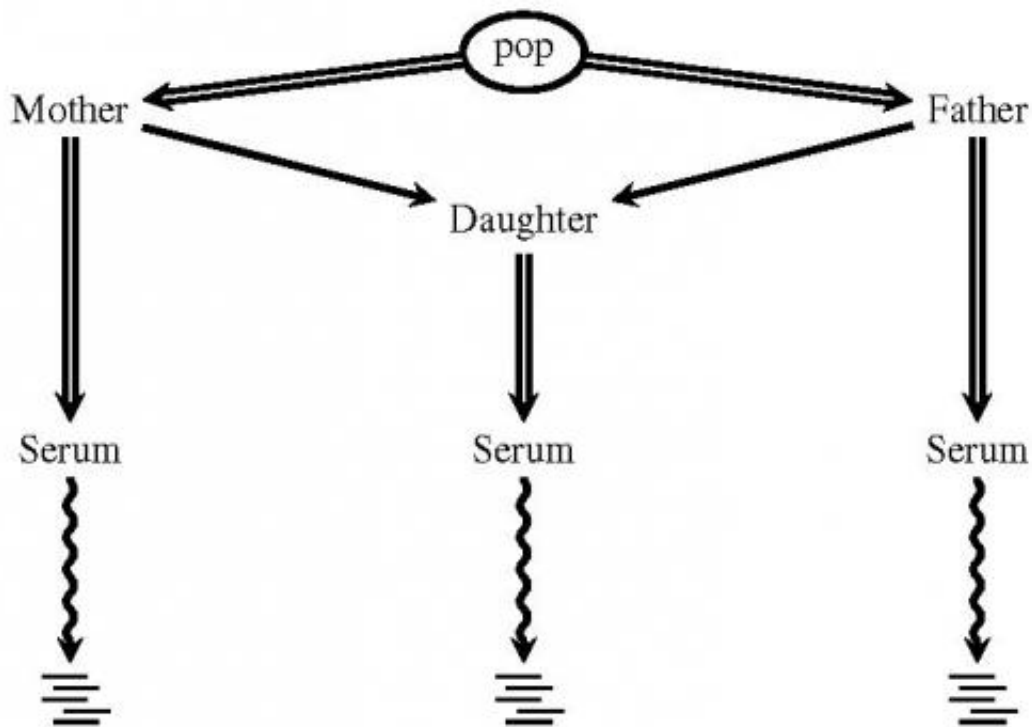# New tool enhances the search for genetic mutations

August 25 2013, by Richard Harth

In this illustration, population sequence data, sequencing error rates and other information are used to evaluate the prior probability of mutations at a given site in the genome. This information is then weighed against direct genome sequencing, using the serum of mother, father and daughter. The daughter receives a single base from each parent at a given genomic site. Multiple sequencing rounds are used to further compensate for error. Credit: Biodesign Institute at Arizona State University

Concealed within the vastness of the human genome, (comprised of some 3 billion base pairs), mutations are commonplace. While the majority of these appear to have neutral effect on human health, many others are associated with diseases and disease susceptibility.

Reed Cartwright, a researcher at Arizona State University's Biodesign Institute, along with colleagues at ASU, Washington University and the Wellcome Trust Sanger Institute, Cambridge, UK, report on a new software tool known as DeNovoGear, which uses statistical probabilities to help identify mutations and more accurately pinpoint their source and their possible significance for health.

Improvements in the accuracy of mutation identification and validation could have a profound impact on the diagnosis and treatment of mutation-related diseases.

"These techniques are being considered in two different realms," Cartwright says. "The first is for pediatric diseases." Here, a child with an unusual genetic disease may undergo genomic sequencing to see if the mutations observed have been acquired from the parents or are instead, unique to the child. "We can identify these mutations and try to detect which gene may be broken," he says.
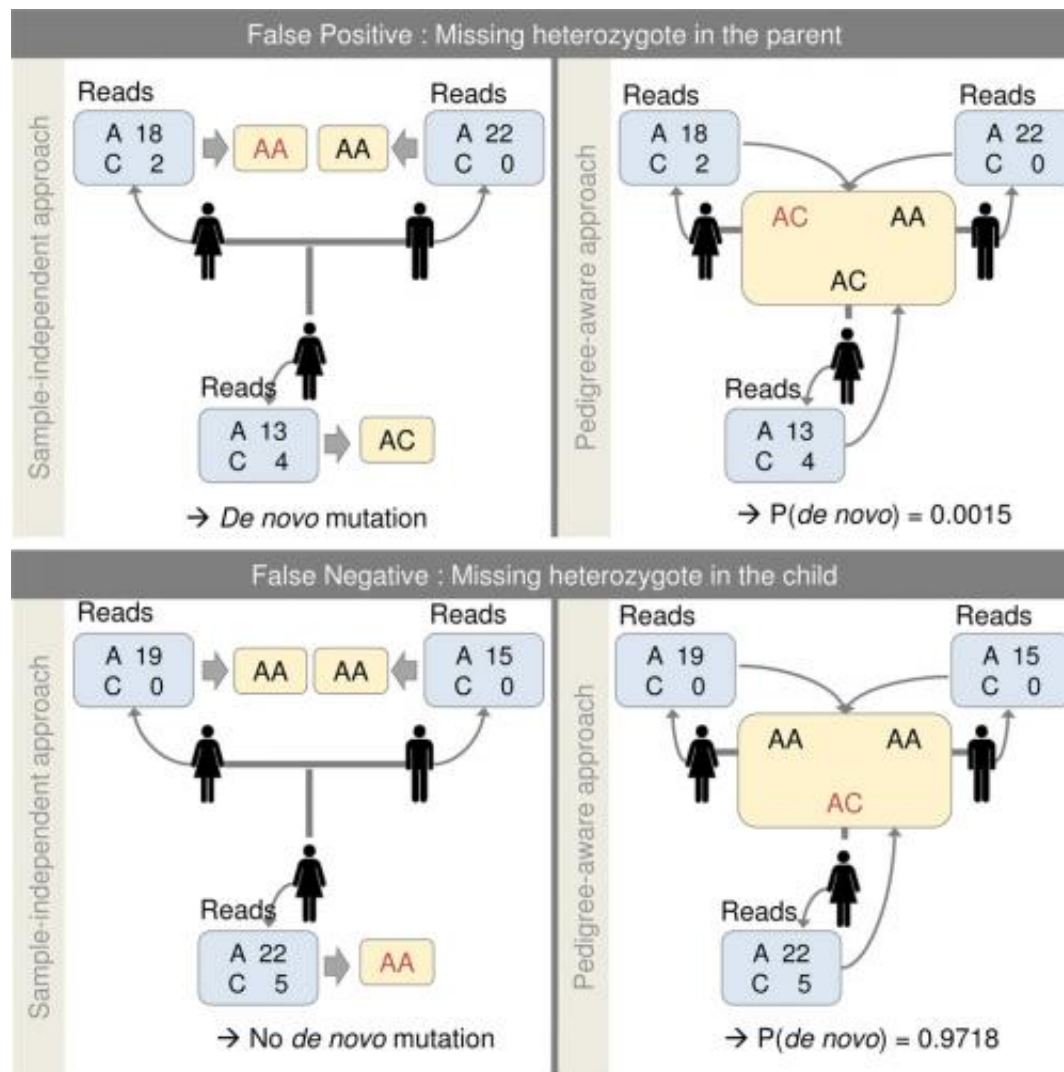
The second application is for cancer research, where tumor tissues are genetically compared with normal tissue. Many now believe that the identification of a specific cancer mutation may eventually permit clinicians to customize a treatment for that tissue type. "We are developing methods to allow researchers to make those types of analyses, using advanced, probabilistic methods," Cartwright says. "We actually model the whole process."

Indeed, the method described provides the first model-based approach for ferreting out certain types of mutations. The group's research results

appear in today's issue of the journal *Nature Methods*.

One of the primary goals in genetics is to accurately characterize genetic variation and the rate at which it occurs. Searching for DNA mutations through genetic sequencing is an important ingredient in this quest, but many challenges exist. The current study focuses on a class of mutations that play a critical role in human disease, namely de novo mutations, which arise spontaneously and are not derived from the genomes of either parent.

Traditionally, two approaches for identifying de novo mutation rates in humans have been applied, each involving estimates of average mutations over multiple generations. In the first, such rates are measured directly through an estimation of the number of mutations occurring over a known number of generations. In the second or indirect method, mutation rates are inferred by estimating levels of genetic variation within or between species.

False Positive : Missing heterozygote in the parent

Sample-independent approach

Reads
A 18
C 2
→ AA   AA ←
Reads
A 22
C 0

Reads
A 13
C 4
→ AC

→ *De novo* mutation

Pedigree-aware approach

Reads
A 18
C 2
AC   AA
AC
Reads
A 22
C 0

Reads
A 13
C 4

→ P(*de novo*) = 0.0015

False Negative : Missing heterozygote in the child

Sample-independent approach

Reads
A 19
C 0
→ AA   AA ←
Reads
A 15
C 0

Reads
A 22
C 5
→ AA

→ No *de novo* mutation

Pedigree-aware approach

Reads
A 19
C 0
AA   AA
AC
Reads
A 15
C 0

Reads
A 22
C 5

→ P(*de novo*) = 0.9718

In the upper left-hand panel, sequencing results for mother, father and daughter are examined, based on multiple reads of a given site in the genome. Each parent is judged to be homozygous -- carrying two copies of the same base, (AA). Sequencing of the daughter identifies a heterozygous state at the same site (AC) indicating a de novo mutation, though the new software indicates a very low probability for a mutation at this site, suggesting a false positive result (upper right-hand panel). The lower left and right panels show a failure to identify a heterozygous site in the child, which results in a false negative (with the new method indicating a very high likelihood of mutation at this site). Credit: National Center for Biotechnology Information

In the new study, a novel approach is used. The strategy, pioneered in part by Donald Conrad, professor in the Department of Genetics at Washington University School of Medicine and corresponding author of the current study, takes advantage of high throughput genetic sequencing to examine whole genome data in search of de novo mutations.

"This collaboration started a few years ago, when Donald and I were both working on mutations for the 1000 genomes project," Cartwright says, referring to an ambitious project to produce a comprehensive map of human variation using next-gen sequencing.

The mutations under study may take the form of either point mutations—individual nucleotide substitutions, or so-called indel (insertion-deletion) mutations. In the latter case, single nucleotides or nucleotide sequences may be either added or subtracted from the genome.

While point mutations and indel mutations can both have adverse affects on health, indels are significantly more difficult to identify and verify. They have a strong potential to cause havoc when they occur in coding portions of the genome as the addition or deletion of nucleotides can disrupt the translation process needed to accurately assemble proteins. (The current study is the first paper to use model-based approaches to detect indel mutations.)

A seemingly simple approach to pinpointing mutations is to compare sequence data from each parent with sequence data from their offspring. Where changes exist at a given site in the offspring, de novo mutations can be inferred and their potential affect on human health, assessed.

In reality, such efforts are complicated by a number of potential sources of error, including insufficient sampling of the genome, mistakes in the gene sequencing process and errors of alignment between sequences.

The new method uses a probabilistic algorithm to evaluate the likelihood of mutation at each site in the genome, comparing it with actual sequence data (See Figure 1).

Human cells contain two copies of the genome—one from each parent. For most positions in the genome, the bases from each parent are the same or homozygous but occasionally, they are different or heterozygous.

Errors derived from conventional methods can take the form of false negatives, particularly when gene sequencing misses heterozygous sites in the genotype of the child. On the other hand, failure to identify a heterozygous site in one of the parents can lead to a false positive result. (See Figure 2)

The new method assesses the chances of an actual de novo mutation for the child in the example as very low, hence revising the false positive result (upper right panel). In the lower right panel, the new method finds a high probability for a de novo mutation, thereby revising the conventional method's false negative.

In the current study, data from the 1000 genomes project was analyzed using DeNovoGear, with markedly improved accuracy. The technique will assist ongoing efforts to better understand which mutations contribute to sporadic disease or cancer in individuals, the distribution of mutations and their characteristics across populations.

The power of this technique comes from its probabilistic model which calculates the probability of a de novo mutation at a site based on estimations of mutation rates, sequencing error rate, and the initial genetic variation in the population from which the parents arise. This model is able to consider multiple explainations for experimental observations and decide between them. The probabilities are used to

indentify candidate loci which are then evaluated using target resequncing.

Given adequate data of genetic pedigree, the method is able to distinguish germline from somatic mutations in an automated manner with high accuracy. "Our goal is to develop software that will allow researchers and clinicians to estimate a range of mutation types, faster, more accurately, and cheaper," Cartwright says.

In addition to further refining the DeNovoGear software, Cartwright's group plans to more closely examine normal human tissue in order to establish rates of somatic mutation. Some of the specific mutations currently associated with cancer for example, may actually be part of normal variability, which appears to be much greater than originally assumed. "No one has really looked at this at the level we're interested in."

More information: DeNovoGear: de novo indel and point mutation discovery and phasing, DOI: 10.1038/nmeth.2611

Provided by Arizona State University