
*DRAGEN 4.2 software includes several targeted callers for detecting copy number variants in high-homology regions. Photo by Illumina*

# Finding the needle in the genomic haystack with DRAGEN

*Illumina scientists tailor-made research solutions for detecting genes in difficult-to-read regions that cause both common and rare diseases*

WHEN SCIENTISTS WANT TO SEQUENCE a DNA sample on an Illumina system, they don't try to read all 4 billion base pairs of the genome at once. Instead, they slice the DNA into short fragments of about 500 base pairs that are easier to work with and faster to read.

DNA samples, in the form of a small amount of tissue or fluid, usually contain many cells, and thus many copies of the organism's genome—so once the system captures images of the fragments, it reassembles the data for one complete sequence by comparing where the fragments overlap.

Think of it like tossing several identical copies of a book into a paper shredder, each one at a random angle. You can't reassemble any individual copy like you would a puzzle, because all the pieces are the same shape. (Especially if you don't have an intact book to reference.) But, since each copy of the book was shredded in different random locations from the others, you can match up fragments from different copies based on where the text overlaps.

If the species of interest has never been sequenced before, scientists must rely on these overlaps, known as contiguous regions, or "contigs," to build a reference genome. Fortunately, human reference genomes are available thanks to The Human Genome Project,[1] completed in 2003, and the ongoing work of the Genome Reference Consortium.[2] Every individual human shares 99.99% of the same base pairs, so scientists can identify an individual's genetic variants by comparing them to existing references.

Unfortunately, in many regions of the human genome, the sequence of base pairs is highly repetitive. Entire genes—many thousands of base pairs long—may be duplicated multiple times, with only a handful of base pair variations to differentiate the copies. Furthermore, the number of duplications of given genes, and the specific differences between the copies, frequently varies from person to person.

These regions of "high homology" are notoriously difficult to analyze, even with a reference genome available. Fragments from them are likely to "fit" in several possible locations, leaving the system with low confidence that it's aligned them correctly.

Unfortunately, many genetic diseases result from

*1. https://www.genome.gov/human-genome-project*
*2. https://www.ncbi.nlm.nih.gov/grc*

having an atypical number of copies of specific genes, or a variant in just one gene of a multigene family with many copies—so in order to screen for these diseases, sequencing systems and data analysis pipelines must be sophisticated enough to accurately detect variants even in high-homology regions.

One way to tackle this is to perform longer reads. If the fragments are long enough to bridge the homologous region, reads can be mapped unambiguously to different copies of that region in the reference genome. For example, Illumina Complete Long Reads[3] can create reads up to about 10,000 base pairs. But some homologous regions are still longer than that.

Luckily, Illumina scientists have developed solutions to this problem even for short reads. Targeted callers are tailor-made to quickly and accurately detect the copy number (and other variants) of genes associated with specific diseases. Read on to learn about three of them—congenital adrenal hyperplasia, alpha thalassemia, and atherosclerosis—and how DRAGEN Secondary Analysis software[4] version 4.2 pierces the fog to locate their genetic origin.

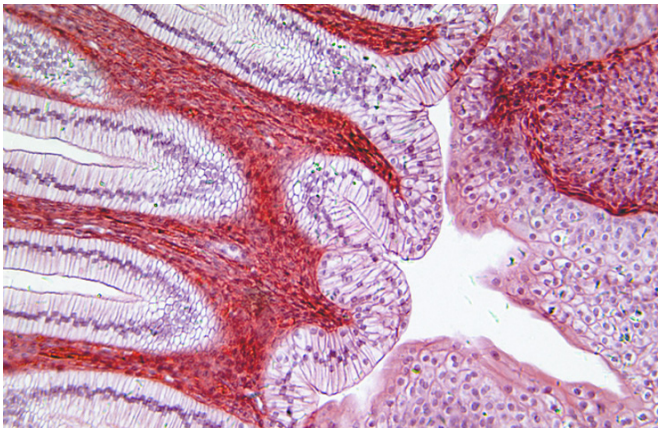

PHOTO: GETTY IMAGES

*Without the gene CYP21A2, our adrenal glands can't synthesize certain hormones, and our kidneys can't retain salt.*

### *CYP21A2* and congenital adrenal hyperplasia

Our adrenal glands, located atop our kidneys, help regulate the levels of sodium and potassium in our blood. They do this by synthesizing the hormones cortisol and aldosterone, with the help of the protein 21-hydroxylase.

That protein is coded by the gene *CYP21A2*, which unfortunately sits in just one of two copies of the highly homologous RCCX region. The other copy of RCCX contains a very similar "pseudogene," *CYP21A1P*, which has no function. This homology confuses not just gene sequencing, but human reproduction: When parental chromosomes recombine to form their child's DNA, they might mistakenly mix genetic material between the functional *CYP21A2* and the nonfunctional *CYP21A1P*, impairing the resulting gene's ability to code for 21-hydroxylase… or they might delete the gene entirely.

As with many inherited diseases, a child usually shows no symptoms as long as they have one functioning copy of *CYP21A2*—they're only a carrier for the condition. But having two nonfunctional copies leads to congenital adrenal hyperplasia (CAH).

CAH caused by 21-hydroxylase deficiency occurs about once in 10,000 to 15,000 live births.[5] If a developing fetus's adrenal glands have a decreased level of 21-hydroxylase, they can't synthesize as much cortisol and aldosterone as their body needs. Then, the excess materials they would've used for synthesis accumulate and instead form androgens, or male sex hormones. This is called "simple virilizing CAH." Girls with excessive androgens may develop ambiguous genitalia; boys may not show any external signs and require targeted screening to diagnose.

Children born without either functional *CYP21A2* gene—and thus a complete lack of 21-hydroxylase—can't synthesize cortisol and aldosterone at all, and their kidneys can't retain salt. This severe "salt-wasting CAH" causes dehydration, diarrhea, vomiting, and adrenal crisis within days or weeks of birth, and is often fatal.

The *CYP21A2* targeted caller in DRAGEN 4.2 is a research-use-only tool specially geared to detect a wide range of variation in the gene: It can discern the number of copies of the RCCX region, gene deletions, gene conversions, and 33 different small variants in the gene or the pseudogene. Illumina scientists tested the caller on a large selection of publicly available data from The 1000 Genomes Project,[6] including healthy individuals, CAH carriers, and 16 CAH cases. The caller successfully detected the RCCX copy number, full gene deletions, and small variants in every case.

*To learn about how the CYP21A2 caller works and the methods used to test it in greater detail, read this article[7] on Illumina's Genomics Research Hub.*

*3. https://www.illumina.com/products/by-brand/complete-long-reads-portfolio.html*
*4. https://www.illumina.com/products/by-type/informatics-products/dragen-secondary-analysis.html*
*5. https://pubmed.ncbi.nlm.nih.gov/15514016 6. http://www.internationalgenome.org*
*7. Belyeu J, Klötzl F, Roller E, Newman E, Onuchic V, Bekritsky M. "Overcoming high homology to detect variation in CYP21A2 with whole-genome sequencing in DRAGEN." Illumina Genomics Research Hub. 2023. https://www.illumina.com/science/genomics-research/articles/CYP21A2.html*

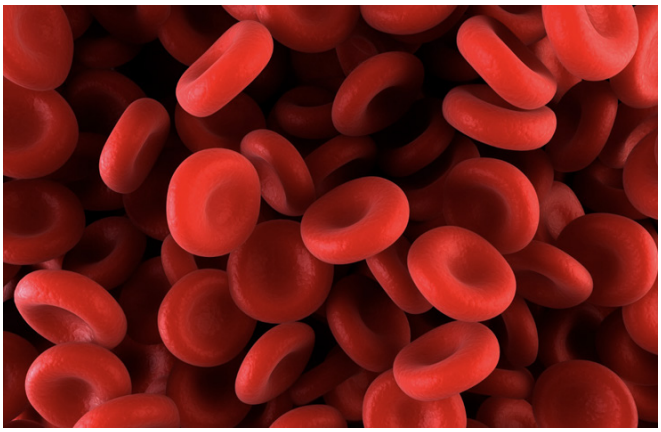

ILLUSTRATION: CANVA/SCIENCE PHOTO LIBRARY

*About 5% of the world's population has some variant of alpha thalassemia, a genetic deficiency of hemoglobin in the red blood cells.*

### *HBA* and alpha thalassemia

By weight, human red blood cells are composed of about 35% hemoglobin, which is responsible for transporting oxygen. (Most of the rest is water.) There are a few different combinations of hemoglobin proteins, but an essential ingredient of them all is alpha hemoglobin, encoded by the genes *HBA1* and *HBA2*.

People typically have four total copies of these genes, two on each copy of chromosome 16. Children who inherit only three copies will be carriers—they still produce enough alpha hemoglobin and don't generally need treatment. But children who inherit two or fewer copies of *HBA1* and *HBA2* will have some form of alpha thalassemia, an autosomal recessive blood disorder.

Insufficient hemoglobin causes anemia, in which red blood cells are undersized or disintegrate entirely—and the number of missing copies of *HBA1* and *HBA2* directly affects alpha thalassemia's severity. A person with two missing copies has "alpha thalassemia trait," which causes mild anemia. Three missing copies entails "hemoglobin H disease," or HbH, which requires blood transfusion therapy. A person without any copies has "hemoglobin Bart's hydrops foetalis," which is usually fatal.

According to the *Orphanet Journal of Rare Diseases*, alpha thalassemia "is probably the most common monogenic disorder in the world and is especially frequent in Mediterranean countries, South-East Asia, Africa, the Middle East and in the Indian subcontinent."[8] In some regions, more than 30% of the population may be carriers. Some scientists theorize that this prevalence arose because it actually confers an evolutionary advantage in these regions, since carriers are less susceptible to malaria.

In all, the CDC estimates that about 5% of the world's population has some variant of alpha thalassemia,[9] and both the American College of Obstetricians and Gynecologists and the American College of Medical Genetics recommend screening for the condition in people who are pregnant or who plan to reproduce.

The genomic region containing *HBA1* and *HBA2* is highly homologous, making copy number detection and accurate read alignment difficult for gene sequencing systems. So Illumina scientists developed the research-use-only *HBA* targeted caller for DRAGEN 4.2, which estimates *HBA* copy number genotype based on several nearby regions that are not homologous. They tested the caller on hundreds of samples from The 1000 Genomes Project and found that it accurately detected 14 copy number genotypes—variations based on exactly which region of the HBA genes was deleted.

In the article referenced below, these scientists report being confident that this research tool will be able to aid large-scale population studies, and in turn "help guide decisions about how to best deploy carrier and newborn screening tests."

*To learn about how the HBA caller works and the methods used to test it in greater detail, read this article[10] on Illumina's Genomics Research Hub.*

### KIV-2 and atherosclerosis

We humans require cholesterol as a vital component of our cell membranes, and low-density lipoproteins (LDL) are the main vehicle our bodies use to transport cholesterol-containing fats. However, elevated LDL levels are a major factor in atherosclerosis—a buildup of fats along arterial walls—and subsequently of cardiovascular disease (CVD). According to the World Health Organization, nearly a third of all deaths are due to CVD, 85% of which involve heart attack and stroke.[11]

Lipoprotein a (LPA) is one kind of LDL. The concentration of LPA in a person's blood is highly heritable from parent to child and varies widely from one population to the next—for instance, an article in *JAMA Cardiology* estimates that 20% of individuals of European ancestry have elevated LPA levels.[12] These elevated levels can be traced to a genetic cause: the number of copies of the

*8. https://ojrd.biomedcentral.com/articles/10.1186/1750-1172-5-13 9. https://www.cdc.gov/mmwr/volumes/69/wr/mm6936a7.htm?s_cid=mm6936a7_w*
*10. Han S, Onuchic V, Rossi M, Roller E, Cameron D. "Genotyping of high homology HBA1 and HBA2 from Illumina whole-genome sequencing." Illumina Genomics Research Hub. 2022https://www.illumina.com/science/genomics-research/articles/HBA-targeted-caller.html*
*11. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)*
*12. https://jamanetwork.com/journals/jamacardiology/fullarticle/2771455*

kringle-IV 2 domain (KIV-2) in the LPA protein.

Named after the twisted Danish pastry, kringle domains are sections of a protein that fold together into loops and help it bind to other proteins. KIV-2 has an astounding range of occurrence: human reference genomes usually record fewer than six copies of it in the *LPA* gene, but some people have 50 copies or more.

Why does this region get copied so much? For now, scientists are unsure. The more copies of KIV-2 a person has, the longer their LPA proteins are… but here's the catch: The longer these proteins are, the longer they take to synthesize, so—counterintuitively—a person with *more* copies of KIV-2 actually has *lower* levels of LPA in their blood.

Regardless, the extremely variable copy number of this region makes it a huge challenge for a gene sequencer to accurately describe and quantify.

Working with colleagues at eight universities and laboratories across the United States, Illumina scientists developed a targeted caller that generates a highly accurate copy number measurement for this difficult region.[13] They tested the research-use-only caller on the genomic data of 120 Mendelian trios (a child and both their parents) recorded in The 1000 Genomes Project, and the results correlated extremely closely with those generated by other methods. In the article referenced below, the Illumina team presents these results as evidence that the KIV-2 targeted caller in DRAGEN software "will provide a valuable tool for enhanced LPA and CVD research." ◆

*To learn about how the KIV-2 caller works and the methods used to test it in greater detail, read this article[14] on Illumina's Genomics Research Hub.*

ILLUSTRATION: SHUTTERSTOCK

*The KIV-2 region of the genome is related to lower cholesterol buildup in the arteries; most people have six copies of it, but some have 50 or more.*

*13. https://www.biorxiv.org/content/10.1101/2023.04.24.538128v1.full.pdf*
*14. Belyeu J, Onuchic V, Bekritsky M. "Using whole-genome sequencing to evaluate copy number variants of the LPA Kringle-IV type 2 domain with DRAGEN." Illumina Genomics Research Hub. 2023. https://www.illumina.com/science/genomics-research/articles/using-whole-genome-sequencing-to-evaluate-copy-number-variants-o.html*