

Sep 29, 2025

Ultrafast WGS protocol

DOI

<https://dx.doi.org/10.17504/protocols.io.j8nlkyn96g5r/v1>

Aditi Vedi^{1,2}, Sean Humphray², Martina Mijuskovic³, Joao Dias²

¹1. Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK;

²University of Cambridge, Cambridge, UK; ³Illumina Inc., Cambridge, UK

Aditi Vedi's protocols



Aditi Vedi

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.j8nlkyn96g5r/v1>

Protocol Citation: Aditi Vedi, Sean Humphray, Martina Mijuskovic, Joao Dias 2025. Ultrafast WGS protocol. **protocols.io**
<https://dx.doi.org/10.17504/protocols.io.j8nlkyn96g5r/v1>

License: This is an open access protocol distributed under the terms of the **[Creative Commons Attribution License](#)**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

We use this protocol and it's working

Created: September 26, 2025

Last Modified: September 29, 2025

Protocol Integer ID: 228301

Keywords: comprehensive molecular profiling in paediatric haematology, test approach in paediatric oncology, paediatric oncology, paediatric haematology, additional technical advantages in variant detection, variant detection, feasible across diverse tumour type, diverse tumour type, support for timely precision therapy, timely precision therapy, ultrafast wgs protocol ultra, oncology, comprehensive molecular profiling, confirmed cancer, targeted treatment

Funders Acknowledgements:

Rosetrees Trust

NIHR Cambridge Biomedical Research

Grant ID: NIHR203312

Addenbrookes Charitable Trust

Isaac Newton Trust

Abstract

Ultra-Fast Whole Genome Sequencing (UF-WGS) offers rapid, comprehensive molecular profiling in paediatric haematology-oncology, enabling critical improvements in diagnosis and patient management. In a prospective study of 54 children with suspected or confirmed cancer, UF-WGS reduced turnaround time from 37 to 3 days versus standard NHS practice and captured 95% of all clinically actionable variants found by conventional methods. UF-WGS also identified 19 additional actionable variants, leading to demonstrable improvements in care for 51% of prospective patients, including avoidance of over-medicalisation and support for timely precision therapy. UF-WGS was feasible across diverse tumour types and sample sources, with additional technical advantages in variant detection and workflow simplicity. These results support UF-WGS as a transformative single-test approach in paediatric oncology, facilitating safer, more rapid allocation of targeted treatment and precise risk stratification.

Materials

QIAmp Blood DNA Mini kit, MagAttract® HMW DNA kit (Qiagen, Germany), Qubit dsDNA BR Assay kit (ThermoFisher Scientific), genomic DNA ScreenTape assay (Agilent Technologies)

KEY RESOURCES TABLE

RESOURCE	SOURCE	IDENTIFIER
----------	--------	------------

Deposited Data

Raw 26 Processed Data	This study	Available upon request
-----------------------	------------	------------------------

Software and Algorithms

R	R-Core Team	v4.5.0
---	-------------	--------

RStudio	Posit	v2025.05.0+496
---------	-------	----------------

readxl	CRAN	v1.4.5
--------	------	--------

dplyr	CRAN	v1.1.4
-------	------	--------

ggplot2	CRAN	v3.5.2
---------	------	--------

ComplexHeatmap	Bioconductor	v2.24.0
----------------	--------------	---------

circlize	CRAN	v0.4.16
----------	------	---------

stringr	CRAN	v1.5.1
---------	------	--------

rvest	CRAN	v1.0.4
-------	------	--------

reshape2	CRAN	v1.4.4
----------	------	--------

grDevices	R-Core Package	v4.5.0
-----------	----------------	--------

grid	R-Core Package	v4.5.0
------	----------------	--------

Analysis Scripts	This study	See METHOD
------------------	------------	------------

DNA extraction

QIAmp Blood DNA Mini kit	Qiagen
--------------------------	--------

MagAttract® HMW DNA kit	Qiagen
-------------------------	--------

Qubit dsDNA BR Assay kit	Thermofisher
--------------------------	--------------

DNA ScreenTape assay	Agilent Technologies
----------------------	----------------------

Sequencing data alignment and variant calling

DRAGEN somatic pipeline version 4.0

DRAGEN somatic pipeline version 4.2

DRAGEN germline pipeline 4.2

Nirvana version 3.9.0

Clinical study

- 1 This study included prospectively and retrospectively ascertained patients with cancer or suspected cancer aged <25 years in accordance with the NHSE guidelines for clinical-diagnostic WGS. The study protocol was approved by the Health Research Authority and Health Care Research Wales (REC reference 22/ A/0336). Patients were recruited between April 2023 – March 2025, with written informed consent obtained from all participants or their parents/guardians. Sample handling and DNA extraction for all solid tumour and archival samples was carried out in the East Genomics Laboratory Hub, Cambridge University Hospitals NHS Foundation Trust (CUH), Cambridge, UK. All Ultra-Fast whole genome sequencing (UF-WGS) data were generated using the Illumina Constellation technology (Cambridgeshire, UK). All clinically actionable genomic variants identified by UF-WGS were orthogonally confirmed via NHS standard-of-care (SOC) molecular testing prior to changing management. In the UF-WGS workflow, all prospectively recruited haematological malignancies (except lymphoma) were sequenced directly from peripheral blood (PB) and/or bone marrow (BM) (or pleural fluid for Patient 66) crude lysate in tumour-only mode. Retrospectively recruited haematological malignancies and all solid tumours were sequenced from an extracted DNA sample and analysed using paired tumour and germline DNA samples. Turn-around time (TAT) was defined in this study as the interval between sample availability for WGS analysis and issuance of a clinically interpretable report. Only clinically actionable variants were reported as defined by the NHSE genomics test directory. All analyses and data visualisation were performed using R Statistical Software (v 2024.1

UF-WGS Workflow

- 2 *DNA extraction*
Genomic DNA from the peripheral blood and bone marrow was extracted using the QIAmp Blood DNA Mini kit and MagAttract® HMW DNA kit (Qiagen, Germany), following the manufacturer's protocols. A total of 200µl of peripheral blood and 100µl of bone marrow (diluted with 100µl of PBS) were used as input for the extractions. Genomic DNA from frozen tumour samples was extracted using the MagAttract® HMW DNA kit (Qiagen, Germany), following the manufacturer's protocols, with a lysis incubation at 56°C for 2 hours. Depending on tissue availability, between 5mg and 20mg of tissue were used as input for the extractions. DNA concentration was

determined using the Qubit dsDNA BR Assay kit (ThermoFisher Scientific) and quality was assessed using the genomic DNA ScreenTape assay (Agilent Technologies).

3 *Constellation sequencing*

Illumina WGS mapped read technology (Constellation) [1] leverages on-flow cell library preparation and

uses proximity information from neighbouring nanowells to generate long-range genomic insights using standard SBS sequencing. In this highly simplified method, separate library preparation is eliminated by the use of flow cell-bound transposomes which capture and tagment long molecules of DNA as they are flowed onto the flow cell surface and ensures that adjacent regions in a sample's genome remain physically proximal on the flow cell.

Clustering and sequencing by synthesis (SBS) are performed as standard, resulting in high quality

polymerase chain reaction (PCR) free short-read sequencing data. No modifications to the instrument hardware are required.

The resulting reads from neighbouring clusters can be reconstructed into an interspersed version of the original DNA template molecule, maintaining the accuracy, depth of coverage, and scalability of standard SBS sequencing while adding phasing, enhanced mapping ability, and improved structural variant detection often associated with long-read methods.

4 *Alignment and Variant Calling*

Sequencing reads were aligned to the linear version of the GRCh38 reference genome. Haematological

cancer samples were analysed in the tumour-only mode using DRAGEN somatic pipeline version 4.0, with likely germline small variants called based on their population frequency and gnomAD count > 50 [2]. Solid cancer samples were analysed using paired tumour/normal mode using DRAGEN somatic pipeline version 4.2 [3, 4] and DRAGEN germline pipeline version 4.2 to call germline variants [5].

For all sample types, copy-number variant (CNV) calling was done in the 'heterogeneous' mode, allowing subclonal CNV calls. Systematic noise files were used to reduce false positive small and structural variant calls. Only variants with VCF PASS status were included in the downstream analyses and reporting.

5 *Tumour Content Estimation*

Tumour purity was initially estimated by the DRAGEN pipeline, fitting the purity/ploidy model

using B-allele frequencies and coverage data across the whole genome. It was then manually re-estimated using genome-wide plots of B-allele frequencies, coverage levels and somatic small variant allele frequency (VAF) distributions. When the two estimates disagreed, which was sometimes observed in heterogeneous samples, manual estimation was preferred. In cases where DRAGEN pipeline couldn't provide a

confident purity estimate, manual estimate was used as above or by utilizing VAF of the known driver mutation.

6 Annotation and Variant Filtering

Germline and somatic variants were annotated by Nirvana version 3.9.0 using the Ensembl 91 transcript reference database. Pertinent variants, other than pharmacogenomic variants, were triaged via consequence annotations (SO terms): feature elongation, non-synonymous coding sequence variant, unidirectional gene fusion, bidirectional gene fusion, transcript ablation, canonical splice acceptor/donor variants, stop gained, transcript truncation, frameshift variant, stop lost, start lost, transcript amplification, in-frame insertion, in-frame deletion, missense variant, protein altering variant, splice region variant, incomplete terminal codon variant, copy number increase, copy number decrease, copy number change, transcript variant. Somatic and germline structural variants, other than inter-chromosomal translocations, were considered only if they would disrupt the exons of one of the genes or were located within its promoter region.

Pertinent germline and somatic variants were triaged further using gene and genomic region of interest list provided for each cancer and/or variant type. In the case of no findings, the list was extended to COSMIC cancer gene census v83 (<https://www.sanger.ac.uk/data/cancer-gene-census/>).

These analyses focused primarily on data aggregation and visualization rather than formal statistical testing. The core of the method involved transforming data from a long format into wide-format matrices suitable for heatmaps. This was achieved using functions like dcast or by transposing the data frame. The resulting matrices contained values indicating the detection of a variant in an assay or platform (1 for presence, 0 for absence in Analysis 1, and 1 for UF and GMS concordance, 0.5 for UF additional variants and 0 for GMS additional variants in Analyses 2).

No statistical clustering was performed. Instead, the rows and columns of the heatmaps were explicitly ordered based on predefined factors such as Variant type and Diagnosis Subgroup to facilitate qualitative comparisons between groups and platforms.

7 DUX4r calling

DUX4 rearrangements were called using the customized caller Pelops (feature available in DRAGEN version 4.4+). SRPB threshold of 10 was used to call samples positive for DUX4::IGH, while a threshold of 15 was used for DUX4-other. False positive regions were filtered using a blacklist as described in Grobecker et al [10].

Variant Allele Frequency (VAF) Analysis

- 8 This analysis was performed using an Excel file containing patient data, diagnostic subgroups, and genetic variant data – including variant type and variant allele frequency (VAF) – for both Solid and Haematological cancer types across the GMS WGS and UF

WGS platforms. - Software and Libraries: The scatter plot was generated using R script with the following libraries: readxl, dplyr, and ggplot2. - Data Processing and Visualisation: The dataset was imported and filtered to include only somatic single nucleotide variants (SNVs) and insertions/deletions (Indels), based on the "Origin" and "Variant type" columns for both solid and haematological cancer types. VAF values were converted to numeric format and visualized as a scatter plot using ggplot2, with points coloured by concordance between GMS WGS and UF WGS platforms.

Comparative Analysis of WGS and Other Assays

- 9 This analysis was performed using an Excel File contains patient data, diagnostic subgroups, and detection across different genomic assays.
 - Environment Setup: The analysis was conducted in R using the readxl, dplyr, ComplexHeatmap, circlize, and grid packages.
 - Data Structuring for Visualization: Data was loaded and converted to a numerical matrix, with patient IDs as row names, the assays detection as columns and missing values imputed to 0. The matrix was transposed so that assays detection were rows and patients were columns. The data was then partitioned into three matrices (Ultrafast WGS data, Routine NHS WGS and all other assays). The top two heatmaps display the detection in Ultrafast WGS and Routine NHS WGS, respectively. The larger heatmap displays all other assays performed in GMS, with rows grouped by assay type. All heatmaps share the same column order, which is determined by the Diagnosis Subgroup and are annotated accordingly. The final figure is a composite of three vertically concatenated heatmaps and was saved as a high-resolution SVG file.

Turnaround Time (TAT) Analysis

- 10 This analysis was performed using an Excel file containing turnaround time (TAT) data for Solid and Haematological cancer types and for GMS WGS and UF WGS platforms.
 - Software and Libraries: The analysis used R with the readxl, dplyr, and ggplot2 packages.
 - Data Processing and Visualization: Data was imported and grouped by Sample Type (Solid and Haem), Platform, and sample processing stage. Summary statistics, including the number of observations, mean TAT, and standard deviation were calculated for each Sample Type. A bar chart was generated using ggplot2 to represent the mean TAT, with error bars indicating the 95% confidence intervals. To visually separate each sample processing stage, the mean values for the "Samples 3e Sendout" stage were negated on the plot. The final plot was faceted by Sample Type and saved as an SVG file.

- 11 For the TAT analysis, 95% confidence intervals (CI) for the mean TAT were calculated for each group using the t-distribution. The formula used was: $CI = \text{mean_TAT} \pm t_{0.975, n-1} \times \text{sd_TAT} / \sqrt{n}$

where $t_{0.975, n-1}$ is the critical value from the t-distribution for a 95% confidence level with $n-1$ degrees of freedom. The calculated means and confidence intervals were then visualized as bar plots with error bars.

Genetic Variant Detection by GMS and UF Platforms

- 12 This analysis was performed using two independent Tables in an Excel File (Solid and Haematological cancer types) contains patient data, genetic variants, diagnostic subgroups, and detection across GMS and UF WGS platforms.
- Environment Setup: The heatmaps were generated in R script using libraries readxl, stringr, rvest, ggplot2, ComplexHeatmap, reshape2, dplyr, grDevices, and circlize.
 - Data Loading and Preprocessing: The dataset was loaded from an Excel file. Data was aggregated to create unique entries for each genetic variant per patient. A custom function converted protein change notations in the Variant column from three-letter to one-letter amino acid codes. "Yes"/"No" values in the GMS and UF WGS columns were converted to numeric 1/0.
 - Heatmap Visualization: Two data matrices were created to represent the presence or absence of each variant for each patient on GMS and UF WGS platforms, respectively. A primary heatmap was built using ComplexHeatmap with a custom cell function to render each cell as a rectangle split by a diagonal line. The upper triangle was coloured based on the GMS value, and the lower triangle by the UF value. A second heatmap was created to display the presence of mutational signatures (MMRD/Hypermutated, AID/APOBEC, UV-light, Signature 6). The two heatmaps were combined vertically and the final figure was exported as an SVG file.

Calculation of Tumour Mutational Burden (TMB) and Mutational Signatures

- 13 TMB was calculated as the sum of somatic SNVs and indels divided by the size of the GRCh38 genome reference in megabases. Variants with a population frequency $3e1\%$ in the 1000 Genomes Project were excluded on the basis that they likely represent germline variants.
- Mutational signature analysis was performed by calculating the fraction of each somatic mutation type in a specific class and then applying non-negative least squares to decompose the fraction data from each sample into the operative mutational signatures. To calculate fractions, single base substitutions were divided into 96 classes, double base substitutions into 78 classes, and indels into 83 classes using SigProfilerMatrixGenerator [7, 8]. Reference mutational signatures used for

decomposition were from the COSMIC 3.3 release [9]. Only signatures supported by 3e1000 single-base substitutions or 3e200 indels were considered significant.

Method Details

- 14 **Data Processing and Visualisation:** The dataset was imported and filtered to include only somatic single nucleotide variants (SNVs) and insertions/deletions (Indels), based on the "Origin" and "Variant type" columns for both solid and haematological cancer types. VAF values were converted to numeric format and visualized as a scatter plot using ggplot2, with points coloured by concordance between GMS WGS and UF WGS platforms.
- 15 **Genetic Variant Detection by GMS and UF Platforms:** This analysis was performed using two independent Tables in an Excel File (Solid and Haematological cancer types) contains patient data, genetic variants, diagnostic subgroups, and detection across GMS and UF WGS platforms.
- 16 **Environment Setup:** The heatmaps were generated in R script using libraries readxl, stringr, rvest, ggplot2, ComplexHeatmap, reshape2, dplyr, grDevices, and circlize.
- 17 **Data Loading and Preprocessing:** The dataset was loaded from an Excel file. Data was aggregated to create unique entries for each genetic variant per patient. A custom function converted protein change notations in the Variant column from three-letter to one-letter amino acid codes. "Yes"/"No" values in the GMS and UF WGS columns were converted to numeric 1/0.
- 18 **Heatmap Visualization:** Two data matrices were created to represent the presence or absence of each variant for each patient on GMS and UF WGS platforms, respectively. A primary heatmap was built using ComplexHeatmap with a custom cell function to render each cell as a rectangle split by a diagonal line. The upper triangle was coloured based on the GMS value, and the lower triangle by the UF value. A second heatmap was created to display the presence of mutational signatures (MMRD/Hypermuted, AID/APOBEC, UV-light, Signature 6). The two heatmaps were combined vertically and the final figure was exported as an SVG file.
- 19 **Comparative Analysis of WGS and Other Assays:** This analysis was performed using an Excel File contains patient data, diagnostic subgroups, and detection across different genomic assays.
Environment Setup: The analysis was conducted in R using the readxl, dplyr, ComplexHeatmap, circlize, and grid packages.
- 20 **Turnaround Time (TAT) Analysis:** This analysis was performed using an Excel file containing turnaround time (TAT) data for Solid and Haematological cancer types and for GMS WGS and UF WGS platforms.
Software and Libraries: The analysis used R with the readxl, dplyr, and ggplot2 packages.

Data Processing and Visualization: Data was imported and grouped by Sample Type (Solid and Haem), Platform, and sample processing stage. Summary statistics, including the number of observations, mean TAT, and standard deviation were calculated for each Sample Type. A bar chart was generated using ggplot2 to represent the mean TAT, with error bars indicating the 95% confidence intervals. To visually separate each sample processing stage, the mean values for the "Samples 3e Sendout" stage were negated on the plot. The final plot was faceted by Sample Type and saved as an SVG file.

Variant Allele Frequency (VAF) Analysis: This analysis was performed using an Excel file containing patient data, diagnostic subgroups, and genetic variant data – including variant type and variant allele frequency (VAF) – for both Solid and Haematological cancer types across the GMS WGS and UF WGS platforms.

Software and Libraries: The scatter plot was generated using R script with the following libraries: readxl, dplyr, and ggplot2.

Quantification and Statistical Analysis

- 21 **Heatmap-Based Visual Analysis (Analyses 1 26 2):** These analyses focused primarily on data aggregation and visualization rather than formal statistical testing. The core of the method involved transforming data from a long format into wide-format matrices suitable for heatmaps. This was achieved using functions like dcast or by transposing the data frame. The resulting matrices contained values indicating the detection of a variant in an assay or platform (1 for presence, 0 for absence in Analysis 1, and 1 for UF and GMS concordance, 0.5 for UF additional variants and 0 for GMS additional variants in Analyses 2).

No statistical clustering was performed. Instead, the rows and columns of the heatmaps were explicitly ordered based on predefined factors such as Variant type and Diagnosis Subgroup to facilitate qualitative comparisons between groups and platforms.

Statistical Analysis of Turnaround Times (Analysis 3): For the TAT analysis, 95% confidence intervals (CI) for the mean TAT were calculated for each group using the t-distribution. The formula used was: $CI = \text{meanTAT} \pm t_{0.975, n-1} \times \text{sdTAT} / \sqrt{n}$ where $t_{0.975, n-1}$ is the critical value from the t-distribution for a 95% confidence level with $n-1$ degrees of freedom. The calculated means and confidence intervals were then visualized as bar plots with error bars.

Protocol references

- [1] Introducing constellation mapped read technology <https://emea.illumina.com/science/genomics-research/articles/constellation-mapped-read-technology.html>
- [2] DRAGEN 4.0: <https://support-docs.illumina.com/SW/DRAGENv40/Content/SW/DRAGEN/GPipelineIntrofDG.htm>
- [3] DRAGEN 4.2: <https://support-docs.illumina.com/SW/dragen42/Content/SW/DRAGEN/GPipelineIntrofDG.htm>
- [4] Konrad Scheffler, Severine Catreux, Taylor O'Connell, Heejoon Jo, Varun Jain, Theo Heyns, Jeffrey Yuan, Lisa Murray, James Han, Rami Mehio. Somatic small-variant calling methods in Illumina DRAGEN™ Secondary Analysis. bioRxiv. doi: <https://doi.org/10.1101/2023.03.23.534011>
- [5] Behera S, Catreux S, Rossi M, Truong S, Huang Z, Ruehle M, Visvanath A, Parnaby G, Roddey C, Onuchic V, Finocchio A, Cameron DL, English A, Mehtalia S, Han J, Mehio R, Sedlaczek FJ. Comprehensive genome analysis and variant detection at scale using DRAGEN. Nat Biotechnol. 2024 Oct 25;10.1038/s41587-024-02382-2. doi: 10.1038/s41587-024-02382-1. Epub ahead of print. PMID: 39455800; PMCID: PMC12022141.
- [6] Michael Stromberg, Rajat Roy, Julien Lajugie, Yu Jiang, Haochen Li, and Elliott Margulies. 2017. Nirvana: Clinical Grade Variant Annotator. In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB '17). Association for Computing Machinery, New York, NY, USA, 596. <https://doi.org/10.1145/3107411.3108204>
- [7] GitHub - AlexandrovLab/SigProfilerMatrixGenerator:
<https://github.com/AlexandrovLab/SigProfilerMatrixGenerator> (accessed 24 Oct 2022).
- [8] Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, Boot A, Covington KR, Gordenin DA, Bergstrom EN, Islam SMA, Lopez-Bigas N, Klimczak LJ, McPherson JR, Morganella S, Sabarinathan R, Wheeler DA, Mustonen V; PCAWG Mutational Signatures Working Group; Getz G, Rozen SG, Stratton MR; PCAWG Consortium. The repertoire of mutational signatures in human cancer. Nature. 2020 Feb;578(7793):94-101. doi: 10.1038/s41586-020-1943-3. Epub 2020 Feb 5. Erratum in: Nature. 2023 Feb;614(7948):E41. doi: 10.1038/s41586-022-05600-5. PMID: 32025018; PMCID: PMC7054213.
- [9] <https://cancer.sanger.ac.uk/signatures/>
- [10] Grobecker P, Berri S, Peden JF, Chow KJ, Fielding C, Armogida I, Northen H, McBride DJ, Campbell PJ, Becq J, Ryan SL, Bentley DR, Harrison CJ, Moorman AV, Ross MT, Mijuskovic M. A dedicated caller for DUX4 rearrangements from whole-genome sequencing data. BMC Med Genomics. 2025 Jan 30;18(1):24. doi: 10.1186/s12920-024-02069-1. PMID: 39885506; PMCID: PMC11783778.

Acknowledgements

DATA AND CODE AVAILABILITY

The original data files and the R scripts used for analysis and visualization are available from the corresponding author upon reasonable request. All software packages used are publicly available from the CRAN or Bioconductor.