The Journal of Molecular Diagnostics, Vol. 24, No. 3, March 2022

ELSEVIER

the Journal of Molecular Diagnostics
jmdjournal.org

Failure to Detect Mutations in *U2AF1* due to Changes in the GRCh38 Reference Sequence

Christopher A. Miller,*† Jason R. Walker,‡ Travis L. Jensen,§ William F. Hooper,§ Robert S. Fulton,‡ Jeffrey S. Painter,¶ Mikkael A. Sekeres,‖ Timothy J. Ley,*† David H. Spencer,*†** Johannes B. Goll,§ and Matthew J. Walter*†

From the Division of Oncology, Department of Internal Medicine, the Siteman Cancer Center,† the McDonnell Genome Institute,‡ and the Department of Pathology and Immunology,** Washington University School of Medicine, St. Louis, Missouri; The Emmes Company,§ Rockville, Maryland; the Moffitt Cancer Center,¶ Tampa, Florida; and the Division of Hematology,‖ Department of Medicine, Sylvester Comprehensive Cancer Center, University of Miami School of Medicine, Miami, Florida*

Accepted for publication October 28, 2021.

Address correspondence to Christopher A. Miller, Ph.D., Division of Oncology, Washington University School of Medicine, 660 S. Euclid Ave, Campus Box 8007, St. Louis, MO 63110. E-mail: c.a.miller@wustl.edu.

The *U2AF1* gene is a core part of mRNA splicing machinery and frequently contains somatic mutations that contribute to oncogenesis in myelodysplastic syndrome, acute myeloid leukemia, and other cancers. A change introduced in the GRCh38 version of the human reference build prevents detection of mutations in this gene, and others, by variant calling pipelines. This study describes the problem in detail and shows that a modified GRCh38 reference build with unchanged coordinates can be used to ameliorate the issue. *(J Mol Diagn 2022, 24: 219—223; https://doi.org/10.1016/j.jmoldx.2021.10.013)*

The *U2AF1* gene encodes a core component of the mRNA splicing machinery that is frequently mutated in myelodysplastic syndrome (MDS) and other cancers.[1—3] Though predominantly associated with hematopoietic cancers (73%), mutations are also recurrent in lung tumors (6.5%) and have been reported in 24 other tumor types. Specifically, mutations at residues S34 and Q157 have been shown to promote exon skipping and are confirmed driver mutations contributing to cancer pathogenesis.[4—7]

Materials and Methods

Genomic data were collected as part of the MDS National History Study or The Cancer Genome Atlas (TCGA) project and appropriate consent was received under those protocols.[8,9] Sequencing reads from the MDS cohort were aligned to both masked and unmasked GRCh38 reference genomes using the BWA-MEM software package version 0.7.15 (*https://github.com/lh3/bwa/releases/tag/v0.7.15*),[10] followed by sorting and deduplication, as detailed in a CWL (Common Workflow Language) workflow archived at *https://git.io/JYbGl* (last accessed December 1, 2021). Variants in the MDS cohort were called in single-sample mode using VarScan software version 2.4.2 with params "–min-coverage 20 –min-reads2 5 –min-var-freq 0.05 –strand-filter 1".[11] Data from TCGA acute myeloid leukemia samples were aligned using the same process, and somatic variants were called using an ensemble approach, described in detail in the CWL workflow at *https://git.io/JYbGM* (GitHub, last accessed December 1, 2021). The modified genome FASTA file used for these analyses is available at *https://zenodo.org/record/4684553* (Zenodo, last accessed December 1, 2021).

Sequence data from the MDS cohort are available in the database of Genotypes and Phenotypes (dbGaP) under accession id phs002714.v1.p1. The study is currently

Supported by National Cancer Institute (NCI) Research Specialist Award R50 CA211782 (C.A.M.), Genomics of Acute Myeloid Leukemia Program Project grant P01 CA101937 (T.J.L.), and the Edward P. Evans Foundation (M.J.W.). The National MDS Natural History Study has been supported by US Federal government contracts HHSN268201400003I and HHSN268201400002I from the National Heart, Lung, and Blood Institute and additional funding by the NCI to the participating member clinical centers in the NCORP and NCTN. This work has been supported in part by the Tissue Core Facility at Moffitt Cancer, an NCI designated Comprehensive Cancer Center (P30-CA076292).

Disclosures: None declared.

https://doi.org/10.1016/j.jmoldx.2021.10.013

available at *https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002714.v1.p1* (dbGaP, last accessed January 23, 2022). TCGA acute myeloid leukemia data are available via the Genomic Data Commons at *https://portal.gdc.cancer.gov* (last accessed December 1, 2021). Pipelines used for variant calling are available at *https://github.com/genome/analysis-workflows* (GitHub, last accessed December 1, 2021). All links in the *Materials and Methods* are to the specific versions of the workflows used.

Results

As part of the National Myelodysplastic Syndrome (MDS) Natural History Study (ClincialTrials.gov, *https://www.clinicaltrials.gov*, identifier: NCT02775383), a targeted gene panel was used to sequence bone marrow samples from 120 patients either diagnosed with MDS or suspected to have MDS.[8] Of these patients, 38 were eventually confirmed to have MDS or a myeloproliferative neoplasm. Initial analyses looking at sequencing quality metrics revealed coverage levels and mutation frequencies that closely matched expectations, with one exception: mutations in the *U2AF1* spliceosome gene are typically observed in nearly 10% of MDS or myeloproliferative neoplasm patients, but only 2 of the 38 MDS or myeloproliferative neoplasm patients (5.2%) in this group had such mutations, both at the Q157 hotspot.[1,12,13] Although this deviation was not significant (compared with the Walter et al[12] cohort; $P = 0.53$ via Fisher's exact test), both mutations had only about 15% of the expected sequence coverage (mean of 204× depth), whereas the full targeted panel had a median coverage of 1337.3× (Supplemental Table S1). Nearly the entirety of the *U2AF1* gene was likewise affected, with a median depth of only 130× and four exons with no coverage at all (Figure 1). Manual inspection of the alignments with Integrated Genomics Viewer (IGV) revealed that this poor coverage extended to a 150-kb region of the genome where the majority of reads had a mapping quality of zero (Figure 1).[14] The sequence aligner BWA-MEM reserves this mapping quality zero score for reads that cannot be uniquely placed in the genome, and it generally indicates a highly repetitive sequence or problems with the underlying assemblies.[10] The region with such problematic alignments spans the *CBS*, *U2AF1*, *FRGCA*, and *CRYAA* genes.

Further investigation revealed that in the GRCh38 reference build, content added to the p-arm of chromosome 21 (chr21:6427259-6580,181) contained sequence that replicated the sequence of the *U2AF1* locus (chr21:43035875-43187577) with 99.0% identity. The same issue does not exist in prior reference builds GRCh36 or GRCh37. After consultation with members of the Genome Reference Consortium (GRC), it was determined that a bacterial artificial chromosome (BAC) clone (Nucleotide, *https://www.ncbi.nlm.nih.gov/nuccore*; accession number FP236240.8) was incorrectly added to the reference genome, creating this duplicate sequence. This resulted in the alignment algorithm, BWA-MEM, splitting the reads among these two loci, thus lowering the overall coverage substantially (Figure 1, Supplemental Figure S1, and Supplemental Table S1). In addition, reads with mapping quality scores of zero are typically excluded or down-weighted during variant calling, due to the increased chance of artefactual calls, and these factors combined explained the paucity of mutations observed in *U2AF1*, especially at the S34F position.

To address this issue, the authors' group created a modified version of GRCh38 that maintains the coordinate system, but masks the new duplicate sequence on chromosome 21p by replacing it with "N" characters. The authors realigned the data to this reference and observed a substantial increase in coverage and mapping qualities across the affected region. Over the exons of *U2AF1*, the coverage of reads with mapping quality >0 rose from a median of 0.3× to a median of 1195×. This enabled the discovery of an additional *U2AF1* mutation (S34F) in this MDS cohort. The data were also aligned to GRCh37, and it was confirmed that this region was not problematic in the older reference, where the median coverage over the exons of *U2AF1* was 1381× (Supplemental Table S1 and Supplemental Figure S2).

To validate this finding in an orthogonal data set, data were retrieved from acute myeloid leukemia patients sequenced for the TCGA paper.[9] The data in this study were originally aligned to genome build GRCh36, and six mutations in the *U2AF1* gene were reported from exome sequencing. After this paper was published, the Genomic Data Commons took this sequence dataset, realigned it (with BWA-MEM) to the stock GRCh38.d1.vd1 reference, and produced variant calls using four different algorithms.[10,15] These variant files [in MAF (Mutation Annotation Format)] reported no *U2AF1* mutations in the six expected samples. Their GRCh38 sequence alignments were then downloaded, and the same coverage and mapping quality issues on chromosome 21 were observed. One sample lacked any *U2AF1* reads at all, due to unrelated sequencing problems (the TCGA consortium only identified the mutation using orthogonal assays). This sample was excluded from further analyses, leaving five evaluable samples.

Running another somatic variant calling pipeline on these alignments also did not reveal any *U2AF1* mutations. However, after realigning the sequence data to the masked version of GRCh38, the same somatic pipeline identified all five expected mutations in *U2AF1* (Supplemental Table S2).

Discussion

The reference genome is essential to modern cancer genomics, but problems with these assemblies have the potential to cause both false positive and negative results. In this

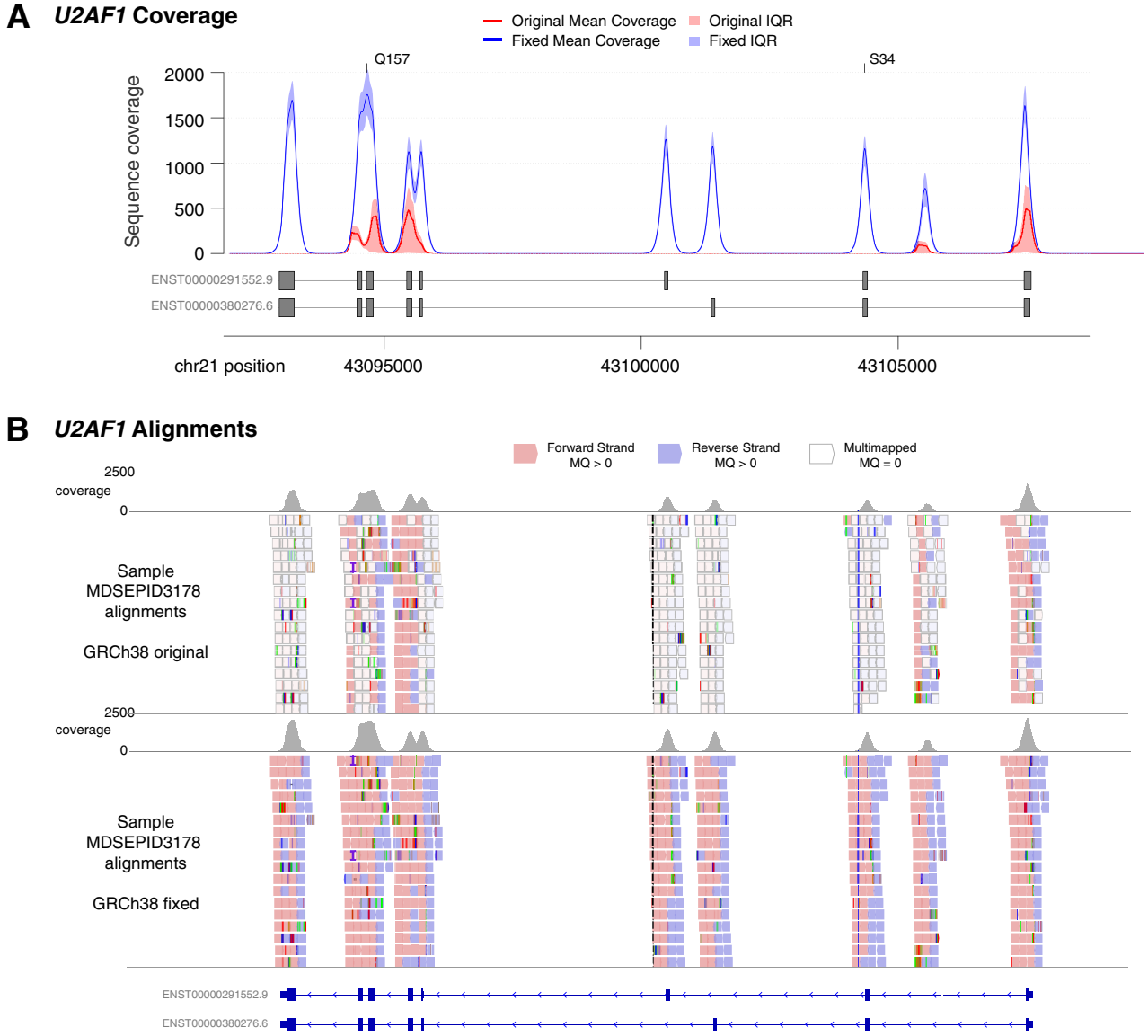

Figure 1 Alignment issues across the U2AF1 locus. **A:** Sequence coverage of reads with mapping quality (MQ) >0 across the *U2AF1* gene for 120 bone marrow samples sequenced after capture with a custom reagent. The mean coverage for realignments to GRCh38 are shown in red, whereas the mean coverage for alignments to the authors' modified GRCh38 reference is shown in blue. Shading indicates the interquartile range (IQR) for each. Exons from the two primary protein-coding isoforms are shown below, and the locations of hotspot mutations at amino acids S34 and Q157 are indicated. **B:** An Integrated Genomics Viewer (IGV) view of sequence reads, with alignments to GRCh38 at top and alignments to the modified reference at bottom. Grey bars at top show overall coverage. Reads in white indicate multimapped reads, with mapping qualities of zero, whereas red and blue reads have higher quality alignments.

study, the authors describe the latter, where changes introduced in the GRCh38 human reference build cause mutations in a cancer driver gene to be missed with standard analysis approaches. GRCh38 was first released in late 2013, but widespread adoption lagged somewhat, so some studies involving myeloid malignancies (TCGA, the Beat AML trial) have been spared this issue because they used older versions of the reference.[9,16] Nonetheless, large cancer databases, including the National Cancer Institute's Genomic Data Commons, are likely missing many *U2AF1* mutations due to this artefact. This also has clinical implications because *U2AF1* mutations have strong associations with prognosis, and clinical trials of splice-modulating drugs are being planned or are underway.[5,17,18]

These findings have been reported to the GRC (*https://www.ncbi.nlm.nih.gov/grc/human/issues/HG-2544*, last accessed December 10, 2021), but as there are no patches or new releases scheduled for the human reference genome, the problem remains unresolved in the current release of GRCh38 (GRCh38.p13). However, in the time since these analyses were performed, the GRC has released a masking file that includes this chr21 region, along with two other contaminating sequences on alternate contigs.[19] Using this file to create a masked genome

mirrors the approach that the authors' analyses used and likewise resolves the issues in *U2AF1* reported in this paper.

To apply this fix, the bed file can be downloaded from NCBI at *https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_GRC_exclusions.bed* (last accessed December 1, 2021), then applied to a GRCh38 FASTA file using the bedtools maskfasta tool, followed by reindexing with the aligner of choice.[20]

Although masking the genome offers better quality alignments at this locus, the next leap forward will come with a new reference sequence, likely based on the draft genome recently produced by the Telomere-to-Telomere consortium.[21] In the longer term, the genomics community can look forward to graph genome approaches capable of representing haplotypes from many different populations. These should further increase the accuracy of short-read genomic alignments, upon which many analyses, and increasingly clinical decisions, are based.

As these genome reference improvements are released, the genomics community will need to validate them before they can be used on clinical cases. Previous data sets will need to be realigned to ensure that changes are understood and problematic portions of assemblies that might alter diagnostic results are identified. Genomic annotations and pipelines will also need to be updated, which can be resource intensive. Hence, GRCh38 will likely remain in use for years. The use of GRC-released bed file is recommended to create a masked reference that is coordinate-compatible to be used interchangeably with the standard GRCh38 reference in cancer genomics applications. Especially, applications where detection of *U2AF1* mutations is critical, including sequencing of hematological cancers or studies of spliceosome dysfunction.

Acknowledgments

We thank Nancy DiFronzo for leadership of the National MDS Natural History Study. The National MDS Natural History Study thanks the study participants, as well as the investigator teams at the participating clinical sites.

Author Contributions

C.A.M., J.R.W., T.L.J., W.F.H., R.S.F., D.H.S., J.B.G., and M.J.W. conceptualized the study; J.S.P. led sample acquisition; C.A.M., J.R.W., T.L.J., and W.F.H. performed data analysis; C.A.M., J.R.W., and T.L.J. visualized the data; M.A.S., T.J.L., J.B.G., and M.J.W. provided supervision and funding; C.A.M., J.B.G., T.L.J., and M.J.W. wrote and edited the manuscript. All authors read and approved the final manuscript. C.A.M. is the guarantor of this work and, as such, had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Supplemental Data

Supplemental material for this article can be found at *http://doi.org/10.1016/j.jmoldx.2021.10.013*.

References

1. Graubert TA, Shen D, Ding L, Okeyo-Owuor T, Lunn CL, Shao J, Krysiak K, Harris CC, Koboldt DC, Larson DE, McLellan MD, Dooling DJ, Abbott RM, Fulton RS, Schmidt H, Kalicki-Veizer J, O'Laughlin M, Grillot M, Baty J, Heath S, Frater JL, Nasim T, Link DC, Tomasson MH, Westervelt P, DiPersio JF, Mardis ER, Ley TJ, Wilson RK, Walter MJ: Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. Nat Genet 2011, 44: 53–57
2. Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, Sato Y, Sato-Otsubo A, Kon A, Nagasaki M, Chalkidis G, Suzuki Y, Shiosaka M, Kawahata R, Yamaguchi T, Otsu M, Obara N, Sakata-Yanagimoto M, Ishiyama K, Mori H, Nolte F, Hofmann W-K, Miyawaki S, Sugano S, Haferlach C, Koeffler HP, Shih L-Y, Haferlach T, Chiba S, Nakauchi H, Miyano S, Ogawa S: Frequent pathway mutations of splicing machinery in myelodysplasia. Nature 2011, 478:64–69
3. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA: The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer 2018, 18:696–705
4. Okeyo-Owuor T, White BS, Chatrikhi R, Mohan DR, Kim S, Griffith M, Ding L, Ketkar-Kulkarni S, Hundal J, Laird KM, Kielkopf CL, Ley TJ, Walter MJ, Graubert TA: U2AF1 mutations alter sequence specificity of pre-mRNA binding and splicing. Leukemia 2015, 29:909–917
5. Shirai CL, Ley JN, White BS, Kim S, Tibbitts J, Shao J, Ndonwi M, Wadugu B, Duncavage EJ, Okeyo-Owuor T, Liu T, Griffith M, McGrath S, Magrini V, Fulton RS, Fronick C, O'Laughlin M, Graubert TA, Walter MJ: Mutant U2AF1 expression alters hematopoiesis and pre-mRNA splicing in vivo. Cancer Cell 2015, 27: 631–643
6. Ilagan JO, Ramakrishnan A, Hayes B, Murphy ME, Zebari AS, Bradley P, Bradley RK: U2AF1 mutations alter splice site recognition in hematological malignancies. Genome Res 2015, 25: 14–26
7. Fei DL, Zhen T, Durham B, Ferrarone J, Zhang T, Garrett L, Yoshimi A, Abdel-Wahab O, Bradley RK, Liu P, Varmus H: Impaired hematopoiesis and leukemia development in mice with a conditional knock-in allele of a mutant splicing factor gene U2af1. Proc Natl Acad Sci U S A 2018, 115:E10437–E10446
8. Sekeres MA, Gore SD, Stablein DM, DiFronzo N, Abel GA, DeZern AE, Troy JD, Rollison DE, Thomas JW, Waclawiw MA, Liu JJ, Al Baghdadi T, Walter MJ, Bejar R, Gorak EJ, Starczynowski DT, Foran JM, Cerhan JR, Moscinski LC, Komrokji RS, Deeg HJ, Epling-Burnette PK: The National MDS Natural History Study: design of an integrated data and sample biorepository to promote research studies in myelodysplastic syndromes. Leuk Lymphoma 2019, 60:3161–3171
9. Cancer Genome Atlas Research Network, Ley TJ, Miller C, Ding L, Raphael BJ, Mungall AJ, Robertson AG, et al: Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med 2013, 368:2059–2074
10. Li H: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-bioGN] 2013, [Preprint], arXiv:13033997

11. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK: VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 2012, 22:568−576
12. Walter MJ, Shen D, Shao J, Ding L, White BS, Kandoth C, Miller CA, Niu B, McLellan MD, Dees ND, Fulton R, Elliot K, Heath S, Grillot M, Westervelt P, Link DC, DiPersio JF, Mardis E, Ley TJ, Wilson RK, Graubert TA: Clonal diversity of recurrently mutated genes in myelodysplastic syndromes. Leukemia 2013, 27: 1275−1282
13. Papaemmanuil E, Gerstung M, Malcovati L, Tauro S, Gundem G, Van Loo P, et al: Clinical and biological implications of driver mutations in myelodysplastic syndromes. Blood 2013, 122:3616−3627. quiz 3699
14. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: Integrative genomics viewer. Nat Biotechnol 2011, 29:24−26
15. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM: Toward a shared vision for cancer genomic data. N Engl J Med 2016, 375:1109−1112
16. Tyner JW, Tognon CE, Bottomly D, Wilmot B, Kurtz SE, Savage SL, et al: Functional genomic landscape of acute myeloid leukaemia. Nature 2018, 562:526−531
17. Tefferi A, Finke CM, Lasho TL, Hanson CA, Ketterling RP, Gangat N, Pardanani A: U2AF1 mutation types in primary myelofibrosis: phenotypic and prognostic distinctions. Leukemia 2018, 32:2274−2278
18. Seiler M, Yoshimi A, Darman R, Chan B, Keaney G, Thomas M, et al: H3B-8800, an orally available small-molecule splicing modulator, induces lethality in spliceosome-mutant cancers. Nat Med 2018, 24: 497−504
19. Wagner J, Olson ND, Harris L, McDaniel J, Cheng H, Fungtammasan A, et al: Towards a comprehensive variation benchmark for challenging medically-relevant autosomal genes. bioRxiv 2021, [Preprint]. doi:10.1101/2021.06.07.444885
20. Quinlan AR, Hall IM: BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010, 26:841−842
21. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al: The Complete Sequence of a Human Genome. bioRxiv 2021, [Preprint]. doi:10.1101/2021.05.26.445798