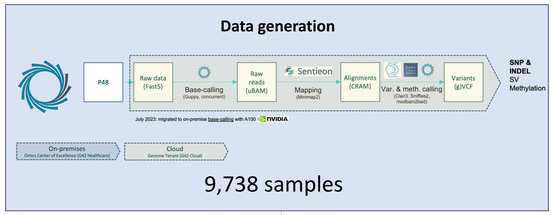
Leveraging long-read Oxford Nanopore Technologies and a national genome program to better understand structural variation inheritance and medical impact

D. Matías Sánchez[1], A. Al Awadhi[1], A. El-Khani[1], A. Al Mannaei[2], F. Al Marzooqi[1], F. Aldhuhoori[1], F. Sedlazeck[3], H. Wu[1], H. Sajad[1], I. Eltantawy[4], J. Arrés[1], J. Quílez[1], J. Mafofo[1], L. Paulin[3], M. Ibrahim[4], M. Al Ameri[2], S. Elavalli[1], S. Zhang[1], T. Magalhães[1], W. Abdulrahman[2]; [1]M42, Abu Dhabi, United Arab Emirates, [2]Dept. of Health of Abu Dhabi, Abu Dhabi, United Arab Emirates, [3]Baylor Coll. Med., Houston, TX, [4]M42; IROS, Abu Dhabi, United Arab Emirates.

Introduction

- The Emirati Genome Program (EGP) is sequencing **1 million individuals** using advanced technologies, with over **100,000 sequenced using Oxford Nanopore (ONT)** becoming the largest long-reads dataset in the world.
- This study provides a **preliminary overview**, including **~10,000 ONT samples**.
- The high-resolution capabilities of **ONT** are pivotal for uncovering **novel structural variants (SVs)** that were previously undetectable, deepening our understanding of genomic complexity.
- The presence **of extensive family networks** in the EGP, allows for the study of **SV inheritance** patterns.
- Understanding genetic inheritance patterns is crucial for elucidating the genetic basis of diseases and improving diagnostic accuracy.

Methodology



Quality Control

- Yield ≥ 90gb
- Coverage > 26X
- Mapping rate ≥ 70%
- Het/hom ratio < 2.5
- Number of SVs < 65,000

Merge into multi-vcf:

- Bcftools for SNVs and INDELs
- Sniffles2 for SVs

8,547 high-quality samples

multi-vcf

Summary cohort statistics generation
(Table 1, Figure 1)

- Single nucleotide variants
- INDELs
- Structural variants

Population and Family structure analysis

- **Principal Component Analysis (PCA + UMAP)**
- **Relationship estimation with KING software**
- **Inference of trios and SV Inheritance Analysis**

♂ (AA) + ♀ (aa) = (Aa??)

Population structure

Figure 2. Principal Component Analysis (PC1 vs PC2) and UMAP (10 PCs) of 8,547 EGP ONT samples vs 3,202 1000genomes project Phase 3 samples.

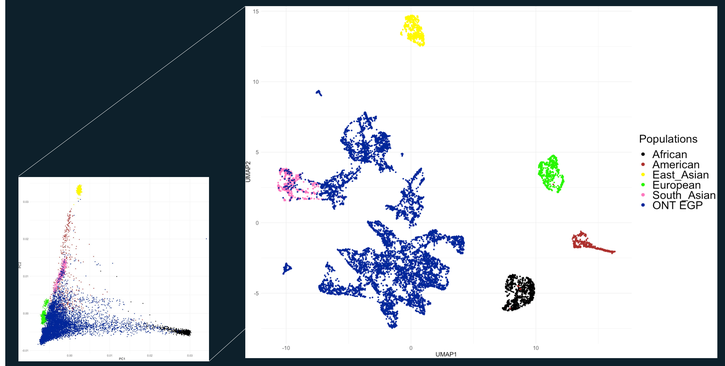


Table 2. Relationship estimation predicted using KING software

Number of SNVs used	529,851
IBD genome covered (cM)	3572.2
Total relations (pairs)	15,342
Parent-offspring	532
Full sibling	737
2nd degree	2,759
3rd degree and distant	8,501
Monozygotic twins	13

40 trios (Father, Mother and Child) found in the cohort

Inheritance Analysis

- **Inherited Variants**: Variants where at least one parent carries the variant (0/1, 1/1) and the child carries the variant with a genotype compatible with the parental genotypes.
- **De novo Variants:** Variants found in the child but absent in both parents, or homozygous (1/1) in the child and heterozygous (0/1) in only one parent.
- **Errors/Lack of Inheritance:** Variants homozygous (1/1) in at least one parent but absent (0/0) in the child.

Figure 3. Box-plot showing the distribution of Inherited and *de novo* variants in family trios

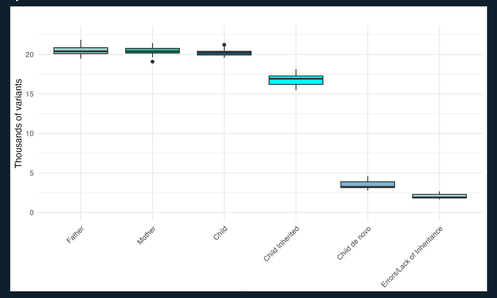


Figure 4. Distribution of inherited and *de novo* variants in family trios by variant length in whole-genome and medically relevant genes

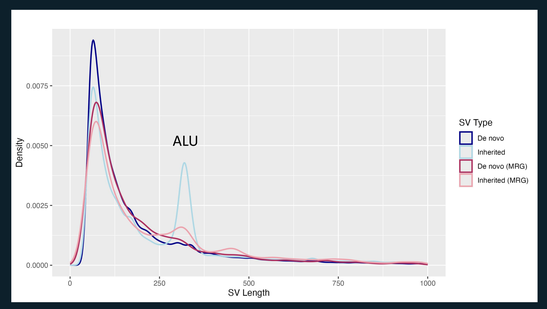


Table 3. Average number of variants in autosomes present in offspring: Inherited vs. *de novo* after merging with Sniffles2

SV type	Average per child	Average Inherited	Average de novo
Deletion	8,060	7,897	163
Duplication	8	7	1
Insertion	12,137	8,850	3,287
Inversion	34	32	2
Total	20,239	16,786	3,453

Figure 5. Box-plot showing the distribution of *de novo* SVs in medically relevant genes by variant type

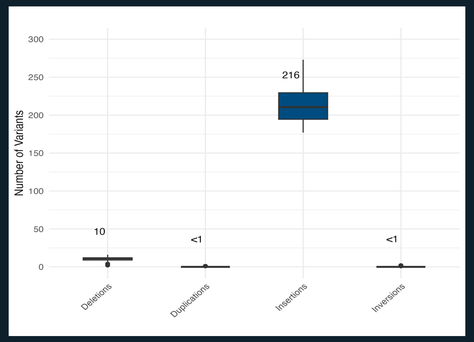
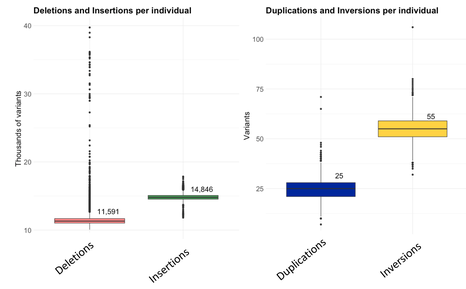


Table 1. Summary cohort statistics by variant type per individual

Variant type	Mean	std
SNVs	4,298,903.3	167,865.5
Indels	524,412.3	77,196.1
SVs	26,516.9	2,890.0

Figure 1. SV type distribution per individual



Conclusions

- **Deletions and insertions** dominate structural variants (SVs), making up over **99%** of the total, while individuals carry only a few dozen duplications and inversions.
- Parents and offspring have comparable SV counts in autosomes, with **80% of offspring SVs being inherited and 20% de novo**.
- **Insertions are the most frequent de novo** SVs, potentially due to **bias** from SV **merging** tools.
- **No significant length differences** are observed between inherited and de novo SVs in both, whole genome and MRG, except for a notable peak at 300 bp, which can correspond to ALUs.
- The initial findings indicate in**candidate pathogenic de novo structural variants (SVs)** medically relevant genes.
- These preliminary results highlight the vast potential of this cohort. With the analysis of all **100,000 samples**, we are positioned to gain **unprecedented insights** into structural variants (SVs), their **inheritance patterns**, and their **implications for human health**.

dsanchez@m42.ae

@m42_health