UNIVERSIDAD DE MURCIA

Escuela de Doctorado

TESIS DOCTORAL

Métodos bioinformáticos para la caracterización molecular a partir de secuenciación de tercera generación: aplicación a la deficiencia de antitrombina

Bioinformatic methods for molecular characterization with third generation sequencing: application to antithrombin deficiency

AUTOR/A Javier Cuenca Guardiola

DIRECTOR/ES Jesualdo Tomás Fernández Breis
Javier Corral de la Calle

2025

UNIVERSIDAD DE MURCIA Escuela de Doctorado

TESIS DOCTORAL

Métodos bioinformáticos para la caracterización molecular a partir de secuenciación de tercera generación: aplicación a la deficiencia de antitrombina

Bioinformatic methods for molecular characterization with third generation sequencing: application to antithrombin deficiency

AUTOR/A Javier Cuenca Guardiola

DIRECTOR/ES Jesualdo Tomás Fernández Breis
Javier Corral de la Calle

2025

UNIVERSIDAD DE MURCIA

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA EN MODALIDAD DE COMPENDIO O ARTÍCULOS PARA OBTENER EL TITULO DE DOCTOR/A

Aprobado por la Comisión General de Doctorado el 19 de octubre de 2022.

Yo, D. Javier Cuenca Guardiola, habiendo cursado el Programa de Doctorado en Informática de la Escuela Internacional de Doctorado de la Universidad de Murcia (EIDUM), como autor/a de la tesis presentada para la obtención del título de Doctor/a titulada:

Bioinformatic methods for molecular characterization with third generation sequencing: application to antithrombin deficiency / Métodos bioinformáticos para la caracterización molecular a partir de secuenciación de tercera generación: aplicación a la deficiencia de antitrombina

y dirigida por:

D.: Jesualdo Tomás Fernández Breis
D.: Javier Corral de la Calle
D.:

DECLARO QUE:

La tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la Ley de Propiedad Intelectual (R.D. legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, modificado por la Ley 2/2019, de 1 de marzo, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita, cuando se han utilizado sus resultados o publicaciones.

Además, al haber sido autorizada como prevé el artículo 29.8 del reglamento, cuenta con:

- *La aceptación por escrito de los coautores de las publicaciones de que el doctorando las presente como parte de la tesis.*
- *En su caso, la renuncia por escrito de los coautores no doctores de dichos trabajos a presentarlos como parte de otras tesis doctorales en la Universidad de Murcia o en cualquier otra universidad.*

Del mismo modo, asumo ante la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad del contenido de la tesis presentada, en caso de plagio, de conformidad con el ordenamiento jurídico vigente.

Murcia, a 9 de julio de 2025

(firma)

Información básica sobre protección de sus datos personales aportados:	
Responsable	Universidad de Murcia. Avenida teniente Flomesta, 5. Edificio de la Convalecencia. 30003; Murcia. Delegado de Protección de Datos: dpd@um.es
Legitimación	La Universidad de Murcia se encuentra legitimada para el tratamiento de sus datos por ser necesario para el cumplimiento de una obligación legal aplicable al responsable del tratamiento. art. 6.1.c) del Reglamento General de Protección de Datos
Finalidad	Gestionar su declaración de autoría y originalidad
Destinatarios	No se prevén comunicaciones de datos
Derechos	Los interesados pueden ejercer sus derechos de acceso, rectificación, cancelación, oposición, limitación del tratamiento, olvido y portabilidad a través del procedimiento establecido a tal efecto en el Registro Electrónico o mediante la presentación de la correspondiente solicitud en las Oficinas de Asistencia en Materia de Registro de la Universidad de Murcia

Esta DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD debe ser insertada en la quinta hoja, después de la portada de la tesis presentada para la obtención del título de Doctor/a.

To my family.

Contents

Preface **III**

***Resumen* and Summary** **V**

Resumen en español **IX**

1 Introduction and State of the Art **1**
1.1 Motivation . . . 3
1.2 Evolution of sequencing methods . . . 5
1.3 Applications of sequencing methods . . . 11
1.4 Bioinformatic software . . . 14

2 Challenges and Objectives **31**
2.1 Overview . . . 31
2.2 Objectives . . . 31

3 Methodology **35**
3.1 Benchmarking tools . . . 35
3.2 Variant calling and subsequent analysis . . . 36

4 Improvement of large copy number variant detection by whole genome nanopore sequencing **39**

5 Detection and annotation of transposable element insertions and deletions on the human genome using nanopore sequencing **41**

6 Advanced analysis of retrotransposon variation in the human genome with nanopore sequencing using RetroInspector **43**

7 Discussion and Conclusions **45**
7.1 Discussion . . . 45
7.2 Further work . . . 48
7.3 Contributions . . . 49
7.4 Conclusions . . . 50

Bibliography **53**

Preface

This thesis is based on three scientific publications. The full articles are included as chapters for ease of reading. All three articles are published under some variant of the Creative Commons License, which allow their redistribution. Below, the articles are presented in chronological order, from the oldest to the most recent one. The full citation, the DOI, and its license are listed.

Indexed publications

Cuenca-Guardiola, J., de la Morena-Barrio, B., García, J. L., Sanchis-Juan, A., Corral, J. & Fernández-Breis, J. T. Improvement of Large Copy Number Variant Detection by Whole Genome Nanopore Sequencing. *Journal of Advanced Research* **50,** 145–158 (Aug. 2023). Doi: `https://doi.org/10.1016/j.jare.2022.10.012`. License: Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (`https://creativecommons.org/licenses/by-nc-nd/4.0/`).

Cuenca-Guardiola, J., de la Morena-Barrio, B., Navarro-Manzano, E., Stevens, J., Ouwehand, W. H., Gleadall, N. S., Corral, J. & Fernández-Breis, J. T. Detection and Annotation of Transposable Element Insertions and Deletions on the Human Genome Using Nanopore Sequencing. *iScience* **26.** (2023) (Nov. 2023). Doi: `https://doi.org/10.1016/j.isci.2023.108214`. License: Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (`https://creativecommons.org/licenses/by-nc-nd/4.0/`).

Cuenca-Guardiola, J., de la Morena-Barrio, B., Corral, J. & Fernández-Breis, J. T. Advanced Analysis of Retrotransposon Variation in the Human Genome with Nanopore Sequencing Using RetroInspector. *Scientific Reports* **15,** 14489 (Apr. 2025). Doi: `https://doi.org/10.1038/s41598-025-98847-7`. License: Creative Commons Attribution-NonCommercial-No Derivatives 4.0 International (`https://creativecommons.org/licenses/by-nc-nd/4.0/`).

Resumen and Summary

Resumen

Las variantes estructurales (VE) han sido difíciles de estudiar y caracterizar por completo hasta la llegada de la secuenciación de tercera generación. Esta permite detectar VE con precisión en sus coordenadas e información del efecto en la secuencia a nivel genómico, gracias a lecturas más largas. Esta tesis estudia cómo mejorar la detección de VE de distintos tipos con secuenciación de nanoporos, con sus resultados recogidos en un compendio de tres publicaciones científicas indexadas. Durante su desarrollo, se presta especial atención a VE relacionadas con la deficienciad de antitrombina, una enfermedad congénita que aumenta el riesgo de sufrir trombosis. Al tratarse de una enfermedad de herencia mendeliana, normalmente causada por mutaciones en un único gen, supone un excelente punto de partida para el desarrollo de métodos que puedan extenderse posteriormente.

Los objetivos de la tesis son la detección y caracterización, usando datos de secuenciación con nanoporos, de variantes de número de copias mayores de 50 kilobases, de inserciones y deleciones de elementos trasponibles, así como el estudio del posible impacto funcional de estas dos, y la creación de métodos bioinformáticos reproducibles y reusables.

La metodología general ha consistido en el alineamiento de las lecturas de nanoporo, detección de variantes, unión de los conjuntos de variantes y análisis posteriores específicos, como cálculo de cobertura o la detección de secuencias de elementos trasponibles, además de programas propios creados para el propósito de cada trabajo. Para el alineamiento y detección de variantes se han hecho análisis de desempeño considerando minimap2, lra y NGMLR como alineadores, y Sniffles, Sniffles2, CuteSV, SVIM y NanoVar como detectores de variantes. Para asegurar la reproducibilidad, se ha usado conda para distribuir una herramienta, disCoverage, y Snakemake, para una nueva *pipeline*, RetroInspector.

Los resultados aparecen en tres publicaciones. Primero, se analiza el rendimiento de la secuenciación de tercera generación con duplicaciones y deleciones de gran tamaño. Tras seleccionar los programas con mejor desempeño, se baja exactitud en los resultados, por lo que se presenta un método para filtrar los resultados de estos programas. Se trata de disCoverage, un programa que analiza la salida de detectores de variantes y filtra grandes mutaciones mediante análisis estadístico de cobertura, aunque no es capaz de detectar variantes por sí mismo. Segundo, se presenta un análisis de inserciones y deleciones de elementos trasponibles a nivel genómico, en

el que se investiga la diversidad de estas mutaciones, su impacto funcional y se comparan los resultados con otros estudios. Se discute la viabilidad de la secuenciación por nanopores para detectar mutaciones patogénicas, aunque no se analiza la presencia de secuencias indicadoras de retrotransposición. Por último, se presenta una herramienta que refina la metodología del segundo artículo, incluyendo el análisis de marcas de transposición, aunque con resultados difíciles de interpretar, y la empaqueta en una *pipeline*, RetroInspector, de sencilla instalación y ejecución, con ficheros documentados y resúmenes gráficos.

La tesis en su conjunto presenta el resultado de una labor investigadora con secuenciación de nanoporos, enmarcada en el estado del arte, presentando nuevos métodos y sus limitaciones, comentadas también como futuro trabajo.

Summary

Structural variants (SVs) have been challenging to study and fully characterize until the advent of third-generation sequencing technologies. These technologies, by providing longer reads, enable accurate detection of SVs with precise genomic coordinates and information about their sequence-level effects. This thesis explores how to improve the detection of various types of SVs using nanopore sequencing, with its findings compiled in a compendium of three peer-reviewed scientific publications. Special attention is given to SVs related to antithrombin deficiency, a congenital disorder that increases the risk of thrombosis. As a Mendelian disease typically caused by mutations in a single gene, it serves as an excellent starting point for the development of methods that may later be generalized.

The objectives of this thesis are the detection and characterization, using nanopore sequencing data, of copy number variants larger than 50 kilobases, insertions and deletions of transposable elements, as well as the assessment of their potential functional impact, and the development of reproducible and reusable bioinformatics tools.

The general methodology involves alignment of nanopore reads, variant calling, merging of variant sets, and specific downstream analyses such as coverage calculation or transposable element sequence detection. Custom software was also developed for the specific needs of each study. Performance analyses were conducted to compare alignment tools (minimap2, lra, and NGMLR) and variant callers (Sniffles, Sniffles2, CuteSV, SVIM, and NanoVar). To ensure reproducibility, conda was used to distribute a software tool, disCoverage, and Snakemake was employed for workflow management for the pipeline RetroInspector.

The results are presented across three publications. The first evaluates the performance of third-generation sequencing for detecting large duplications and deletions. After identifying the best-performing tools, low precision was observed, prompting the development of a filtering method. This led to the creation of disCoverage, a program that filters large mutations based on coverage statistics, using the output of variant callers. While not a variant caller itself, it

improves result quality. The second publication presents a genome-wide analysis of insertions and deletions of transposable elements, investigating their diversity and functional impact, and comparing the findings to previous studies. The feasibility of using nanopore sequencing to detect pathogenic mutations is discussed, although the presence of retrotransposition hallmarks is not analyzed. Finally, the third publication introduces a tool that refines the methodology from the second study, incorporating the detection of transposition signatures, though with results that are difficult to interpret, and packages the process into an easy-to-use pipeline, RetroInspector, featuring documented output files and graphical summaries.

Overall, this thesis presents a body of research grounded in the state of the art in nanopore sequencing, offering new methodological approaches alongside a critical discussion of their limitations and future directions.

Resumen en español

Introducción

La genómica moderna ha experimentado una transformación radical con la aparición de las tecnologías de secuenciación de tercera generación (STG), particularmente la secuenciación por nanoporos. Su capacidad de generar lecturas largas y ultralargas ha abierto nuevas vías para el estudio de variantes genéticas complejas, especialmente las variantes estructurales (VE), que han sido históricamente difíciles de caracterizar. Las VE son reordenamientos del material genético de más de 50 pares de bases (pb) que comprenden una amplia gama de mutaciones. A pesar de ser menos numerosas que las variantes de un solo nucleótido (SNP) o las inserciones y deleciones cortas (indels), las VE representan la mayor fracción de la diferencia de longitud entre dos genomas individuales, lo que les confiere gran importancia en la diversidad genética y funcionalidad genómica. Han sido asociadas a enfermedades multifactoriales, entre las que se incluyen factores neurológicos, enfermedades autoinmunes y cáncer. En relación a la deficiencia de antitrombina, central en este trabajo, estudios previos han mostrado que distintos tipos de VE tienen efecto patogénico. Algunas de estas variantes escapan la detección de otras técnicas, mientras que las que ya se podían detectar se caracterizan mejor mediante STG.

Esta tesis doctoral se centra en la mejora de la detección y caracterización de VE utilizando datos de secuenciación por nanoporos, con variantes relacionadas con la deficiencia de antitrombina (AT) como punto de partida. La deficiencia de AT es un trastorno congénito hereditario autosómico dominante que aumenta el riesgo de padecer trombosis, por la disminución en el nivel de AT o producción de una proteína no funcional. Aunque la mayoría de las mutaciones patogénicas conocidas en el gen *SERPINC1*, que codifica la AT, son variantes pequeñas, aproximadamente el 10% son VE. Se han asociado distintos tipos de VE a la deficiencia de AT, incluidas duplicaciones, deleciones e inserciones de elementos transponibles (ET), por lo que las VE son relevantes para el estudio de esta enfermedad. La complejidad de estas VE y las limitaciones de las tecnologías de secuenciación tradicionales indican la necesidad de nuevos enfoques bioinformáticos.

Objetivos

Los objetivos principales de esta tesis se formularon para abordar las lagunas existentes en la caracterización de VE mediante secuenciación por nanoporos y para desarrollar metodologías

bioinformáticas robustas y reutilizables. Estos objetivos, que surgen del estudio de deficiencia de AT y se extienden al análisis genómico, se detallan a continuación:

- **OB1: Examinar la eficacia de la secuenciación por nanoporos para la detección de VE grandes.**
  - **OB1.1**: Comparar la llamada de variantes con nanoporos y la detección de variantes de número de copias (CNV) con aCGH (hibridación genómica comparada en array).
  - **OB1.2**: Desarrollar métodos para examinar los datos de secuenciación y refinar los resultados de nanoporos, si es necesario.
- **OB2**: Estudiar la diversidad de inserciones de elementos transponibles (ET) en el genoma humano.
  - **OB2.1**: Crear y calibrar un flujo de trabajo adaptado para detectar inserciones de ET.
  - **OB2.2**: Estudiar el impacto genético de las inserciones de ET, identificando qué genes son afectados y si ocurren en secuencias codificantes, intrónicas, promotores, u otras.
  - **OB2.3**: Estudiar el impacto funcional de las inserciones de ET, realizando análisis de enriquecimiento para determinar los procesos biológicos que podrían verse afectado.
- **OB3**: Estudiar la diversidad de deleciones de elementos transponibles (ET) en el genoma humano. Realizar un análisis análogo al de las inserciones de ET, pero centrado en las deleciones que corresponden a secuencias de ET presentes en el genoma de referencia, pero ausentes en las muestras de estudio.
- **OB4**: Generar herramientas reproducibles y reutilizables para la comunidad científica.

Metodología

La metodología empleada en esta tesis sigue un flujo de trabajo bioinformático estándar para el descubrimiento de variantes, que incluye el alineamiento de lecturas, la llamada de variantes y análisis posteriores específicos según el propósito de cada estudio. Se ha prestado especial atención a la evaluación comparativa de las herramientas existentes y al desarrollo de software propio para abordar las limitaciones identificadas. Como referencia se han usado conjuntos de datos producidos por el Human Genome Structural Variant Consorcium (HGSV). Los alineadores empleados a lo largo de la tesis han sido NGMLR, lra y minimap2, que ha sido el seleccionado. Para detectar variantes, se ha usado NanoVar, CuteSV, SVIM, Sniffles y Sniffles2. Para unir los resultados de estos programas se ha empleado SURVIVOR y, para calcular la profundidad de cobertura, mosdepth. Para procesos específicos de cada estudio se

ha recurrido a herramientas como Repeat Masker, para detectar secuencias de ET; librerías especializadas de R y python, como annotatR para anotación genómica, ClusterProfiler para análisis de enriquecimiento, o pysam para acceder y manipular ficheros; y programas propios implementados en R y Python.

Resultados y publicaciones

Artículo 1: Mejora de la detección de grandes variantes de número de copias mediante secuenciación de genoma completo por nanoporos

Este artículo (Capítulo 4) aborda el objetivo OB1, centrándose en la eficacia de la secuenciación por nanoporos para detectar CNVs grandes (mayores de 50 kb), puesto que la literatura indica que es una deficiencia en los programas existentes. Se comparó el rendimiento de diferentes combinaciones de alineadores y programas de llamada de variantes utilizando datos de pacientes con deficiencia de AT y un conjunto de datos públicas estandarizados usados como referencia. Se observó que, para CNVs grandes, el software existente mostraba un rendimiento inferior en términos de sensibilidad y especificidad en comparación con variantes más pequeñas. Las combinaciones de Sniffles y NGMLR mostraron los mejores resultados, aunque la sensibilidad para CNVs grandes seguía siendo baja. Además, se identificó un número considerable de falsos positivos, incluyendo deleciones que afectaban fracciones significativas de cromosomas enteros, lo que indicaba la necesidad de un refinamiento.

Para abordar estas limitaciones (OB1.2), se desarrolló un método personalizado, implementado en la herramienta disCoverage, que utiliza el análisis estadístico del valor de cobertura de secuenciación en las regiones afectadas por las presuntas VE. Este enfoque permitió pulir los resultados de los programas de llamada de variantes, mejorando la especificidad hasta en un 62% y la sensibilidad en un 15%. DisCoverage presenta la limitación de que no es un programa de llamada de variantes, sino una herramienta de filtrado que valida los resultados de detectores de variantes y elimina resultados erróneos, pero no puede generar nuevos conjuntos de datos. La herramienta fue empaquetada como un paquete conda, relacionado con el objetivo OB4 de generar herramientas reproducibles y reutilizables.

Artículo 2: Detección y anotación de inserciones y deleciones de elementos transponibles en el genoma humano mediante secuenciación por nanoporos

Este artículo (Capítulo 5) aborda los objetivos OB2 y OB3, mediante la investigación de la diversidad de inserciones y deleciones de ET en el genoma humano. Se analizaron los genomas de 24 pacientes con deficiencia de AT utilizando secuenciación por nanoporos. Se identificaron 7.344 inserciones de ET y 3.056 deleciones de ET, de las cuales 2.926 no habían sido descritas previamente en bases de datos públicas. Para cumplir con el subobjetivo OB2.1 (creación de

un flujo de trabajo para detectar inserciones de ET), se seleccionaron programas de llamada de variantes basándose en su sensibilidad para inserciones y se desarrollaron programas propios para reensamblar secuencias insertadas a partir de lecturas. Se estudió la distribución de las inserciones de ET a lo largo de los cromosomas, encontrando regiones enriquecidas. La comparación con otras cohortes reveló la novedad de muchas de las inserciones reportadas.

En cuanto al impacto genético (OB2.2), se encontró que casi 4.000 genes contenían al menos una inserción de ET, la mayoría en intrones (incluyendo una inserción patogénica en *SERPINC1*), y seis inserciones en exones. El análisis de enriquecimiento funcional (OB2.3) de los genes afectados reveló una fuerte relación con funciones neuronales y el autismo, lo que concuerda con estudios previos que vinculan los polimorfismos de ET con el autismo. Un proceso análogo de detección y anotación génica se realizó para las deleciones de ET, identificando aproximadamente 3.000 polimorfismos de ET presentes en el genoma de referencia, pero no universalmente en la cohorte.

Artículo 3: Análisis avanzado de la variación de retrotransposones en el genoma humano con secuenciación por nanoporos utilizando RetroInspector

El tercer artículo (Capítulo 6) refina y extiende la metodología del segundo, con un enfoque en el objetivo OB4: la creación de una herramienta reproducible y fácil de usar. Se presenta RetroInspector, un pipeline configurable basado en Snakemake que automatiza la detección, anotación, enriquecimiento y genotipado de variantes de ET.

RetroInspector mejora la metodología previa al incluir Sniffles2, una versión actualizada y más precisa de Sniffles. Se comparó el rendimiento de RetroInspector con otras herramientas como PALMER y GraffiTE, demostrando ser más rápido y preciso, además de ofrecer funcionalidades adicionales como la anotación génica y el análisis de enriquecimiento. RetroInspector también permite la comparación de resultados entre dos muestras y la opción de omitir el proceso de anotación y enriquecimiento para genomas de referencia que carecen de esta información. Esta herramienta representa una contribución significativa a la comunidad científica al proporcionar un flujo de trabajo robusto y accesible para el análisis de la variación de retrotransposones.

Discusión y conclusiones

Esta tesis doctoral ha abordado la detección y el análisis de diversas variantes estructurales en el contexto del diagnóstico molecular, con especial énfasis en la deficiencia de antitrombina, utilizando datos de secuenciación por nanoporos. Los trabajos presentados muestran las fortalezas y debilidades de esta tecnología y describen su potencial para caracterizar esta fuente previamente inexplorada de diversidad genética. El primer artículo reveló las limitaciones del software existente para la detección de CNVs grandes y propuso una solución para refinar los resultados, mejorando la especificidad y la sensibilidad. Esto subraya la necesidad de métodos

complementarios y la importancia de la validación de las llamadas de variantes, especialmente en aplicaciones clínicas.

El segundo artículo demostró la capacidad de la secuenciación por nanoporos para un análisis exhaustivo de las inserciones y deleciones de ET a nivel genómico. La identificación de miles de variantes de ET, muchas de ellas novedosas, y la exploración de su impacto genético y funcional, abren nuevas vías para comprender su papel en la salud y la enfermedad, incluyendo su conexión con trastornos como el autismo.

Finalmente, el tercer artículo culminó en el desarrollo de RetroInspector, una herramienta bioinformática integral que empaqueta las metodologías desarrolladas en un formato reproducible y reutilizable. Este pipeline facilita el análisis de la variación de retrotransposones para la comunidad científica, alineándose con los principios de la ciencia abierta.

En conjunto, esta tesis ha contribuido al campo de la bioinformática al:

- Evaluar y mejorar la detección de grandes CNVs con secuenciación por nanoporos.
- Proporcionar un análisis detallado de las inserciones y deleciones de ET en el genoma humano.
- Desarrollar herramientas bioinformáticas de código abierto que son reproducibles y reutilizables, facilitando la investigación futura.

Trabajo futuro

Los temas abordados en esta tesis son de gran interés actual en las comunidades de secuenciación por nanoporos y diagnóstico molecular. Sin embargo, no está completo. A partir del trabajo aquí presentado se proponen varias líneas para continuarlo.

En el ámbito de las CNVs, la continua evolución de las herramientas de detección sugiere la necesidad de una reevaluación periódica del estado del arte. Sería valioso comparar el rendimiento de las herramientas actuales, incluyendo Spectre (adoptada por ONT para la detección de CNVs por análisis de cobertura), con tecnologías como el mapeo óptico del genoma (OGM) de Bionano, una tecnología de alta precisión y mayor *recall* que el *array* de hibridación comparativa usado en este trabajo, para identificar las limitaciones persistentes y desarrollar herramientas sucesoras o mejoras para disCoverage.

En relación con los ETs, los esfuerzos actuales para clasificar la diversidad genética humana podrían beneficiarse enormemente de la aplicación de RetroInspector para crear una base de datos exhaustiva de inserciones y deleciones de ET. La integración de datos de diversas cohortes, más allá del Proyecto 1000 Genomas, permitiría construir un catálogo robusto de variantes de ET. Con un catálogo así, sería posible avanzar hacia la creación de un pangenoma que no solo represente las variantes conocidas, sino que también incorpore las frecuencias alélicas y las

anotaciones clínicas. Además del interés clínico del catálogo de varaintes de TE, se podría usar para crear una aproximación a un pangenoma humano centrado en estas variantes, lo que es de gran interés, por considerarse la adopción del análisis basado en pangenomas la siguiente etapa de la bioinformática genómica.

1

Introduction and State of the Art

Structural variants (SVs) are defined as DNA rearrangements that are larger than 50 base pairs (bp)[1]. SVs are classified into different classes depending on how they alter DNA sequences. First, there are deletions, which are losses of genetic material. Second, duplications, which are uninterrupted sequences of DNA that appear repeated across the genome. They are also referred to as segmental duplications to differentiate them from whole genome duplications[2], rare evolutionary events in which the entire genome is doubled[3]. Duplications may be interspersed, if the duplicated material is separated by another sequence, or tandem, in case the new sequence is juxtaposed to the original one. Third, inversions are mutations in which a region changes its orientation. Fourth, translocations, which occur when a fragment of a chromosome changes its position on the same chromosome or moves to a different one. Fifth, insertions of novel sequences. Last, SVs that are a combination of any of the former are called complex SVs[4]. The first two types can also be referred to as copy number variants (CNVs) in polymorphic regions, in which a repetitive sequence appears in different numbers in different individuals[5]. However, this is reserved for SVs with a length of at least ten kilobases[6].

Another factor in SV grouping is whether they affect the total amount of genetic material. Unbalanced SVs (deletions, duplications, insertions, and interchromosomal translocations) result in a change in genetic material, while balanced ones (intrachromosomal translocations and inversions) do not. Deletions result in a net loss of material; duplications and insertions, in a net gain; while translocations between chromosomes cause a loss of DNA in the donor, which is subsequently gained by the receptor[4].

As with any genetic variation, SVs are defined with a reference sequence[7]. This implies that they are relative to that reference; for example, if sequence A has a deletion when using sequence B as the reference, sequence B will have an insertion (or a duplication) when using A as reference.

Unlike shorter variants, such as single nucleotide variants (SNVs) and short insertions and deletions (indels), SV detection poses many challenges, so the current gold standard for sequencing, Illumina's Next Generation Sequencing (NGS), has severe limitations. However, despite its shortcomings, it has been used to detect CNVs[5], insertions[8], or all types of SVs

of the same type[9], but it has limitations that other methods can solve[10–12]. The limitations are both quantitative and qualitative: short reads are not enough to detect most SVs (>70%), particularly those smaller than 2 kb, with atypical GC content[12], or those that occur within repetitive sequences[10].

The difficulty in SV detection arises mainly due to their larger size, which cannot be encompassed within Illumina reads. Additionally, their diversity renders non-sequencing technologies unable to detect all types of variants, or to locate their coordinates accurately[12].

There are also techniques that are not based on sequencing that are used to detect SVs, usually when short variants are not found with NGS[13]. These methods can be either specific to a variant or groups of variants, or genome-wide. Most of the specific methods are based on polymerase chain reaction (PCR) amplification. First, long-range PCR uses a high-fidelity, highly processive polymerase, i.e., one that introduces few to no errors during amplification and yields long products. This polymerase produces an amplified sequence in the order of kilobases that can be sequenced[14]. Second, real-time quantitative PCR (qPCR) uses cases and a control or standard curve for comparison, allowing to find the number of copies of a sequence that the cases carry[15]. Third, MLPA (Multiple Ligning Probe Amplification) uses several labeled probes to target more than one sequence, and the product is compared with a control to quantify the copies of the sequence of interest. Fourth and last, fluorescence *in situ* hybridization (FISH), which does not depend on PCR, employs DNA probes with labels designed for specific targets. The presence of the target sequence is confirmed by observing fluorescence from the corresponding label. It was designed for very large defects (1-3 Mb), although its resolution is better if applied to the cell nucleus during the interphase (50 kb to 2 Mb). A variant of FISH, fiber-FISH, is applied to chromosome fibers that have been mechanically stretched, allowing to study smaller variants (5-500 kb). FISH, in addition to CNVs, can also detect translocations[15] and large inversions[16].

Non-specific, genome-wide probes are not limited to a single region per experiment. First, FISH can be made non-specific by using different labels, resulting in multicolored FISH[15]. Second, karyotyping Giemsa-banded chromosomes can detect chromosomal defects, such as aneuploidies or large rearrangements[17]. Third, comparative genome hybridization (CGH), also employs fluorescent probes to quantify copies of a sequence against a control. It was integrated into microarrays to analyze the whole genome in a single experiment as array CGH (aCGH)[18]. Fourth, and last, optical genome mapping (OGM) consists in labeling the motif "CTTAAG". The labeled molecules hybridize in an array, where changes in labeling patterns and spacing are analyzed to detect CNVs, insertions, and translocations[19].

In the 2010s, two early long-read sequencing technologies, Pacific Biosciences'single-molecule real-time sequencing (SMRT) and Oxford Nanopore Technologies (ONT) nanopore sequencing, were announced and released, which, among other capabilities, allowed for a better characterization of SVs at the genomic level[20]. Throughout that decade and beyond, companies

that owned these technologies and independent research groups would continually introduce improvements in equipment and analysis software[21]. These technologies, the third generation of sequencing, are discussed in more detail in Section 1.2.1, and the application of sequencing and the use of variant knowledge is explored in the following sections, with particular attention to existing software (Sections 1.4)..

LRS allows to report SVs with single-base precision genome-wide and, since it is a sequencing method, it reports the modified sequencing, which is useful for allele-specific studies[22,23], better discovery of polymorphic variants[24,25], or examination of the sequence of the variant, in the case of transposable elements[10].

This thesis studies the current situation with regard to the discovery and characterization of SVs using nanopore sequencing. Focus will be on covering aspects pertaining to different types of SV, as well as on the importance of producing reusable software for reproducible experiments.

1.1 — Motivation

SVs are of great clinical and scientific interest. Currently, both international[11,26] and national[27–30] studies have been conducted to catalog the genetic diversity caused by SVs in the human genome. There is also interest in the study of SVs in other species[31,32].

The most extreme case of SV is aneuploidy, in which a complete chromosome is present in a number different from two. Their study has been possible with karyotyping techniques since their discovery in 1956. It allowed to connect, for example, Down syndrome and the trisomy of chromosome 21 in 1959[33]. Other diseases have also been associated with large SVs, such as deletions or duplications causing Prader-Willi, Angelman, Williams, and Charcot-Marie-Tooth syndromes[34].

In addition to these rare examples, SVs are an important source of genetic variation between genomes. A human genome carries on average 3.3 million SNVs and 492 000 indels[35], and these figures have been revised to be actually larger: 4 to 5 million and 700 000-800 000, respectively[36]. On the other hand, the amount of SVs is smaller: 23 000-28 000[36]. These numbers have been validated by later work[11]. Despite this, the smaller categoriesare responsible for the length of the variation equivalent to 0. 147% of the genome, whereas the SVs do so for 0.397%, and represent the largest fraction of the difference between two given genomes; that is, despite being smaller, they are responsible for a larger fraction of genetic diversity[36].

Due to their size, their effect on gene function is often more deleterious than shorter variants and, therefore, they are less frequent on genes[37]. For this greater impact, SVs have been linked to many multifactorial diseases, such as Parkinson's, Chron's[38], autism[39], or susceptibility to AIDS[40].

All the previous information concerning SVs describes germline mutations, the ones that are inherited and affect all cells in an organism. If we shift our focus to somatic mutations (those that appear during the life of an individual and are present only in some cells), SVs are also relevant. For example, more than half of all cancer driver mutations are SVs[41]. Cancer inducing SVs seem to act by either altering the 3D structure of the genome, altering the regulation of genes[42], or by either adding or deleting entire copies of a gene[43].

The disease most relevant to this work is inherited antithrombin (AT) deficiency, which increases the risk of suffering thrombosis (clot formations within blood vessels)[13,44]. AT deficiency is the first form of thrombophilia (a state of the blood prone to excessive coagulation or thrombosis[45]) known to be inherited, discovered when studying the hereditary predisposition to venous thromboembolism. Inherited AT deficiency is relatively rare, with a prevalence reported as high as 1 out of 500 people and as low as 1 out of 5 000 people[46]. Despite these figures dating from the 1990 and 1987, respectively, there is no updated, and they are what the literature currently mentions[45].

Antithrombin is a protein encoded by the gene *SERPINC1*. Its main biological function is to stop blood coagulation. It does this by efficiently inhibiting the actions of procoagulant serine proteases by a suicide mechanism common to all serin protease inhibitors (*SERPIN*) superfamily, mainly FXa and thrombin, which are the last effectors in coagulation.[13]. AT has two functional domains; a heparin binding site, and a reactive center. The first binds to glycosaminoglycanes such as heparin, which activates AT by inducing a conformational change that increases its affinity for thrombin by 1 000 times. The second domain binds to thrombin, or any other target protease, as a pseudosubstrate, preventing further activity from the bound thrombin molecule[47].

In addition to its origin, AT deficiency can be classified into two types. Type I has a quantitative effect, meaning that there is less AT circulating in the blood. Type II, on the other hand, has a qualitative impact, which means that there is a normal concentration of AT in the blood, but part of it is not functional. Type II deficiencies can be subclassified according to the functional defect; type II heparin binding site (HBS), reactive site (RS) or pleiotropic (PE) if it has affected both functions[48], which is indeed the result of a conformational transition to a noninhibitory conformation with low heparin affinity (latent)[49]. Finally, congenital disorders of glycosylation may also cause AT deficiency, supporting the key role of glycosylation for this anticoagulant[50].

Although most of the currently known pathogenic *SERPINC1* mutations are small variants, approximately 10% are SVs, ranging from 193 bp to 8 Mb, including duplications, deletions, and TE insertions. Pathological variants can remain undetected using the previously described methods. However, previous work by our research group has identified novel pathogenic SVs that cause AT deficiency using nanopore sequencing, which has enabled the molecular diagnosis of several patients. Nanopore sequencing has single-base resolution and is not limited

to specific regions selected by probes, so it can, for example, uncover SVs within introns[51,52]. However, current methods have gaps that need to be addressed, such as false positives[53] or negatives[54]. Furthermore, SVs have complex boundaries with small breakpoint and length deviations, making them difficult to characterize[55].

Currently, half of patients with a suspected mendelian disease have no molecular diagnosis[56], and in some cases have had to wait more than 50 years to get one, going through the so-called "diagnostic odysseys", which can end with LRS[29]. LRS is unique for clinical applications because, at least in theory, a single sequencing experiment could replace all the genetic tests currently performed[57]. However, concerns about accuracy limit its adoption, although there are ongoing clinical studies for its adoption in medical centers[58].

With this information, it is clear that there is a need to explore the current state of nanopore sequencing for detecting SVs, particularly pathogenic ones. This thesis aims to present new methods that improve the results of existing ones, as well as novel analysis that can shed new light on the features of SVs and how to study them.

1.2 — Evolution of sequencing methods

Genomic sequencing is a valuable tool for biological research. Sequence analysis is responsible for the development of several bioinformatic fields. The relevance of its applications has gone beyond the scientific or clinical communities, gaining general social interest, as shown with the human genome. This has resulted in continuous use, which in turn has shown each generation's limitations, and new methods have been developed with either technical or monetary advantages.

The sequencing of the human genome was one of the most widely publicized scientific advances at the turn of the century. Built from a decade of international collaboration, with competition from a private enterprise, after the complete genomes from several model organisms. The release of the human genome draft was publicly announced on the same date by the US president and British prime minister at the time. Today, sequencing a human genome can be performed in a few hours, with a cost approaching or below a thousand dollars[11,59].

The first method for sequencing DNA, chain-termination sequencing, was invented by Sanger and collaborators in the mid-1970s. Its molecular principle is the separation of DNA molecules by length with single-base resolution using polyacrylamide gel electrophoresis. This method was used to sequence whole genomes, including the human one, but is nowadays mostly used to sequence small DNA molecules[60].

In 2006, about 30 years later, a new sequencing method was presented by a company now assimilated by Illumina. This next-generation sequencing (NGS) created millions of small molecules that could be sequenced in parallel, yielding more data[61]. Although other NGS

methods were created, such as Roche's 454 pyrosequencing and IonTorrent[62], Illumina's method remains the most versatile and widely used, although new NGS technologies are still being developed[61].

Although NGS was an advancement in sequencing, its main limitation was the length of produced reads, which do not exceed 300 bases. Even in this limited range, the read length matters. For example, software for SNV discovery has a false positive rate when using 150 bp reads lower than when using 125 bp reads, apparently because longer reads result in alignments with fewer reads aligned to multiple locations[63].

The use of Sanger and NGS allowed for two great breakthroughs in genomics: the assembly of the first draft of the human genome and the sequencing of any individual human genome at an economic cost of less than USD 1 000 (United States dollars).

The assembly of the human genome was achieved by the Human Genome Project (HGP), an international consortium that despite its name aimed to map the human and mouse genomes, as well as other model organisms with smaller and less complex genomes[64]. The International Human Genome Sequencing Consortium (IHGSC) was created as a consortium between 20 research centers from six countries[64]. They published a first draft in 2001[65], in parallel with a private enterprise[66]. A few years later, the HGP would release a more complete version of the human genome, addressing the limitations of both 2001 releases[64]. This showcases one of the issues of studying repetitive sequences, such as TEs: it is difficult to pinpoint the location of a sequence similar to many others in the same genome.

The cost of sequencing a single genome is also a relevant factor. The HGP was a resource intensive project, requiring more than ten years for 90% of the genome, and about three for an additional 9%, and had an approximate cost of USD 3 billion (not adjusted for inflation)[67]. Before its arrival, Applied Biosystems had had the monopoly on automated Sanger sequencing[59], which costed USD 2 400 per megabase[68], which amounts to approximately USD 10 million per genome (the human genome is approximately 3 Gb long). The release of NGS, a high-throughput method more appropriate for larger-scale projects technology, lowered costs considerably, in a manner that was nowhere near the previous biannual halving. Roche's pyrosequencing lowered the cost 240-fold, to USD 10 per megabase. The SOLiD platform would reduce the price by another two orders of magnitude to USD 0.13, and Illumina's technology would be the cheapest at USD 0.07. All NGS technologies self-report a per-base accuracy of at least 99.9%[68,69].

Regardless of the possibly hidden costs of the "USD 100 000 analysis", accounting for infraestructure and labour[70], sequencing became 1 000 times cheaper in 5 years, which made it possible for many more laboratories to adopt it. With more high-throughput data available, identifying variability and correlating it with traits, including clinical ones, reached a new scale[71]. Despite its near-perfect single base accuracy, the short reads of NGS could not detect

SVs as easily as SNVs[12].

1.2.1. Third generation sequencing

To improve this, two third-generation methods were introduced in the last decade: Pacific Biosciences' SMRT and ONT's nanopore sequencing[20].

Single-Molecule Real-Time (SMRT) sequencing, developed by Pacific Biosciences, produces long reads (approximately 150 kb long, which actually contain repetitions of a 5-60kb insert). It relies on a DNA polymerase that amplifies the product, as previous technologies[12]. Its workflow begins with the formation of SMRTbell templates through the circularization of double-stranded DNA fragments (Fig. 1.1). These templates are subsequently introduced into zero-mode waveguides (ZMWs), which are nanoscale reaction chambers integrated onto an SMRT Cell. Within each ZMW, a DNA polymerase incorporates nucleotides with fluorescent labels. The resulting fluorescence emissions for each added nucleotide are detected in real time, generating continuous long reads (CLRs). The circular template structure facilitates multiple passes of the polymerase over the same DNA molecule, enabling the generation of high-accuracy consensus sequences, which may be circular consensus sequences (CCS), or reads of insert (ROI). They are different in that CCSs require two full passes of the insert, while ROIs may be generated from even a partial pass, although they may be used interchangeably omitting this difference[72].

Despite using DNA polymerization for sequencing, library preparation for SMRT does not require amplification, which removes amplification bias as a source of distortion. Furthermore, the technology permits the detection of epigenetic modifications by analyzing the interpulse duration (IPD), which reflects the polymerase kinetics during nucleotide incorporation. This is possible due to DNA polymerases holding several nucleotides at the same time, and epigenetic changes present in one of them can alter the incorporation rate of the surrounding ones.[72]. Each modification has a different effect that needs to be characterized individually, and patterns for known events may also be detected[74].

The raw, continuous reads generated by SMRT have a high error rate (85-92% vs. 99.9%[12]), particularly in sequences containing homopolymers and repetitions of the same base. Creating CCSs with computational methods alleviates this problem, which allowed PacBio technology to be the first technology that offered reads that were both long (over 10 kb) and accurate (over 99% accuracy). However, this came at the cost of using shorter inserts, which allow for more passes, but result in shorter reads overall[12]. This is an acceptable exchange, and while continuous long reads were the standard in 2020[12], consensus reads are now the prevalent datatype for SMRT[72].

The distinctive characteristics of SMRT reads require specialized bioinformatics tools for downstream analysis, encompassing alignment and assembly procedures. First efforts included

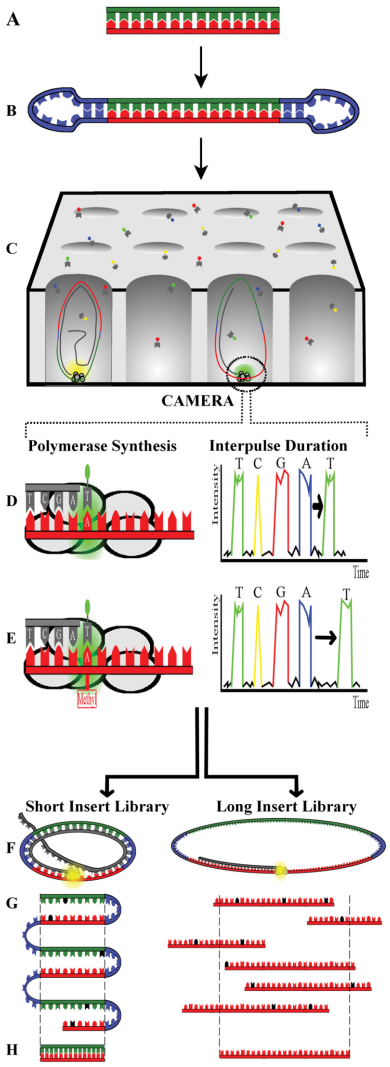

Figure 1.1. Molecular mechanism of SMRT sequencing. A) Initial double-strand DNA to sequence. B) Single strand hairpin adapters are added to circularize the DNA. C) The library of circularized DNA is loaded onto chambers with a DNA polymerase at the bottom (some chambers may not contain a polymerase). D) The polymerase adds fluoresecently labelled nucleotides. The emitted fluorescence is recorded. Both the fluorescence color and duration between each added nucleotide (interpulse duration, IPD) are registered. E) If the polymerase encounters an epigenetically modified base, the IPD changes. F-H) On the left, a small insert in sequenced with multiple passes, each conataining errors, which can be assembled into a circular consensus sequence (CCS) with high accuracy. On the right, a larger insert is sequenced with single passes. After sequencing, overlapping reads can be assembled into consensus sequences to reduce error rates[73]. Reproduced from Ardui et al.[73], which was published under the terms of the Creative Commons Attribution License, which allows non-commercial use with attribution (`http://creativecommons.org/licenses/by-nc/4.0/`).

adapting existing algorithms for alignment (such as BWA, Burrow-Wheeler alignment), which were first developed for short reads and did not perform well with much longer reads; additionally, they were not prepared to handle low accuracy reads[75], which were used for SNV discovery with tools for short reads[76]. Nowadays, there is much software designed for long read analysis compatible with both SMRT and ONT data, as discussed later in this thesis.

The other long-read technology is ONT's nanopore sequencing, which is the main sequencing technology used in the work that led to this dissertation. With nanopore sequencing,

a dual-stranded DNA molecule is linked to sequencing adapters, which are bound to a motor protein. Then, it is mixed with tethering proteins and loaded onto a flow cell that contains thousands of protein nanopores (which give the technique its name) placed into a synthetic membrane (Fig. 1.2). The tethering proteins lead the DNA to the pores, where the adapters enter the nanopore, after which the motor protein starts unwinding the DNA helix. The application of an electric current helps the motor protein move the DNA, which has a negative charge, through the pore. When DNA goes through the pore, it disrupts the current, with different groups of nucleotides presenting characteristic patterns[12]. Although this is the description for DNA sequencing, the same process is true for RNA sequencing, which can be sequenced directly by nanopores, while NGS and SMRT require DNA to be produced from it[21].

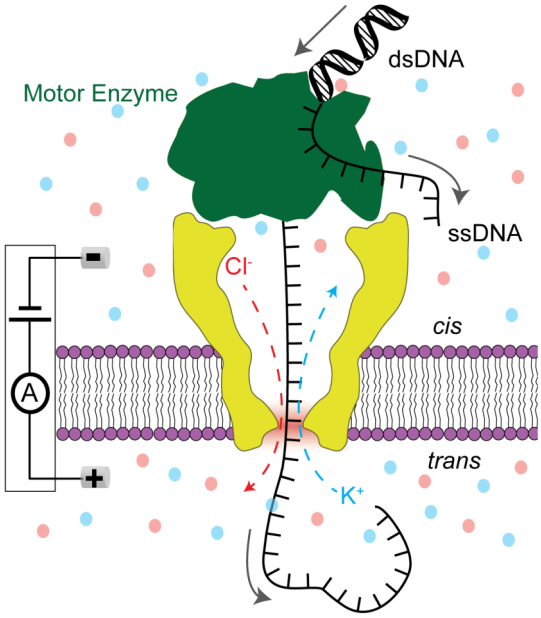

Figure 1.2. Molecular mechanism of nanopore sequencing. The nanopore protein (yellow) goes through a membrane. An electric current is applied to the membrane. DNA, which is charged negatively, moves from the *cis* side of the membrane, positively charged, to the *trans* side, negative charged, aided by a motor protein. The motor protein ensures less variation in the speed at which the DNA goes trhough the nanopore and uncoils dual-strand DNA into single-strand DNA. Charged ions (Cl^- and K^+, in this case) also move through the nanopore according to the charge[21,77]. Note that the membrane used in ONT devices is not a lipid bilayer, but an electrically resistant polymer, and ions are not necessarily the ones shown here[21]. Reproduced with modifications from Nova et al.[77], which was published under the terms of the Creative Commons Attribution License, which allows use with attribution (`https://creativecommons.org/licenses/by/4.0/`).

Despite reaching commercialization in the last decade, nanopore sequencing appeared in the 1980s, when it was discovered that certain bacterial membrane proteins could be used to detect variations in an ionic current when the channel was blocked by a nucleotide chain[21]. During decades, efforts have focused on characterizing nucleotide chains to differentiate them apart[78,79]. The first protein used as a nanopore membrane for DNA sequencing was an engineered α-Hemolysin originated from *Staphylococcus aureus*, although it was not able to resolve complex sequences. Later, similar performance would be achieved with porin A from *Mycobacterium smegmati*. However, it would not be possible to discriminate individual nucleotides in a sequence until the incorporation of a processive enzyme (i.e., an enzyme that catalyzes several

reactions consecutively without releasing its substrate[80], which in this case would be the nucleic acid) that would both reduce the speed at which DNA passed through the nanopore and make it more uniform. The protein used in this manner that produced the best results was the DNA polymerase of the bacteriophage Φ29. Changes in translocation speed add noise that hinders nucleotide identification, which limited the technology until the issue was reduced. This was achieved in 2012, when ONT announced the first nanopore sequencing device. The first nanopore sequencers would become commercially available in 2015, after having been released to early users in 2014. Although other nanopore sequencing technologies have been developed, ONT's products remain the most widely used in research[21]. Although alleviated, the technology's two main obstacles (non-constant speed, and multiple nucleotides contributing to the same signal) persisted[81], which result in low accuracy (87%-98%, from a work that reports 87%-92% for PacBio, and >99% for PacBio HiFi). However, the most recent chemistry toolkits produce reads with accuracy of 99%[82].. In exchange, it can produce longer reads (10-100 kb), and over 100 kb with a particular protocol for ultra-long reads, which can reach megabase length, although in both applications the longest reads are outliers[12]. The aforementioned ultralong reads have enable genome assembly and variant phasing[11]. As discussed in Section 1.3.2, evaluating the performance of variant detection with such data is complex, and the literature uses several criteria to present results. The performance of nanopore sequencing is of particular interest in clinical applications[29,54], as these applications also need to face limitations that arise in practical cases[54].

ONT sequencers combine a flow cell and a sequencer. The flow cell determines the number of channels available. Each channel can be measured individually and its current inverted if necessary, for example, to remove the molecule that is being sequenced[83]. A single channel is connected to four wells, structures that contain the nanopores (one well contains zero or more nanopores, ideally one). Before sequencing starts, pore activity is checked, and wells are ranked and sorted into four groups. In this manner, only a quarter of the wells are available at a time, and each channel communicates only to one well, switching groups every 24 fours of sequencing[84].

There are three types of flow cells, functionally differentiated by the number of channels (a higher number of pores available results in higher DNA yield, which ensures higher depth coverage), indicated by distinct shapes and sizes. The MinION cells hold 512 channels (2 048 wells), PromethION cells contain 3 000 channels (12 000 wells), and Flongle cells contain 126 channels[85].

The sequencing machine determines the type of flow cell employed and how many are accepted at a single time. The first device launched was the MinION, which accepts one MinION cell at a time. A second device is the Flongle adapter, which can be connected to a MinION sequencer to use Flongle cells. It produces fewer reads, but it is suitable for small and rapid applications, such as viral diagnoses[86]. Third, PromethION devices use their respective

flow cells and are fit for uses that require higher DNA yield, since they can accept 24 or 48 flow cells simultaneously[87].

Other models balance the yield of individual cells with that of the machine. For example, GridION X5 can accept five MinION cells at the same time, for applications where MinION depth coverage is acceptable, but a large number of sequencing experiments are still needed[88]. On the other hand, P2 is a smaller PromethION that accepts up to two flow cells, allowing applications that produce more DNA in individual runs, but do not need 24 simultaneous runs[87].

1.3 — Applications of sequencing methods

The sequencing chain of nucleic acids has an economic cost and has required the development of more than a dozen different technologies, each trying to solve the issues present in the previous ones, as has been discussed above. In this section, we will review the applications of sequencing that motivate these efforts.

1.3.1. Assembly of genomes

With LRS, it is possible to assembly a genome from a single sequencing experiment, generating assemblies that span more than 90% of the current reference, although methods still struggle to resolve repetitive sequences, with >97% of assemblies ending in such sequences[11].

Read length facilitates enormously genome assembly, to the point that Sanger reads with depth coverage in the range of 10-20X yield better results than the shorter Illumina reads at 100X. However, Sanger sequencing is more expensive, even in these coverage ranges, than high-throughput Illumina technology[89].

A measure of completeness of an assembly can be obtained by sequencing mRNA from different tissues to generate a catalog of genes whose presence can be checked in the assembled genome[90,91].

More efficient bioinformatic methods have allowed genomes larger than the human genome to be assembled[91]. Such advances include the efficient creation of string graphs[92], detailed in 1.4.3, part of the next section. An example of such larger genome is the axolotl's, spanning 32 Gb (more than 10 times the size of the human genome), 18.6 of which are repetitive sequences (65.6%).

In this manner, the use of a reference sequence has determined the design of experiments and bioinformatic tools[7]. However, this approach may have limitations. First, most references are derived from a single or a few individuals. This is the case with the Neanderthal Genome, although for that particular case DNA is very difficult to find, since they disappeared approximately 30 000 years ago, so the use of only three individuals[93] is justified. On the other hand,

the human reference originates from only 20 samples, although 70% of it derives from a single donor and does not represent a species-wide consensus, but more closely resembles an individual genome[94]. The T2T assembly was produced from a cell line derived from a complete hydatidiform mole (CHM), which has a completely homozygous genome[95] (a CHM is a tumor that originates from an ovum without nucleus by a sperm cell, with post-fecundation duplication of the genetic material, resulting in a homozygous genome[96]). As a result of using a reference genome, any sequenced DNA will have differences and similarities with the reference genome. This creates a bias in alignment that penalizes variants during alignment[97] and overrepresents reference alleles in high diversity regions, as has been observed with the *HLA locus*[98].

The proposed approach to solve this situation is the use of pangenome graphs, in which the genome is represented as a graph that can describe the diversity of the species or population. In such a graph, the nodes are sequences connected by edges. Following a path of edges and joining the visited sequences results in the full haplotype of a chromosome[7]. Several human pangenomes have already been built, usually reported notable amounts of genomic sequences not present in the reference, including protein-coding genes. In 2019, one was released from 285 sequencing experiments and 90 previously assembled genomes from Han Chinese samples, reporting nearly 30 Mb of newly discovered genomic sequences and almost 200 protein-coding genes. The project used NGS, which resulted in very short contigs, in the order of kilobases[99]. Pangenomes are also of interest for studying other species; for example, the Bovine Pangenome Consortium aims to create a graph genome that includes both different breeds of cattle and other bovine species[100].

1.3.2. Detection of variants

The knowledge of genetic variation has multiple applications. It can be used in studies of populations and demographic history, molecular mechanisms of variant genesis, and clinical impact of variants[28]. In other species, genetic variety can be useful in creating more productive breeds of crops or cattle[101].

The origin of CNVs has been linked to different factors, according to the size of the CNV. The smaller ones appear near VNTR (variable number of tandem repeats) and dispersed duplications, while the larger ones appear near NAHR (non-allelic homologous recombination) that occurs during meiosis[102].

Large deletions, which can span multiple genes, have been linked to different diseases. The deletion of two adjacent genes, *LCE3B* and *LCE3C*, is correlated with psoriasis, an inflammatory skin disease that affects 2-3% of Caucasian people. This implies either a factor that requires a specific allele of HLA (human leukocyte antigens, a cluster of highly variable genes related to the immune system) or by multiplying the effects of other mutations[103]. Another case is the *FCGR3B* gene, which is found with a median of two copies in the population, but carrying fewer copies results in a higher risk of glomerulonephritis and lupus, two autoimmune

diseases[104]. There are reports that CNVs increase the risk of intellectual disability associated with autism and congenital malformations[105], or intellectual disability with no other known molecular cause[106]. CNVs are also linked diseases of complex inheritance such as schizophrenia[107–109], bipolar disorder[107], or epilepsy[110–112]. The regions related to syndromic genomic disorders are rich in polymorphic CNVs, which may represent a factor in their inheritance[34,113].

Inversions that affect pathogenic genes also have a clinical impact. They can span a fraction of the gene, stopping its expression as a protein; or they may cover the whole gene, in this case, leaving it regulated by different promoters that normally control other genes[29].

Last, insertions of genetic material within a gene sequence can also impact its function. There is an association between TE insertions and autism[39], but also with monogenic diseases. The insertion of an *Alu* element within an exon of *BRCA1*, associated to breast and ovarian cancers, yields a transcript with a frame shift and a premature stop codon that does not result in a functional protein. This mutation causes hereditary breast and ovarian cancer[114]. TE insertions can also function as affect genes when occurring in introns by altering mRNA splicing. This has been observed for mutations that cause Lynch syndrome, which is an autosomal, dominant, hereditary predisposition to many types of cancers[115]; *BRCA2*, linked to breast cancers as *BRCA1* is[116,117]; and AT deficiency, by altering *SERPINC1*'s sequence[52].

In addition to its clinical importance, knowledge of genetic variance also has industrial applications. The economic impact includes the use of QTLs (quantitative traits *loci*), which are groups of mutations that affect complex traits, such as resistance to stress and diseases or crop yield. It is possible to establish correlation between mutations that have a small quantitative effect and perform selective breeding to create better-performing variants faster than using traditional breeding[118,119]. The same principle is also applied to livestock, with regard to milk production and composition in cattle[120]; quality of meat, weight and body composition in pigs[101]; or egg, feed efficiency, meat quality and disease resistance in chickens[121], for example. Another commercial use is the selection of a group of variants in a vegetal species to identify cultivars, hybrids for the purpose of registration, classification, and legal protection[122].

1.3.3. Confirmation of reported sequences

There are several instances in molecular genetics in which it is necessary to confirm a given nucleotide sequence. One case is the confirmation of detected variants, as discussed in Section 1.3.2. Since third-generation reads may contain errors, which can in turn affect alignment and variant calling, an independent technique is used to confirm that the variant does exist, and that its sequence is as described by the calling software. Sanger sequencing, although not easily scalable, is very cheap (USD 4 for experiment, covering 800 bp[68]) and can be used to confirm variants to assert the accuracy of a study, paired with PCR that will amplify the region of interest[123]. This can be achieved either by designing PCR primers that will cover the whole sequence of interest[123], or by designing primers that will only result in amplification if the called

insertion exists and consists of the reported sequence[124]. In both cases, sequencing reports a sequence that either verifies or denies the putative variant.

1.3.4. Study of methylation

In cells, there are changes in gene expression that are not caused by a change in a nucleotide sequence. These changes are known as epigenetic modifications[125]. One of the most studied epigenetic mechanisms is the methylation of DNA, which consists of the addition of methyl groups to different positions within the nucleotide bases.[126].

Epigenetic modifications vary by cell tissue, are different in cell cultures compared to non-cultured cells, are responsible for the inactivation of one of the X chromosomes in mammalian females, among other factors. They are of clinical relevance, as they are also related to cancer and DNA methylation sites are targets of some cancer drugs[127].

The classical manner to study the state of methylation is bisulfite analysis, performed with either microarrays or SRS, a technique that requires DNA preparation with limited conversion efficiency and allows for PCR bias, as well as other limitations. Nanopore sequencing can detect methylation comparably to bisulfite sequencing, without the issues of the latter[127]. Nanopore sequencing can be used to correlate changes in sequences with changes in methylation, as well as to examine differences in methylation between populations[11].

1.4 — Bioinformatic software

This section provides a review of existing tools for genomic analysis of data relevant for genomic analysis with nanopore sequencing. It also presents publicly available datasets and databases as resources that are used for the validation of methods developed. Special attention will be paid to those used or considered in the work presented in this dissertation.

The processes have been ordered as they are normally executed in a variant calling workflow. However, the first two subsections correspond to prior work that is necessary for the workflow: data repositories and benchmark datasets help developers of bioinformatic software developers design, test, and benchmark their tools; and variant calling requires an available assembled genome to be used as a reference. With these resources, pipelines start with alignment of sequenced reads, which may have been filtered in a quality assurance step, against the reference genome. Then, variant callers analyze the differences with the reference present in the alignment to report variants, which may be SNV and small indels, or SVs, with different programs usually dedicated to one of the two types. Finally, specific studies may require additional processing, such as gene annotation, genotyping, haplotyping, and coverage calculation, discussed in the last subsection.

1.4.1. Data, repositories, and datasets

Sequencing experiments generate large amounts of data[11]. Its management can prove difficult, and some works deal with compression and the design of more efficient file formats. However, they are codified, sequencing data need to be stored physically. Public repositories can make the data available to third parties, making it more reusable, and the research based on it easier to verify. These principles conflict with the privacy of the donors of the genetic material, in the case of human samples. Given that a genome is unique to a person (except for identical twins, which can also carry small differences[128,129]),it cannot be effectively anonymized. In addition, it is sensitive clinical information, as it may include mutations that are pathogenic or increase the risk of contracting a disease.

Data formats

In bioinformatics, information and data are coded in several formats. This section summarizes the most prominent ones that are used in this thesis, also organized in the order of a bioinformatic workflow.

Nanopore raw data correspond to the electric signals detected during sequencing. It was originally encoded in FAST5, a Hierarchical Data Format 5 (HDF5) format that used a schema created by ONT. For a typical experiment with human data, they had a total file size of approximately 1.3 TB, which made them difficult to store and slows down downstream analysis[130]. To improve raw data storage for nanopore sequencing, the ironically named SLOW5 was created, a tab-delimited format that contains the same information as FAST5 in a different structure. Like other bioinformatic data formats, it can exist in human-readable, ASCII-encoded format, or a compressed, binary indexed format (called BLOW5). BLOW5 reduced file size by 25% compared to FAST5 and enabled parallel access, which speeds up data access by a factor of 32, while the analysis time was reduced by 15 to 30 times[130].

Parallel to this academic effort, ONT developed their own format, pod5, based on APACHE arrow tables[131], which allows the storage of hierarchical data and the efficient access to them.Having been developed by ONT, it is generated by their software[131], and does not need to be converted from FAST5. Newer software that deals with raw data seems to be compatible with FAST5 and pod5 for basecalling[132] and methylation[127]. SLOW5 and BLOW5, on the other hand, have seen less adoption, mainly by RawHash2, a signal analysis tool focused on analysis of raw reads in real time[133].

To further process it, raw data are converted into nucleotide sequences in the basecalling process. Sequences are stored in FASTA or FASTQ format. In FASTA files, the entries are formed by a one-line header and one or more lines containing the sequence, with nucleotides encoded as letters using the IUPAC standard. FASTQ files are designed for high-throughput sequencing and contain 4 lines per entry: a header that contains a read ID, in a format deter-

mined by the manufacture of the sequencer or the database storing the sequence; the nucleotide chain, also using the IUPAC nomenclature; a line consisting of the addition symbol "+", and, optionally, the read ID again; and a last line that includes the quality of each base in the sequence. The quality values are limited to a range of 0-93 (both included), although 33 is added to them to convert numerical values into visible ASCII characters. Quality values are Phred-score values (Eq. 1.1), which makes the range of possible qualities range from 1 (a base that is known to be incorrect) to $10^{-9.3134}$. Since it is stored as ASCII characters, only integer values are allowed, which limits the amount of information; however, this is alleviated by including the first decimal value by multiplying by 10.

The Phred-score is calculated by multiplying by -10 the logarithm of the probability that the base is erroneously determined[134].

$$PHRED\text{-}score = -10 * log(P_e) \tag{1.1}$$

The FASTQ format is used to store sequencing reads from high-throughput sequencing experiments[134], while FASTA is used for other purposes, such as reference genomes and sequences[135]. FASTA dates from 1988[134], when it was used as the format of an eponymous software suite designed to align sequences and perform searches in protein and nucleic acid databases, while FASTQ was created at the Sanger Institute in 2000. Both formats were not formally defined, with the consequence of variants that look identical yet encode information in different ways, such as a different score metric still stored as ASCII characters[136]. Both files can be compressed and indexed to allow random access[137].

To store the alignment of reads against a reference sequence, the SAM format was created in the late 2000s. Unlike FASTA/FASTQ, it has a formal specification that lists possible values and metadata fields. A file header can be used to store the definitions of the metadata[138]. SAM is a tab-delimited format that uses ASCII characters (with UTF-8 characters allowed as exceptions). It consists of an optional header at the beginning and a list of the alignments, one per line. The header includes information relevant to the format version used to create the file, information on the execution of the alignment program, and the reference sequence. For the reference sequence, information can be stored such as the length of each sequence and their checksum. For alignments, coordinates in the reference sequence, as well as sequence and query names, are stored among other fields. Of particular interest is the CIGAR string, a summary of the alignment that can be used to locate a sequence within the read[138], to allow applications to retrieve insertion sequences, as done in Chapter 5. The CIGAR string consists of operations defined by an operation code and a length; for example, "M8" means that 8 bases match the reference sequence. Matches are for alignment, they mean that there are no insertions or deletions, and can, in fact, correspond to both sequence matches and mismatches[139].

Although SAM is stored as text, it can be encoded in binary form as a BAM file, which

has reduced file sizes and increases performance when reading. Both files can be sorted by genomic coordinates and, when sorted, BAM files can be indexed to allow random access[138]. A difference between SAM and BAM relevant for LRS is that the CIGAR string has a maximum length defined in the BAM specification (65 535 operations[139]), while SAM has no limit. Long reads with error may require more than this number of operations. Some aligners, such as the more recent versions of minimap2[140] store them as optional metadata, which has no size limits, as a workaround[141].

The CRAM format is an alternative to BAM that was created to reduce file sizes by storing only the difference with the reference sequence, instead of complete sequences. Additionally, instead of storing data by rows, like BAM, it compresses them by column, which is often the main factor in size reduction due to data in the same column being more similar, which is more effective in SRS than LRS[142].

Despite SAM, BAM, and CRAM having been designed to store alignments, the three formats support unaligned data, as an alternative to FASTQ[142].

For variations in a genome with respect to a reference, the VCF (variant call format) format is used. It also consists of a header with metadata and a body with information in a row-per-entry manner. Each row contains eight fixed fields with coordinates, an ID, quality information, and other optional information. In addition to that, a ninth column can define the format of subsequent columns, which are used to store information of different samples[143]. Other biological information can be stored in BED, another tab delimited format which was designed, initially to store data that data browsers with a graphical user interface (GUI) could load[144]. A variant of BED, bedGraph, stores quantitative data related to genomic intervals. The first three columns store genomic coordinates on the chromosome, start (0-based, included) and end (0-based, not included), and the fourth contains the variable, such as depth coverage of the sequencing[145]. In the case of this work, genomic coverage was stored in bedGraph files for the purposes described in Chapter 4.

Public datasets

New software is usually tested and benchmarked with a combination of simulated and real data[10,146–148], although some researchers also choose to use their own data[148]. Real world datasets come from different projects. This section discusses data that have either been used for this thesis, or were considered for it, but eventually were deemed inadequate.

The Human Genome Structural Variant Consortium (HGSVC) has used multiple technologies to create reliable SV datasets. In 2019, they released a study of three mother-father-child trios of different ancestry (Han Chinese, Puerto Rican and Yoruban). These samples offered different levels of genomic diversity and population admixture. The parents were part of the 1000 genomes project, a previous effort to sequence varied genomes with SRS. The children

were sequenced with a combination of SRS (Illumina) and LRS (PacBio, ONT); in combination with other genomic technologies, such as OGM. With the combination of technologies, they reported 818 054 indels smaller than 50 bp, and 27 622 SVs[149]. Both the sequencing data and the variant sets are publicly available and have been used during the design and evaluation of the works presented in this thesis.

In 2021, the second phase of the project released a more extensive project that focused on 35 samples from several populations. They reported 107 590 SVs, including 316 inversions and 9 453 TE insertions. Approximately 42% of the reported SVs are novel, i.e., not described in other LRS studies. A key improvement in this study is that they provide genotypes for variants[150], which can be used to evaluate tools in more detail[151].

Another notable initiative is the Genome in a Bottle (GIAB), a consortium hosted by NIST. They have characterized one sample from the HapMap project, an early study focused on detecting, genotyping and phasing SNVs[152], and two parent/mother/child trios. The trios are relevant, because in addition to the biological insight that can be gained from them in relation to genotypes or haplotypes, they have granted broad medical consent to use their genome data for academic and commercial redistribution, which makes their DNA easily available[26]. They have created several benchmarks[148,153] and guidelines[154] throughout the years for small mutations and SVs by characterizing these samples[155]. They have also released a set of relevant genes in autosomal hereditary diseases, although it contains few SVs (200) compared to 17 000 SNVs and 3 600 small indels[156].

A project similar to GIAB is the Chinese Quartet, consisting of two twin daughters and their biological parents. It studies small variants and SVs, and they report a total of 3 962 453 SNVs, 886 648 indels, 9 726 large deletions (≥50 bp), 15 600 large insertions (≥50 bp), 40 inversions, 31 complex SVs, and 68 *de novo* mutations, which are those present in the twins but not in the parents[157]. The data is available, but, unlike GIAB, it requires registration and filling out a form as principal investigator to be granted access.

Every dataset presents strengths and limitations that must be taken into account when deciding whether to include them as a benchmark. The datasets used in this thesis have been the ones produced by the two phases of the HGSVC (referred to in this section as HGSVC1 and HGSVC2). HGSVC1 is a well-rounded dataset with structural variants discovered by several tools, including LRS. It contains all types of SVs, including large CNVs, relevant for Chapter 4, and TE insertions, for Chapters 5 and 6. The latter is particularly relevant because it labels insertions that are classified as TE insertions. HGSVC2, which also contains all types of SVs, does not differentiate TE insertions from others, but does include genotype information and has a larger sample size. The HGSVC1 dataset was used for Chapters 4-6, while HGSVC2 was used only in Chapter 6.

Both HGSVC datasets are mapped against the current version of the genome (hg38), a

relevant consideration, since translating coordinates between different versions of the reference introduces a degree of error[158] unfit for a truth set. This made us decide against using the GIAB dataset, which is focused on small variants and has SV data from a single sample, aligned using hg19. GIAB also offers the dataset for clinically relevant genes, including SVs, although very few[156], which limits its usability for this work. Last, the Chinese quartet was discarded because of the access limitations.

Unlike benchmarks, there are public datasets that contain sequencing data from many individuals. One such example is the 1000 Genomes Project (which now leads the previously discussed HGSVC), aiming to provide a detailed description of human genetic variation. In 2015, it was considered complete after analyzing 2 504 genomes with SRS WGS and discovering more than 88 million different mutations, 84.7 million SNV s, 3.6 million indels and 60 000 SVs[159]. It started as a much smaller project, with low-coverage WGS of 197 samples originating from four different populations, high-coverage WGS of two mother/father/child trios, and exome sequencing for another 697 samples from seven populations, the three groups with SRS. At this initial stage, they reported 15 million SNVs, 1 million indels, and 20 000 SVs[160]. It was expanded from its 2015 state to 3 202 samples, 602 of which belong to father/mother/child trios. This version also used SRS WGS to generate the data[161]. However, there are current plans to sequence at least 800 of these samples with LRS, using both PacBio and ONT technologies. The results for the first 100 samples, from 5 different populations, have been published. In addition to releasing data and results to the scientific community, they communicate a series of claims, such as the readiness of LRS to detect SNVs and indels, for genome assembly, and to study SVs[11]. Since variants in samples from 1000 Genomes have been extensively haplotyped[159,161], this dataset presents itself as a large benchmark with fully described variants.

While these datasets are useful and broadly used, they are generated by a consensus of different techniques[149]. As a consequence, if a tool performs poorly on a type of variant, it may cause true calls to be missing in the dataset. This case is particularly relevant for SVs, which, as explained previously, are difficult to detect. This implies that it is difficult to generate complete variants, resulting in true calls being classified as false positives, and conversely, false variants in the dataset can result in false negatives. erroneous[25]. The Chinese Quartet Project claims to be a more true dataset because its design leverages Mendelian inheritance to discard erroneous variants[157], which does not necessarily address completion.

The IndiGen project has produced a map of *Alu* insertions found on the Indian population by analyzing 1 021 genomes[123]. This dataset was used in Chapters 5 and 6 to demonstrate the novelty of the data produced and the comparison capabilities of the presented software, respectively.

Another resource that hosts SV data is the the Genome Aggregation Database (gnomAD), a database of genetic variants, which contains 153 894,851 mutations (including 258 975 SVs) from 4 150 SRWGS samples generated by 1000 Genomes and HGDP. GnomAD has been released

incrementally and SVs aligned to the current reference (hg38) became available in 2024[162]. Other databases, run by the NCBI, are dbSNP[163] and dbVar[164], for SNVs and SVs, respectively. Although dbVar host files from HGSVC, the variants that it describes are submitted by the scientific community in general[164], so using it as a reference is not a possibility.

Last, it is worth mentioning the SRA[165], a repository for sequencing data that includes, for example, nanopore data from HGSVC2. It allows data to be searched by access status and experiment type, among other parameters.

1.4.2. Basecalling

The first step in a nanopore workflow is basecalling, the interpretation of the electric signals produced by the nanopore sequencer into nucleotide sequences[166,167]. There are several different programs that implement algorithms that do this, which can also be automatically executed by ONT stock software as part of the sequencing experiment, if the sequencer is connected to a computer or includes its own. Guppy was the former toolkit for perform basecalling separately with custom parameters[168], which has been deprecated in favor of Dorado, which runs different basecalling models depending on the origin of data and model of flow cells; in addition to utilities such as primer detection and removal from the sequence[169]. For developers to train models, bonito is used instead[170]. Both tools are not open source[127].

Detection of methylation is also done from raw data, thus its inclusion in this section. It can be done with ONT's tools (Guppy, Dorado), as well as with different ones from the scientific community: Nanopolish[171], Tombo[172], DeepSignal[173], Rockfish[174], or DeepMod[175]. Each tool is trained using neural networks, other machine learning algorithms or Hidden Markov Models (HMM) to detect one or more types of methylation, the majority focusing on 5-methylcytosine (5mC) (the most common DNA methylation in the human genome)[127].

The choice of basecalling software usually defaults to the newest model available at the time of the sequencing experiments[11]. For this thesis, model availability, compatibility with flow cells used, as well as availability of a GPU to run the optimal models, were considered.

1.4.3. Genome assembly

LRS allows assembling genomes from sequencing experiments as if it was the creation of a reference genome, similar to the projects explained in the previous section. During assembly, graphs are used. A graph is a representation of information consisting of points called vertices connected by lines termed edges[176]. Using graphs to represent multiple sequence alignments reduces noise in consensus[177].

The first efforts in assembly from LRS reads tackled both the need for error correction and the opportunity for *de novo* assembly. PBcR (PacBio corrected reads) was designed to work with PacBio data (which were released earlier). It polished the high error rate reads with

Illumina reads (requiring both technologies), making it possible to generate better assemblies than with SRS alone[89]. This procedure was quickly improved to require only an LRS experiment by using a hierarchical genome assembly process (HGAP). It consisted of using the largest reads as seeding data, to which shorter reads would be recruited and preassembled using a directed acyclic graph (DAG), in which nodes are connected with direction without forming loops. Then, the pre-assembled data could be chained with any long read enabled assembler. Finally, the total dataset would be reused to polish the final result[177]. This would be refined once more into MHAP (MinHash Alignment Process) by incorporating MinHash, a hashing algorithm designed to compare websites that had been proved to be applicable in other fields, to compress sequences before comparing them. The minimized reads would then be compared and progressively overlapped with each other[178]. In 2015, hybridSPAdes was published to assemble nanopore reads (or PacBio reads) with assistance of NGS data using de Bruijn graphs[179]. In a de Bruijn graph, the edges represent overlaps between sequences, while nodes are k-mers, words of k-nucleotides, and have been used to assemble with NGS data, also[180]. FALCON was presented as an optimized HGAP, once again for PacBio data and using a string graph[181], in which edges are the overlap between sequences (they represent a string of nucleotides), and nodes are one of the ends of a sequence[182].

Miniasm[183] used assembly graphs, similar to string graphs, with the set of vertices being forward- and reverse-complemented sequences, instead of the end of the reads.

Assemblers that use overlap are more robust against errors in sequences, although assemblers using de Bruijn graphs handle repetitive regions better. It produced better assemblies than miniasm and the others by handling repeats different if they were "bridged", i.e., completely covered by a single read, or "unbridged", the opposite case[184]. Canu, on the other hand, reunited the first methods for PacBio tools[89,177,178] with added support for nanopore and improved methods. It was reported to be better than FALCON and hybrid methods (the ones that require SRS and LRS), but it did not benchmark itself against newer options[185].

After running genome assemblers, it is possible to run certain polishers to reduce errors. This is performed often[21], despite articles reporting new assemblers seldom using polishers[179,183], posing them as optional[184], or providing their own[181]; yet studies use them[186,187]. Flye improved the situation on more complex genomes, while staying on par with Canu on simpler ones[188].

Shasta is a more recent assembler that optimizes assembly of homopolymers[189]. It is part of the pipeline Shasta-Hapdup, which yields diploid references. The phasing is relative to local contigs, however[190]. Both Flye and Shasta-Hapdup produce very accurate assemblies, but Flye is reported to produce longer contigs[11].

The importance of genome assembly is two-fold for this thesis. First, as commented, it is used to create reference genomes, required for the workflow used in this thesis. Second, it is possible to use *de novo* assemblies to call variants, by comparing the assembly with a reference.

Assembly-based methods require higher depth coverage and more computational power than alignment-based ones. However, they offer better for some variant types, but not for the more complex ones (translocations, duplications, and inversions)[191]. In the field of TE analysis, alignment and assembly-based methods have been used with *Drosophila melanogaster* data.

Assembly-based variant analysis on humans is currently limited by yield, in addition to the factors mentioned below. The assembly of a human genome with nanopore data requires a sequence of each sample using more than one flow cell, which incurs additional cost and sequencing time[11].

1.4.4. Alignment

Sequencing reads can be aligned against a reference genome to compare them and later process the differences. SRS reads can be aligned with tools such as BWA (Burrow-Wheelers Alignment), which implements three different algorithms depending on the type of sequencing equipment used. The first one, backtrack, is designed for very short reads up to 100 bp. The other two, SW (Smith-Waterson) and MEM (Maximal Exact Matches), accept reads in a length range from 70 to several thousands bp, beating backtrack in accuracy and speed in the overlapping range[75]. The best algorithm of the three, MEM, has been reimplemented to obtain identical results in a shorter time[192]. Prior to this optimization, it was also used to align long reads, but was replaced by programs specifically designed for LRS.

There exist three major long-read aligners: minimap2, lra, and NGMLR. Minimap2 is the one currently recommended by ONT and used in their stock workflow[193], including the human variation workflow[194], in addition to being the most recently updated. Minimap2 is also the only considered in recent benchamrks[147,195] and in newer pipelines[146].

In 2012, BLASR (Basic Local Alignment with Successive Refinement) was released to address the two key differences between SRS and LRS: read length and, at the time, much higher error rate. It generates a rough alignment first, which is refined via dynamic programming[196]. In 2013, BWA-MEM was published as a preprint. It was compatible with LRS, although the accompanying article[192], never went through peer review. Although nowadays long reads are those produced by newer technologies in the range of kilobases to hundreds of kilobases, initial SRS were designed for 36 bp reads, lengths which would become diminutive even by SRS standards, requiring the possibility of local alignments and gaps[192].

After the midpoint of the decade, new tools would appear focused solely on LRS. One was rHAT (Regional Hashing-based Alignment Tool) in 2016, designed for PacBio's SMRT, which used a seeded-and-extension approach. This method, which was shared by the previous state-of-the-art tools, consists in creating initial alignments with partial reads, the nominal "seeds", with the reference genome, which are then extended to full matches. The difference from its predecessors was the creation of a hash table with different windows of the genome, to be used

in the seeding step to select the window with most matches, then used as candidate sites for extension[197]. The same research group would release a year later LAMSA (Long approximate matches-based split aligner). It took into consideration that long reads could more frequently contain the breakpoint of SVs, relatively rare in the genome. LAMSA used read fragment as seeds longer than what the other method used, and build alignment skeletons from the approximate matches of those fragments in two steps. In the first step, it would assume the absence of SVs, and align the fragments against the reference, building a series of skeleton alignments consisting of matches and gaps. In the second one, it filled the gaps by classifying them into different types of split reads and processing them accordingly. It should be noted that LAMSA is indicated to be suited to PacBio and ONT data[198]. GraphMap, also from 2016, was an aligner tailored to nanopore reads, using a five-stage process that would incrementally reduce candidate numbers. It would start with gapped seeds, refine the approximate alignments, and finally construct alignments from the remaining candidates.

Another aligner released in 2016 is worth mentioning: minimap. It took different algorithms previously developed for long reads and created a program that created fast and accurate read-to-genome and genome-to-genome alignments. It performed well compared to BWA-MEM while being 50 times faster[183].

2018 would see the release of NGMLR, another long read aligner designed to reliably discover SVs. It created 254 bp seeds and aligned them against the reference, group the matches, and align the long segments with their own scoring method, which allowed for both relatively long indels and shorter ones (1-10 bp), likely erroneous ones. The highest scoring set of segments without overlap would be the resulting alignment[199].

Minimap2 was released the same year. Starting with a novel alignment algorithm that could fully leverage modern processors' capabilities[200], it would refine minimap's process to be 30 times faster than LRS aligners and more accurate alignments, as well as being optimized to align LRS mRNA sequences[201].

Two more aligners were built to solve particular shortcomings of minimap2[140]: Winnowmap2 and lra. Winnowmap2 was built on top of minimap2 (and for some cases it defaults to the minimap2 algorithm). It aimed to improve alignment to highly repetitive sequences, such as centromeres, by diminishing reference bias. This bias occurs when an aligner offers lower scores to a read that contains non-reference bases[97]. Reference bias is of special importance for variant calling in highly polymorphic regions. For example, the frequencies of reference alleles have been observed to be overestimated at the HLA locus[98]. Lra was designed by implementing a better chaining alignment for partial matches than minimap2's heuristic approach or NGMLR's SW-like dynamic programming solution. Lra was at one point the aligner recommended by ONT[53], but it has been replaced by minimap2[193,194].

The current state of the art in nanopore read alignment was reached after gradual im-

provements in minimap2 for situations in which it had underperformed, achieving the same performance as Winnowmap2, but being faster[140].

For this thesis, lra, NGMLR, and minimap2 (both before and after the improvements previously mentioned) were evaluated. For the specific purpose of CNV discovery, NGMLR performed better (Chapter 4). However, even at the time of those experiments, it was not updated and suffered from incompatibilities with newer versions of samtools. Particularly, it would produce few malformed alignments that would not pass samtool's assertions, resulting in runtime errors. For the purpose of detecting insertions, minimap2 is better (Chapter 5), particularly the newest versions, which solve issues with repetitive sequences and large insertions and deletions[140].

1.4.5. Variant calling

In this section, we present tools to detect genetic variation with sequencing experiments.

LRS is a relevant advance in sequencing technology because it improves SV detection[11]; however, it is also interesting to consider its capabilities for smaller variants.

An early methodology for detecting and genotyping selected SNVs consists of performing PCR of the regions containing them, sequencing the PCR products, and performing a SNV search using NGS tools such as bcftools[202,203].

Longshot[204] supposed a jumpstart in SNV detection with long reads, but a series of tools would further improve the situation. It started with Clairvoyante (released before Longshot in early 2019), a neural network SNV and indel caller. It achieved F_1-score near or above 90% for LRS, and above 98% for SRS. The results were better with PacBio data than with nanopores, with a difference of approximately 5%. Interestingly, they also reported variants detected by LRS but not SRS, although no confirmation was performed with other techniques[205].

The next year, Clair, Clairvoyante's successor, would outperform its predecessor Clairvoyante, Longshot, as well as the multipurpose Medaka in performance and speed, while staying marginally behind Google's updated DeepVariant (by $<0.2\%$ in F_1-score), yet being several times faster[206].

DeepVariant was integrated into a pipeline, PEPPER-Margin-DeepVariant that reportedly surpassed other SRS tools in SNV calling with short reads, as well as being better than Clair and Longshot for LRS, with the additional function of haplotyping variants. Metrics were better with PacBio reads than with nanopore reads, however[207].

Finally, Clair3, another iteration on Calirvoyante, was published. It outperformed all tools at low coverages (10x-20x)[208], which were realistic depth ranges to find in sequencing experiments[53]. At higher coverage, it performed comparably to PEPPER, but was faster. At this point, precision and recall for SNVs surpassed 99% for SNVs, and were 89% and 62,

respectively, for indels.[208]. Clair3 is the current method of choice for SNV discovery in the human variation workflow offered by ONT's epi2me pipeline[193].

To call SVs, the main focus of this work and one of the strengths of LRS, completely different tools are needed to interpret the data contained in long read alignments[199].

Prior to Sniffles, some SV detection methods for PacBio[209] were incorporated into SMRT-SV. Another program, PBHoney, was designed to detect SVs with PacBio and tested on *E. coli* samples[210].

Sniffles was released alongside NGMLR. By processing alignments against a reference, it could call all types of SVs. For this purpose, it followed three different criteria. First, with insertion or deletions within long reads, it could detect relatively small SVs entirely encompassed within a read. For example, deletions and insertions in the range of hundreds of base pairs (which, by their size, are SVs and not indels) can be called in this manner. Second, by considering split alignments, in which a read has chimeric alignments (i.e., it has more than one good quality, noncontiguous partial alignment). For example, a read that has two alignments separated by 100 kb would indicate a 100 kb deletion. Third, examining noisy regions, Sniffles was also able to report SVs. For all three cases, a certain RE (read evidence) threshold would be needed to report it, unlike the examples listed here (although it could be configured to report SVs with any minimum RE, including one read)[199].

SVIM came out after Sniffles. Similarly, it analyzed signatures within read alignments for deletions and insertions, as well as features between alignments, for larger SVs. These "SV signatures" would then be grouped into clusters by genomic location, to be finally combined and classified into SVs. It reported high accuracy and precision, performing better than other tools with runtimes similar to Sniffles[211]. Afterward, Sniffles would be updated to use multiple threads, while SVIM was not, making it slower than the alternative.

More SV callers would be released in the following years. For example, pbsv, PacBio's proprietary caller that works exclusively with PacBio data[212]. NanoVar, published in 2020, analyzes split reads and processes alignment characteristics through a neural network to characterize SVs[213]. It performed very poorly in a benchmark run by an independent team[195], which raised questions from the developers of NanoVar, citing another independent study (and their own initial publication) in their defense[214]. In the other benchmark, NanoVar performance was mixed, more frequently underperforming when compared to other callers[215]. In one of the works presented in this thesis, it was benchmarked for general SV and insertion recall and performed better than Sniffles and SVIM with minimap2 alignments[124].

CuteSV, another SV caller, also uses large insertions and deletions, as well as split alignments, to cluster them and classify into different types of SVs[216]. It performed comparably to Sniffles (and better than other alternatives)[195], although it was better for certain types of SV, for example, duplications[216]. Another tool, NanoSV[98], performed poorly with real data[215] and

when benchmarked in another study[195].

After some improvements, Sniffles was updated, became more accurate and faster, and the updated version was released as Sniffles2, published in 2024. It was reported to be more than ten times faster and 29% more accurate than its competitors on a broad range of depth coverage, on both PacBio and nanopore data, as well as including utilities for population genomics[147].

Related to this thesis, it is also worth mentioning PALMER[217], which detects exclusively mobile TE insertions, although it has a very slow runtime[151,218]. Additionally, it is limited to selected types of TE selected. These limitations are in part the reason why we developed a tool for TE detection, Retroinspector (Chapter 6).

To aid researchers in tool selection, benchmark studies are independently conducted and provide a concise summary of the state of the art[195,215,219]. They can also provide additional guidelines for the process, such as using the union of two variant callers for increased sensitivity and the intersection for greater specificity[219], or experimental factors that influence the calling process[195]. Nonetheless, researchers currently seem content with running Sniffles2 after minimap2[146,193], although it does not yield the most accurate results[151,220], and large projects, such as HGSV, do not follow this practice[11]. In the case of population analyses, using only Sniffles2, for best speed at good performance, is indeed recommended[221].

Benchmarks are usually performed with publicly available datasets[146,222,223]. Real data can also be used alongside simulated data[222,223]. However, while simulated benchmarks report great metrics[222], real applications find low accuracy for CNVs[54]. CNVs can also be detected using an older methodology based on SNV detection. THis method is based on estimating SNV genotypes. For this purpose, three different alleles are considered: the reference allele, the alternative one, or a deletion, which leads to 9 different genotypes. The deletion can be observed in two ways: if it is homozygous, it will be detected as missing data, which can be associated to a deleted sequence. If it is heterozygous, one of the other will be reported as homozygous. Using the frequency of alleles or the genotype of the parents, it is possible to calculate the probability of each genotype, taking into account the error margins. This enables selection of the most likely genotype[224]. Although this method was developed for SNV arrays, ONT's Human variation workflow also uses SNP calls produced with Clair3 to complement coverage data[194].

These studies show that accuracy is similar between callers, but there exist some differences between tools, as explained in the next paragraphs.

Currently, Sniffles2 is the fastest and perform better or as well as other callers across all types of insertions. It has addiotional functionality, too, such as finding somatic mutations and analyzing variants at population levels[147]. Its performance and versatility have earned it the position as the only general purpose SV caller in ONT's human variation workflow[194].

It is worth mentioning that, while early tools were sometimes limited to one or few SV types[225], current ones detect all types[211,216,222]. However, each caller may perform better for specific purposes; for example, SVIM performs worse for insertions when paired with NGMLR[55]. Other factors, such as read length or coverage, are not relevant at the values produced by current equipment[195]. Regarding coverage, there is only a 15% increase in recall when comparing 10x depth to 50x[25]. Even at 5x coverage, precision and recall can surpass 80% with Sniffles2 or CuteSV[226]. While this is true, higher coverage values result in better performance, with diminishing returns after 20X. Depth coverage has a greater impact on the detection of large SVs than relatively smaller ones[195].

Another aspect that differentiate variant callers is the type of SV. These programs perform poorly for CNVs[54], for example, and are substituted by other approaches in ONT's official methods[194], as explained previously. Most callers, paired with any aligner, accurately detect insertions and deletions[55], which are the most common and most reported SVs[53], but struggle with more complex SVs, such as inversions and duplications[55].

Variant callers perform other functions in addition to variant calling, such as genotyping variants[227]. Genotyped variants can be phased to determine which ones are found on the same chromosomes on a given sample. The length of nanopore reads allows phasing without pedigree information and can be done by genome assembly[190], with manual reconstruction[228], using SNPs called by Clair3[229], or simply from the reads[230].

A last consideration for SV detection is the search for polymorphic variants. Several *loci* in the genome are highly polymorphic, such as the *HLA*[22], CNVs[25], and tandem repeats[24]. Since LRS reports the sequence, it is possible to perform allele-specific studies, for gene expression[22], methylation[23], or to study TEs[10].

1.4.6. Annotation and other analyses

After alignment and variant calling, data may be interrogated in different ways to answer different biological questions. This subsection presents a list of miscellaneous software that has been useful in the development of this thesis, as well as other tools relevant to nanopore sequencing.

First, as mentioned in the previous section, it is useful to combine variant sets generated by different callers. For this, SURVIVOR[231], its wrapper surpyvor[219]. Other options also exist, Jasmine[232], which has been used by HGSVC[11], or combiSV, which requires the inclusion of certain variant callers in the workflow to work properly[215]. Truvari is designed to compare call sets and truth sets[233], and sees use in benchmarks[157,195].

Calculating depth coverage from a sequencing experiment can be a quality metric[166,234], but it may also be used during variant calling for genotyping; either by simply setting thresholds for the fraction of supporting reads relative to total reads in the SV region, such as Sniffles[222], or

with more refined statistical methods[147]. It can also be used as an approach to detect duplicates and deletions, particularly with NGS data, where SV detection is more limited[235–237]. Samtools can calculate coverage, but mosdepth is faster and supports more file types as output[238].

Separately, annotation of variants with gene information is useful in predicting their clinical or biological impact[239]. Standalone tools can perform gene annotation, such as AnnotSV[240], but it can also be done within the R programming language through packages such as annotatr[241]. Annotatr reports which variants overlap genes and in more detail whether they fall into exons, coding sequences, promoters, among other features. It can also detect overlap with non-gene features such as CpG islands and their surroundings. We have used annotatr to annotate TE insertions and deletions (Chapters 5 and 6). It allowed us to take advantage of the extensive set of annotations it provides while keeping the process contained in R, which simplifies building the workflow.

After annotation, enrichment analysis is an interesting possibility. With ClusterProfiler[242], another R package, it is possible to know if a set of genes that is, for example, affected by one type of SVs in a sample or cohort, is enriched with a particular feature, that is, the said feature is overrepresented in that set of genes. Such features can be molecular functions, biological processes or cell locations from Gene Ontology (GO)[243,244], metabolic pathways from KEGG (Kyoto Encyclopedia of Genes and Genomes)[245], cancers from the Network of Cancer Genes (NCG)[246], or other diseases from Disease Ontology (DO)[247] (with its own R library, DOSE[248]). Enrichment can help to interpret the results of annotation. On a genomic level, producing a list of genes affected by TE insertions is of limited use (if one is only interested in a single gene, or few genes, checking if they are affected can be the objective, however). However, obtaining a list of biological processes affected by genes significantly affected by TE can be more informative and can be used to connect the results to other studies, as done in Chapter 5, or to formulate a new hypothesis based on these observations.

Another task is the visualization of data. As has been explained before, many file formats used in bioinformatics follow a tabular design and are, to some extent, human-readable, even if they may be stored compressed. However, seeing genomic data, such as aligned reads or variants spread across genome coordinates, can be helpful to illustrate localized changes in characteristics such as methylation[11] or alignments[151,222]. IGV (Integrative Genomics Viewer) can load SAM, BAM, or VCF files, among others, and display them alongside chromosome coordinates. It can also load data such as repeats in the reference genome, gene annotations, and chromosomal bands[249]. Other alternatives exist, but IGV is the most widely used (and most usable) option for genomic visualization. Using a visualization tool is useful to provide examples on small differences between alignment produced by different methods[199], to show alignment features relevant for designing tools or fine-tuning them for specific purposes, such as somatic variants[147]; or, as has been the case in this thesis, to highlight alignment anomalies related to variant detection and analysis (Chapter 6).

Finally, installation of dependencies is not a problem particular to Bioinformatics, but it still needs to be managed. For this, the package manager can be used to create virtual environments with both libraries and tools, ensuring that the versions are compatible. These environments are reproducible and can be installed into any compatible machine without superuser permissions. Conda packages are downloaded from repositories called channels. Among them, there is one specific to bioinformatics, bioconda[250]. Conda channels contain libraries for several programming languages, such as C, C++, Java, Python, or R.

1.4.7. Pipelines

Performing all the steps described in this section can be complex and the experimental design can be time consuming. If one wants to see if a gene of interest is affected by any mutation in a given sample with a nanopore experiment, they would need to basecall their data, select a reference genome, choose an aligner and one or more variant callers based on their performance, run the annotation, and finally inspect the data either by reading annotated VCF files or displaying them in IGV. To ameliorate this and reduce the bioinformatic expertise required for some experiments, researchers and manufacturers develop pipelines.

We have already mentioned that ONT develops and publishes some pipelines for nanopore analysis. They can be installed from public GitHub repositories or launched by their desktop application[193]. These pipelines cover different applications for nanopore sequencing, the one closest to the topic of this thesis is the Human variation workflow, a one-stop tool to characterize SNVs and indels, SVs, and CNVs (separated from other SVs). It uses a selection of the tools previously discussed, namely minimap2 for alignment, Clair3 for SNPs and indels, and only Sniffles2 for SVs. For CNV reporting, it uses a fork of Spectre, an (as of now) unpublished tool for CNVs, aided by Clair3's results.

Pipelines connect tools and launch them with preset options, and they may include their own scripts for added functionality[146]. They can be implemented in different languages; in the scope of bioinformatics, several prevail: Nextflow[251], which uses Groovy (a dialect of Java); Snakemake[252], which uses a dialect of Python; WDL (Workflow Description Language), with a focus on portability to high-performance computers and cloud systems[253]; and Galaxy[254], which provides a GUI, as it is targeted to users without programming knowledge, and has its own package manager. On the other hand, Nextflow and Snakemake require programming instructions for the pipeline. The method each framework requires to define pipelines may condition its userbase: Galaxy is aimed to people without programming knowledge, while Snakemake uses a language commonly taught in Bioinformatic courses, while Java is more commonly used by software engineers. Pipelines can also be distributed as combinations of scripts, managing their dependencies with conda[255]. Packaging a workflow as a pipeline makes it very reproducible, as the frameworks manage the dependencies, reducing installation to writing few lines in the terminal. Then, by a single command, it is possible to run the entire

pipeline with controlled settings, achieving comparable results and reproducibility with no effort on the side of the user.

Examples of Snakemake pipelines that support nanopore data (either exclusively or among other technologies) include ONTdeCIPHER, for pathogen population genomics[256], RESCUE, for bacteria phylogeny[257], or RetroInspector, for detection and characterization of TE insertions and deletions[151], showing great adoption. Examples made with Nextflow include GraffiTE, also for TE detection[146]. WDL has been adopted by clinical and research institution, such as the NIH, or hospitals, and has been used for Shasta-Hapdup[190]. The use of pipelines in clinical environments is very interesting, as it allows to create standardized methods that can be applied by different centers. To find well-performing methods, several institutions, such as the FDA (Food and Drug Administration, from the USA), create competitions where pipelines are evaluated[258]. These competitions are based on truth sets similar to the ones, or the same ones, discussed in Section 1.4.1.

2

Challenges and Objectives

2.1 — Overview

The previous section has shown the current state of nanopore data analysis, the relevance of SVs, and their clinical impact with respect to AT deficiency. This section itemizes the goals of this thesis, which aim to examine the degree at which nanopore sequencing can answer questions about human genetics, both at the genomic level and focused on *SERPINC1*. AT deficiency is a hereditary disease with clear symptoms, usually caused by alterations in *SERPINC1*. This makes it a great starting point to study SVs, considering that these mutations explain a fraction of cases previously without molecular diagnosis[51].

After the introduction of AT deficiency genetics, it is clear that neither the traditional SV detection method nor NGS offers a complete solution on a genomic scale, while LRS may. These techniques, despite their limits, show the existence of large genetic variants in the human genome, which calls for the question whether they are also found with nanopore sequencing, to what extent and if the results leave room for improvement. With the knowledge of AT deficiency genetics[13] and the impact of SVs on *SERPINC1*[51], it is possible to define concrete objectives to improve SV detection methods. AT deficiency is a great case for genetic studies (it was the first trombophilia to be discovered as hereditary[48]) which eases method design. The methods can be checked on a controlled genomic region and extended to the rest of the genome.

Two groups of SVs appear more relevant for this research: CNVs, which can cause AT deficiency[51], and are difficult to detect with LRS[54]; and TEs, recently discovered to cause this disease[52].

The objectives described next were designed from a study of the scientific literature and were updated during the research process to answer new, specific questions.

2.2 — Objectives

The main objective of this thesis is the development of bioinformatic methods to detect and study SVs using data from nanopore sequencing experiments. Different types of SVs present

unique challenges and opportunities, and will be explored separately. The objectives stem from studies on AT deficiency[13,51]. The disease serves as a starting point for research questions, which are later extended to a genomic level, as presented in this section.

The following technical objectives and sub-objectives are defined according to the factors detailed in Chapter 1, named as *OBX*, which may be divided into subobjectives (*OBX.Y*).

- To examine the efficacy of nanopore sequencing for the detection of large SVs, which are known to be responsible for some AT deficiencies[51] (OB1).
  - *To compare variant calling with nanopore and CNV detection with aCGH.* With samples from a cohort of patients with AT deficiency, some of them carrying gross genetic defects affecting *SERPINC1*, find which CNVs with length over 50 kb are found by nanopore, and the agreement with the array. The size range is determined from the detection limits of the aCGH.
  - *To find methods to examine the sequencing data and refine nanopore results, if necessary.* Considering that nanopore technology is not fully mature, practical applications do not produce perfect results.
- To study the diversity of TE insertions in the human genome (OB2), since TEs are a novel source of pathogenic mutations in *SERPINC1*[52], they may be relevant for other diseases.
  - *To create and calibrate a workflow tailored to detect TE insertions (OB2.1).* The features of TE sequences, such as high repetition, may make them harder to detect. It is necessary to find which tools are best suited to detect a wide range of complete and incomplete TE insertions.
  - *To study the genetic impact of TE insertions (OB2.2).* With the knowledge that TE insertions can be responsible for monogenic diseases, even if they occur within introns, it is interesting to find how many TE insertions affect genes, which genes are affected, in which type of sequence (coding, intron, promoter, etc.).
  - *To study the functional impact of TE insertions (OB2.3).* After having verified that TE insertions are present in genes, even in coding sequences, we performed enrichment analysis to investigate the biological process that could be affected by polymorphic TE insertions.
- To study the diversity of TE deletions in the human genome (OB3). Considering that the reference genome describes mostly one genome, the sequence it contains is not a reference for the most common alleles. When deletions that span only the length of a TE, they can be considered polymorphic insertions present on the genome donor(s), but not on samples. This requires the same analysis described in OB2, except for OB2.3, which

would imply studying the enrichment of genes affected by TEs present in the reference but not in the cohort.

- To generate reproducible and reusable tools that can be widely used by the scientific community (OB4). There has been a shift in the scientific community in the recent years, with more emphasis on reproducible software[259], which makes research more reusable.

Chapter 4 focuses on CNV analysis and addresses objectives OB1 and OB4. Chapter 5 targets TE insertions and deletions, and tackles objectives OB2 and OB3. Chapter 6, which improves the methods from the previous one and implements them in a pipeline, revisits both objectives OB2 and OB3. Since its main focus is on the creation of a pipeline, it is also connected to the objective OB4.

3

Methodology

This thesis aims to improve the detection and analysis of structural variants with special attention to their clinical impact, focused on AT deficiency, using nanopore sequencing data. For this, the methodology follows the alignment-variant calling-analysis methodology reviewed in Section 1.4. The thesis includes the development of methods for CNV analysis in Chapter 4, and for TE insertions and deletions in Chapters 5 and 6. Both applications share the alignment and variant calling steps, as well as benchmarking and coverage calculation, and diverge in the end steps.

This chapter contains a summary of these common steps. However, the specifics are detailed in the Methods sections of the open source articles corresponding to Chapters 4, 5, and 6.

3.1 — Benchmarking tools

As showcased in the previous section, the development of software for the analysis of nanopore sequencing data is a continuous process. Thus, it is necessary to reevaluate which tools to use for alignment and variant calling.

Although the literature already offers updated benchmarks[195,215], they may be flawed[214] or not suitable for the particular purpose, such as large CNVs[53].

For benchmarking, we selected the truth set produced by the HGSVC[149], with sequencing data from either the HGSVC (the three samples) or one sample[219]. We used the more recent dataset, also by the HGSVC[150] to benchmark genotyping.

Benchmarks usually use both simulated and real data. Simulations are much easier to produce than other real truth sets and are both complete and error-free. However, simulation benchmark results rarely compare to real data[215,227], and occasionally produce near-perfect data, raising concerns about its validity[260]. There are several explanations for these discrepancies. First, simulated reads may be simpler than real ones[213,227]. Second, real SVs can be more complex than simulated ones[55]. Third, simulations reproduce what has been first detected by variant callers, so they are generated following their biases, which ultimately inflates their

performance[226]. Last, truth sets have also been proposed to be incomplete and to contain false positives, as discussed previously[25].

The process described in the next section was run for the truth set(s) and performance metrics were calculated. To compare or evaluate tools, three metrics are often used[147,195,215]: recall, also known as sensitivity and true positive rate, which quantifies the true positives recovered (Eq. 3.1); precision, also known as positive predictive value, which quantifies false negatives (Eq. 3.2); and F_1-score, which summarizes the other two (Eq. 3.3). These metrics have been reviewed elsewhere[261]. For Chapter 4, only recall was considered, since the comparison with aCGH presented a smaller and incomplete set of SVs, which made false positives difficult to quantify. This is common when dealing with large CNVs[54]. To address this limitation, the public truth set was used to determine false positives, which allowed us to calibrate a method that was able to reduce both false positives and negatives. This is a compromise made in similar studies[54]. Comparably, the benchmark for Chapter 5 aimed to maximize the recall of insertions, so that was the chosen metric. This did not create a risk for an excess of false positives, since all of the considered tools had performed very well on previous benchmarks[195,199,211,213,214,216]. Chapter 6 uses the three metrics mentioned here to evaluate the novel pipeline and compare it with its alternatives.

In addition to benchmarks with datasets, orthogonal validations was performed with our own data with aCGH for Chapter 6, with some of the CNVs have previously been validated[51]. PCR validation was performed for TE insertions as described in detail in Chapters 5 and 6.

$$Recall = Sensitivity = TPR = \frac{TP}{TP + FN} \tag{3.1}$$

$$Precision = FPR = \frac{TP}{TP + FP} \tag{3.2}$$

$$F_1 - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{3.3}$$

3.2 — Variant calling and subsequent analysis

This section lists tools used in the three articles for the variant discovery workflow, as well as others for later analysis. Each article part of this thesis has a more detailed Methods section, as well as an accompanying GitHub repository (linked within the article) with full code and commands used. A summary of the process is shown in Fig. 3.1.

Alignment was performed against the human genome reference hg38[262] using minimap2[140,201] with the recommended settings for nanopore data, and with the flag to store long CIGAR strings as metadata (once it became available), as discussed in Chapter 1, Section 1.4.1. Alignment files

were sorted and indexed with samtools[137]. Variants were called by an iteration of Sniffles[147,199] and SVIM[211], considering only reads with at least a mapping quality of 20 (30 for Chapter 5). SURVIVOR[231] or its wrapper surpyvor[219] were used to create unions of call sets, to analyze later with R[263] and Python scripts. Libraries for these programming languages usually have specific scopes and are not present in all articles, with some exceptions, such as data.table[264].

Tools were selected by evaluating state-of-the-art and new (at the time while we worked on Chapter 4) tools for the purpose of each article, with the benchmarks summarized in the previous section and described in more detail in Chapters 4-6. We normally used the default parameters of each tool, selecting presets for nanopore data for the pieces of software that offer them. One of the most changed parameters during this work is the minimum read thresholds to select a variant, set to 3 reads, and changed in Chapter 6 to 5 and 7 in addition to 3, to evaluate how it affected performance. The value of 3 was selected by examining mutations in *SERPINC1* that we had previously confirmed. Second, dealing with large CNVs in Chapter 4 required removing the size limit of reported variants, which CuteSV normally uses. Third, for the purpose of TE analysis, we enabled reporting the IDs of supporting reads, to latter assemble a consensuns from them. Last, we generally set callers to ignore reads with a mapping quality lower than 20 (the default value, except NanoVar). Since NanoVar does not support a mapping quality threshold, we filtered alignment files to simulated this option. This threshold was changed to 30 for Chapter 5 because the idea of incomplete TE insertions raised concerns during peer-review, particularly for those found within other TE sequences. It should be noted that TEs are commonly found within other TE sequences in eukaryote genomes[265]. After proving that these insertions, which we also validated orthogonally, were also reported with the more stringent value, we returned to the default value for subsequent experiments.

Coverage was calculated with mosdepth[238], and used to check the presence of CNVs (Chapter 4) or to genotype TE insertions and deletions (Chapters 5 and 6).

To address errors nanopore sequences may find with repeats within TEs, we reassembled the isnerted sequences from aligned reads as described in Chapters 5 and 6.

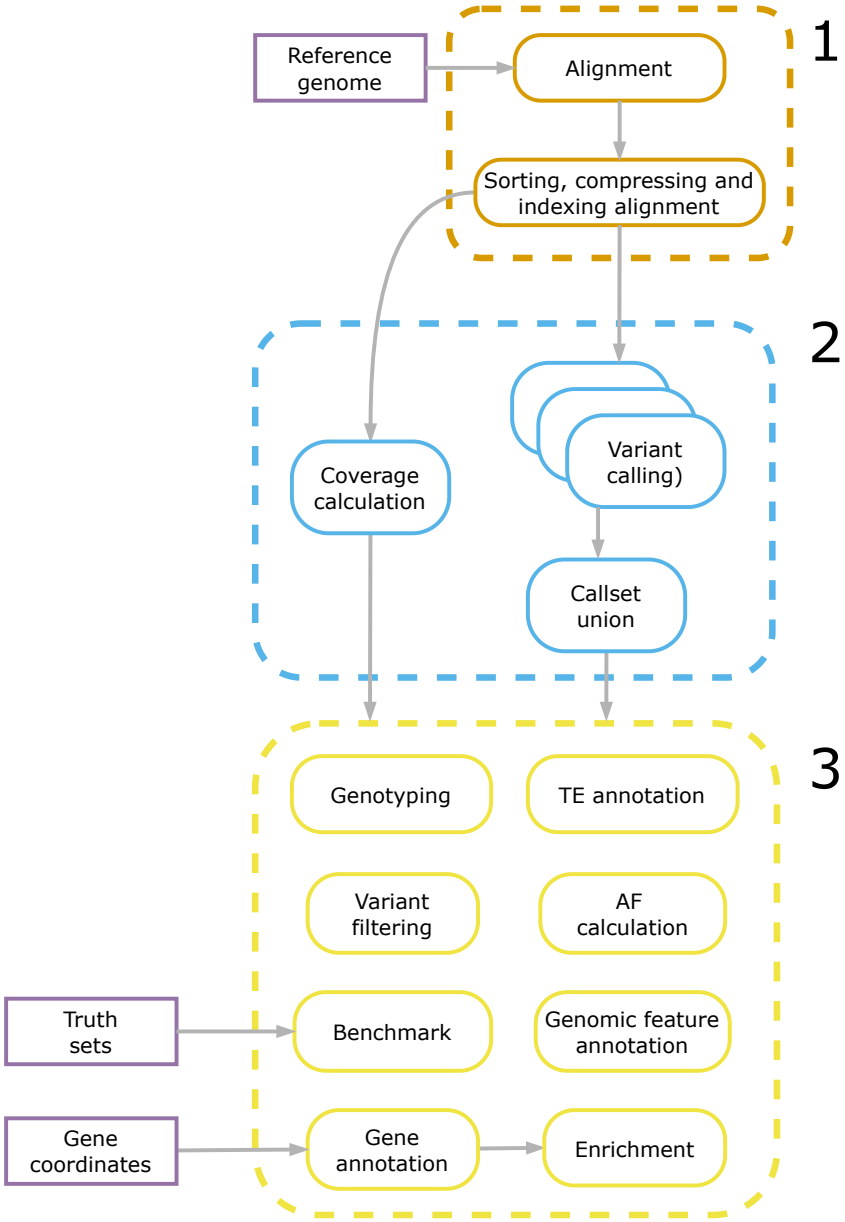

Figure 3.1. Summary of bioinformatics workflow used for the works in this compendium. Boxes 1, 2 and 3 group steps as seen on section 1.4 (they correspond to subsections sections 1.4.4–6). Boxes with angular corners are input files normally generated by an external project, such as the reference genome and its annotation. Boxes with round corners represent analysis steps. Color is used to restate information provided in other form: purple, externally produced input; orange, blue and yellow, steps 1, 2 and 3, respectively. Arrows indicate the order of the steps. 1) Alignment and postprocessing of alignment file (compression, sorting and indexing save storage space and allow for arbitrary access). 2) Processes that derive information for the alignment file: coverage calculation, and variant calling, usually done with several programs whose results are merged into an union set. 3) Specific analysis to variant callsets (and in some cases, also use the alignment files): genotyping (which may be performed during variant calling instead), filtering of variants with credibility thresholds or statistical parameters, calculation of allele frequencies (AF), benchmarking of tools and pipelines, gene annotation and subsequent enrichment, annotation of other biological features (such as TE sequences or genomic features like CpG islands).

4

Improvement of large copy number variant detection by whole genome nanopore sequencing

Article Information

Title: Improvement of large copy number variant detection by whole genome nanopore sequencing
Authors: Javier Cuenca-Guardiola, Belén de la Morena-Barrio, Juan L. García, Alba Sanchis-Juan, Javier Corral, and Jesualdo T. Fernández Breis
Published in: Journal of Advanced Research
ISSN: 2090-1232
Publication date: 2023
Volume: 50
Pages: 11152–11165
Article type: Open Access Research Article
DOI: https://doi.org/10.1016/j.jare.2022.10.012
Contributions of Javier Cuenca Guardiola: contributed to the experimental design during the conceptualization stage, designed and implemented the bioinformatic methodology, performed the data analysis and wrote the manuscript.

Abstract

Introduction: Whole-genome sequencing using nanopore technologies can uncover structural variants, which are DNA rearrangements larger than 50 base pairs. Nanopore technologies can also characterize their boundaries with single-base accuracy, owing to the kilobase-long reads that encompass either full variants or their junctions. Other methods, such as next-generation short read sequencing or PCR assays, are limited in their capabilities to detect or characterize structural variants. However, the existing software for nanopore sequencing data analysis still reports incomplete variant sets, which also contain erroneous calls, a considerable obstacle for the molecular diagnosis or accurate genotyping of populations.

Methods: We compared multiple factors affecting variant calling, such as reference genome version, aligner (minimap2, NGMLR, and lra) choice, and variant caller combinations (Sniffles, CuteSV, SVIM, and NanoVar), to find the optimal group of tools for calling large (>50 kb) deletions and duplications, using data from seven patients exhibiting gross gene defects on *SERPINC1* and from a reference variant set as the control. The goal was to obtain the most complete, yet reasonably specific group of large variants using a single cell of PromethION sequencing, which yielded lower depth coverage than short-read sequencing. We also used a custom method for the statistical analysis of the coverage value to refine the resulting datasets.

Results: We found that for large deletions and duplications (>50 kb), the existing software performed worse than for smaller ones, in terms of both sensitivity and specificity, and newer tools had not improved this. Our novel software, disCoverage, could polish variant callers' results, improving specificity by up to 62% and sensitivity by 15%, the latter requiring other data or samples.

Conclusion: We analyzed the current situation of >50-kb copy number variants with nanopore sequencing, which could be improved. The methods presented in this work could help to identify the known deletions and duplications in a set of patients, while also helping to filter out erroneous calls for these variants, which might aid the efforts to characterize a not-yet well-known fraction of genetic variability in the human genome.

5

Detection and annotation of transposable element insertions and deletions on the human genome using nanopore sequencing

Article Information

Title: Detection and annotation of transposable element insertions and deletions on the human genome using nanopore sequencing

Authors: Javier Cuenca-Guardiola, Belén de la Morena-Barrio, Esther Navarro-Manzano, Jonathan Stevens, Willem H. Ouwenhand, Nicholas S. Gleadall, Javier Corral, and Jesualdo T. Fernández Breis

Published in: iScience

ISSN: 2589-0042

Publication date: 2023

Volume: 26

Pages: 108214

Article type: Open Access Research Article

DOI: https://doi.org/10.1016/j.isci.2023.108214

Contributions of Javier Cuenca Guardiola: contributed to the experimental design during the conceptualization stage, designed and implemented the bioinformatic methodology, performed the data analysis and wrote the manuscript.

Abstract

Repetitive sequences represent about 45% of the human genome. Some are transposable elements (TEs) with the ability to change their position in the genome, creating genetic variability both as insertions or deletions, with potential pathogenic consequences. We used long-read nanopore sequencing to identify TE variants in the genomes of 24 patients with antithrombin deficiency. We identified 7 344 TE insertions and 3 056 TE deletions, 2 926 were not previously described in publicly available databases. The insertions affected 3 955 genes, with 6 insertions located in exons, 3 929 in introns, and 147 in promoters. Potential functional impact was evalu-

ated with gene annotation and enrichment analysis, which suggested a strong relationship with neuron-related functions and autism. We conclude that this study encourages the generation of a complete map of TEs in the human genome, which will be useful for identifying new TEs involved in genetic disorders.

6

Advanced analysis of retrotransposon variation in the human genome with nanopore sequencing using RetroInspector

Article Information

Title: Advanced analysis of retrotransposon variation in the human genome with nanopore sequencing using RetroInspector

Authors: Javier Cuenca-Guardiola, Belén de la Morena-Barrio, Javier Corral, and Jesualdo T. Fernández Breis

Published in: Scientific Reports

ISSN: 2045-2322

Publication date: 2025

Volume: 15

Pages: 14489

Article type: Open Access Research Article

DOI: https://doi.org/10.1038/s41598-025-98847-7

Contributions of Javier Cuenca Guardiola: contributed to the experimental design during the conceptualization stage, designed and implemented the bioinformatic methodology, performed the data analysis and wrote the manuscript.

Abstract

Transposable elements (TEs) make up 45% of the human genome, are a source of genetic variability difficult to detect, and involved in processes related to gene regulation and disease. Nanopore sequencing is recognized as one of the best technologies for detecting TEs; however, tools for analyzing of human TE insertions and deletions with nanopore-based data can be improved.

RetroInspector is an easy to use, configurable Snakemake pipeline that performs detection, annotation, enrichment, and genotyping of TEs. RetroInspector requires the FASTQ files of

the samples and the reference genome to start the identification and analysis of TEs.

The user can also set the threshold for the number of supporting reads for the variant filtering. RetroInspector also allows users to compare the results of two samples. Different versions of the reference genome can be used and the presence of retrotransposition features can be annotated. RetroInspector has been run on three nanopore sequencing datasets and validated experimentally using proprietary and public data with over 80% precision.

7

Discussion and Conclusions

7.1 — Discussion

Sequencing technologies have enabled the discovery of genetic causes in many diseases, as well as traits of economic interest in crops and livestock. LRS enables studies of genomic features, such as methylation and, the focus of this thesis, the detection and characterization of SVs. SVs, mutations over 50 bp, are a previously untapped source of genetic diversity that represents most of the genetic differences, in sequence length, between individuals. This thesis reunites the results of three articles that achieve the objectives defined in Chapter 2. Here, we summarize the discussion of the three works. Furthermore, since we are writing this after the publication of the individual works, we will also briefly comment on their impact in other scientific works. We do this only for Chapters 4 and 5, since Chapter 6 was published shortly before this document.

The first article (Chapter 4) addresses objective OB1 (to examine the efficacy of nanopore sequencing for the detection of large SVs). In that work, we compared aCGH and nanopore sequencing to examine the accuracy of the results of different variant callers. Four different variant callers were tested with alignments by three different aligners, resulting in 12 different callsets. Sniffles and NGMLR yielded the better results for large CNVs. Regardless of the combination of tools, recall and precision were very low in both our data and the public dataset, which agrees with similar work that shows how limited variant callers at detected CNVs[54]. As an additional problem, many large CNVs were detected. The nature of some of them, for example, large deletions affecting a sizeable fraction of entire chromosomes (which would probably have pathological consequences), pointed that they were not real variants, but false positives from the calling process. Although the correctness of truth sets has been questioned for CNVs (when using a different dataset)[25], these findings indicate that variant callers do report errors. This was checked with one of the samples from HGSVC that already had a curated set of SVs. After completing subobjective OB1.1 (to compare variant calling with nanopore and CNV detection with aCGH), it was undertood that it was not enough to complete the work regarding the discovery of CNVs with nanopore, which resulted in the the addition of OB1.2 (to find methods to examine the sequencing data and refine nanopore results, if necessary), consisting of improving the methods existing at the moment. We selected the approach of checking depth

coverage on the regions affected by putative SVs, based on the fact that if a region has a different number of copies, it will affect the number of reads aligned to it. For example, a homozygous deletion should result in no reads from the deleted region, while a heterozygous one would result in approximately half the reads normally expected. Conversely, a duplication should increase coverage, as a longer genomic region is represented by a shorter fragment in the reference. A method to compare coverage of putative SVs and their surroundings was implemented, and package into the tool disCoverage. This tool is not a variant caller, but a more limited program, as it needs a list of variants to filter. It is intended to run after variant calling, to remove spurious results and validate true positives; or it can also accept a list of polymorphic insertions as input, to check if they are present on another sample. However, even after running disCoverage, the results contained errors, meaning that, although the problem is alleviated, it is not solved. While having less spurious calls can help with molecular diagnosis of AT deficiency or other diseases, there is still work to be done. We packaged the tool into a conda package, to ensure that the installation process is straightforward, which is related to OB4 (to generate reproducible and reusable tools that may be used broadly by the scientific community). The article in Chapter 4 has been cited by other researchers because it shows the poor performance of variant callers for CNV detection[266], as well for its systematic benchmark that suggests Sniffles as the best tool among existing options[267], although researchers opted to use its update, Sniffles2, which was not available at the time of our work. It has also been cited for displaying methods that can be used to shed light into genetic alterations[268,269].

The second article (Chapter 5) addresses objectives OB2 (to study the diversity of TE insertions in the human genome) and OB3 (to study the diversity of TE deletions in the human genome). That publication showcased the capabilities of nanopore sequencing to study TEs. We investigated TE insertions and deletions in a cohort of 24 patients diagnosed with AT deficiency. Variant callers were selected by their insertion recall with another benchmark. Using both the selected tools and our own programs, we detected over 7 000 TE different insertions, with coordinates, sequence, and allele frequency within the cohort. To fulfill subobjective OB2.1 (to create and calibrate a workflow tailored to detect TE insertions), in addition to the previous analysis, we studied the distribution of insertions across chromosomes, finding that some chromosomes have regions richer in TE insertions; compared our results with other cohorts[123,149], discovering that many of the TE insertions we reported were novel ones; and found that TE most TE insertions occur into other TE sequences. We performed experimental validation with PCR, confirming the exonic insertions, incomplete insertions, and the ones found within other elements, discarding the possibility of them being sequencing artifacts. Gene annotation was performed to fulfill subobjective OB2.2 (to study the genetic impact of TE insertions). Nearly 4 000 genes were found to contain at least one insertion, most of them on introns (including a pathogenic one on *SERPINC1*), although we did find 6 TE insertions in coding sequences. We conducted enrichment analysis on the affected genes, and found significant present of neuron development genes with Gene Ontology[244] and genes related

to autism with Disease Ontology[247], autism having been connected to TE polymorphisms[39]. The enrichment process, which supposed tackling subobjective OB2.3 (to study the functional impact of TE insertions), also included the ontologies KEGG[245], which showcased synapsis, and NCG[246]. An analogous process of detection and gene annotation was performed for TE deletions, i.e., deletions carried by samples in the cohort that have similar coordinates and length to TE sequences in the reference. This showed ∼3 000 TE polymorphisms that are present in the reference, but not found universally. This work showed how TE variants can affect introns, such as the *SERPINC1* mutation, which is pathological[52], and in exons, where functional impact is clearer. While it shows the clinical potential of TE studies, backed up by other works[39], it also points a technical limitation. Two patients carried the TE insertion in *SERPINC1*, yet it was only reported in one of them. It remained a false negative in the other due to sequencing circumstances: only a read with the insertion was sequenced, which stopped it from passing the minimum read threshold. If coverage on a normal sequencing experiment improved, this would be less likely to occur. Other researchers have cited this article as an example of how TE deletions are being studied, and how doing so would improve their own work[270]. It was also used as reference for which distance limit to set when merging variant calls[271], a factor which is not standardized[151], and was further explored in our next study. The benchmark for TE detection has also been informative for another SV work[272].

The third article built upon the process described in the second one. It focused into objective OB4, presenting Retrosnake, a Snakemake[252] pipeline that, with minimal input from the user, performs what has been described in the previous work. To further address objectives OB2 and OB3, we included Sniffles2[147], which was in preprint during the process of the second article. We added it as an option and as one of the default callers. We also compared our workflow with other tools. In our benchmark, Retrosnake performed better than PALMER[217] and similar to GraffiTE[146]. In addition, RetroInspector detects more TE types than PALMER and is faster (PALMER is known to need over 48 hours per sample[273]). Unlike GraffiTE, it performs gene annotation and enrichment analysis, as an additional feature. A final improvement we included was the option to skip the annotation and enrichment process, so other versions of the reference genome, currently lacking annotation, such as T2T[95], could be used. RetroInspector generates both reports with plots and tabulated information, accessible to users with little bioinformatic experience, as well as standard formats (such as VCF files), for users able to include them in their own downstream analysis. GraffiTE only produces VCFs (and graph-based output, if enabled), while PALMER generates poorly documented tabulated text files. Creating a tool that is both easy to use and install should ease its adoption in molecular diagnosis, for either AT deficiency or any other genetic disease. Another contribution to the literature in this work is using a pair of identical twins to explore the effects that distance limit when merging variants has in accuracy, which, as discussed before, is not an obvious choice[271].

Last, it is worth noting the sample size of our experiments. The cohort for Chapter 4 was smaller, with seven patients. However, it consists of patients from an infrequent disease[13]

who carry rare mutations, which complicates patient recruitment. The cohort for Chapters 5 and 6 are larger, and comparable to other studies, as discussed there. Despite this, they are smaller than national and international efforts that include thousands of participants[11,123]. Since RetroInspector is efficient and easy to deploy, it should be possible to expand it to a larger cohort, if we find groups with whom to collaborate.

7.2 — Further work

The topics of this thesis are of current interest to the nanopore and molecular diagnosis communities.

In the field of CNVs, ONT has adopted the tool Spectre, which detects large CNVs by coverage analysis aided with SNV calls, and lacks, at the time of writing this, an accompanying peer-reviewed publication[274]. This suggests that the current Sniffles2 based approach has room for improvement in real-world applications dealing with CNVs. In Chapter 4, we showcased how tools underperform with large CNVs, and after applying disCoverage, false positives and negatives remained present in the results. A reevaluation of the state of the art with this specific application, including comparison with Bionano's OGM, could be interesting to address limitations, and update disCoverage or create a successor tool for CNV analysis. Bionano would be of particular interest, since it is as very precise method[260], that is, it has very few false positives, yet it discovers many more variants than aCGH[53,260].

Regarding TEs, there are already efforts to classify human genetic diversity[11,150]. It is compelling to use RetroInspector to create a database of TE insertions and deletions that includes samples in addition to the extensively used 1000Genomes Project. With a robust catalog of TE variants, it would also be possible to create an approach to a pangenome. Pangenomes are proposed as the next stage of bioinformatic genetics[7,100,146]. Creating a pangenome from such a TE database would allow not only to represent known variants but also the cataloged allele frequencies (which RetroInspector generates) and even clinical annotations. An extensive catalog of TE variants has clinical interests. Currently, it is difficult or not possible to check the clinical annotation of a TE variant, or even its allele frequency. When looking for new variants related to complex diseases, or with no molecular explanation, many candidates are present in the human genome. A simple method to reduce candidates is to check their frequency: if a variant is common in any given population, it is unlikely to be pathogenic.

In addition to this, TEs are relevant in health for several reasons. For example, *Alu*-rich genes show increased expression under interferon presence[275] and some *Alu* sequences can bind to an estrogen receptor that acts as a transcription factor[276]. When transcribed, *Alu* RNA can affect other RNA's transcription and translation via several mechanisms, including splicing and miRNA production[277]. *Alu* elements are normally hypermethylated, and thus, its expression is inhibited, probably due to being very rich in CpG sequences, although they appear demethy-

lated in tumors[278]. A comparison of newborns and elderly people found more *Alu* and MIrs in the latter[279]. SVAs are also hypermethylated, but they contain ceratin motifs that can act as enhancers of regions nearby their 5' end[280], an effect that MIRs also have[277]. Since methylation can be analyzed from routine nanopore sequencing, updating RetroInspector to include methylation analysis would allow testing whether, for example, somatic insertions can act as cancer drivers, either via producing nonfunctional transcripts or proteins, or by regulating expression. Some work has been conducted exploring these ideas[281], but a reproducible framework that acts at the genomic level could reveal more information, using, for example, the Network of Cancer Genes[246].

Besides further research, there is also the possibility of presenting the tools produced in public forums, such as ONT's annual conference, London Calling. RetroInspector has already caught ONT's attention, and inclusion into epi2me has been discussed. This would also help with the recruitment of collaborators for the goals described above.

7.3 — Contributions

7.3.1. Publications Indexed in Journal Citation Report (JCR)

Part of this thesis

- Cuenca-Guardiola, J., de la Morena-Barrio, B., García, J. L., Sanchis-Juan, A., Corral, J. & Fernández-Breis, J. T. Improvement of Large Copy Number Variant Detection by Whole Genome Nanopore Sequencing. *Journal of Advanced Research* **50,** 145–158 (Aug. 2023). Doi: `https://doi.org/10.1016/j.jare.2022.10.012`.

- Cuenca-Guardiola, J., de la Morena-Barrio, B., Navarro-Manzano, E., Stevens, J., Ouwehand, W. H., Gleadall, N. S., Corral, J. & Fernández-Breis, J. T. Detection and Annotation of Transposable Element Insertions and Deletions on the Human Genome Using Nanopore Sequencing. *iScience* **26.** (2023) (Nov. 2023). Doi: `https://doi.org/10.1016/j.isci.2023.108214`.

- Cuenca-Guardiola, J., de la Morena-Barrio, B., Corral, J. & Fernández-Breis, J. T. Advanced Analysis of Retrotransposon Variation in the Human Genome with Nanopore Sequencing Using RetroInspector. *Scientific Reports* **15,** 14489 (Apr. 2025). Doi: `https://doi.org/10.1038/s41598-025-98847-7`.

Other publications

- De la Morena-Barrio, B., Orlando, C., Sanchis-Juan, A., García, J. L., Padilla, J., de la Morena-Barrio, M. E., Puruunen, M., Stouffs, K., Cifuentes, R., Borràs, N., Bravo-Pérez, C., Benito, R., Cuenca-Guardiola, J., Vicente, V., Vidal, F., Hernández-Rivas,

J. M., Ouwehand, W., Jochmans, K. & Corral, J. Molecular Dissection of Structural Variations Involved in Antithrombin Deficiency. *The Journal of molecular diagnostics: JMD,* S1525-1578(22)00042–3 (Feb. 2022). Doi: 10.1016/j.jmoldx.2022.01.009.

7.3.2. Software contributions

- Variant calling pipeline for nanopore data, related to Chapter 4, and available at `https://github.com/javiercguard/nanopore_pipeline_21`.
- DisCoverage, a tool for polishing callsets for large CNVs, which can be installed as a conda package. Ir is related to Chapter 4, and is available at `https://github.com/javiercguard/disCoverage`.
- Scripts for TE analysis. It includes instructions to run the analysis and generate the callsets for the tool benchmark, but paths require to be edited across scripts. This code is related to Chapter 5, and available at `https://github.com/javiercguard/teNanoporePipeline`.
- RetroInspector, a pipeline for detection and analysis of TE insertions and deletions. It is implemented as a Snakemake pipeline, which handles paths and dependencies. All parameters are listed in a configuration file. Includes documentation of usage, output, and how to run with a real dataset. It is available at `https://github.com/javiercguard/retroinspector`.

7.3.3. Acknowledgments

Javier Cuenca Guardiola has been funded by the Ministerio de Universidades through grant FPU19/03662.

7.4 — Conclusions

This thesis has dealt with efforts to improve the detection and analysis of various types of SVs in the context of molecular diagnosis, particularly AT deficiency. For this purpose, nanopore sequencing has been used, exploring its strengths and weaknesses. In addition to addressing questions of clinical and biological relevance, great effort has been put into making the developed methods reproducible and reusable, starting with an individual conda package, and ending with the development of a full pipeline.

During this research, we have made decisions on methodology that have helped other researchers[267,271,272] and have been cited as relevant examples of data examination in the context of SV analysis[268,269]. We have also been metioned for performing research that shows and addresses a gap in the field of discovering and cataloging TE variants[270].

A key factor is the effort we have put into making our research reproducible and reusable. From conducting benchmarks with public datasets[53,124,151] that can be rerun by others, to using standard methods to produce resusable software. This has led to the creation of disCoverage, which improves CNV detection, a standing limitation in nanopore analysis[53], and RetroInspector, a full and versatile pipeline that can perform a complete genomic analysis of TE variants, with accuracy and speed on par or better than existing alternatives[151]. These methods and the result of our work are published as open-access articles in peer-reviewed, high impact, journals, so others can make full use of what has been presented here.

Bibliography

1. Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., *et al.* An Integrated Map of Structural Variation in 2,504 Human Genomes. *Nature* **526,** 75–81 (2015).
2. Bailey, J. A. & Eichler, E. E. Primate segmental duplications: crucibles of evolution, diversity and disease. en. *Nature Reviews Genetics* **7,** 552–564 (2006).
3. Van De Peer, Y., Maere, S. & Meyer, A. The evolutionary significance of ancient genome duplications. en. *Nature Reviews Genetics* **10,** 725–732 (2009).
4. Escaramís, G., Docampo, E. & Rabionet, R. A Decade of Structural Variants: Description, History and Methods to Detect Structural Variation. *Briefings in Functional Genomics* **14,** 305–314 (2015).
5. Sudmant, P. H., Mallick, S., Nelson, B. J., Hormozdiari, F., Krumm, N., Huddleston, J., *et al.* Global Diversity, Population Stratification, and Selection of Human Copy Number Variation. *Science (New York, N.Y.)* **349,** aab3761 (2015).
6. Nakagawa, H. & Fujita, M. Whole Genome Sequencing Analysis for Cancer Genomics and Precision Medicine. *Cancer Science* **109,** 513–522 (2018).
7. Miga, K. H. & Wang, T. The Need for a Human Pangenome Reference Sequence. *Annual Review of Genomics and Human Genetics* **22,** 81–102 (2021).
8. Gardner, E. J., Lam, V. K., Harris, D. N., Chuang, N. T., Scott, E. C., Pittard, W. S., *et al.* The Mobile Element Locator Tool (MELT): Population-Scale Mobile Element Discovery and Biology. *Genome Research* **27,** 1916–1929 (2017).
9. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., *et al.* Manta: Rapid Detection of Structural Variants and Indels for Germline and Cancer Sequencing Applications. *Bioinformatics* **32,** 1220–1222 (2016).
10. Chu, C., Borges-Monroy, R., Viswanadham, V. V., Lee, S., Li, H., Lee, E. A., *et al.* Comprehensive Identification of Transposable Element Insertions Using Multiple Sequencing Technologies. *Nature Communications* **12,** 3836 (2021).
11. Gustafson, J. A., Gibson, S. B., Damaraju, N., Zalusky, M. P., Hoekzema, K., Twesigomwe, D., *et al.* High-Coverage Nanopore Sequencing of Samples from the 1000 Genomes Project to Build a Comprehensive Catalog of Human Genetic Variation. *Genome Research* **34,** 2061–2073 (2024).

12. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-Read Human Genome Sequencing and Its Applications. *Nature reviews. Genetics* **21,** 597–614 (2020).

13. Corral, J., de la Morena-Barrio, M. E. & Vicente, V. The Genetics of Antithrombin. *Thrombosis Research* **169,** 23–29 (2018).

14. Davies, P. A. & Gray, G. in *PCR Mutation Detection Protocols* 051–055 (Humana Press, New Jersey, 2002). (2021).

15. Ceulemans, S., van der Ven, K. & Del-Favero, J. in *Genomic Structural Variants* (ed Feuk, L.) 311–328 (Springer, New York, NY, 2012).

16. Bravo-Perez, C., Cifuentes-Riquelme, R., Padilla, J., de la Morena-Barrio, M. E., Ortuño, F. J., Garrido-Rodríguez, P., *et al.* The whole is greater than the sum of its parts: Long-read sequencing for solving clinical problems in haematology. *Journal of Cellular and Molecular Medicine* **28,** e17961 (2024).

17. Hu, Q., Maurais, E. G. & Ly, P. Cellular and Genomic Approaches for Exploring Structural Chromosomal Rearrangements. *Chromosome Research* **28,** 19–30 (2020).

18. Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., *et al.* High Resolution Analysis of DNA Copy Number Variation Using Comparative Genomic Hybridization to Microarrays. *Nature Genetics* **20,** 207–211 (1998).

19. Mantere, T., Neveling, K., Pebrel-Richard, C., Benoist, M., van der Zande, G., Kater-Baats, E., *et al.* Optical Genome Mapping Enables Constitutional Chromosomal Aberration Detection. *American Journal of Human Genetics* **108,** 1409–1422 (2021).

20. Van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. The Third Revolution in Sequencing Technology. *Trends in Genetics* **34,** 666–681 (2018).

21. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore Sequencing Technology, Bioinformatics and Applications. *Nature biotechnology* **39,** 1348–1365 (2021).

22. Hughes, A. E. O., Montgomery, M. C., Liu, C. & Weimer, E. T. Allele-specific quantification of human leukocyte antigen transcript isoforms by nanopore sequencing. *Frontiers in Immunology* **14.** Publisher: Frontiers (18, 2023).

23. O'Neill, K., Pleasance, E., Fan, J., Akbari, V., Chang, G., Dixon, K., *et al.* Long-read sequencing of an advanced cancer cohort resolves rearrangements, unravels haplotypes, and reveals methylation landscapes. *Cell Genomics* **4,** 100674 (14, 2024).

24. English, A. C., Dolzhenko, E., Jam, H. Z., McKenzie, S. K., Olson, N. D., De Coster, W., *et al.* Analysis and benchmarking of small and large genomic variants across tandem repeats. *Nature biotechnology* **43,** 431–442 (2025).

25. Yuan, N. & Jia, P. Comprehensive assessment of long-read sequencing platforms and calling algorithms for detection of copy number variation. *Briefings in Bioinformatics* **25,** bbae441 (2024).

26. Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., *et al.* Extensive Sequencing of Seven Human Genomes to Characterize Benchmark Reference Materials. *Scientific Data* **3,** 160025 (2016).

27. Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H. P., Bjornsson, E., Jonsson, H., *et al.* Long-Read Sequencing of 3,622 Icelanders Provides Insight into the Role of Structural Variants in Human Diseases and Other Traits. *Nature Genetics* **53,** 779–786 (2021).

28. Hehir-Kwa, J. Y., Marschall, T., Kloosterman, W. P., Francioli, L. C., Baaijens, J. A., Dijkstra, L. J., *et al.* A High-Quality Human Reference Panel Reveals the Complexity and Distribution of Genomic Structural Variants. *Nature Communications* **7,** 12989 (2016).

29. Pagnamenta, A. T., Yu, J., Walker, S., Noble, A. J., Lord, J., Dutta, P., *et al.* The Impact of Inversions across 33,924 Families with Rare Disease from a National Genome Sequencing Project. *American Journal of Human Genetics* **111,** 1140–1164 (2024).

30. Reis, A. L. M., Rapadas, M., Hammond, J. M., Gamaarachchi, H., Stevanovski, I., Ayuputeri Kumaheri, M., *et al.* The Landscape of Genomic Structural Variation in Indigenous Australians. *Nature* **624,** 602–610 (2023).

31. Funk, E. R., Mason, N. A., Pálsson, S., Albrecht, T., Johnson, J. A. & Taylor, S. A. A Supergene Underlies Linked Variation in Color and Morphology in a Holarctic Songbird. *Nature Communications* **12,** 6833 (2021).

32. Luo, H., Jiang, X., Li, B., Wu, J., Shen, J., Xu, Z., *et al.* A High-Quality Genome Assembly Highlights the Evolutionary History of the Great Bustard (*Otis tarda*, Otidiformes). *Communications Biology* **6,** 1–11 (2023).

33. Mégarbané, A., Ravel, A., Mircher, C., Sturtz, F., Grattau, Y., Rethoré, M.-O., *et al.* The 50th Anniversary of the Discovery of Trisomy 21: The Past, Present, and Future of Research and Treatment of Down Syndrome. *Genetics in Medicine* **11,** 611–616 (2009).

34. Gil, E., Bosch, A., Lampe, D., Lizcano, J. M., Perales, J. C., Danos, O., *et al.* Functional Characterization of the Human Mariner Transposon Hsmar2. *PLoS ONE* **8,** e73227 (2013).

35. Shen, H., Li, J., Zhang, J., Xu, C., Jiang, Y., Wu, Z., *et al.* Comprehensive Characterization of Human Genome Variation by High Coverage Whole-Genome Sequencing of Forty Four Caucasians. *PLoS ONE* **8,** e59494 (2013).

36. Eichler, E. E. Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. *The New England Journal of Medicine* **381,** 64–74 (2019).

37. Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E. & Pritchard, J. K. A High-Resolution Survey of Deletion Polymorphism in the Human Genome. *Nature Genetics* **38,** 75–81 (2006).

38. Xu, Z., Li, Q., Marchionni, L. & Wang, K. PhenoSV: Interpretable Phenotype-Aware Model for the Prioritization of Genes Affected by Structural Variants. *Nature Communications* **14,** 7805 (2023).

39. Borges-Monroy, R., Chu, C., Dias, C., Choi, J., Lee, S., Gao, Y., *et al.* Whole-Genome Analysis Reveals the Contribution of Non-Coding de Novo Transposon Insertions to Autism Spectrum Disorder. *Mobile DNA* **12,** 28 (2021).

40. Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., *et al.* The Influence of CCL3L1 Gene-Containing Segmental Duplications on HIV-1/AIDS Susceptibility. *Science* **307,** 1434–1440 (2005).

41. Cosenza, M. R., Rodriguez-Martin, B. & Korbel, J. O. Structural Variation in Cancer: Role, Prevalence, and Mechanisms. *Annual Review of Genomics and Human Genetics* **23,** 123–152 (2022).

42. Dubois, F., Sidiropoulos, N., Weischenfeldt, J. & Beroukhim, R. Structural Variations in Cancer and the 3D Genome. *Nature reviews. Cancer* **22,** 533–546 (2022).

43. Nieboer, M. M., Nguyen, L. & de Ridder, J. Predicting Pathogenic Non-Coding SVs Disrupting the 3D Genome in 1646 Whole Cancer Genomes Using Multiple Instance Learning. *Scientific Reports* **11,** 14411 (2021).

44. Bucciarelli, P., Passamonti, S. M., Biguzzi, E., Gianniello, F., Franchi, F., Mannucci, P. M., *et al.* Low Borderline Plasma Levels of Antithrombin, Protein C and Protein S Are Risk Factors for Venous Thromboembolism. *Journal of Thrombosis and Haemostasis* **10,** 1783–1791 (2012).

45. Ehrhardt, J. D., Boneva, D., McKenney, M. & Elkbuli, A. Antithrombin Deficiency in Trauma and Surgical Critical Care. *Journal of Surgical Research* **256,** 536–542 (2020).

46. Tait, R. C., Walker, I. D., Davidson, J. F., Islam, S. I. A. & Mitchell, R. Antithrombin III Activity in Healthy Blood Donors: Age and Sex Related Changes and the Prevalence of Asymptomatic Deficiency. *British Journal of Haematology* **75,** 141–142 (1990).

47. Luxembourg, B., Delev, D., Geisen, C., Spannagl, M., Krause, M., Miesbach, W., *et al.* Molecular Basis of Antithrombin Deficiency. *Thrombosis and Haemostasis* **105,** 635–646 (2011).

48. Patnaik, M. M. & Moll, S. Inherited Antithrombin Deficiency: A Review. *Haemophilia* **14,** 1229–1239 (2008).

49. De la Morena-Barrio, M., Sandoval, E., Llamas, P., Wypasek, E., Toderici, M., Navarro-Fernández, J., *et al.* High levels of latent antithrombin in plasma from patients with antithrombin deficiency. eng. *Thrombosis and Haemostasis* **117,** 880–888 (2017).

50. De la Morena-Barrio, M. E., Martínez-Martínez, I., de Cos, C., Wypasek, E., Roldán, V., Undas, A., *et al.* Hypoglycosylation Is a Common Finding in Antithrombin Deficiency in the Absence of a SERPINC1 Gene Defect. *Journal of Thrombosis and Haemostasis* **14,** 1549–1560 (2016).

51. De la Morena-Barrio, B., Orlando, C., Sanchis-Juan, A., García, J. L., Padilla, J., de la Morena-Barrio, M. E., *et al.* Molecular Dissection of Structural Variations Involved in Antithrombin Deficiency. *The Journal of molecular diagnostics: JMD,* S1525-1578(22)00042–3 (2022).

52. De la Morena-Barrio, B., Stephens, J., de la Morena-Barrio, M. E., Stefanucci, L., Padilla, J., Miñano, A., *et al.* Long-Read Sequencing Identifies the First Retrotransposon Insertion and Resolves Structural Variants Causing Antithrombin Deficiency. *Thrombosis and Haemostasis* **122,** 1369–1378 (2022).

53. Cuenca-Guardiola, J., de la Morena-Barrio, B., García, J. L., Sanchis-Juan, A., Corral, J. & Fernández-Breis, J. T. Improvement of Large Copy Number Variant Detection by Whole Genome Nanopore Sequencing. *Journal of Advanced Research* **50,** 145–158 (2023).

54. Greer, S. U., Botello, J., Hongo, D., Levy, B., Shah, P., Rabinowitz, M., *et al.* Implementation of Nanopore sequencing as a pragmatic workflow for copy number variant confirmation in the clinic. *Journal of Translational Medicine* **21,** 378 (2023).

55. Liu, Z., Xie, Z. & Li, M. Comprehensive and deep evaluation of structural variation detection pipelines with third-generation sequencing data. *Genome Biology* **25,** 188 (2024).

56. Oehler, J. B., Wright, H., Stark, Z., Mallett, A. J. & Schmitz, U. The application of long-read sequencing in clinical settings. *Human Genomics* **17,** 73 (8, 2023).

57. Damaraju, N., Miller, A. L. & Miller, D. E. Long-Read DNA and RNA Sequencing to Streamline Clinical Genetic Testing and Reduce Barriers to Comprehensive Genetic Testing. *The Journal of Applied Laboratory Medicine* **9,** 138–150 (1, 2024).

58. Nyaga, D. M., Tsai, P., Gebbie, C., Phua, H. H., Yap, P., Le Quesne Stabej, P., *et al.* Benchmarking nanopore sequencing and rapid genomics feasibility: validation at a quaternary hospital in New Zealand. *NPJ Genomic Medicine* **9,** 57 (8, 2024).

59. Check Hayden, E. Technology: The $1,000 Genome. *Nature* **507,** 294–295 (2014).

60. Brown, T. A. *Genomes 4* 4th (Garland Science, New York, NY, 2017).

61. Lee, H., Min, J. W., Mun, S. & Han, K. Human Retrotransposons and Effective Computational Detection Methods for Next-Generation Sequencing Data. *Life* **12,** 1583 (2022).

62. Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., *et al.* Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology* **12,** 997 (2023).

63. Zhao, Y., Wang, K., Wang, W.-l., Yin, T.-t., Dong, W.-q. & Xu, C.-j. A High-Throughput SNP Discovery Strategy for RNA-seq Data. *BMC Genomics* **20,** 160 (2019).

64. International Human Genome Sequencing Consortium. Finishing the Euchromatic Sequence of the Human Genome. *Nature* **431,** 931–945 (2004).

65. International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research, Center for Genome Research: Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., *et al.* Initial Sequencing and Analysis of the Human Genome. *Nature* **409,** 860–921 (2001).

66. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., *et al.* The Sequence of the Human Genome. *Science* **291,** 1304–1351 (2001).

67. Van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten Years of Next-Generation Sequencing Technology. *Trends in Genetics* **30,** 418–426 (2014).

68. Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., *et al.* Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology* **2012,** 251364 (2012).

69. Fox, E. J., Reid-Bayliss, K. S., Emond, M. J. & Loeb, L. A. Accuracy of Next Generation Sequencing Platforms. *Next generation, sequencing & applications* **1,** 1000106 (2014).

70. Schwarze, K., Buchanan, J., Fermont, J. M., Dreau, H., Tilley, M. W., Taylor, J. M., *et al.* The Complete Costs of Genome Sequencing: A Microcosting Study in Cancer and Rare Diseases from a Single Center in the United Kingdom. *Genetics in Medicine* **22,** 85–94 (2020).

71. Yang, T.-H., Kon, M. & DeLisi, C. Genome-Wide Association Studies. *Methods in Molecular Biology (Clifton, N.J.)* **939,** 233–251 (2013).

72. Scarano, C., Veneruso, I., De Simone, R. R., Di Bonito, G., Secondino, A. & D'Argenio, V. The Third-Generation Sequencing Challenge: Novel Insights for the Omic Sciences. *Biomolecules* **14,** 568 (2024).

73. Ardui, S., Ameur, A., Vermeesch, J. R. & Hestand, M. S. Single Molecule Real-Time (SMRT) Sequencing Comes of Age: Applications and Utilities for Medical Diagnostics. *Nucleic Acids Research* **46,** 2159–2168 (2018).

74. Schadt, E. E., Banerjee, O., Fang, G., Feng, Z., Wong, W. H., Zhang, X., *et al.* Modeling Kinetic Rate Variation in Third Generation DNA Sequencing Data to Detect Putative Modifications to DNA Bases. *Genome Research* **23,** 129–141 (2013).

75. Li, H. & Durbin, R. Fast and Accurate Long-Read Alignment with Burrows–Wheeler Transform. *Bioinformatics* **26,** 589–595 (2010).

76. Carneiro, M. O., Russ, C., Ross, M. G., Gabriel, S. B., Nusbaum, C. & DePristo, M. A. Pacific Biosciences Sequencing Technology for Genotyping and Variation Discovery in Human Data. *BMC Genomics* **13,** 375 (2012).

77. Nova, I. C., Derrington, I. M., Craig, J. M., Noakes, M. T., Tickman, B. I., Doering, K., *et al.* Investigating asymmetric salt profiles for nanopore DNA sequencing with biological porin MspA. en. *PLOS ONE* **12,** e0181599 (2017).

78. Meller, A., Nivon, L., Brandin, E., Golovchenko, J. & Branton, D. Rapid Nanopore Discrimination between Single Polynucleotide Molecules. *Proceedings of the National Academy of Sciences of the United States of America* **97,** 1079–1084 (2000).

79. Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of Individual Polynucleotide Molecules Using a Membrane Channel. *Proceedings of the National Academy of Sciences of the United States of America* **93,** 13770–13773 (1996).

80. *Biochemistry* 8. ed (eds Berg, J. M., Tymoczko, J. L., Gatto, G. J. & Stryer, L.) (W.H. Freeman/Macmillan, New York, NY, 2015).

81. Deamer, D., Akeson, M. & Branton, D. Three Decades of Nanopore Sequencing. *Nature Biotechnology* **34,** 518–524 (2016).

82. Sereika, M., Kirkegaard, R. H., Karst, S. M., Michaelsen, T. Y., Sørensen, E. A., Wollenberg, R. D., *et al.* Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nature Methods* **19,** 823–826 (2022).

83. Martin, S., Heavens, D., Lan, Y., Horsfield, S., Clark, M. D. & Leggett, R. M. Nanopore Adaptive Sampling: A Tool for Enrichment of Low Abundance Species in Metagenomic Samples. *Genome Biology* **23,** 11 (2022).

84. Ip, C. L., Loose, M., Tyson, J. R., de Cesare, M., Brown, B. L., Jain, M., *et al.* MinION Analysis and Reference Consortium: Phase 1 Data Release and Analysis. *F1000Research* **4,** 1075 (2015).

85. Hall, C. L., Zascavage, R. R., Sedlazeck, F. J. & Planz, J. V. Potential Applications of Nanopore Sequencing for Forensic Analysis. *Forensic Science Review* **32,** 23–54 (2020).

86. Avershina, E., Frye, S. A., Ali, J., Taxt, A. M. & Ahmad, R. Ultrafast and Cost-Effective Pathogen Identification and Resistance Gene Detection in a Clinical Setting Using Nanopore Flongle Sequencing. *Frontiers in Microbiology* **13,** 822402 (2022).

87. Munro, R., Payne, A., Holmes, N., Moore, C., Cahyani, I. & Loose, M. Enhancing Nanopore Adaptive Sampling for PromethION Using Readfish at Scale. *Genome Research* **35,** 877–885 (2025).

88. Tshiabuila, D., Choga, W., San, J. E., Maponga, T., Van Zyl, G., Giandhari, J., *et al.* An Oxford Nanopore Technology-Based Hepatitis B Virus Sequencing Protocol Suitable for Genomic Surveillance Within Clinical Diagnostic Settings. *International Journal of Molecular Sciences* **25,** 11702 (2024).

89. Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J., Ganapathy, G., *et al.* Hybrid Error Correction and de Novo Assembly of Single-Molecule Sequencing Reads. *Nature biotechnology* **30,** 693–700 (2012).

90. Nowoshilow, S., Schloissnig, S., Fei, J.-F., Dahl, A., Pang, A. W. C., Pippel, M., *et al.* The Axolotl Genome and the Evolution of Key Tissue Formation Regulators. *Nature* **554,** 50–55 (2018).

91. Warren, R. L., Keeling, C. I., Yuen, M. M. S., Raymond, A., Taylor, G. A., Vandervalk, B. P., *et al.* Improved White Spruce (Picea Glauca) Genome Assemblies and Annotation of Large Gene Families of Conifer Terpenoid and Phenolic Defense Metabolism. *The Plant Journal* **83,** 189–212 (2015).

92. Simpson, J. T. & Durbin, R. Efficient Construction of an Assembly String Graph Using the FM-index. *Bioinformatics* **26,** i367–i373 (2010).

93. Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., *et al.* A Draft Sequence of the Neandertal Genome. *Science* **328,** 710–722 (2010).

94. Ballouz, S., Dobin, A. & Gillis, J. A. Is It Time to Change the Reference Genome? *Genome Biology* **20,** 159 (2019).

95. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., *et al.* The Complete Sequence of a Human Genome. *Science* **376,** 44–53 (2022).

96. Fan, J.-B., Surti, U., Taillon-Miller, P., Hsie, L., Kennedy, G. C., Hoffner, L., *et al.* Paternal Origins of Complete Hydatidiform Moles Proven by Whole Genome Single-Nucleotide Polymorphism Haplotyping. *Genomics* **79,** 58–62 (2002).

97. Jain, C., Rhie, A., Hansen, N. F., Koren, S. & Phillippy, A. M. Long-Read Mapping to Repetitive Reference Sequences Using Winnowmap2. *Nature Methods* **19,** 705–710 (2022).

98. Cretu Stancu, M., van Roosmalen, M. J., Renkens, I., Nieboer, M. M., Middelkamp, S., de Ligt, J., *et al.* Mapping and Phasing of Structural Variation in Patient Genomes Using Nanopore Sequencing. *Nature Communications* **8,** 1326 (2017).

99. Duan, Z., Qiao, Y., Lu, J., Lu, H., Zhang, W., Yan, F., *et al.* HUPAN: A Pan-Genome Analysis Pipeline for Human Genomes. *Genome Biology* **20,** 149 (2019).

100. Smith, T. P. L., Bickhart, D. M., Boichard, D., Chamberlain, A. J., Djikeng, A., Jiang, Y., *et al.* The Bovine Pangenome Consortium: Democratizing Production and Accessibility of Genome Assemblies for Global Cattle Breeds and Other Bovine Species. *Genome Biology* **24,** 139 (2023).

101. Ernst, C. W. & Steibel, J. P. Molecular Advances in QTL Discovery and Application in Pig Breeding. *Trends in Genetics* **29,** 215–224 (2013).

102. Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., *et al.* Origins and Functional Impact of Copy Number Variation in the Human Genome. *Nature* **464,** 704–712 (2010).

103. De Cid, R., Riveira-Munoz, E., Zeeuwen, P. L., Robarge, J., Liao, W., Dannhauser, E. N., *et al.* Deletion of the Late Cornified Envelope (LCE) 3B and 3C Genes as a Susceptibility Factor for Psoriasis. *Nature genetics* **41,** 211–215 (2009).

104. Fanciulli, M., Norsworthy, P. J., Petretto, E., Dong, R., Harper, L., Kamesh, L., *et al.* FCGR3B Copy Number Variation Is Associated with Susceptibility to Systemic, but Not Organ-Specific, Autoimmunity. *Nature genetics* **39,** 721–723 (2007).

105. Girirajan, S., Brkanac, Z., Coe, B. P., Baker, C., Vives, L., Vu, T. H., *et al.* Relative Burden of Large CNVs on a Range of Neurodevelopmental Phenotypes. *PLoS genetics* **7,** e1002334 (2011).

106. Koolen, D. A., Pfundt, R., De Leeuw, N., Hehir-Kwa, J. Y., Nillesen, W. M., Neefs, I., *et al.* Genomic Microarrays in Mental Retardation: A Practical Workflow for Diagnostic Applications. *Human Mutation* **30,** 283–292 (2009).

107. Grozeva, D., Kirov, G., Ivanov, D., Jones, I. R., Jones, L., Green, E. K., *et al.* Rare Copy Number Variants. *Archives of general psychiatry* **67,** 318–327 (2010).

108. Stone, J. L., O'Donovan, M. C., Gurling, H., Kirov, G. K., Blackwood, D. H. R., Corvin, A., *et al.* Rare Chromosomal Deletions and Duplications Increase Risk of Schizophrenia. *Nature* **455,** 237–241 (2008).

109. Walsh, T., McClellan, J. M., McCarthy, S. E., Addington, A. M., Pierce, S. B., Cooper, G. M., *et al.* Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia. *Science* **320,** 539–543 (2008).

110. Helbig, I., Mefford, H. C., Sharp, A. J., Guipponi, M., Fichera, M., Franke, A., *et al.* 15q13.3 Microdeletions Increase Risk of Idiopathic Generalized Epilepsy. *Nature Genetics* **41,** 160–162 (2009).

111. Mefford, H. C., Muhle, H., Ostertag, P., von Spiczak, S., Buysse, K., Baker, C., *et al.* Genome-Wide Copy Number Variation in Epilepsy: Novel Susceptibility Loci in Idiopathic Generalized and Focal Epilepsies. *PLoS Genetics* **6,** e1000962 (2010).

112. Mizuguchi, T., Suzuki, T., Abe, C., Umemura, A., Tokunaga, K., Kawai, Y., *et al.* A 12-Kb Structural Variation in Progressive Myoclonic Epilepsy Was Newly Identified by Long-Read Whole-Genome Sequencing. *Journal of Human Genetics* **64,** 359–368 (2019).

113. Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., *et al.* Global Variation in Copy Number in the Human Genome. *Nature* **444,** 444–454 (2006).

114. Bouras, A., Leone, M., Bonadona, V., Lebrun, M., Calender, A. & Boutry-Kryza, N. Identification and Characterization of New Alu Element Insertion in the BRCA1 Exon 14 Associated with Hereditary Breast and Ovarian Cancer. *Genes* **12,** 1736 (2021).

115. Li, Y., Salo-Mullen, E., Varghese, A., Trottier, M., Stadler, Z. K. & Zhang, L. Insertion of an Alu-like Element in MLH1 Intron 7 as a Novel Cause of Lynch Syndrome. *Molecular Genetics & Genomic Medicine* **8,** e1523 (2020).

116. Qian, Y., Mancini-DiNardo, D., Judkins, T., Cox, H. C., Brown, K., Elias, M., *et al.* Identification of Pathogenic Retrotransposon Insertions in Cancer Predisposition Genes. *Cancer Genetics* **216–217,** 159–169 (2017).

117. Peixoto, A., Santos, C., Rocha, P., Pinheiro, M., Príncipe, S., Pereira, D., *et al.* The c.156_157insAlu BRCA2 Rearrangement Accounts for More than One-Fourth of Deleterious BRCA Mutations in Northern/Central Portugal. *Breast Cancer Research and Treatment* **114,** 31–38 (2009).

118. My, W. & Hm, G. QTL Mapping in Sesame (Sesamum Indicum L.): A Review. *Journal of biotechnology* **376** (2023).

119. Raj, S. R. G. & Nadarajah, K. QTL and Candidate Genes: Techniques and Advancement in Abiotic Stress Resistance Breeding of Major Cereals. *International Journal of Molecular Sciences* **24,** 6 (2022).

120. Lopdell, T. J. Using QTL to Identify Genes and Pathways Underlying the Regulation and Production of Milk Components in Cattle. *Animals : an Open Access Journal from MDPI* **13,** 911 (2023).

121. Abasht, B., Dekkers, J. C. M. & Lamont, S. J. Review of Quantitative Trait Loci Identified in the Chicken. *Poultry Science* **85,** 2079–2096 (2006).

122. Yuan, X., Li, Z., Xiong, L., Song, S., Zheng, X., Tang, Z., *et al.* Effective Identification of Varieties by Nucleotide Polymorphisms and Its Application for Essentially Derived Variety Identification in Rice. *BMC Bioinformatics* **23,** 30 (2022).

123. Prakrithi, P., Singhal, K., Sharma, D., Jain, A., Bhoyar, R. C., Imran, M., *et al.* An Alu Insertion Map of the Indian Population: Identification and Analysis in 1021 Genomes of the IndiGen Project. *NAR Genomics and Bioinformatics* **4,** lqac009 (2022).

124. Cuenca-Guardiola, J., de la Morena-Barrio, B., Navarro-Manzano, E., Stevens, J., Ouwehand, W. H., Gleadall, N. S., *et al.* Detection and Annotation of Transposable Element Insertions and Deletions on the Human Genome Using Nanopore Sequencing. *iScience* **26** (2023).

125. Dupont, C., Armant, D. R. & Brenner, C. A. Epigenetics: Definition, Mechanisms and Clinical Perspective. *Seminars in reproductive medicine* **27,** 351–357 (2009).

126. Peixoto, P., Cartron, P.-F., Serandour, A. A. & Hervouet, E. From 1957 to Nowadays: A Brief History of Epigenetics. *International Journal of Molecular Sciences* **21,** 7571 (2020).

127. Ahsan, M. U., Gouru, A., Chan, J., Zhou, W. & Wang, K. A Signal Processing and Deep Learning Framework for Methylation Detection Using Oxford Nanopore Sequencing. *Nature Communications* **15,** 1448 (2024).

128. Jukarainen, S., Heinonen, S., Rämö, J. T., Rinnankoski-Tuikka, R., Rappou, E., Tummers, M., *et al.* Obesity Is Associated With Low NAD(+)/SIRT Pathway Expression in Adipose Tissue of BMI-Discordant Monozygotic Twins. *The Journal of Clinical Endocrinology and Metabolism* **101,** 275–283 (2016).

129. Yuan, L., Chen, X., Liu, Z., Liu, Q., Song, A., Bao, G., *et al.* Identification of the Perpetrator among Identical Twins Using Next-Generation Sequencing Technology: A Case Report. *Forensic Science International. Genetics* **44,** 102167 (2020).

130. Gamaarachchi, H., Samarakoon, H., Jenner, S. P., Ferguson, J. M., Amos, T. G., Hammond, J. M., *et al.* Fast Nanopore Sequencing Data Analysis with SLOW5. *Nature Biotechnology* **40,** 1026–1029 (2022).

131. Oford Nanopore Technologies. *POD5 README* `https://github.com/nanoporetech/pod5-file-format/blob/master/docs/README.md`. (2025).

132. Perdomo, J. E., Ahsan, M. U., Liu, Q., Fang, L. & Wang, K. LongReadSum: A Fast and Flexible Quality Control and Signal Summarization Tool for Long-Read Sequencing Data. *Computational and Structural Biotechnology Journal* **27,** 556–563 (2025).

133. Firtina, C., Soysal, M., Lindegger, J. & Mutlu, O. RawHash2: Mapping Raw Nanopore Signals Using Hash-Based Seeding and Adaptive Quantization. *Bioinformatics* **40,** btae478 (2024).

134. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ File Format for Sequences with Quality Scores, and the Solexa/Illumina FASTQ Variants. *Nucleic Acids Research* **38,** 1767–1771 (2010).

135. NCBI Resource Coordinators. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **44,** D7–19 (2016).

136. Pearson, W. R. & Lipman, D. J. Improved Tools for Biological Sequence Comparison. *Proceedings of the National Academy of Sciences of the United States of America* **85,** 2444–2448 (1988).

137. Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., *et al.* Twelve Years of SAMtools and BCFtools. *GigaScience* **10,** giab008 (2021).

138. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., *et al.* The Sequence Alignment/Map Format and SAMtools. *Bioinformatics (Oxford, England)* **25,** 2078–2079 (2009).

139. The SAM/BAM Format Specification Working Group. *Sequence Alignment/Map Format Specification* (The SAM/BAM Format Specification Working Group, 2024).

140. Li, H. New Strategies to Improve Minimap2 Alignment Accuracy. *Bioinformatics* **37,** 4572–4574 (2021).

141. Li, H. *Minimap2 GitHub Repository* `https://github.com/lh3/minimap2`. 2025. (2025).

142. Bonfield, J. K. CRAM 3.1: Advances in the CRAM File Format. *Bioinformatics (Oxford, England)* **38,** 1497–1503 (2022).

143. Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., *et al.* The Variant Call Format and VCFtools. *Bioinformatics (Oxford, England)* **27,** 2156–2158 (2011).

144. UCSC. *Genome Browser FAQ: BED Format* `https://genome.ucsc.edu/FAQ/FAQformat.html#format1`. (2025).

145. UCSC. *Genome Browser bedGraph Track Format* `https://genome.ucsc.edu/goldenpath/help/bedgraph.html`. (2025).

146. Groza, C., Chen, X., Wheeler, T. J., Bourque, G. & Goubert, C. A Unified Framework to Analyze Transposable Element Insertion Polymorphisms Using Graph Genomes. *Nature Communications* **15,** 8915 (2024).

147. Smolka, M., Paulin, L. F., Grochowski, C. M., Horner, D. W., Mahmoud, M., Behera, S., *et al.* Detection of Mosaic and Population-Level Structural Variants with Sniffles2. *Nature Biotechnology,* 1–10 (2024).

148. Olson, N. D., Wagner, J., Dwarshuis, N., Miga, K. H., Sedlazeck, F. J., Salit, M., *et al.* Variant calling and benchmarking in an era of complete human genome sequences. *Nature Reviews Genetics* **24.** Publisher: Nature Publishing Group, 464–483 (2023).

149. Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., *et al.* Multi-Platform Discovery of Haplotype-Resolved Structural Variation in Human Genomes. *Nature Communications* **10,** 1784 (2019).

150. Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., *et al.* Haplotype-Resolved Diverse Human Genomes and Integrated Analysis of Structural Variation. *Science* **372,** eabf7117 (2021).

151. Cuenca-Guardiola, J., de la Morena-Barrio, B., Corral, J. & Fernández-Breis, J. T. Advanced Analysis of Retrotransposon Variation in the Human Genome with Nanopore Sequencing Using RetroInspector. *Scientific Reports* **15,** 14489 (2025).

152. International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–796 (2003).

153. Zook, J. M., Hansen, N. F., Olson, N. D., Chapman, L., Mullikin, J. C., Xiao, C., *et al.* A Robust Benchmark for Detection of Germline Large Deletions and Insertions. *Nature Biotechnology* **38,** 1347–1355 (2020).

154. Krusche, P., Trigg, L., Boutros, P. C., Mason, C. E., De La Vega, F. M., Moore, B. L., *et al.* Best Practices for Benchmarking Germline Small-Variant Calls in Human Genomes. *Nature Biotechnology* **37,** 555–560 (2019).

155. Jarvis, E. D., Formenti, G., Rhie, A., Guarracino, A., Yang, C., Wood, J., *et al.* Semi-Automated Assembly of High-Quality Diploid Human Reference Genomes. *Nature* **611,** 519–531 (2022).

156. Wagner, J., Olson, N. D., Harris, L., McDaniel, J., Cheng, H., Fungtammasan, A., *et al.* Curated Variation Benchmarks for Challenging Medically-Relevant Autosomal Genes. *Nature biotechnology* **40,** 672–680 (2022).

157. Jia, P., Dong, L., Yang, X., Wang, B., Bush, S. J., Wang, T., *et al.* Haplotype-Resolved Assemblies and Variant Benchmark of a Chinese Quartet. *Genome Biology* **24,** 277 (2023).

158. Genovese, G., Rockweiler, N. B., Gorman, B. R., Bigdeli, T. B., Pato, M. T., Pato, C. N., *et al.* BCFtools/liftover: an accurate and comprehensive tool to convert genetic variants across genome assemblies. *Bioinformatics* **40,** btae038 (2024).

159. The 1000 Genomes Project Consortium. A Global Reference for Human Genetic Variation. *Nature* **526,** 68–74 (2015).

160. Durbin, R. M., Altshuler, D., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., *et al.* A Map of Human Genome Variation from Population-Scale Sequencing. *Nature* **467,** 1061–1073 (2010).

161. Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., *et al.* High-Coverage Whole-Genome Sequencing of the Expanded 1000 Genomes Project Cohort Including 602 Trios. *Cell* **185,** 3426–3440.e19 (2022).

162. Koenig, Z., Yohannes, M. T., Nkambule, L. L., Zhao, X., Goodrich, J. K., Kim, H. A., *et al.* A Harmonized Public Resource of Deeply Sequenced Diverse Human Genomes. *Genome Research* **34,** 796–809 (2024).

163. Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., *et al.* dbSNP: The NCBI Database of Genetic Variation. *Nucleic Acids Research* **29,** 308–311 (2001).

164. Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., *et al.* dbVar and DGVa: Public Archives for Genomic Structural Variation. *Nucleic Acids Research* **41,** D936–D941 (2013).

165. Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., *et al.* Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **50,** D20–D26 (2022).

166. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: Visualizing and Processing Long-Read Sequencing Data. *Bioinformatics* **34,** 2666–2669 (2018).

167. Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C. & Sedlazeck, F. J. Structural Variant Calling: The Long and the Short of It. *Genome Biology* **20,** 246 (2019).

168. Petersen, L. M., Martin, I. W., Moschetti, W. E., Kershaw, C. M. & Tsongalis, G. J. Third-Generation Sequencing in the Clinical Laboratory: Exploring the Advantages and Challenges of Nanopore Sequencing. *Journal of Clinical Microbiology* **58,** e01315–19 (2019).

169. Oxford Nanopore Technologies. *Dorado* `https://github.com/nanoporetech/dorado`. 2025. (2025).

170. Chris Seymour, Oxford Nanopore Technologies Ltd. *Bonito: A PyTorch Basecaller for Oxford Nanopore Reads* `https://github.com/nanoporetech/bonito`. 2019.

171. Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J. & Timp, W. Detecting DNA Cytosine Methylation Using Nanopore Sequencing. *Nature Methods* **14,** 407–410 (2017).

172. Stoiber, M., Quick, J., Egan, R., Lee, J. E., Celniker, S., Neely, R. K., *et al. De Novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing* `https://www.biorxiv.org/content/10.1101/094672v2`. 2017. (2025).

173. Ni, P., Huang, N., Zhang, Z., Wang, D.-P., Liang, F., Miao, Y., *et al.* DeepSignal: Detecting DNA Methylation State from Nanopore Sequencing Reads Using Deep-Learning. *Bioinformatics* **35,** 4586–4595 (2019).

174. Stanojević, D., Li, Z., Bakić, S., Foo, R. & Šikić, M. Rockfish: A Transformer-Based Model for Accurate 5-Methylcytosine Prediction from Nanopore Sequencing. *Nature Communications* **15,** 5580 (2024).

175. Liu, Q., Fang, L., Yu, G., Wang, D., Xiao, C.-L. & Wang, K. Detection of DNA Base Modifications by Deep Recurrent Neural Network on Oxford Nanopore Sequencing Data. *Nature Communications* **10,** 2449 (2019).

176. Wilson, R. J. *Introduction to graph theory* Fifth edition. eng (Prentice Hall, Pearson, Harlow, England London New York, 2010).

177. Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., *et al.* Nonhybrid, Finished Microbial Genome Assemblies from Long-Read SMRT Sequencing Data. *Nature Methods* **10,** 563–569 (2013).

178. Berlin, K., Koren, S., Chin, C.-S., Drake, J. P., Landolin, J. M. & Phillippy, A. M. Assembling Large Genomes with Single-Molecule Sequencing and Locality-Sensitive Hashing. *Nature Biotechnology* **33,** 623–630 (2015).

179. Antipov, D., Korobeynikov, A., McLean, J. S. & Pevzner, P. A. hybridSPAdes: An Algorithm for Hybrid Assembly of Short and Long Reads. *Bioinformatics (Oxford, England)* **32,** 1009–1015 (2016).

180. Zerbino, D. R. & Birney, E. Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs. *Genome Research* **18,** 821–829 (2008).

181. Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., *et al.* Phased Diploid Genome Assembly with Single-Molecule Real-Time Sequencing. *Nature Methods* **13,** 1050–1054 (2016).

182. Myers, E. W. The Fragment Assembly String Graph. *Bioinformatics (Oxford, England)* **21 Suppl 2,** ii79–85 (2005).

183. Li, H. Minimap and Miniasm: Fast Mapping and de Novo Assembly for Noisy Long Sequences. *Bioinformatics* **32,** 2103–2110 (2016).

184. Kamath, G. M., Shomorony, I., Xia, F., Courtade, T. A. & Tse, D. N. HINGE: Long-Read Assembly Achieves Optimal Repeat Resolution. *Genome Research* **27,** 747–756 (2017).

185. Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H. & Phillippy, A. M. Canu: Scalable and Accurate Long-Read Assembly via Adaptive k-Mer Weighting and Repeat Separation. *Genome Research* **27,** 722–736 (2017).

186. Xu, F., Ge, C., Luo, H., Li, S., Wiedmann, M., Deng, X., *et al.* Evaluation of Real-Time Nanopore Sequencing for Salmonella Serotype Prediction. *Food Microbiology* **89,** 103452 (2020).

187. Mirza, J. D., de Oliveira Guimarães, L., Wilkinson, S., Rocha, E. C., Bertanhe, M., Helfstein, V. C., *et al.* Tracking Arboviruses, Their Transmission Vectors and Potential Hosts by Nanopore Sequencing of Mosquitoes. *Microbial Genomics* **10,** 001184 (2024).

188. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of Long, Error-Prone Reads Using Repeat Graphs. *Nature Biotechnology* **37,** 540–546 (2019).

189. Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., *et al.* Nanopore Sequencing and the Shasta Toolkit Enable Efficient de Novo Assembly of Eleven Human Genomes. *Nature Biotechnology* **38,** 1044–1053 (2020).

190. Kolmogorov, M., Billingsley, K. J., Mastoras, M., Meredith, M., Monlong, J., Lorig-Roach, R., *et al.* Scalable Nanopore Sequencing of Human Genomes Provides a Comprehensive View of Haplotype-Resolved Variation and Methylation. *Nature methods* **20,** 1483–1492 (2023).

191. Liu, Y. H., Luo, C., Golding, S. G., Ioffe, J. B. & Zhou, X. M. Tradeoffs in alignment and assembly-based methods for structural variant detection with long-read sequencing data. en. *Nature Communications* **15.** Publisher: Nature Publishing Group, 2447 (2024).

192. Li, H. *Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM* `https://doi.org/10.48550/arXiv.1303.3997`. 2013. arXiv: `1303.3997 [q-bio]`. (2025).

193. Oxford Nanopore Technologies. *Epi2me Alignment Workflow* `https://github.com/epi2me-labs/wf-alignment`. (2025).

194. Oxford Nanopore Technologies. *epi2me-labs -Human variation workflow* en. `https://github.com/epi2me-labs/wf-human-variation/releases` (2025).

195. Jiang, T., Liu, S., Cao, S., Liu, Y., Cui, Z., Wang, Y., *et al.* Long-Read Sequencing Settings for Efficient Structural Variation Detection Based on Comprehensive Evaluation. *BMC bioinformatics* **22,** 552 (2021).

196. Chaisson, M. J. & Tesler, G. Mapping Single Molecule Sequencing Reads Using Basic Local Alignment with Successive Refinement (BLASR): Application and Theory. *BMC Bioinformatics* **13,** 238 (2012).

197. Liu, B., Guan, D., Teng, M. & Wang, Y. rHAT: Fast Alignment of Noisy Long Reads with Regional Hashing. *Bioinformatics* **32,** 1625–1631 (2016).

198. Liu, B., Gao, Y. & Wang, Y. LAMSA: Fast Split Read Alignment with Long Approximate Matches. *Bioinformatics* **33** (ed Hancock, J.) 192–201 (2017).

199. Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., *et al.* Accurate Detection of Complex Structural Variations Using Single-Molecule Sequencing. *Nature Methods* **15,** 461–468 (2018).

200. Suzuki, H. & Kasahara, M. Introducing Difference Recurrence Relations for Faster Semi-Global Alignment of Long Sequences. *BMC Bioinformatics* **19,** 45 (2018).

201. Li, H. Minimap2: Pairwise Alignment for Nucleotide Sequences. *Bioinformatics* **34,** 3094–3100 (2018).

202. Bonfield, J. K., Marshall, J., Danecek, P., Li, H., Ohan, V., Whitwham, A., *et al.* HTSlib: C Library for Reading/Writing High-Throughput Sequencing Data. *GigaScience* **10,** giab007 (2021).

203. Cornelis, S., Gansemans, Y., Vander Plaetsen, A.-S., Weymaere, J., Willems, S., Deforce, D., *et al.* Forensic Tri-Allelic SNP Genotyping Using Nanopore Sequencing. *Forensic Science International: Genetics* **38,** 204–210 (2019).

204. Edge, P. & Bansal, V. Longshot Enables Accurate Variant Calling in Diploid Genomes from Single-Molecule Long Read Sequencing. *Nature Communications* **10,** 4660 (2019).

205. Luo, R., Sedlazeck, F. J., Lam, T.-W. & Schatz, M. C. A Multi-Task Convolutional Deep Neural Network for Variant Calling in Single Molecule Sequencing. *Nature Communications* **10,** 998 (2019).

206. Luo, R., Wong, C.-L., Wong, Y.-S., Tang, C.-I., Liu, C.-M., Leung, C.-M., *et al.* Exploring the Limit of Using a Deep Neural Network on Pileup Data for Germline Variant Calling. *Nature Machine Intelligence* **2,** 220–227 (2020).

207. Shafin, K., Pesout, T., Chang, P.-C., Nattestad, M., Kolesnikov, A., Goel, S., *et al.* Haplotype-Aware Variant Calling with PEPPER-Margin-DeepVariant Enables High Accuracy in Nanopore Long-Reads. *Nature Methods* **18,** 1322–1332 (2021).

208. Zheng, Z., Li, S., Su, J., Leung, A. W.-S., Lam, T.-W. & Luo, R. Symphonizing Pileup and Full-Alignment for Deep Learning-Based Long-Read Variant Calling. *Nature Computational Science* **2,** 797–803 (2022).

209. Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., *et al.* Resolving the Complexity of the Human Genome Using Single-Molecule Sequencing. *Nature* **517,** 608–611 (2015).

210. English, A. C., Salerno, W. J. & Reid, J. G. PBHoney: Identifying Genomic Variants via Long-Read Discordance and Interrupted Mapping. *BMC Bioinformatics* **15,** 180 (2014).

211. Heller, D. & Vingron, M. SVIM: Structural Variant Identification Using Mapped Long Reads. *Bioinformatics* **35,** 2907–2915 (2019).

212. PacificBiosciences. *Pbsv* `https://github.com/PacificBiosciences/pbsv`. 2025. (2025).

213. Tham, C. Y., Tirado-Magallanes, R., Goh, Y., Fullwood, M. J., Koh, B. T., Wang, W., *et al.* NanoVar: Accurate Characterization of Patients' Genomic Structural Variants Using Low-Depth Nanopore Sequencing. *Genome Biology* **21,** 56 (2020).

214. Tham, C. Y. & Benoukraf, T. Correspondence on NanoVar's Performance Outlined by Jiang T. et al. in "Long-read Sequencing Settings for Efficient Structural Variation Detection Based on Comprehensive Evaluation". *BMC Bioinformatics* **24,** 350 (2023).

215. Dierckxsens, N., Li, T., Vermeesch, J. R. & Xie, Z. A Benchmark of Structural Variation Detection by Long Reads through a Realistic Simulated Model. *Genome Biology* **22,** 342 (2021).

216. Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., *et al.* Long-Read-Based Human Genomic Structural Variation Detection with cuteSV. *Genome Biology* **21,** 189 (2020).

217. Zhou, W., Emery, S. B., Flasch, D. A., Wang, Y., Kwan, K. Y., Kidd, J. M., *et al.* Identification and Characterization of Occult Human-Specific LINE-1 Insertions Using Long-Read Sequencing Technology. *Nucleic Acids Research* **48,** 1146–1163 (2020).

218. Yu, Q., Zhang, W., Zhang, X., Zeng, Y., Wang, Y., Wang, Y., *et al.* Population-Wide Sampling of Retrotransposon Insertion Polymorphisms Using Deep Sequencing and Efficient Detection. *GigaScience* **6,** 1–11 (2017).

219. De Coster, W., De Rijk, P., De Roeck, A., De Pooter, T., D'Hert, S., Strazisar, M., *et al.* Structural Variants Identified by Oxford Nanopore PromethION Sequencing of the Human Genome. *Genome Research* **29,** 1178–1187 (2019).

220. Ferrari, R., de Llobet Cucalon, L. I., Di Vona, C., Le Dilly, F., Vidal, E., Lioutas, A., *et al.* TFIIIC Binding to Alu Elements Controls Gene Expression via Chromatin Looping and Histone Acetylation. *Molecular Cell* **77,** 475–487.e11 (2020).

221. LoTempio, J., Delot, E. & Vilain, E. Benchmarking long-read genome sequence alignment tools for human genomics applications. en. *PeerJ* **11.** Publisher: PeerJ Inc., e16515 (2023).

222. Smolka, M., Paulin, L. F., Grochowski, C. M., Mahmoud, M., Behera, S., Gandhi, M., *et al.* Comprehensive Structural Variant Detection: From Mosaic to Population-Level (2022).

223. Helal, A. A., Saad, B. T., Saad, M. T., Mosaad, G. S. & Aboshanab, K. M. Benchmarking long-read aligners and SV callers for structural variation detection in Oxford nanopore sequencing data. en. *Scientific Reports* **14.** Publisher: Nature Publishing Group, 6160 (2024).

224. Kohler, J. R. & Cutler, D. J. Simultaneous Discovery and Testing of Deletions for Disease Association in SNP Genotyping Studies. *American Journal of Human Genetics* **81,** 684–699 (2007).

225. Shao, H., Ganesamoorthy, D., Duarte, T., Cao, M. D., Hoggart, C. J. & Coin, L. J. M. npInv: accurate detection and genotyping of inversions using long read sub-alignment. *BMC Bioinformatics* **19,** 261 (2018).

226. Yildiz, G., Zanini, S. F., Afsharyan, N. P., Obermeier, C., Snowdon, R. J. & Golicz, A. A. Benchmarking Oxford Nanopore read alignment-based insertion and deletion detection in crop plant genomes. en. *The Plant Genome* **16.** _eprint: https://acsess.onlinelibrary.wiley.com/doi/pdf/1 e20314 (2023).

227. Duan, X., Pan, M. & Fan, S. Comprehensive evaluation of structural variant genotyping methods based on long-read sequencing data. *BMC Genomics* **23,** 324 (23, 2022).

228. Nazarenko, M. S., Sleptcov, A. A., Zarubin, A. A., Salakhov, R. R., Shevchenko, A. I., Tmoyan, N. A., *et al.* Calling and Phasing of Single-Nucleotide and Structural Variants of the LDLR Gene Using Oxford Nanopore MinION. *International Journal of Molecular Sciences* **24,** 4471 (24, 2023).

229. Zhou, Y., Leung, A. W.-S., Ahmed, S. S., Lam, T.-W. & Luo, R. Duet: SNP-assisted structural variant calling and phasing using Oxford nanopore sequencing. *BMC Bioinformatics* **23,** 465 (7, 2022).

230. Martin, M., Patterson, M., Garg, S., Fischer, S. O., Pisanti, N., Klau, G. W., *et al.* WhatsHap: Fast and Accurate Read-Based Phasing. *bioRxiv,* 085050 (2016).

231. Jeffares, D. C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., *et al.* Transient Structural Variations Have Strong Effects on Quantitative Traits and Reproductive Isolation in Fission Yeast. *Nature Communications* **8,** 14061 (2017).

232. Kirsche, M., Prabhu, G., Sherman, R., Ni, B., Battle, A., Aganezov, S., *et al.* Jasmine and Iris: Population-scale Structural Variant Comparison and Analysis. *Nature methods* **20,** 408–417 (2023).

233. English, A. C., Menon, V. K., Gibbs, R. A., Metcalf, G. A. & Sedlazeck, F. J. Truvari: Refined Structural Variant Comparison Preserves Allelic Diversity. *Genome Biology* **23,** 271 (2022).

234. De la Morena-Barrio, B., Borràs, N., Rodríguez-Alén, A., de la Morena-Barrio, M. E., García-Hernández, J. L., Padilla, J., *et al.* Identification of the First Large Intronic Deletion Responsible of Type I Antithrombin Deficiency Not Detected by Routine Molecular Diagnostic Methods. *British Journal of Haematology* **186,** e82–e86 (2019).

235. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: An Approach to Discover, Genotype, and Characterize Typical and Atypical CNVs from Family and Population Genome Sequencing. *Genome Research* **21,** 974–984 (2011).

236. Lee, W.-P., Zhu, Q., Yang, X., Liu, S., Cerveira, E., Ryan, M., *et al.* JAX-CNV: A Whole-genome Sequencing-based Algorithm for Copy Number Detection at Clinical Grade Level. *Genomics, Proteomics & Bioinformatics* **20,** 1197–1206 (2022).

237. Rapti, M., Zouaghi, Y., Meylan, J., Ranza, E., Antonarakis, S. E. & Santoni, F. A. CoverageMaster: Comprehensive CNV Detection and Visualization from NGS Short Reads for Genetic Medicine Applications. *Briefings in Bioinformatics* **23,** bbac049 (2022).

238. Pedersen, B. S. & Quinlan, A. R. Mosdepth: Quick Coverage Calculation for Genomes and Exomes. *Bioinformatics* **34,** 867–868 (2018).

239. Chakravorty, S. & Hegde, M. Gene and Variant Annotation for Mendelian Disorders in the Era of Advanced Sequencing Technologies. *Annual Review of Genomics and Human Genetics* **18,** 229–256 (2017).

240. Geoffroy, V., Herenger, Y., Kress, A., Stoetzel, C., Piton, A., Dollfus, H., *et al.* AnnotSV: An Integrated Tool for Structural Variations Annotation. *Bioinformatics* **34,** 3572–3574 (2018).

241. Cavalcante, R. G. & Sartor, M. A. Annotatr: Genomic Regions in Context. *Bioinformatics* **33,** 2381–2383 (2017).

242. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS : a Journal of Integrative Biology* **16,** 284–287 (2012).

243. Consortium, T. G. O., Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., *et al.* Gene Ontology: Tool for the Unification of Biology. *Nature genetics* **25,** 25 (2000).

244. The Gene Ontology Consortium, Carbon, S., Douglass, E., Good, B. M., Unni, D. R., Harris, N. L., *et al.* The Gene Ontology Resource: Enriching a GOld Mine. *Nucleic Acids Research* **49,** D325–D334 (2021).

245. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: Integrating Viruses and Cellular Organisms. *Nucleic Acids Research* **49,** D545–D551 (2021).

246. D'Antonio, M., Pendino, V., Sinha, S. & Ciccarelli, F. D. Network of Cancer Genes (NCG 3.0): Integration and Analysis of Genetic and Network Properties of Cancer Genes. *Nucleic Acids Research* **40,** D978–D983 (2012).

247. Schriml, L. M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., *et al.* Human Disease Ontology 2018 Update: Classification, Content and Workflow Expansion. *Nucleic Acids Research* **47,** D955–D962 (2019).

248. Yu, G., Wang, L.-G., Yan, G.-R. & He, Q.-Y. DOSE: An R/Bioconductor Package for Disease Ontology Semantic and Enrichment Analysis. *Bioinformatics (Oxford, England)* **31,** 608–609 (2015).

249. Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., *et al.* Integrative Genomics Viewer. *Nature biotechnology* **29,** 24–26 (2011).

250. Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., *et al.* Bioconda: Sustainable and Comprehensive Software Distribution for the Life Sciences. *Nature Methods* **15,** 475–476 (2018).

251. Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E. & Notredame, C. Nextflow Enables Reproducible Computational Workflows. *Nature Biotechnology* **35,** 316–319 (2017).

252. Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., *et al. Sustainable Data Analysis with Snakemake* `https://doi.org/10.12688/f1000research.29032.2`. 2021. (2024).

253. *Workflow Description Language (WDL) Documentation* `https://docs.openwdl.org/`. (2025).

254. Goecks, J., Nekrutenko, A., Taylor, J. & The Galaxy Team. Galaxy: A Comprehensive Approach for Supporting Accessible, Reproducible, and Transparent Computational Research in the Life Sciences. *Genome Biology* **11,** R86 (2010).

255. Charron, P. & Kang, M. VariantDetective: An Accurate All-in-One Pipeline for Detecting Consensus Bacterial SNPs and SVs. *Bioinformatics* **40,** btae066 (2024).

256. Cherif, E., Thiam, F. S., Salma, M., Rivera-Ingraham, G., Justy, F., Deremarque, T., *et al.* ONTdeCIPHER: An Amplicon-Based Nanopore Sequencing Pipeline for Tracking Pathogen Variants. *Bioinformatics* **38,** 2033–2035 (2022).

257. Petrone, J. R., Rios Glusberger, P., George, C. D., Milletich, P. L., Ahrens, A. P., Roesch, L. F. W., *et al.* RESCUE: A Validated Nanopore Pipeline to Classify Bacteria through Long-Read, 16S-ITS-23S rRNA Sequencing. *Frontiers in Microbiology* **14,** 1201064 (2023).

258. Olson, N. D., Wagner, J., McDaniel, J., Stephens, S. H., Westreich, S. T., Prasanna, A. G., *et al.* PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. *Cell Genomics* **2,** 100129 (2022).

259. Nature Biotechnology Editorial Team. Changing coding culture. *Nature Biotechnology* **37.** Publisher: Nature Publishing Group, 485–485 (2019).

260. Pei, Y., Tanguy, M., Giess, A., Dixit, A., Wilson, L. C., Gibbons, R. J., *et al.* A Comparison of Structural Variant Calling from Short-Read and Nanopore-Based Whole-Genome Sequencing Using Optical Genome Mapping as a Benchmark. *Genes* **15,** 925 (2024).

261. Taha, A. A. & Hanbury, A. Metrics for Evaluating 3D Medical Image Segmentation: Analysis, Selection, and Tool. *BMC Medical Imaging* **15,** 29 (2015).

262. Genome Reference Consortium. *Genome Reference Consortium Human Build 38 (GRCh38)* `ftp://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_no_alt_analysis_set.fna.gz`. 2013. (2022).

263. R Core Team. *R: A Language and Environment for Statistical Computing* Manual (Vienna, Austria, 2023).

264. Dowle, M. & Srinivasan, A. *Data.Table: Extension of 'data.Frame'* (2021).

265. Gao, C., Xiao, M., Ren, X., Hayward, A., Yin, J., Wu, L., *et al.* Characterization and functional annotation of nested transposable elements in eukaryotic genomes. *Genomics* **100,** 222–230 (2012).

266. Cogan, G., Daida, K., Blauwendraat, C., Billingsley, K. & Brice, A. Exploration of Neurodegenerative Diseases Using Long-Read Sequencing and Optical Genome Mapping Technologies. *Movement Disorders* **40,** 996–1008 (2025).

267. Geysens, M., Huremagic, B., Souche, E., Breckpot, J., Devriendt, K., Peeters, H., *et al.* Clinical evaluation of long-read sequencing-based episignature detection in developmental disorders. *Genome Medicine* **17,** 1 (10, 2025).

268. Bækgaard, C. H., Lester, E. B., Møller-Larsen, S., Lauridsen, M. F. & Larsen, M. J. NanoImprint: A DNA methylation tool for clinical interpretation and diagnosis of common imprinting disorders using nanopore long-read sequencing. *Annals of Human Genetics* **88.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ahg.12556, 392–398 (2024).

269. Trinh, J., Schaake, S., Gabbert, C., Lüth, T., Cowley, S. A., Fienemann, A., *et al.* Optical genome mapping of structural variants in Parkinson's disease-related induced pluripotent stem cells. *BMC Genomics* **25,** 980 (19, 2024).

270. Lee, J.-O., Lee, S., Lee, D., Hwang, T., Joe, S., Yang, J. O., *et al.* KTED: a comprehensive web-based database for transposable elements in the Korean genome. *Bioinformatics Advances* **4,** vbae179 (5, 2024).

271. Padilla-García, N., Le Veve, A., Čermák, V., İltaş, Ö., Contreras-Garrido, A., Legrand, S., *et al.* The Demographic History of Populations and Genomic Imprinting have Shaped the Transposon Patterns in Arabidopsis lyrata. *Molecular Biology and Evolution* **42,** msaf093 (1, 2025).

272. Chen, Q., Wu, B., Li, C., Ding, L., Huang, S., Wang, J., *et al.* Deciphering male influence in gynogenetic Pengze crucian carp (Carassius auratus var. pengsenensis): insights from Nanopore sequencing of structural variations. *Frontiers in Genetics* **15.** Publisher: Frontiers (9, 2024).

273. Bilgrav Saether, K. & Eisfeldt, J. Detecting Transposable Elements in Long-Read Genomes Using sTELLeR. *Bioinformatics* **40,** btae686 (2024).

274. Oxford Nanopore Technologies. *ONT-Spectre* `https://github.com/nanoporetech/ont-spectre`. 2025. (2025).

275. Hung, T., Pratt, G. A., Sundararaman, B., Townsend, M. J., Chaivorapol, C., Bhangale, T., *et al.* The Ro60 autoantigen binds endogenous retroelements and regulates inflammatory gene expression. *Science* **350,** 455–459 (23, 2015).

276. Norris, J., Fan, D., Aleman, C., Marks, J. R., Futreal, P. A., Wiseman, R. W., *et al.* Identification of a New Subclass of Alu DNA Repeats Which Can Function as Estrogen Receptor-dependent Transcriptional Enhancers. *Journal of Biological Chemistry* **270.** Publisher: Elsevier, 22777–22782 (29, 1995).

277. Zhou, S. & Van Bortle, K. The Pol III transcriptome: Basic features, recurrent patterns, and emerging roles in cancer. *Wiley interdisciplinary reviews. RNA* **14,** e1782 (2023).

278. Stenz, L. The L1-dependant and Pol III transcribed Alu retrotransposon, from its discovery to innate immunity. *Molecular Biology Reports* **48,** 2775–2789 (2021).

279. Cardelli, M. The epigenetic alterations of endogenous retroelements in aging. *Mechanisms of Ageing and Development. Epigenetics in Ageing and Development* **174,** 30–46 (1, 2018).

280. Hancks, D. C. & Kazazian, H. SVA retrotransposons: Evolution and genetic instability. *Seminars in cancer biology* **20,** 234–245 (2010).

281. Clayton, E. A., Wang, L., Rishishwar, L., Wang, J., McDonald, J. F. & Jordan, I. K. Patterns of Transposable Element Expression and Insertion in Cancer. *Frontiers in Molecular Biosciences* **3,** 76 (16, 2016).