Visualizing Genetic Mutations Occurrence using Big Data

Silvana Albert

Prof. Dr. Gabriela Czibula

Babes-Bolyai University, Cluj Napoca, Romania
Contact: albert.silvana@cs.ubbcluj.ro

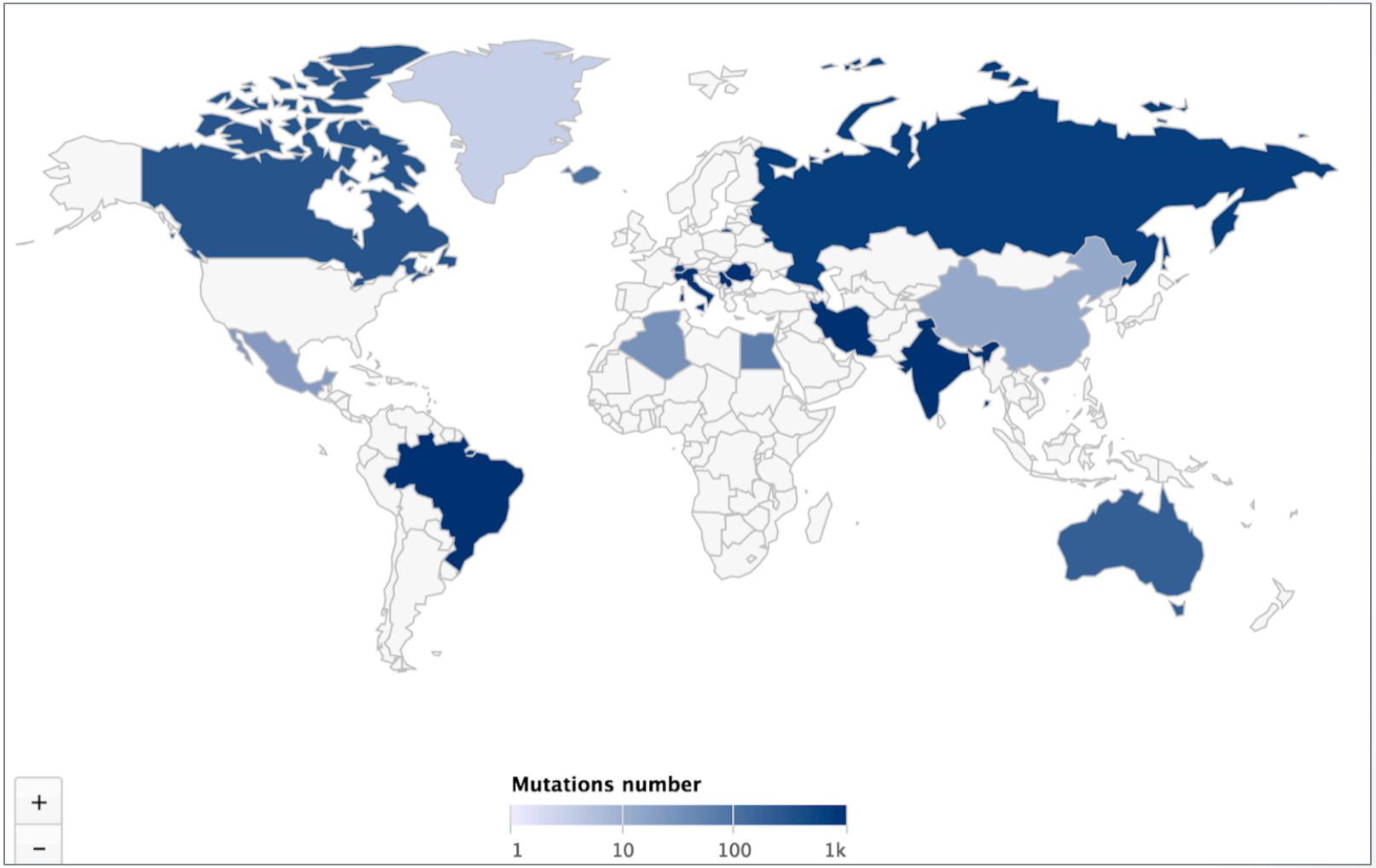
Introduction

- Often a mutation on its own can not provide enough information about a disorder or disease
- Many genes can cause a single disease and in the same time a single gene can cause multiple diseases
- **Big data** is the term used for describing large collections of data sets that will be analyzed, visualized and transferred
- In the vast context of genetic mutations, Big Data constitutes all the information related to a gene and it's possible anomalies that is stored and will be used for further analysis
- We propose an open source solution for visualizing genetic mutations

Methods

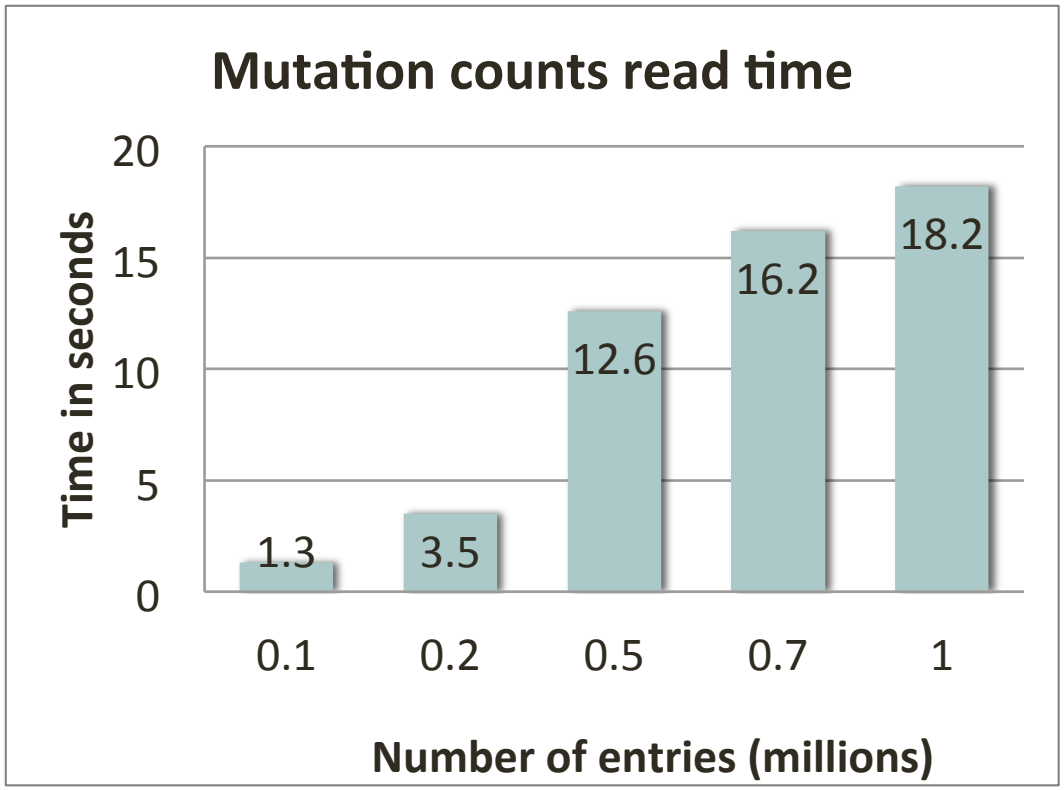
- Provide demographics and metrics about diagnostics and mutations
- See the occurrence of a mutation in a particular geographic region
- Information is stored regarding the patient's gender, location, age, the age when diagnosed, date of death if applicable, all discovered disorders with afferent mutations and for each mutation, its locus, professional exposure substances and exposure time to each of them
- DBMS Apache Cassandra
- To ensure replication and high availability, a number of three Cassandra nodes are deployed in a cluster through Docker

Example (using mock data): The number of women that were diagnosed at the age of 35 with Breast Cancer, have the mutation c.3548A>G on gene BRCA1, is currently under treatment and were professionally exposed to chemical agent ethylene oxide.



Results

Using the proposed Data Model, correlations can be done about how the person's environment and exposure to harmful chemicals can cause mutations or accentuate the effects of existing inherited ones.



Available on Github

https://github.com/albusilvana/genehive

Bibliography

1. S. Mukherjee, The Gene: An Intimate History, First Scribner hardcover edition. New York: Scribner, 2016
2. B. Wang, L. Ruowang, and W. Perrizo. Big Data Analytics in Bioinformatics and Healthcare. IGI Global, Hershey, PA, USA, 1st edition, 2014.
3. R. Hecht and S. Jablonski. NoSQL evaluation: A use case oriented survey. In 2011 International Conference on Cloud and Service Computing, pages 336–341, Dec 2011.
4. R. Hecht and S. Jablonski. NoSQL evaluation: A use case oriented survey. In 2011 International Conference on Cloud and Service Computing, pages 336–341, Dec 2011.
5. MongoDB. Internet: www.mongodb.com, [01.09.2017]
6. Cassandra, Internet: http://cassandra.apache.org [01.09.2017]
7. Membase, Internet: http://www.couchbase.com/ [01.09.2017]
8. Allegro, Internet: http://franz.com/agraph/allegrograph/ [01.09.2017]
9. S. Albert, A Big Data Approach in Mutation Analysis and Prediction, Studia Universitatis Babes-Bolyai Series Informatica, 62(1), 75-89, 2017
10. Github Internet https://github.com/albusilvana/genehive [03.10.2017]
11. Docker Internet https://www.docker.com [03.10.2017]