Initial analysis of 160,000 whole genomes from the Emirati Genome Program (EGP)

J. Quilez[1], D. Sanchez [1], C. Cremin[1], H. Wu[1], H. Sajad[1], J. Mafofo[1], G. Katagi[1], H. Al Mabrazi[1], V. Kusuma[1], S. Pejathaya [1], S. Elavalli [1], F. Aldhuhoori [1], I. Eltantawy[1,2], M. Ibrahim[1,2], A. Alsuwaidi[1], K. Thangarajan[1], A. M. Yousif[3], M. Olbrich[4], M. S. Al Ameri[3], I. Chishty[1], S. A. Al Marzooqi[3], W. M. Abdulrahman[3], A. I. Al Mannaei[3], M. Mousa[4], H. Al Safar[4], I. Iqbal[3], F. Al Marzooqi[1], E-K. Albarah[1], A. Al Awadhi[1], T. R. Magalhaes[1]

Introduction

- The EGP has used three sequencing technologies to perform whole-genome sequencing (WGS) of approximately 600,000 participants so far.
- DNA is extracted from either blood or buccal swab (LRS only on blood-derived DNA).
- For the primary and secondary analysis (i.e., from the sequencing raw data to the individual's genetic variants across the genome) we have implemented optimized sequencing platform specific analysis workflows (**Fig. 1**).
- A key question is: **are the EGP's population-scale genomics datasets of good quality?**

Methods

- Selected 7 parameters and set cutoffs to define high-quality individual WGS datasets (**Table 1**).
- Evaluated those parameters in an initial data freeze comprising ~160,000 WGS (**Table 2**).

Results

- WGS samples for over 150,000 EGP participants (~95%) passed the high-quality criteria.
- All three sequencing platforms performed similarly well.
- As expected, blood performed better than buccal but the latter still was very good.
- Amongst failed samples, most were buccal and did so because of mapping rate only slightly below the 70% cutoff

Table 2. QC summary. Illumina and MGI have lower pass rate because they include buccal swabs while Oxford Nanopore Technologies (ONT) does not.

	Total Samples	Samples passing QC	Pass rate (%)
Illumina	107,739	100,530	93.3
MGI	44,305	43,115	97.3
ONT	7,755	7,602	98.0
All	**159,799**	**151,247**	**94.6**

The Emirati Genome Program (EGP) aims to **sequence the entire Emirati population** (~1 million people) by 2025.

So far, over **600,000 genomes** have been **sequenced** using short- and long-read sequencing.

Implemented **optimized technology-specific analysis workflows** for all three.

High quality of the genomics datasets, with about 98% of blood and 90% buccal swab samples passing high-quality criteria.

jquilez@m42.ae

@m42_health

1 M42 2 IROS Brought to you by M42

3 DEPARTMENT OF HEALTH 4 Khalifa University

Figure 1. Analysis workflows across sequencing platforms.

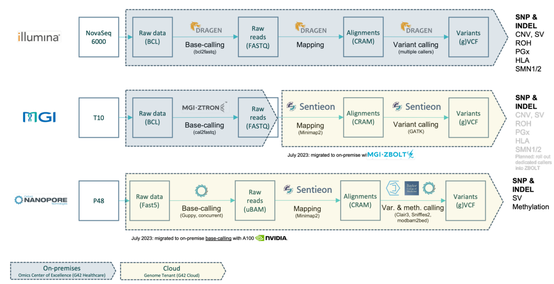

Table 1. 7 QC metrics and associated cutoffs defining high-quality WGS samples.

QC metric	High-quality cutoff
Yield (Gb)	≥70
% bases >Q30 (>Q10 for ONT)	≥85
Mapping rate (%)	≥70
Mean coverage (X)	≥25X
% genome >10X	≥95
Het/hom ratio	<2.5

Based on UK Biobank, 1000 Genomes, GATK Best Practices and internal benchmarks.

Acknowledgements

This work would have not been possible without the commitment of all the Emirati citizens who have voluntarily enrolled into the EGP and the support from the Abu Dhabi Department of Health.

DEPARTMENT OF HEALTH

الجينوم الإماراتي EMIRATI GENOME