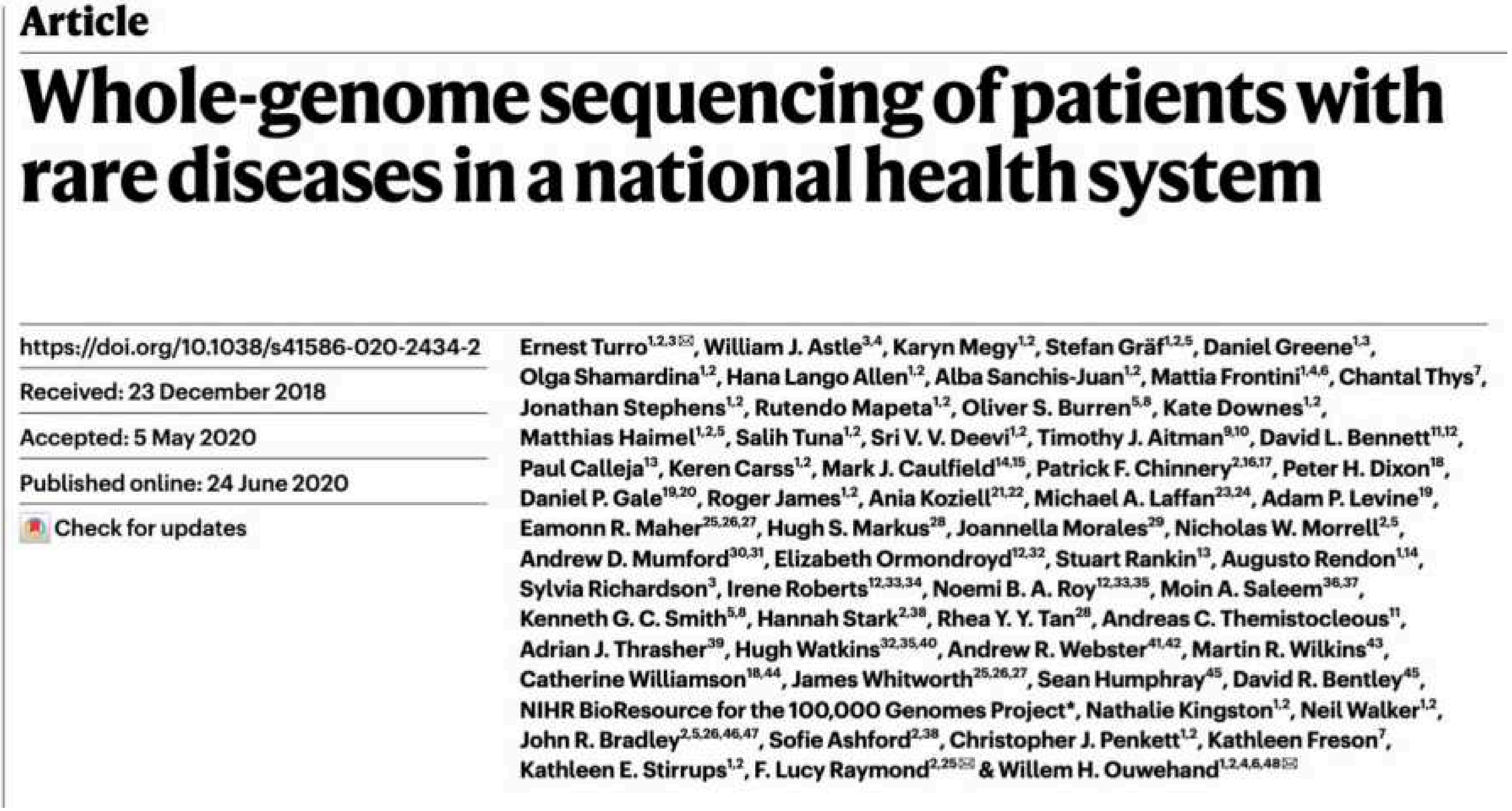
The interplay between histone modifications and DNA methylation drives the establishment and maintenance of the cellular epigenomic landscape, but it remains challenging to investigate the complex relationship between these epigenetic marks across the genome. Here we describe a nanopore-sequencing-based-method, nanoHiMe-seq, for interrogating the genome-wide localization of histone modifications and DNA methylation from single DNA molecules. nanoHiMe-seq leverages a nonspecific methyltransferase to exogenously label adenine bases proximal to antibody-targeted modified nucleosomes in situ. The labelled adenines and the endogenous methylated CpG sites are simultaneously detected on individual nanopore reads using a hidden Markov model, which is implemented in the nanoHiMe software package. We demonstrate the utility, robustness and sensitivity of nanoHiMe-seq by jointly profiling DNA methylation and histone modifications at low coverage depths, concurrently determining phased patterns of DNA methylation and histone modifications, and probing the intrinsic connectivity between these epigenetic marks across the genome.

Histone modifications and DNA methylation are fundamental epigenetic marks that contribute to distinct gene expression programs and biological functions[1]. Instead of acting independently of each other, these different modifications are closely connected and the crosstalk between them plays a crucial role in the establishment of chromatin diversity, with complex modification patterns resulting in distinct functional outcomes[2]. In the past decades, various approaches, such as chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq), cleavage under targets and tagmentation (CUT&Tag) and whole-genome bisulfite sequencing (WGBS), have been developed to map epigenetic marks across the genome[3–5]. Although they are powerful techniques, individual methods are only capable of profiling either histone modifications or DNA methylation, and the relationship between these two types of epigenetic marks uncovered by such techniques is complicated by many factors, such as cell population heterogeneity and allele-specific chromatin marking. Recently, several promising techniques have been developed to jointly analyze chromatin features in the same DNA molecules in a single assay. These techniques include sequential ChIP-bisulfite-sequencing (ChIP-BS-seq), CUT&Tag coupled with tagmentation-based bisulfite sequencing (CUT&Tag-BS), and nucleosome occupancy and methylome-sequencing in single cells (scNOMe-seq)[6–8]. However, these methods all rely on bisulfite conversion to distinguish methylated from unmethylated cytosines, and hence, they have the


chr1
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240
chr2
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240
chr3
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200
chr4
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190
chr5
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180
chr6
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170
chr7
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160
chr8
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140
chr9
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150
chr10
0 10 20 30 40 50 60 70 80 90 100 110 120 130
chr11
0 10 20 30 40 50 60 70 80 90 100 110 120 130
chr12
0 10 20 30 40 50 60 70 80 90 100 110 120 130
chr13
0 10 20 30 40 50 60 70 80 90 100 110
chr14
0 10 20 30 40 50 60 70 80 90 100
chr15
0 10 20 30 40 50 60 70 80 90
chr16
0 10 20 30 40 50 60 70 80 90
chr17
0 10 20 30 40 50 60 70 80
chr18
0 10 20 30 40 50 60 70 80
chr19
0 10 20 30 40 50 60
chr20
0 10 20 30 40 50 60
chr21
0 10 20 30 40
chr22
0 10 20 30 40 50
chrX
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150
EUR
SAS
EAS
AMR
Ancestry
Centromeric satellites
Segmental duplications
CHM13 exclusive gene density
GRCh38 gene density
Fixed GRCh38 gaps and issues


A
Xp
MT
1p
2p
2p11
3p
4p
5p
6p
7p
8p
10p
11p
12p
16p
17p
18p
19p
20p
9q12
9p
9q
22q
14q
14/15/21/22p
rDNAs
21q
13p
15q
13q
B
2p11
1
19
2/18
17
3
16
10
15
8
7
4
11/13
14
6
5
12
C
2x
1x
9q12
1x
1x
2x
D
21p
14p
22p
15p
rDNA DJs

Long read genomes will be less “reference based“ and have new challenges for visualization



Article

Whole-genome sequencing of patients with rare diseases in a national health system

https://doi.org/10.1038/s41586-020-2434-2

Received: 23 December 2018

Accepted: 5 May 2020

Published online: 24 June 2020

Check for updates

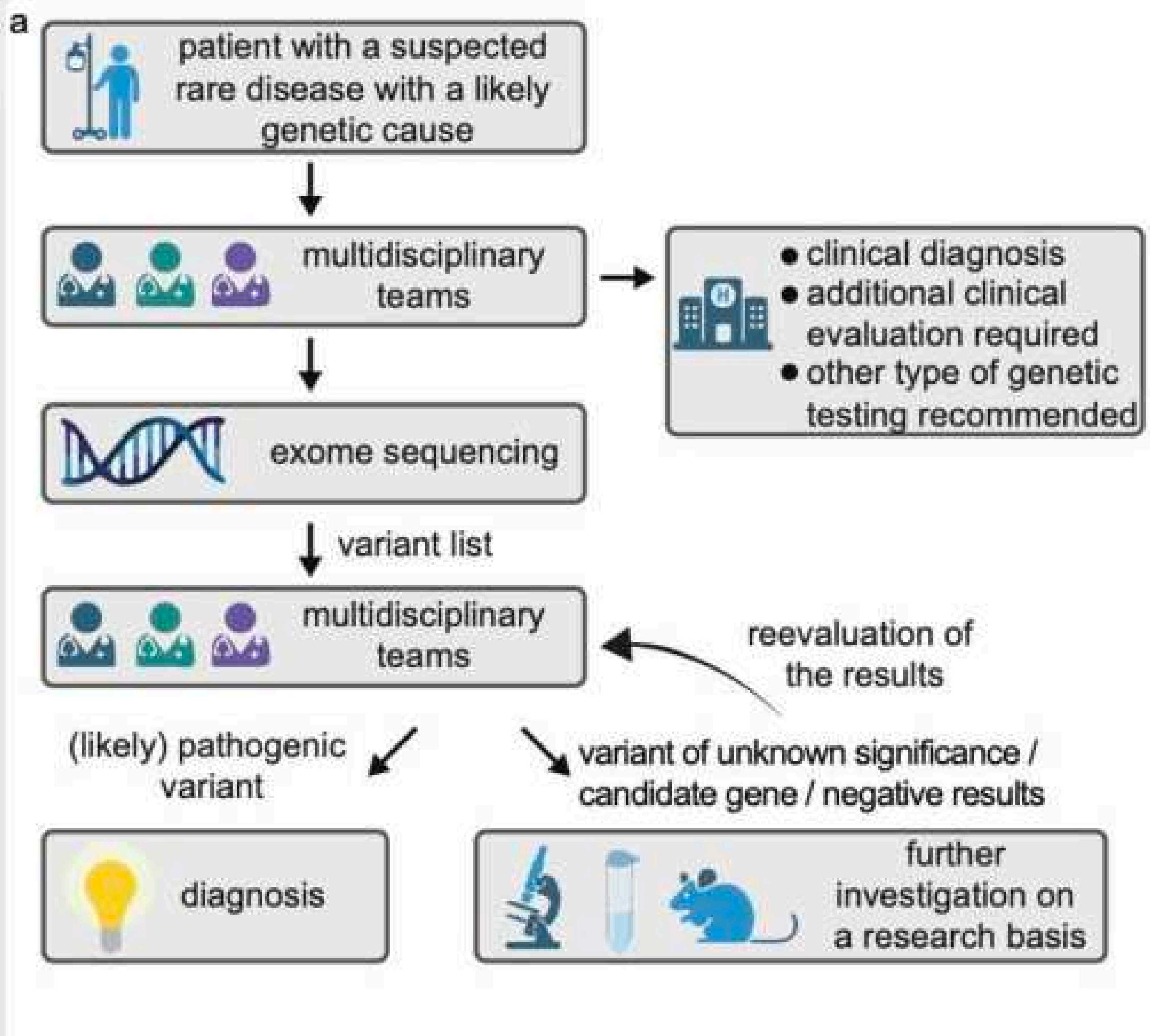
Ernest Turro[1,2,3]✉, William J. Astle[3,4], Karyn Megy[1,2], Stefan Gräf[1,2,5], Daniel Greene[1,3], Olga Shamardina[1,2], Hana Lango Allen[1,2], Alba Sanchis-Juan[1,2], Mattia Frontini[1,4,6], Chantal Thys[7], Jonathan Stephens[1,2], Rutendo Mapeta[1,2], Oliver S. Burren[5,8], Kate Downes[1,2], Matthias Haimel[1,2,5], Salih Tuna[1,2], Sri V. V. Deevi[1,2], Timothy J. Aitman[9,10], David L. Bennett[11,12], Paul Calleja[13], Keren Carss[1,2], Mark J. Caulfield[14,15], Patrick F. Chinnery[2,16,17], Peter H. Dixon[18], Daniel P. Gale[19,20], Roger James[1,2], Ania Koziell[21,22], Michael A. Laffan[23,24], Adam P. Levine[19], Eamonn R. Maher[25,26,27], Hugh S. Markus[28], Joannella Morales[29], Nicholas W. Morrell[2,5], Andrew D. Mumford[30,31], Elizabeth Ormondroyd[12,32], Stuart Rankin[13], Augusto Rendon[1,14], Sylvia Richardson[3], Irene Roberts[12,33,34], Noemi B. A. Roy[12,33,35], Moin A. Saleem[36,37], Kenneth G. C. Smith[5,8], Hannah Stark[2,38], Rhea Y. Y. Tan[28], Andreas C. Themistocleous[11], Adrian J. Thrasher[39], Hugh Watkins[32,35,40], Andrew R. Webster[41,42], Martin R. Wilkins[43], Catherine Williamson[18,44], James Whitworth[25,26,27], Sean Humphray[45], David R. Bentley[45], NIHR BioResource for the 100,000 Genomes Project*, Nathalie Kingston[1,2], Neil Walker[1,2], John R. Bradley[2,5,26,46,47], Sofie Ashford[2,38], Christopher J. Penkett[1,2], Kathleen Freson[7], Kathleen E. Stirrups[1,2], F. Lucy Raymond[2,25]✉ & Willem H. Ouwehand[1,2,4,6,48]✉

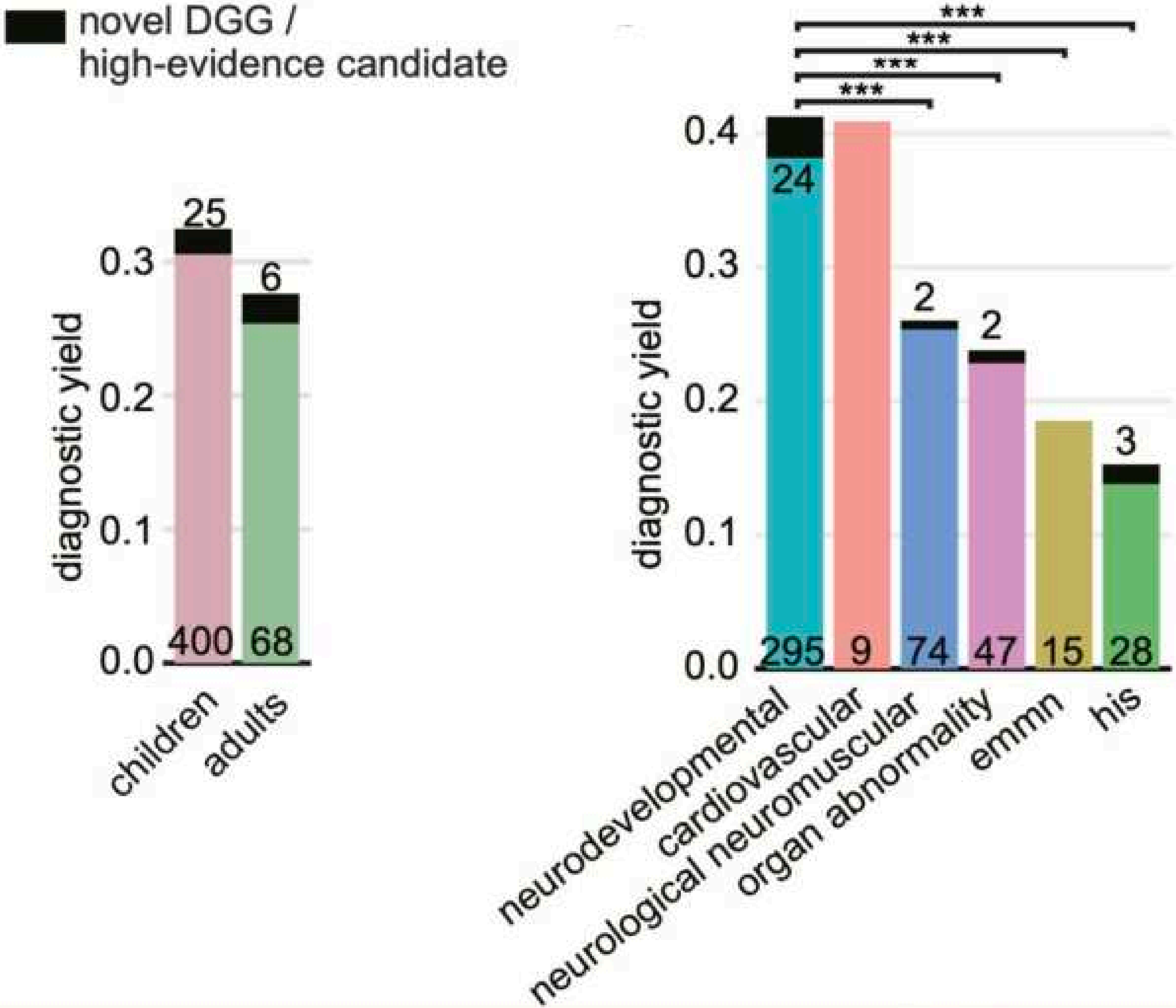
Title:

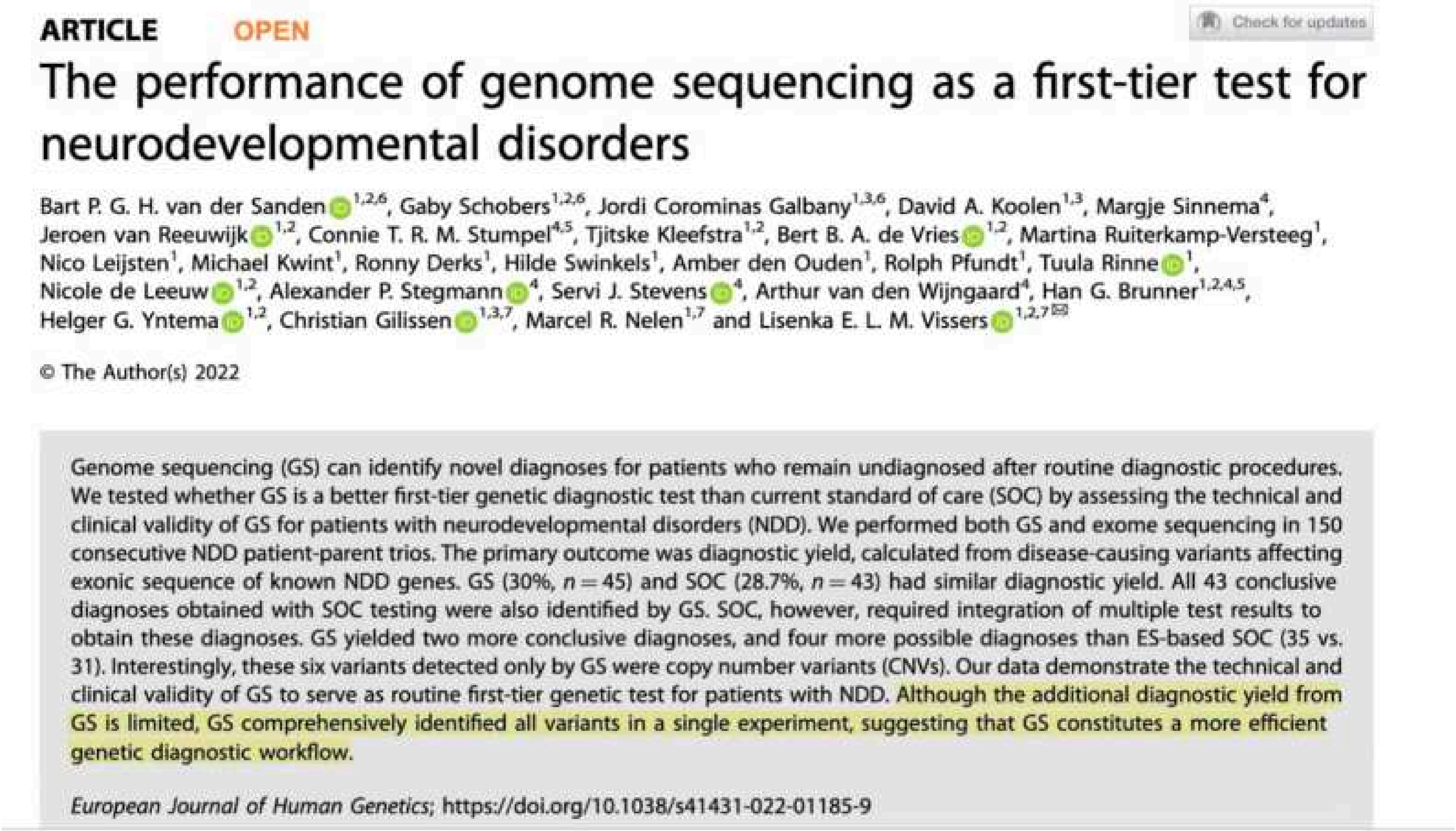
Next-generation phenotyping integrated in a national framework for patients with ultra-rare disorders improves genetic diagnostics and yields new molecular findings

Authors:

Axel Schmidt[1,*], Magdalena Danyel[2,3,*], Kathrin Grundmann[4,*], Theresa Brunel[5,*], Hannah Klinkhammer[6,7], Tzung-Chien Hsieh[6], Hartmut Engels[1], Sophia Peters[1], Alexej Knaus[6], Shahida Moosa[8], Luisa Averdunk[9], Felix Boschann[2,3], Henrike Sczakiel[2,3], Sarina Schwartzmann[2,3], Martin Atta Mensah[2], Jean Tori Pantel[2,10], Manuel Holtgrewe[11], Annemarie Bösch[12], Claudia Weiß[12], Natalie Weinhold[12], Aude-Annick Suter[12], Corinna Stoltenburg[12], Julia Neugebauer[12], Tillmann Kallinich[14], Angela M. Kaindl[13,14,15], Susanne Holzhauer[17], Christoph Bührer[17], Philip Bufler[17], Uwe Kornak[2], Claus-Eric Ott[2], Markus Schülke[2], Hoa Huu Phuc Nguyen[16], Sabine Hoffjen[16], Corinna Grasemann[17], Tobias Rothoeft[17], Folke Brinkmann[17], Nora Matar[17], Sugirthan Sivalingam[1], Claudia Perne[1], Elisabeth Mangold[1], Martina Kreiss[1], Kirsten Cremer[1], Regina C. Betz[1], Tim Bender[1], Martin Mücke[18], Lorenz Grigull[18], Thomas Klockgether[19], Spier Isabel[1], Heimbach André[1], Bender Tim[18], Fabian Brand[6], Christiane Stieber[10], Alexandra Marzena Morawiec[18], Pantelis Karakostas[20], Valentin S. Schäfer[20], Sarah Bernsen[19], Patrick Weydt[19], Sergio Castro-Gomez[19], Ahmad Aziz[19], Marcus Grobe-Einsler[19], Okka Kimmich[19], Xenia Kobeleva[19], Demet Önder[19], Hellen Lesmann[1], Sheetal Kumar[1], Pawel Tacik[19], Min Ae Lee-Kirsch[21,22], Reinhard Berner[21,22], Catharina Schuetz[21,22], Julia Körholz[21,22], Tanita Kretschmer[21,22], Nataliya Di Donato[21,23], Evelin Schröck[21,23], André Heinen[21,22], Ulrike Reuner[21,24], Amalia-Mihaela Hanßke[21], Frank J. Kaiser[25], Eva Manka[26], Martin Munteanu[25], Alma Kuechler[25], Kiewert Cordula[26], Raphael Hirtz[26], Elena Schlapakow[27], Christian Schlein[28], Jasmin Liefeld[28], Christian Kubisch[28,29], Theresia Herget[28], Maja Hempel[28,29,30], Christina Weiler-Normann[28,31], Kurt Ullrich[28], Christoph Schramm[31,29], Cornelia Rudolph[29], Franziska Rillig[29], Maximilian Groffmann[29], Ania Muntau[32], Alexandra Tibelius[32], Eva M. C. Schwaibold[30], Christian P. Schaaf[30], Michal Zawada[30], Lilian Kaufmann[30], Katrin Hinderhofer[30], Pamela M. Okun[33], Urania Kotzaeridou[33], Georg F. Hoffmann[33], Daniela Choukair[33], Markus Bettendorf[33], Malte Spielmann[34], Annekatrin Ripke[35], Martje Pauly[36,37], Alexander Münchau[35,36], Katja Lohmann[38], Irina Hüning[37], Britta Hanker[40], Tobias Bäumer[35,39], Rebecca Herzog[36,35], Yorck Hellenbroich[41], Dominik S. Westphal[5], Tim Strom[5], Reka Kovacs[42], Korbinian M. Riedhammer[5,43], Katharina Mayerhanser[5], Elisabeth Graf[5], Melanie Brugger[5], Julia Hoefele[42], Konrad Oexle[44], Nazanin Mirza-Schreiber[44], Riccardo Berutti[44], Ulrich Schatz[42], Martin Krenn[5,46], Christine Makowski[46], Heike Weigand[47], Sebastian Schröder[47], Meino Rohlfs[47], Vill Katharina[47], Fabian Hauck[47], Ingo Borggraefe[47], Wolfgang Müller-Felber[47], Ingo Kurth[10], Miriam Elbracht[10], Cordula Knopp[10], Matthias Begemann[10], Florian Kraft[10], Johannes R. Lemke[48,49], Julia Hentschel[48], Konrad Platzer[48], Vincent Strehlow[48], Rami Abou Jamra[48], Martin Kehrer[4], German Demidov[4], Stefanie Beck-Wödl[4], Holm Graessner[50], Marc Sturm[4], Lena Zeltner[50], Ludger J. Schöls[51], Janine Magg[52], Andrea Bevot[52], Christiane Kehrer[52], Nadja Kaiser[52], Denise Horn[2], Annette Grüters-Kieslich[53], Christoph Klein[47], Stefan Mundlos[7], Markus Nöthen[1], Olaf Riess[4], Thomas Meitinger[5], Heiko Krude[53], Peter M. Krawitz[6,*,#], Tobias Haack[4,*], Nadja Ehmke[2,3,*], Matias Wagner[5,46,47,*]


a
patient with a suspected rare disease with a likely genetic cause
multidisciplinary teams
clinical diagnosis
additional clinical evaluation required
other type of genetic testing recommended
exome sequencing
variant list
multidisciplinary teams
reevaluation of the results
(likely) pathogenic variant
variant of unknown significance / candidate gene / negative results
diagnosis
further investigation on a research basis


■ novel DGG / high-evidence candidate
diagnostic yield
0.0
0.1
0.2
0.3
0.4
25
6
400
68
children
adults
***
***
***
***
24
2
2
3
295
9
74
47
15
28
neurodevelopmental
cardiovascular
neurological neuromuscular
organ abnormality
emmn
his



ARTICLE OPEN

Check for updates

The performance of genome sequencing as a first-tier test for neurodevelopmental disorders

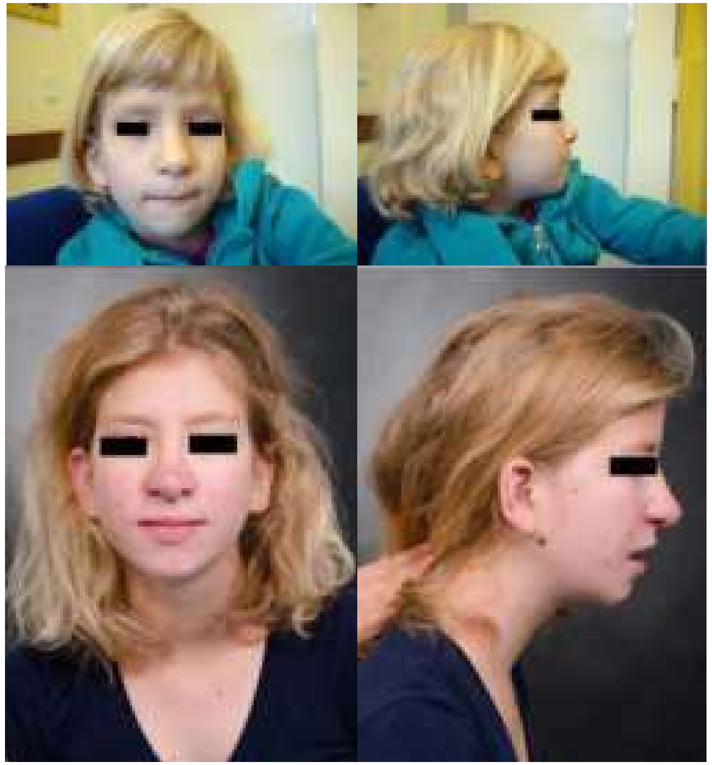
Bart P. G. H. van der Sanden[1,2,6], Gaby Schobers[1,2,6], Jordi Corominas Galbany[1,3,6], David A. Koolen[1,3], Margje Sinnema[4], Jeroen van Reeuwijk[1,2], Connie T. R. M. Stumpel[4,5], Tjitske Kleefstra[1,2], Bert B. A. de Vries[1,2], Martina Ruiterkamp-Versteeg[1], Nico Leijsten[1], Michael Kwint[1], Ronny Derks[1], Hilde Swinkels[1], Amber den Ouden[1], Rolph Pfundt[1], Tuula Rinne[1], Nicole de Leeuw[1,2], Alexander P. Stegmann[4], Servi J. Stevens[4], Arthur van den Wijngaard[4], Han G. Brunner[1,2,4,5], Helger G. Yntema[1,2], Christian Gilissen[1,3,7], Marcel R. Nelen[1,7] and Lisenka E. L. M. Vissers[1,2,7]✉

© The Author(s) 2022

Genome sequencing (GS) can identify novel diagnoses for patients who remain undiagnosed after routine diagnostic procedures. We tested whether GS is a better first-tier genetic diagnostic test than current standard of care (SOC) by assessing the technical and clinical validity of GS for patients with neurodevelopmental disorders (NDD). We performed both GS and exome sequencing in 150 consecutive NDD patient-parent trios. The primary outcome was diagnostic yield, calculated from disease-causing variants affecting exonic sequence of known NDD genes. GS (30%, $n = 45$) and SOC (28.7%, $n = 43$) had similar diagnostic yield. All 43 conclusive diagnoses obtained with SOC testing were also identified by GS. SOC, however, required integration of multiple test results to obtain these diagnoses. GS yielded two more conclusive diagnoses, and four more possible diagnoses than ES-based SOC (35 vs. 31). Interestingly, these six variants detected only by GS were copy number variants (CNVs). Our data demonstrate the technical and clinical validity of GS to serve as routine first-tier genetic test for patients with NDD. Although the additional diagnostic yield from GS is limited, GS comprehensively identified all variants in a single experiment, suggesting that GS constitutes a more efficient genetic diagnostic workflow.

European Journal of Human Genetics; https://doi.org/10.1038/s41431-022-01185-9

DeepGestalt triggered genome sequencing



Clinical features:
- Developmental delay / moderate ID (IQ 43)
- Muscular hypotonia
- Dysmorphic facial features (long face, slightly upslanting palpebral fissures, ptosis of the left eye, prominent bulbous nasal tip, low-hanging columella)

Perne et al.
P-ClinG-88

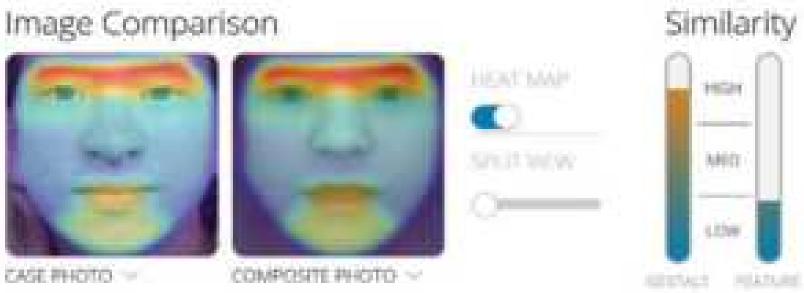
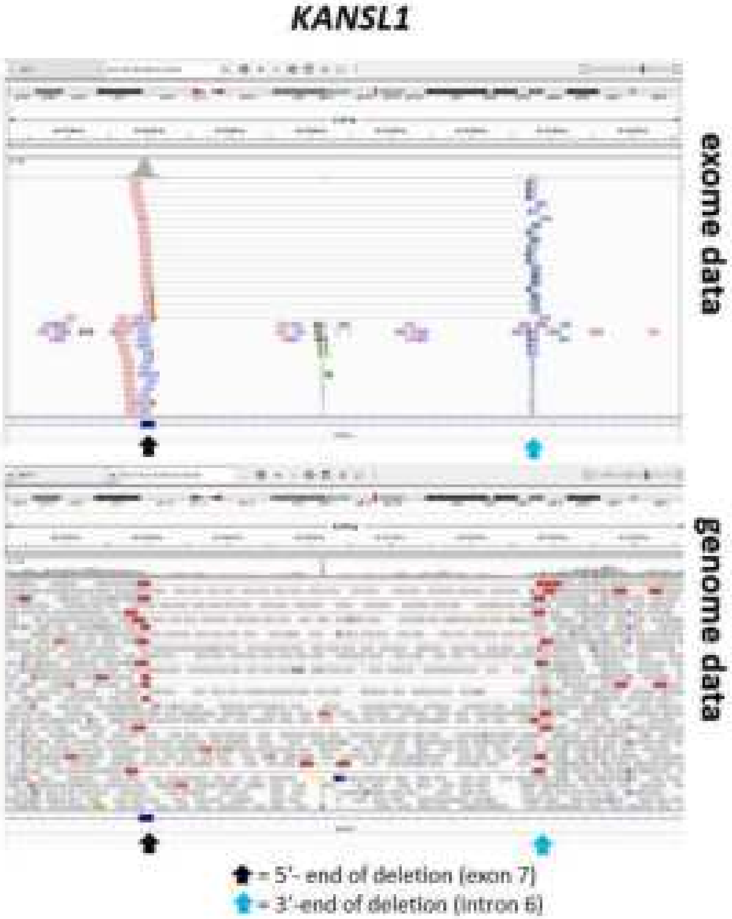


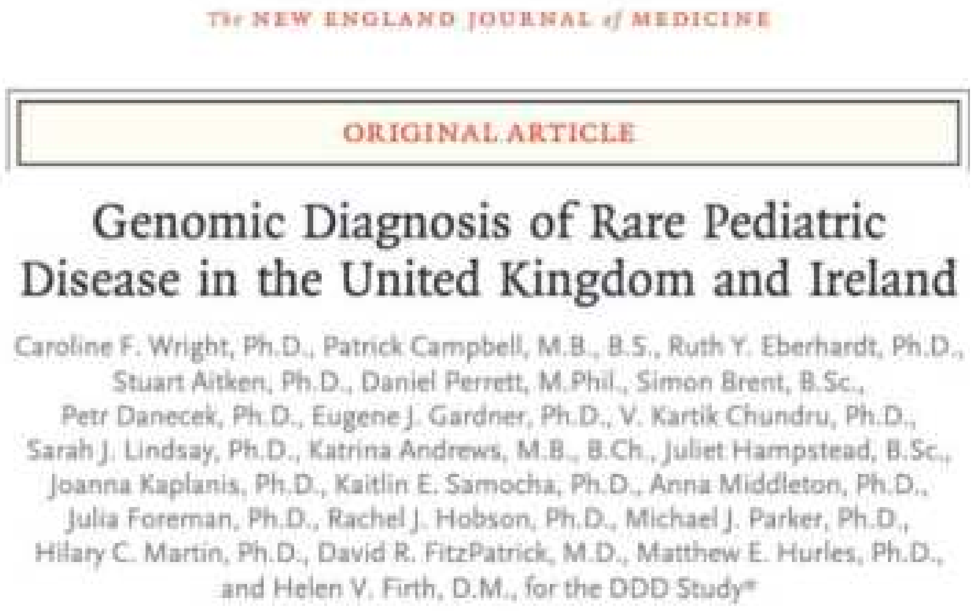
Image Comparison
Similarity
HEAT MAP
SPLIT VIEW
HIGH
MED
LOW
GESTALT
FEATURE
CASE PHOTO
COMPOSITE PHOTO


KANSL1
exome data
genome data
= 5'- end of deletion (exon 7)
= 3'-end of deletion (intron 6)

Computer-assisted image analysis led to identification of a intragenic deletion in *KANSL1* (*de novo*, ACMG class V) associated with **Koolen-de Vries syndrome**

TRANSLATE
NAMSE

2023

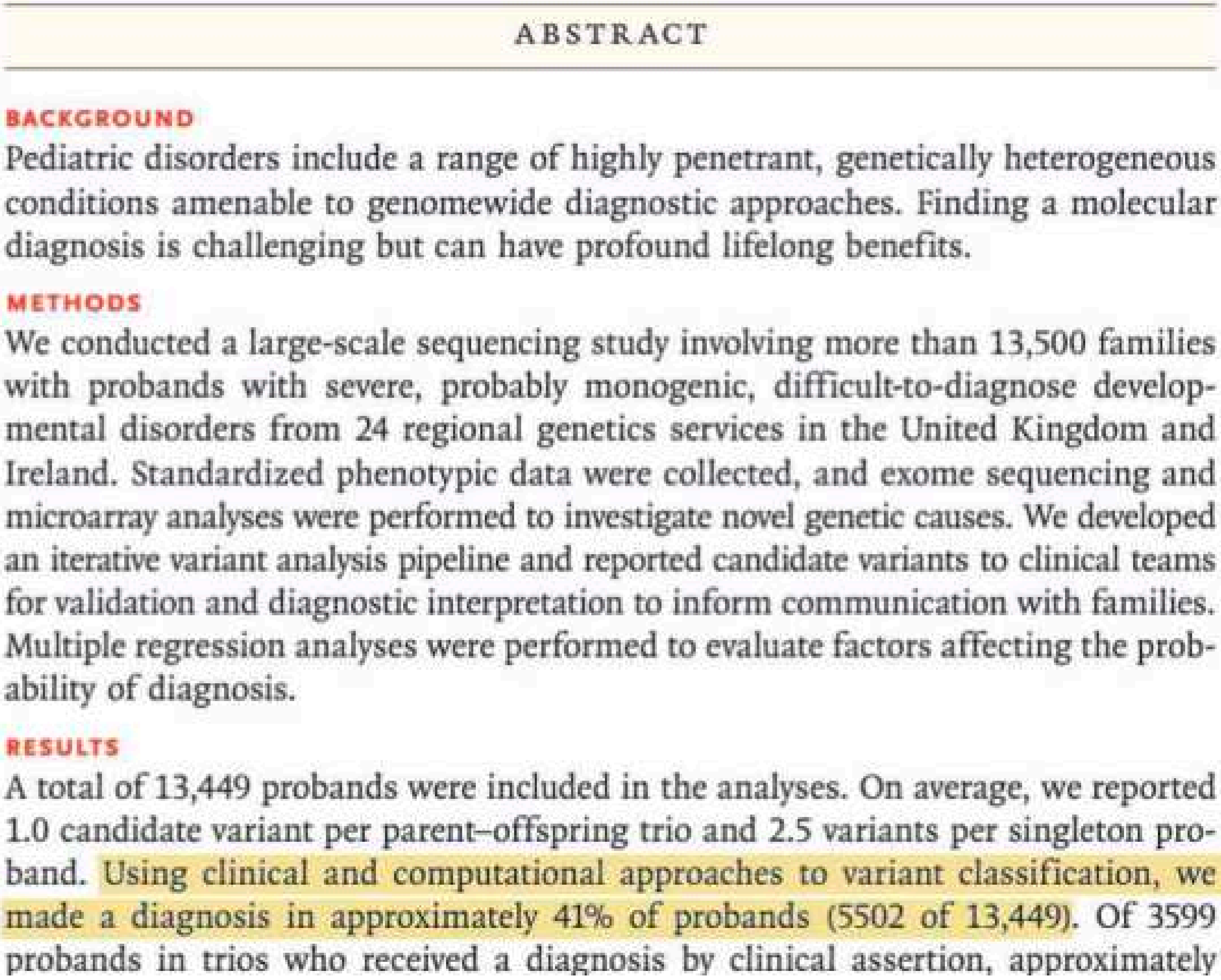


The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Genomic Diagnosis of Rare Pediatric Disease in the United Kingdom and Ireland

Caroline F. Wright, Ph.D., Patrick Campbell, M.B., B.S., Ruth Y. Eberhardt, Ph.D.,
Stuart Aitken, Ph.D., Daniel Perrett, M.Phil., Simon Brent, B.Sc.,
Petr Danecek, Ph.D., Eugene J. Gardner, Ph.D., V. Kartik Chundru, Ph.D.,
Sarah J. Lindsay, Ph.D., Katrina Andrews, M.B., B.Ch., Juliet Hampstead, B.Sc.,
Joanna Kaplanis, Ph.D., Kaitlin E. Samocha, Ph.D., Anna Middleton, Ph.D.,
Julia Foreman, Ph.D., Rachel J. Hobson, Ph.D., Michael J. Parker, Ph.D.,
Hilary C. Martin, Ph.D., David R. FitzPatrick, M.D., Matthew E. Hurles, Ph.D.,
and Helen V. Firth, D.M., for the DDD Study*

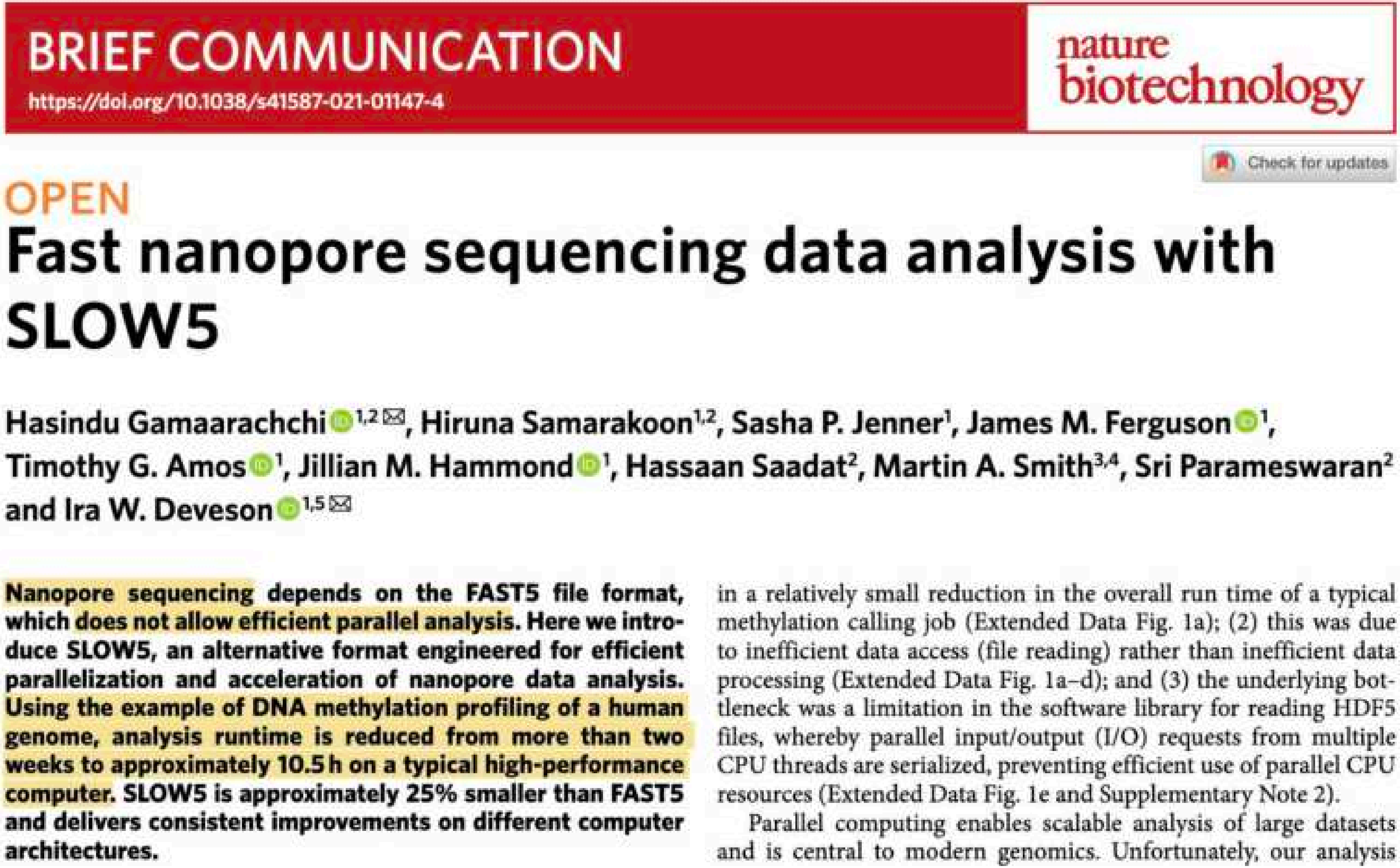


ABSTRACT

BACKGROUND
Pediatric disorders include a range of highly penetrant, genetically heterogeneous conditions amenable to genomewide diagnostic approaches. Finding a molecular diagnosis is challenging but can have profound lifelong benefits.

METHODS
We conducted a large-scale sequencing study involving more than 13,500 families with probands with severe, probably monogenic, difficult-to-diagnose developmental disorders from 24 regional genetics services in the United Kingdom and Ireland. Standardized phenotypic data were collected, and exome sequencing and microarray analyses were performed to investigate novel genetic causes. We developed an iterative variant analysis pipeline and reported candidate variants to clinical teams for validation and diagnostic interpretation to inform communication with families. Multiple regression analyses were performed to evaluate factors affecting the probability of diagnosis.

RESULTS
A total of 13,449 probands were included in the analyses. On average, we reported 1.0 candidate variant per parent–offspring trio and 2.5 variants per singleton proband. Using clinical and computational approaches to variant classification, we made a diagnosis in approximately 41% of probands (5502 of 13,449). Of 3599 probands in trios who received a diagnosis by clinical assertion, approximately



BRIEF COMMUNICATION
https://doi.org/10.1038/s41587-021-01147-4

nature biotechnology

Check for updates

OPEN

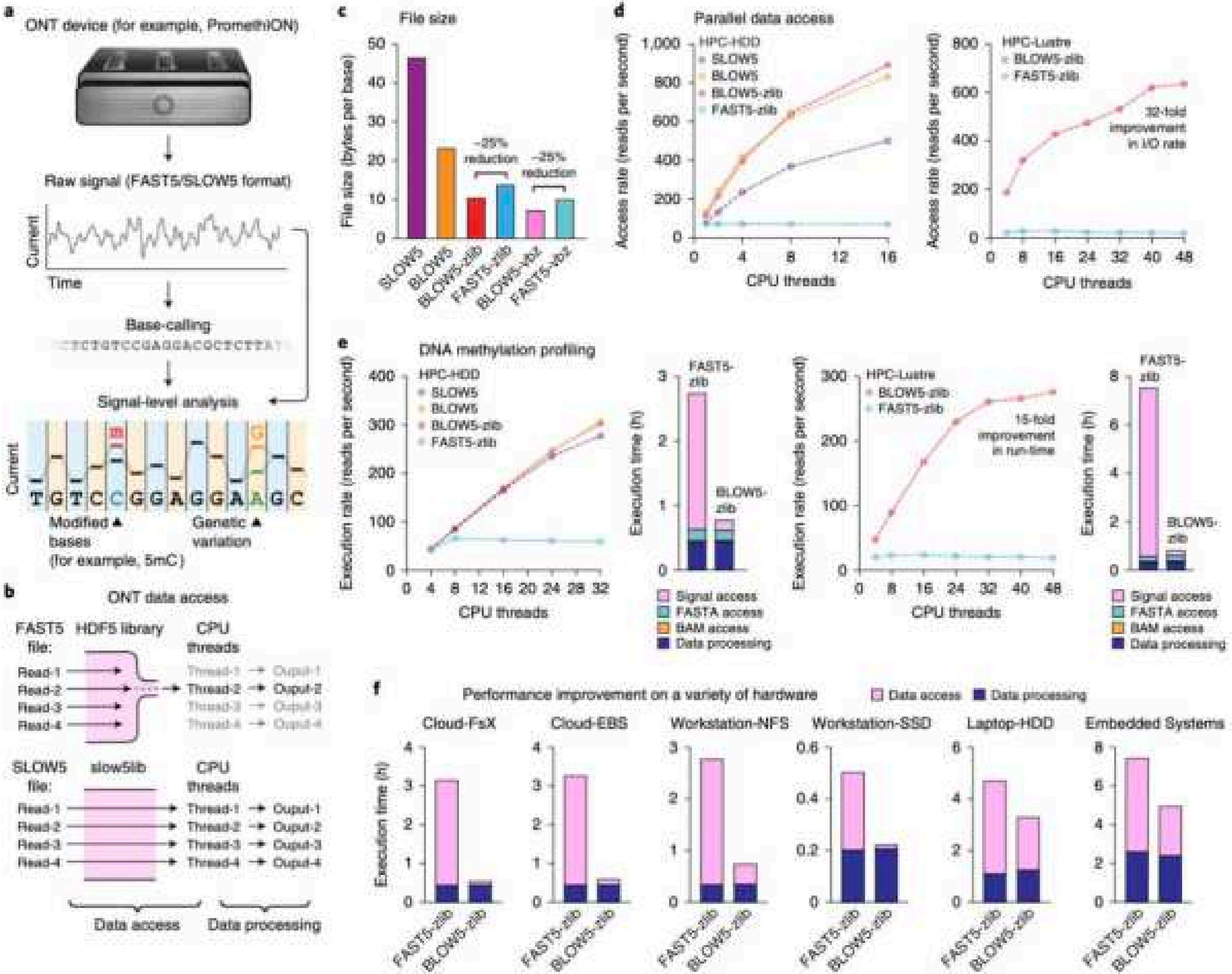
Fast nanopore sequencing data analysis with SLOW5

Hasindu Gamaarachchi[1,2]✉, Hiruna Samarakoon[1,2], Sasha P. Jenner[1], James M. Ferguson[1], Timothy G. Amos[1], Jillian M. Hammond[1], Hassaan Saadat[2], Martin A. Smith[3,4], Sri Parameswaran[2] and Ira W. Deveson[1,5]✉

Nanopore sequencing depends on the FAST5 file format, which does not allow efficient parallel analysis. Here we introduce SLOW5, an alternative format engineered for efficient parallelization and acceleration of nanopore data analysis. Using the example of DNA methylation profiling of a human genome, analysis runtime is reduced from more than two weeks to approximately 10.5 h on a typical high-performance computer. SLOW5 is approximately 25% smaller than FAST5 and delivers consistent improvements on different computer architectures.

in a relatively small reduction in the overall run time of a typical methylation calling job (Extended Data Fig. 1a); (2) this was due to inefficient data access (file reading) rather than inefficient data processing (Extended Data Fig. 1a–d); and (3) the underlying bottleneck was a limitation in the software library for reading HDF5 files, whereby parallel input/output (I/O) requests from multiple CPU threads are serialized, preventing efficient use of parallel CPU resources (Extended Data Fig. 1e and Supplementary Note 2).

Parallel computing enables scalable analysis of large datasets and is central to modern genomics. Unfortunately, our analysis


a
ONT device (for example, PromethION)
Raw signal (FAST5/SLOW5 format)
Current
Time
Base-calling
Signal-level analysis
m
G
T G T C C G G A G G A A G C
Modified bases (for example, 5mC)
Genetic variation
b
ONT data access
FAST5 file:
HDF5 library
CPU threads
Read-1
Read-2
Read-3
Read-4
Thread-1 → Ouput-1
Thread-2 → Ouput-2
Thread-3 → Ouput-3
Thread-4 → Ouput-4
SLOW5 file:
slow5lib
CPU threads
Read-1
Read-2
Read-3
Read-4
Thread-1 → Ouput-1
Thread-2 → Ouput-2
Thread-3 → Ouput-3
Thread-4 → Ouput-4
Data access
Data processing
c
File size
File size (bytes per base)
0
10
20
30
40
50
~25% reduction
~25% reduction
SLOW5
BLOW5
BLOW5-zlib
FAST5-zlib
BLOW5-vbz
FAST5-vbz
d
Parallel data access
HPC-HDD
SLOW5
BLOW5
BLOW5-zlib
FAST5-zlib
Access rate (reads per second)
0
200
400
600
800
1,000
0
4
8
12
16
CPU threads
HPC-Lustre
BLOW5-zlib
FAST5-zlib
32-fold improvement in I/O rate
Access rate (reads per second)
0
200
400
600
800
0
8
16
24
32
40
48
CPU threads
e
DNA methylation profiling
HPC-HDD
SLOW5
BLOW5
BLOW5-zlib
FAST5-zlib
Execution rate (reads per second)
0
100
200
300
400
0
4
8
12
16
20
24
28
32
CPU threads
FAST5-zlib
BLOW5-zlib
Execution time (h)
0
1
2
3
Signal access
FASTA access
BAM access
Data processing
HPC-Lustre
BLOW5-zlib
FAST5-zlib
15-fold improvement in run-time
Execution rate (reads per second)
0
100
200
300
0
8
16
24
32
40
48
CPU threads
FAST5-zlib
BLOW5-zlib
Execution time (h)
0
2
4
6
8
Signal access
FASTA access
BAM access
Data processing
f
Performance improvement on a variety of hardware
Data access
Data processing
Cloud-FsX
Cloud-EBS
Workstation-NFS
Workstation-SSD
Laptop-HDD
Embedded Systems
Execution time (h)
0
1
2
3
4
0
1
2
3
4
0
1
2
3
0
0.2
0.4
0.6
0
2
4
6
0
2
4
6
8
FAST5-zlib
BLOW5-zlib

Wie viel Speicherplatz und Rechenleistung ?

Die ehrlichere Antwort:

Wir haben immer noch keine komplette Genomsequenz und werden nie eine haben.

Mit long read geht vielleicht von Telomer zu Telomer (T2T), aber danach werden die Basenmodifikationen interessant (epigenetic signatures)

Für Nanopore Daten gibt sind wir erst am Anfang in der Hard/Software Optimierung

⇒ Das ist der Grund für Genomrechenzentren*

*) Finanzierung mindestens mit 20% der für das GS vergüteten Summe

Wie viel Speicherplatz und Rechenleistung?

Die pragmatische Antwort:

Für Illumina short read 30x WGS
- 100 GB
- Bis zu 1 MWh (life cycle costs mit Kühlung über 10 Jahre)

Für long read and everything more advanced
- > 1 TB
- > 5 EUR/h (Vergleich GPU A100 in Azure)