An automated real-time solution for genetic disorders detection and classification

Gjorgji Madjarov[a],

Lukasz Krych[b],

David Galevski[c],

Anne Kristine Schack[b,d],

Aleksandar Nikov[c], and

Chris Kyriakidis[d]

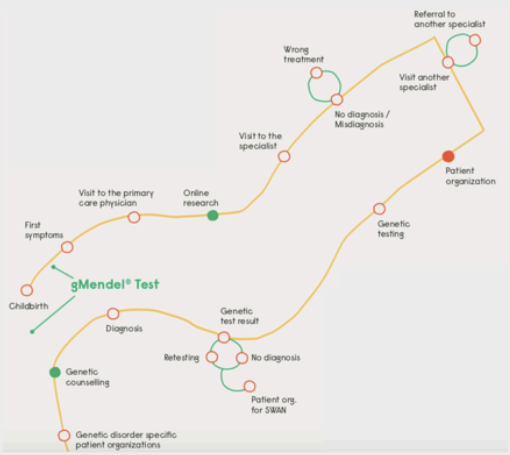
a) University Ss Cyril & Methodius, Skopje 1000, N. Macedonia

b) University of Copenhagen, Rolighedsvej 26, 1958 Frederiksberg, Denmark

c) Netcetera, Zypressenstrasse 71, 8040 Zürich, Switzerland

d) gMendel, Fruebjergvej 3, 2100 Copenhagen, Denmark

Background



- Diagnosis of genetic disorders (GDs) presents a great challenge, impacting the health, survival, and well-being of >350M people worldwide [1], [2]
- Currently, families with children affected with GDs often receive multiple misdiagnoses. The patients are faced with a lengthy and burdensome path to diagnosis, creating a heavy human, economic and societal burden, with lifetime costs surpassing €2.5M [3],[4]
- On average it takes six to eight years before a person with a GD receives the correct diagnosis and more than 40% of GDs patients are misdiagnosed at initial presentation
- Currently, diagnosis of GDs rely on a set of conventional techniques (e.g., Sanger sequencing, MLPA- and PCR-based techniques)
- Advanced genomic technologies have been changing the landscape in GDs. However, the translational gap between these and their clinical implementation has not been bridged, mainly due to:
  i) lack of clear demonstration & real-world evidence
  ii) demanding, time-consuming & complex analysis
  iii) technological limitation linked to the read length iv) lack of scalability due to costs, making it only available to few [5]

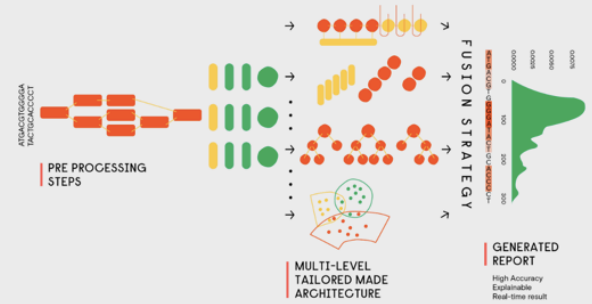
Tests and validation [7], [8]

- Several synthetically prepared samples from healthy persons and aberration positive persons
- The sequences that met the quality criteria are 81% of the total number of sequences generated by the Oxford Nanopore Technologies device.
- Demultiplexing of the barcodes in the sequence data is performed with macro-average precision of 98.2% and unclassified reads of 21%.
- In terms of computational efficiency, the proposed algorithm for demultiplexing is an order of magnitude faster than the guppy barcoder
- The ground truth chromosome classification is defined using the Smith–Waterman algorithm. In comparison to this algorithm, the proposed algorithm showed significant improvement in terms of computational efficiency.
- The chromosome classification is performed with macro-average precision of 99.1% and unclassified reads of 4%.
- The proposed algorithm manages to align a sequence in 1.1 milliseconds which is more than 50 times faster than the Smith–Waterman algorithm.
- The initial classification on the selected genetic disorders (Klinefelter, Turner, Down, Edwards, Patau and Prader-Willi/Angelman syndromes) obtained on a small subset (20%) of the synthetically prepared samples showed very promising results too (macro-average precision of 98% and computational efficiency of 0.4 milliseconds per sequence).
- All the experiments are performed on one referent hardware architecture (Intel i7 10th generation, 8 cores, 32 GB RAM, no CUDA) using thread parallelism of 10.

Process - step by step

1. Preparation of the library (multiplexing) [6], [7]
2. Oxford Nanopore Technologies device GridION x5 that provides max output of 420 bases / second is used for the DNA sequencing (2,560 channels across the device are sequenced at once)
3. The base-called .fastq files are stored and a custom tailor-made analysis pipeline for chromosomal aberration detection is triggered
4. Only the sequences that meet the quality criteria
   - Phred quality score above 10 and sequence length between 900 and 1200
   - Experimenting with sequence length between 500 and 600 (exploring the effects on the demultiplexing speed and the efficiency of chromosomal aberrations detection)
5. Demultiplexing of barcodes (person classification) [8]
6. Chromosome classification and chromosomal aberration detection [8]
7. Multiple base and conceptual models are built utilizing data driven sequence modeling for the specific chromosomal aberration (experimental phase)
   - CNN (Convolutional Neural Networks), RNN (Recurrent Neural Networks) and transformers
8. Following a multi model fusion strategy, the outcomes of the individual models are combined into a single and accurate decision.

The Phivea™ Platform



References

1. Nguengang Wakap, S., Lambert, D.M., Olry, A. et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet 28, 165–173 (2020). doi:10.1038/s41431-019-0508-0

2. Ronicke, S., Hirsch, M., Türk, E. et al. Can a decision support system accelerate rare disease diagnosis? Evaluating the potential impact of Ada DX in a retrospective study. Orphanet J Rare Dis 14, 69 (2019). doi:10.1186/s13023-019-1040-6

3. Xiang Yan, Shenjing He, Dong Dong. Determining How Far an Adult Rare Disease Patient Needs to Travel for a Definitive Diagnosis: A Cross-Sectional Examination of the 2018 National Rare Disease Survey in China. Int. J. Environ. Res. Public Health 2020, 17, 1757. doi:10.3390/ijerph17051757

4. Chediak, L. The Bad Economics of the U.S. Health Care System Shows Up Starkly in Its Approach to Rare Diseases. Time magazine, Feb 29, 2020

5. Schwarze, K., Buchanan, J., Fermont, 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. Genet Med 22, 85–94 (2020). doi.org. /10.1038/s41436-019-0618-7

6. Krych, Ł., Castro-Mejía, J.L., Forero-Junco, L.M. et al. DNA enrichment and fragmentation method for species-level identification and strain-level differentiation using ON-rep-seq. Commun Biol 2, 369 (2019). doi.org/10.1038/s42003-019-0617-x

7. Schack A.K. & Krych, Ł., unpublished results (2021)

8. Galevski, D., Madjarov, G. & Nikov A. unpublished results (2021)

for more information, please write to chris@g-mendel.com or visit www.g-mendel.com