Analyzing the Mutant Gene

AI in the Genetics of Disease Prevention and Cure

TATA CONSULTANCY SERVICES

Rajgopal Srinivasan

IN BRIEF

Deoxyribonucleic acid (DNA) is the data cells read to sustain life. But when DNA mutates, this modified data can cause serious diseases and even death.

However, not all mutations are deleterious. Of the three million mutations that every human carries, only a very small fraction is disease-causing. Identifying deleterious mutations early, as in newborns, can prevent disease onset or make it treatable.

Joint research by TCS, the University of California, Berkeley (UCB), and the University of California, San Francisco (UCSF) shows this is now possible through advanced computing techniques such as parallel computing, machine learning, and deep learning applied to next generation sequencing and newborn genome sequencing. The trio has analyzed, interpreted, ranked, and modeled genomic mutations, making it possible to predict, explain, prevent, and treat diseases with unprecedented speed, accuracy, power, and cost savings.

The human body has 37 trillion cells, each of which contains 100 trillion atoms. Our DNA is made up of 200 billion atoms. Although that is just 0.2% of all the atoms in our cells, those atoms are crucial. DNA is the data our cell machinery reads to carry out every cellular process that sustains our life.

What happens when our DNA mutates? Even mutations that change as few as 20 atoms (10^{-13}% of a cell's atoms) can lead to cancer, untreatable diseases, and, worst of all, death.

Every one of us has around three million mutations.

Thankfully, however, our bodies have cellular mechanisms that can repair the damaged DNA and make exact copies of our DNA during cell reproduction. DNA can also accommodate mutations without significantly affecting the viability of an individual. Humans carry millions of mutations that do not harm their ability to live and reproduce.

Thus, mutant DNA could be harmless or deadly, not only after the birth of the individual, but even

Fact File

TCS Research: Genomics

Outcomes: Over 25 cases of unknown SCID solved, First ever large scale evaluation of Genomics as a tool for New born screening

Principal Investigators: Rajgopal Srinivasan

Academic Partners: University of California Berkeley, University of California San Francisco

Techniques Used: Machine Learning, Network Analysis

Industries Benefited: Healthcare and Life Sciences

Patents: 3 filed, 1 granted

Papers: 11 Conference Presentations

while we are embryos. Mutations—or variations, as they are also called—in DNA can be inherited, and can occur in the embryonic stage or later in life. What is more, since all future cells are made from the embryonic cells, these can also carry these mutations and make the individual susceptible to diseases.

How can we find out if a genetic mutation is likely to be benign or deleterious?

Next generation sequencing

The first step is to identify the list of mutations an individual carries. This is achieved through next generation sequencing (NGS). NGS is a wet lab procedure that generates huge volumes of data for each individual that have to be analyzed using sophisticated high performance parallel computing. Once the mutations in an individual are identified, specialized computational methods, including machine learning, are used to assess the impact of the mutation on the functioning of the cell and, hence, the organism. Because of its power to resolve diseases at the genetic level, NGS is increasingly finding use in the clinic, for understanding a patient's clinical symptoms, i.e., phenotypes, which it can do in just hours what earlier took years.

One common application of NGS is for children with severe disease phenotypes who are born to healthy parents. This genetic condition happens when each parent has one damaged and one working copy of a gene, but passes on the damaged copy to their offspring. The child, inheriting only damaged genes, manifests the disease.

If we can identify such children right at birth, we can catch the disease early, perhaps treat it, and, in some cases, may also prevent its onset.

Newborn screening and NGS

Newborn screening (NBS) is a commonly used process to identify the children at birth who may be at risk for a set of diseases. These tests are usually easy to administer and economical, but have to be followed up with more detailed and expensive tests to confirm the presence of disease. Often, the follow-up tests will be

Specialized computational methods are used to assess the impact of mutation on the cell function.

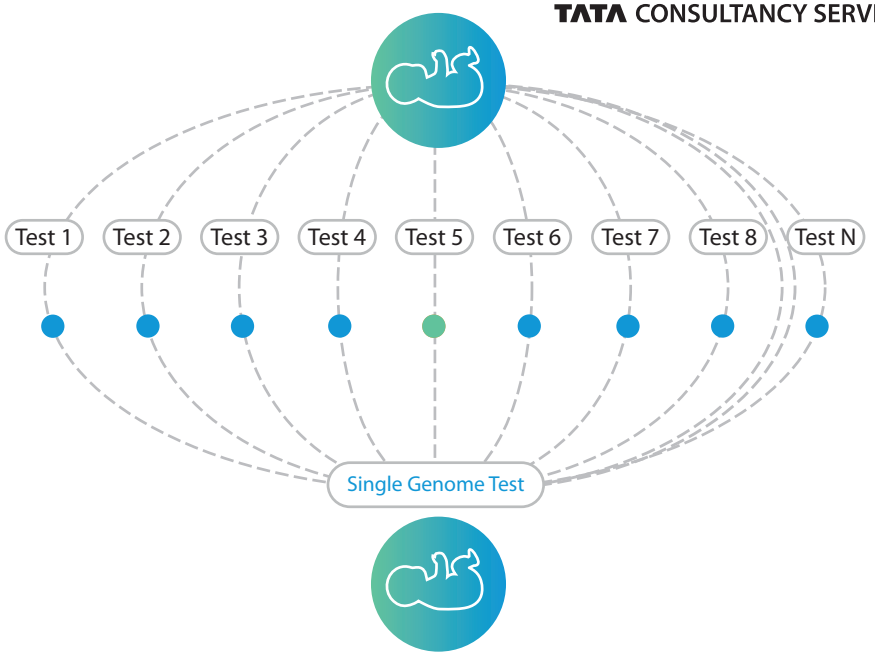

Figure 1: *Newborn Screening Using Genome Sequencing*

inconclusive and such cases are good candidates for NGS analysis. In such cases, genomes of the healthy parents, the affected child, and any siblings of the child are sequenced and analyzed for mutations that may be explanatory of the disease.

Harnessing AI and machine learning for advanced NGS strategies

Working with scientists at UCB and UCSF, TCS has developed accurate and sensitive tools to analyze such genomes. Using machine learning and, more recently, deep learning the team has been able to pin the disease-causing mutations down from thousands of mutations to just one or two gene variants in a child at risk. The mutations are first analyzed, next interpreted, and then ranked.

Analysis: A typical analysis includes mapping the NGS-derived DNA sequence of the patient to a reference genome, identifying variants, conducting quality assessments, annotating the variants, and prioritizing approximately 100-variant shortlist according to the variants' relevance to the phenotype(s) under investigation. This results in precise diagnosis and patient management (around 25%, according to a recent estimate, with slightly better percentages when trios are studied).

Interpretation: The complex part of the disease gene identification analysis is interpreting the NGS data. This is typically done through progressive filtering, based on existing information—chiefly, frequency of the variant in the general population, inheritance patterns, and predicted pathogenicity.

Gene ranking: In typical clinical settings, extensive lab investigations arm the physician with detailed knowledge of a set of disease genes. This helps the physician identify the variants that have caused the disease. However, that well-studied (target) gene-set is pitifully small, when compared with all possibly deleterious mutations. This presents a daunting problem. For genes not well-understood, we used computational methods to rank the gene, based on its

similarity with a well-studied gene. This has been successful.

Network-based NLP algorithm use: Many computational approaches have been developed to prioritize candidate genes. But they use prioritization algorithms based on existing lists of either target genes or disease- and phenotype-relevant key terms. TCS, on the other hand, has developed a comprehensive network-based natural language processing (NLP) approach that mines known gene-disease or phenotype associations in the published literature. In 2014 we tested this to successfully identify a known human oncogene as having caused severe combined immunodeficiency (SCID) in a male infant (See Figure 2).

Predicting phenotypes from genotypes

Success in solving the problem of identifying mutations that explain a set of phenotypes brings us to an inverse—and much more challenging—problem: predicting phenotypes from mutations.

In our analysis of mutations, we have found that even the genomes of healthy humans carry dozens of mutations that most interpretation algorithms classify as disease-causing. How does one explain that? The answer is that such algorithms presently lack the needed accuracy.

An accurate algorithm for predicting disease-causing mutations would replace other tests for screening for metabolic disorders. And the higher the accuracy, the larger should be the number of tests replaced.

Algorithms for NBSeq

TCS, collaborating with UCB and UCSF, has therefore been developing newborn genomic sequencing (NBSeq) algorithms. This collaborative NBSeq work is aimed at checking if the suite of NBS tests can be replaced by a single test on the genome sequence of the newborn. Initial results show that these accurately predict the mutation-caused diseases in up to 90% of the cases. Although that is still not good enough to replace current screening methods, our NBSeq algorithms have, in some cases, diagnosed not only previously undiagnosed disorders but also diagnosed, far more accurately than other comparable algorithms, even disease subtypes.

Further work is underway on improving the predictive power of our NBSeq framework.

DNA editing for the nonprotein coding genome

Over 95% of genome does not code for proteins, but regulates where, when, and how much protein is produced. Mutations in the non-coding DNA can also have disease-producing consequences. Unfortunately, the lack of data on the effects of mutation hampers the interpretation of noncoding regions of the genome.

To develop improved interpreters for mutations in the noncoding genome, TCS has been using deep learning methods on the data available. In addition to that, together with UCB and UCSF, we have been using a recently developed DNA editing technique, CRISPR-Cas9, to create large numbers of well-defined mutations in single-cell DNA for observing the phenotypes manifested. This data will be used by conventional machine learning and deep learning methods to develop improved models for genome interpretation.

Finding the causal gene for a set of clinical symptoms

In 2014, a T-cell excision circle newborn screening, done in (location of screening) a male infant, showed severe T-cell lymphopenia. The parents were non-consanguineous and healthy, with no family history of immune deficiency. On imaging, the patient also presented an absent corpus callosum.

To find the causal gene mutation, chromosome breakage studies, a copy number array, and sequencing of SCID genes were performed, without any success. None of the shortlisted variants occurred in our target list of reported human primary immunodeficiency genes.

So we accessed human phenotype T-cell-related terms from the Mouse Genome Informatics (MGI) and the Human Phenotype Ontology databases, expanding the target list. First, we included the interacting partners of 48 primary immune deficiency genes known to be involved in T-cell development and primary T-cell deficiencies in human. Second, we expanded the list further to include genes with T-cell-related phenotype terms. From this list, we chose five high-quality variants of three genes—*BCL11B*, *NRLP1*, and *USO1*—for further analysis.

Lab experiments confirmed *BCL11B* as the gene responsible for the observed phenotypes.

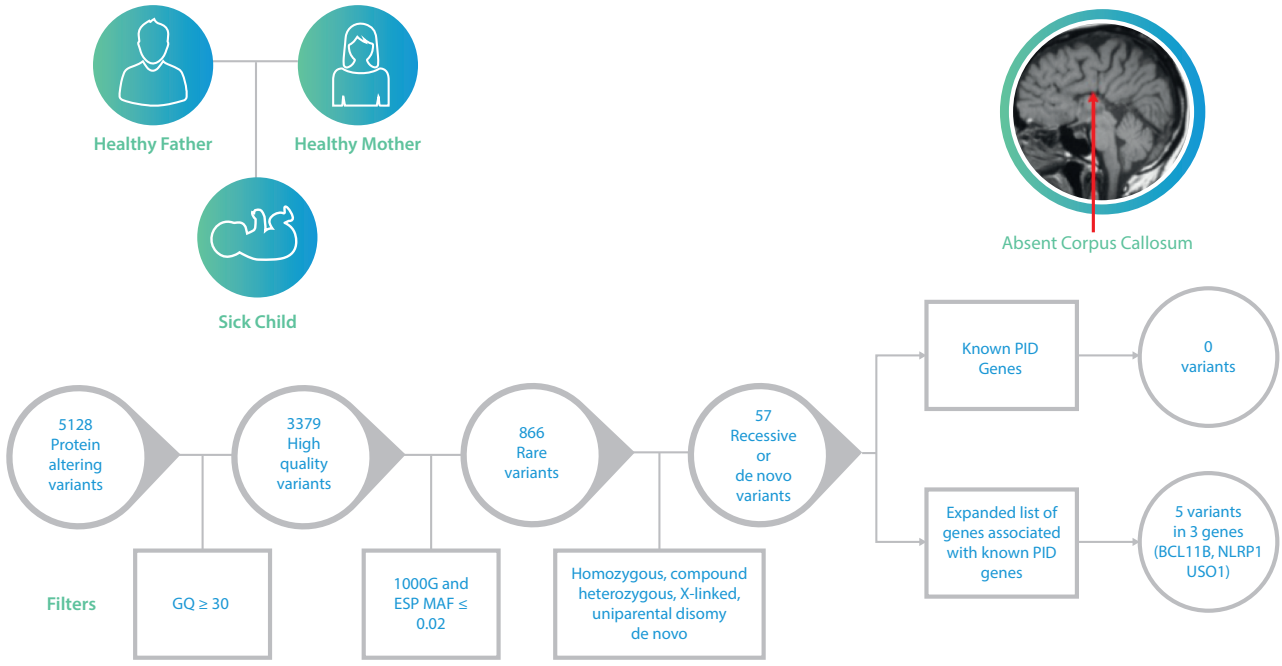

Figure 2: *Finding the Causal Gene for a set of Clinical Symptoms*

The future: longer lives and personalized healthcare

As the cost of gene sequencing continues to become more affordable, we will gain unprecedented ability to use genomic information to personalize our well-being, cure diseases, and live longer and healthier lives. Through genomic knowledge, we will be able to accurately estimate the biological—and not merely chronological—ageing of our cells. This, in turn, will help us develop specific tissue- and organ-targeted health interventions.

As our ability to interpret genomic variations improves with DNA editing techniques, we will see a whole slew of methods to edit our way out of diseases, rather than resort to drugs for cures. Our genome sequences, determined at birth or even earlier (at the embryonic stage), will serve as a lifelong reference to better plan our future.

In parallel, advancements in understanding the microbiome (the microbial communities that reside in our bodies) will enable earlier diagnosis of sub-clinical diseases and, perhaps, even prevent disease through pro- and prebiotic interventions.

About 25 years from now, we will look back at current-day treatments for diseases, such as cancer, as medieval.

As genome interpretation improves, we will edit disease causing genes, rather than use drugs for cures.

Rajgopal Srinivasan

Rajgopal Srinivasan is a Chief Scientist at TCS Research and Innovation, and heads the Life Sciences Research Area. A graduate in Chemistry from The Indian Institute of Technology Madras, Srinivasan holds a PhD in Chemistry from the University of Illinois at Urbana-Champaign in the USA. Following post-doctoral stints at Eli Lilly and Co at Indianapolis, Washington University at St. Louis, and Johns Hopkins Medical School, he worked as a research professor at the Johns Hopkins University in the department of Biophysics. He joined TCS in 2003 as part of its Corporate R&D Center in Hyderabad, India.