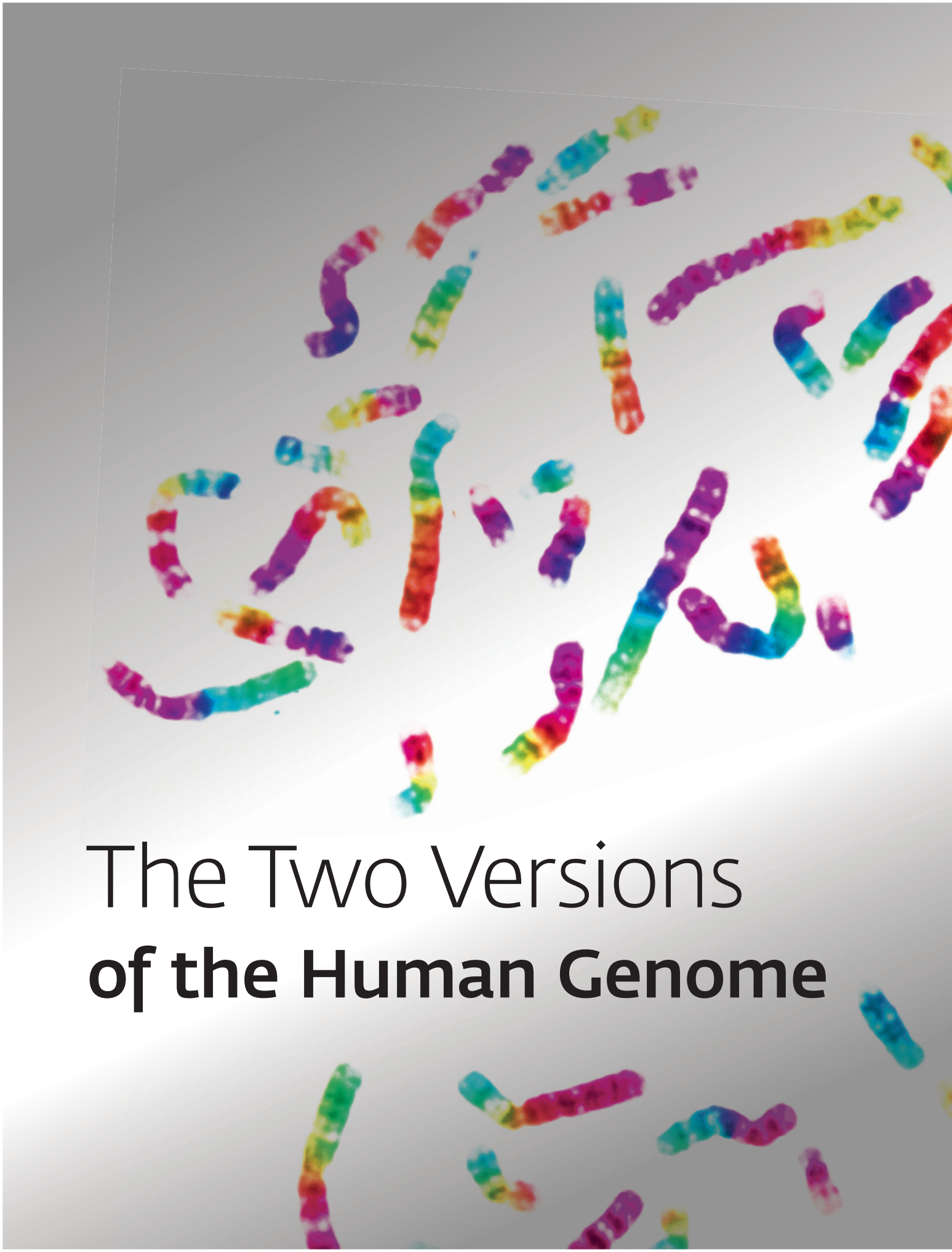


The Two Versions **of the Human Genome**

Chromosomes at the moment of cell division: The human genome consists of two sets of chromosomes, one from the father and one from the mother. There are thus two versions of each chromosome. A special staining method creates a characteristic band pattern for each chromosome, so that the pairs can be identified on the basis of their comparable size, form and color.

Photo: SPL-Agentur Focus

Sequenced, yes – but decoded? We still don't fully understand our human genetic make-up. The answer to many of its mysteries lies in the diploid nature of the genome, which contains two sets of chromosomes: one inherited from the father and one from the mother. **Margret Hoehe** from the **Max Planck Institute for Molecular Genetics** in Berlin has sequenced both versions of the human genome separately for the first time and, in the process, discovered that individuals are even more individual than we thought.

TEXT **CATARINA PIETSCHMANN**

When Margret Hoehe explains her work, you quickly realize how logical her approach is, how obvious – so much so that you wonder why someone else didn't do what she has done a long time ago. Maybe the reason is that, when one is too close to something, it's easy to lose sight of the big picture. Looking through a magnifying glass, one might see a craggy grey landscape and have no idea what it is. Taking a few steps back, you want to shout to them, "Put the magnifying glass down! Then you'll see that what you're looking at is a full-grown elephant."

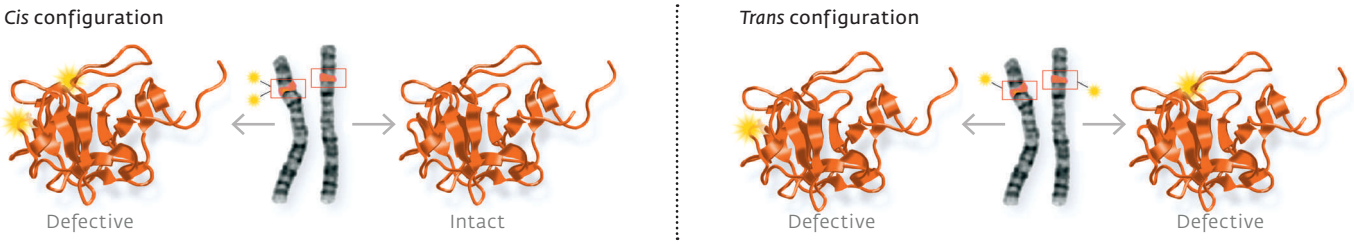
Margret Hoehe is one of the few people who did just that: took a few steps back to get a better view. She reflected on the genetic principles, without which Gregor Mendel's laws of the inheritance of traits can't be understood, and without which no biology lesson is complete. Mendel crossed homozygous pea plants that had either purple or white flowers. The plants in the daughter generation all had purple flowers, because the gene for purple flower color was dominant. In the third generation, purple and white flowers occurred again in particular proportions – a phenomenon that can be explained only by the existence of two sets of chromosomes.

GENES FROM FATHER AND MOTHER

Even if the crosses are not always as clearly predictable in humans, we have one thing in common with many other organisms: every individual has a genome that consists of two different sets of chromosomes. As a result, all genes arise in duplicate: one on the maternal chromosome and the other on the paternal one. Scientists refer to the different parental sequence versions of each chromosome as haplotypes.

A less well known fact is that, as far back as 1957, American biophysicist Seymour Benzer demonstrated in a simple experiment that it makes a difference whether two different mutations of a gene are located on the same chromosome or whether they reside on opposite chromosomes. But Benzer's discovery was forgotten.

So wouldn't it make sense to look at the two versions of the genome separately? Up to now, only "one" genome



The distribution of mutations in a pair of chromosomes can dictate health and disease: In the *cis* configuration, two mutations arise in one and the same gene form, and the associated protein becomes perturbed. The second copy and the protein originating from it remain unaffected. In the *trans* configuration, in contrast, both gene copies are mutated and produce two damaged proteins.

has been read, a sort of composite of the maternal and paternal sequences. There are historical reasons for this: When the Human Genome Project was carried out in the 1990s, it was almost technically impossible to analyze the two versions of the approximately three billion base pairs in the human genome separately. It simply would have been too labor- and cost-intensive.

GENETIC MAKE-UP IN DUPLICATE

Habits that have become entrenched are difficult to change. "Until recently, even high-ranking experts thought of the genome as a 'one-fold' object, not a 'two-fold' one," says Margret Hoehe, head of the Genetic Variation, Haplotypes & Genetics of Complex Disease group at the Max Planck Institute in Berlin, and the first scientist to sequence the genome of a human being – a German – entirely separately based on the haplotypes. She analyzed both sets of chromosomes completely and at a previously unattained level of detail. The approach that she developed with her colleagues to do this was singled out by the journal Nature as one of the most promising methods of 2011.

"Just a minute, we'll be ready soon. It must be possible ... !" Hoehe is looking for space on her large desk for two teacups and her visitor's notes. Not an easy task, as the desk is almost entirely covered with towers of publications and data printouts. The researcher has done practically nothing but write since August to get all of the analyses down on paper.

Why is it important to know how certain mutations are distributed between the two parts of the genome? "Because, it can mean, for instance, the difference between cancer and no cancer," she says. "If there are two mutations – for example, of the BRCA1 risk gene associated with breast cancer – they need not necessarily cause the disease." There are two possibilities: the mutations affect both forms of the gene *(trans)*, in which case the chances of developing breast cancer are very high; or they are located on only one of the two chromosomes *(cis)*, so the person will also have a completely "healthy" form of the gene.

Cis or *trans*? The answer to this question can mean the difference between good health or illness – and even life or death. It can also raise hope. After all, a "healthy" form of the gene can be passed on to the next generation.

A good ten years have passed since Craig Venter and Francis Collins – initially colleagues and later competitors – announced the decoding of the human genome side by side with Bill Clinton in the Oval Office. The end of ignorance, as Venter said. What a triumph! Once we know the genetic code, we will soon also know the function of all genes. We will understand the origin of diseases and be able to heal them. So far, so good.

In the years that followed, medicine clearly made progress, but the great breakthrough never came. This may change now with Margret Hoehe's haplotype approach.

The DNA alphabet consists of only the letters ACGT. They stand for the base molecules adenine, cytosine, guanine and thymine. Around 90 percent of the genetic differences between people consist of sites where one letter of the genetic code – that is, one base – was replaced by another one. The researchers also refer to these base exchanges as single nucleotide polymorphisms (SNPs). "Everyone has an average of one base exchange every 1,700 letters," explains Hoehe.

CHANGES IN THE GENETIC CODE

The information for the production of proteins from individual amino acids is contained in the order or sequence of the bases. Thus, for example, the replacement of a cytosine by a thymine can lead to the use of a different amino acid. This can result in the production of a malfunctioning protein. When bases in the control regions of the DNA are exchanged, biochemical processes in the body may be blocked or accelerated.

To sequence the genome, the researchers isolate DNA from white blood cells, shred it using ultrasound, and sort the fragments by size. The DNA snippets are then fixed to a surface using

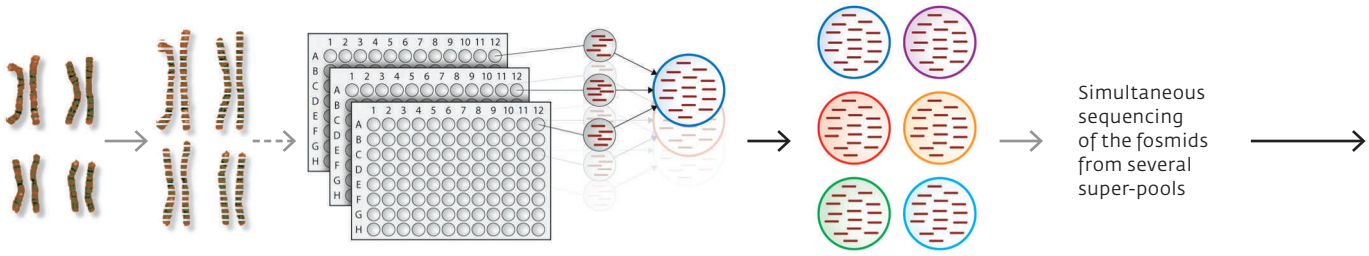
Graphic: Art 4 Science

Library in the freezer: The human genome fragments inserted into bacterial DNA molecules are stored at minus 80 degrees. Each of the aluminum-foil-packed genome libraries originates from one person. The genomes of 100 people fill an entire freezer.

small anchor molecules and copied. The reading of the base sequence is carried out simultaneously for millions of DNA fragments. Today, it takes just a few days to read a complete genome using a second-generation sequencing machine. “The very latest machines reach a throughput of up to three hundred billion bases,” says Hoehe. “The early ones could manage only 900 bases.” Billions of these sequence reads are stored on a supercomputer. Then the work on the vast jigsaw puzzle begins: the computer compares the base sequences of the individual snippets and overlaps them to form bigger pieces until the DNA is complete.

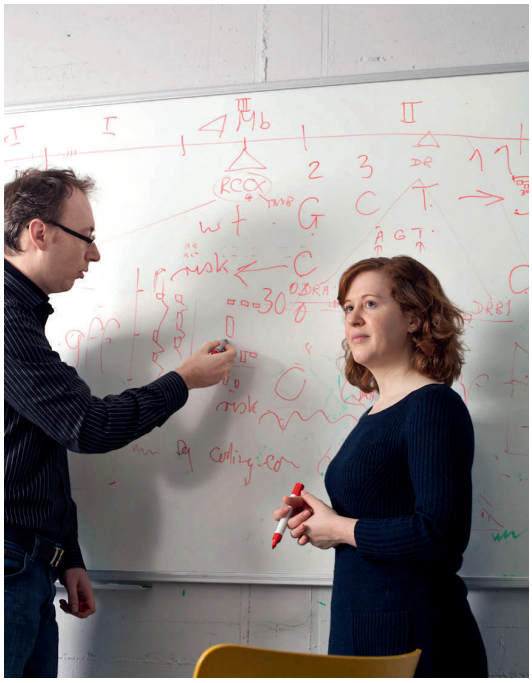
But how does the software know which piece of the puzzle belongs where? “It constantly compares the base sequence of the snippets with those of the reference sequence,” explains Hoehe, “and this is the critical point.” The human genome of 2001, which all scientists still use as a reference, is an artificial product, a cocktail of the DNA of at least six different people. But each person is unique, so this reference sequence as such doesn’t exist naturally. It doesn’t reflect the genetic code of any particular living person. This was the intention, but many experts now view it as an error, as this makes it an unreliable roadmap for exploring the causes of health and disease. “Because sometimes we don’t know whether the exchanged bases we find occur frequently or rarely in reality.” >

Photo: Nina Lüth



Graphic above: Sequencing of the haplotypes of a genome: (1) The DNA of the chromosomes is mechanically sheared into smaller fragments. Fragments with a length of 40,000 base pairs are inserted into bacterial transport DNA and these so-called fosmids (representing now haploid DNA fragments) are reproduced in bacteria. Approximately 1.44 million fosmids are distributed on three plates, each containing 96 wells, representing the fosmid library of an individual. Each well contains a pool of 5,000 fosmids and each pool contains, as the result of dilution, almost without exception fragments of one haplotype. For reasons of efficiency, three pools are always combined to form a super-pool. These pools are decoded using the very latest sequencing technology (2) and the individual fosmid sequences are assembled (3). The relevant base exchanges (perpendicular lines) are identified in the mixture of fragments from the paternal and maternal genome (4). This allows the individual sequences to be assigned to one of the two haplotypes and combined to form the chromosomal haplotypes through identification and tiling of overlapping fragments (5). In this way, the genetic code of the chromosome pairs is being deciphered (6).

below: Bioinformaticians put their heads together: Thomas Hübsch and Gayle McEwen discuss how they can analyze particularly complex DNA sections by haplotype.

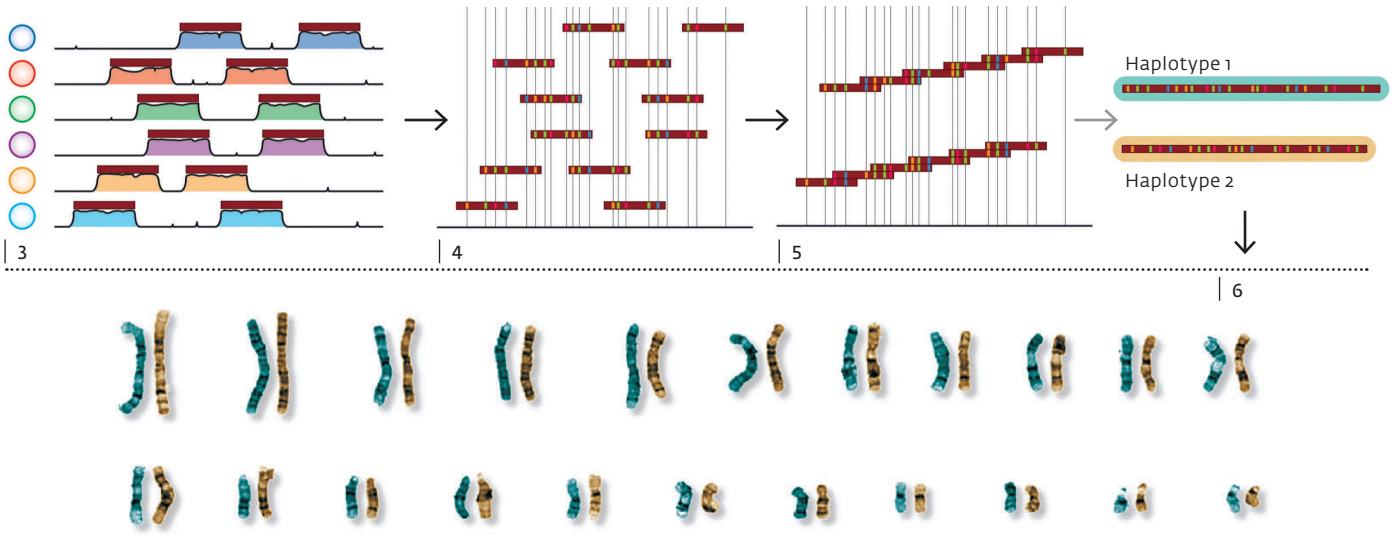


A total of around 2,800 individual human genomes have been sequenced throughout the world – all based on a composite of the two chromosome sets. Craig Venter started it all when he sequenced his own genome. Years later, he tried to separate his genome into its haplotypes, but only partially succeeded. James Watson, the co-discoverer of the DNA double helix, also knows the contents of his genome. The first genome of a Chinese man was published in 2007, and was followed recently by an Aborigine genome. The genes of Archbishop Tutu from the Bantu ethnic group were analyzed last year at the same time as the DNA of the members of four other South African ethnic groups.

According to the findings, Africans are far more variable than Europeans. Modern humans developed over 100,000 years ago in Africa, and their genomes had a long time to intermix there. For this reason, the genomes of Africans contain many old, common variants, as well as rarer and newer ones.

The greater variability of African genomes is also evident in their organization: "In Africans, the shortest length of the DNA blocks, in which the exchanged bases are inherited over generations, is approximately 10,000 bases on average," says Margret Hoehe. The corresponding figure in Europeans is between 20,000 and 120,000

Graphic: Art 4 Science; photos: Nina Lüth (2)



bases and more. "Obviously, just a small group left its continent of origin, Africa, to populate the rest of the world, and few people followed them. Consequently, genetic diversity was significantly reduced – a genetic bottleneck, so to speak."

THE AFRICAN GENOME AS A REFERENCE

Up to now, all human genome research has been based on the reference sequence, a genome with European roots. This was due to the simple fact that the countries involved in the 2001 Human Genome Project were the US, France, Germany, England and Japan. "Biomarkers, receptor variants, disease regions and drug targets – none of these can be applied to the African genome on a one-to-one basis." The availability of an African reference genome to enable the development of more effective drugs for people with African genetic roots is thus long overdue.

To be able to sequence both chromosome sets of a human genome separately, Margret Hoehe and her team had to develop a new molecular genetic method along with all of the necessary bioinformatics tools. An important difference between their method and the traditional technology lies in the fact that the DNA segments are not 25 to 40 bases long, but around 40,000. Because they display characteristic base exchange patterns and are not cut in exactly the same way – for example, at bases 128 and 40,200, or at 14,000 and 55,030 – the computer can easily identify, during overlapping, whether a snippet belongs to part A or B of the genome. However, whether A originates from the father or mother can be established only through further comparison with at least one parent.

In this way, it was possible to resolve the two versions of almost all of the German subject's 17,861 genes that code for proteins. Of those, 90 percent arise in two different molecular forms. The two chromosome sets differ in around two million sites. Comparison with the reference sequence also shows that 60 to 70 percent of the given molecular gene forms exist as such only in this individual. Thus, we are far more individual than was believed to be the case. Moreover, the scientists identified 159 genes with two or more potential disease-predisposing mutations, 86 of which occur in only one gene copy.

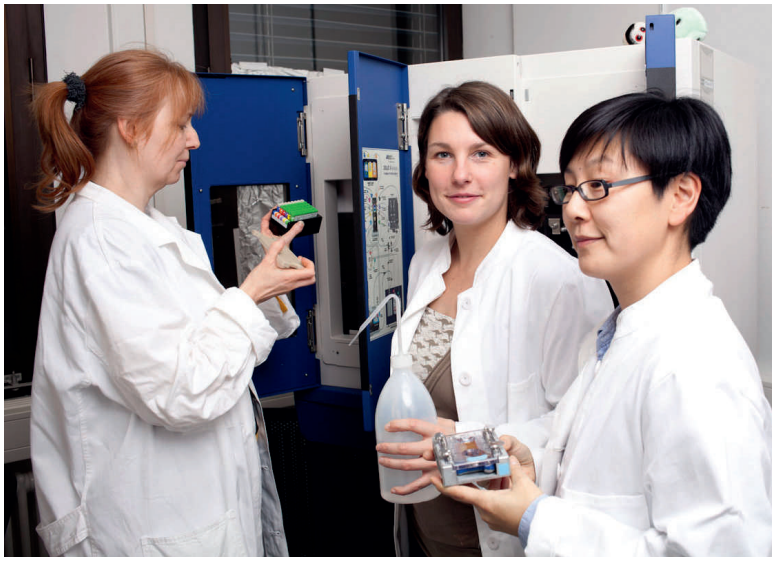
What could the relationship between these two molecular gene forms be? Do they cooperate or do they counteract each other? Or do they alternate? Which one is dominant – and why? A disease arises only when one defective gene overrules its intact counterpart. Hoehe is certain that a new, more in-depth view of biology will arise. And it could actually provide the missing key to understanding the genome, and enable the development of the long-awaited individualized therapy in medicine. The door to a completely new research field is being opened here.

UNKNOWN REGULATION OF GENES

Inheriting good genes – this expression is now taking on a new molecular meaning. "It sometimes seems to me as though my maternal and paternal genes antagonize each other," says Margret Hoehe laughing. No, seriously, there's a grain of truth in this. "You just have to think about it again. Perhaps the genes are regulated differently at different stages in life? Perhaps one version of the chromosome is completely switched off?" We know almost nothing about this. The buffer capacity of the system appears to be high. There are numerous small intermediary stages between healthy and sick. "If one doesn't work, the other one does. But when too many things coincide, at some point, everything is derailed." Ultimately, it is probably a question of the cumulative effect of all of the different individual predispositions.

In order to obtain a better catalog of the diversity of genetic variants, the Human Genome Project was followed up in 2008 by the 1,000 Genomes Project, the aim of which is to decode 2,500 genomes. Whereas many researchers all over the world are cooperating on this

Graphic: Art 4 Science



left: Stefanie Palczewski (left), Sabrina Schulz (center) and Eun-Kyung Suk (right) in front of a high-throughput sequencing system for DNA analysis. right: Each of the device's two flow cells can accommodate over 600 million DNA fragments.

project and contribute partial sequences to it, Hoehe and her team, consisting of a handful of scientists, have completely sequenced, haplotype-resolved and analyzed an entire genome in a matter of a few months. She is aiming to have completed twelve further individual haploid-analyzed genomes by the end of the year.

Individuality – this theme runs through Margret Hoehe's life as a researcher. She studied medicine and psychology in Munich and completed her doctorate on "the effects of opiates on neuroendocrine and psychological parameters" in 1986. During a study on depressive and schizophrenic patients, she quickly noticed that some did not respond at all to drugs; and neither did a proportion of the healthy subjects. "The traditional mean value method ultimately leveled out all of the individual differences." What needed to be examined, instead, is how the individual's DNA influences the effect of drugs.

So she included a chapter on the inter-individual variability of drugs in her doctoral thesis. "There was an uproar," she recalls, and she is still infuriated by the fact that she was forced to remove the chapter. But the idea itself could not be deleted. Today, particular attention is paid in medical studies to the differences between responders and non-responders.

The next step was to carry out family studies. Hoehe wanted to examine the possible genetic roots of the individual markers she had discovered for susceptibility to diseases and the effectiveness of drugs. To this end, she went to the National Institute of Mental Health in the US in 1987. "The human genome project was already on the horizon." When her boss asked her whether she wanted to continue working in the area of neuroendocrinology or with DNA, without hesitating, she opted for the genetic material.

"I knew absolutely nothing about it and had to teach myself the basic tools and methods," she says, smiling. The genes for the opiate and adrenergic receptors, which she had previously studied pharmacologically, were discovered soon after that. "Using methods that would be dismissed as Stone Age today, we found the first variations, and then searched the genome for regions for depression and schizophrenia."

A NEW APPROACH TO GENOME ANALYSIS

At the time, it was still believed that only a few genes in the genome were mutated, and that these mutations had serious effects – a mistaken belief, as we now know today. Very few diseases are actually due to a single gene. And regarding schizophrenia: In 2009, the journal Nature published a study according to which thousands of mutations on many genes can cause the disease, and one of these mutations alone causes only a very slight increase in the risk of developing it. The

Photos: Nina Lüth (2)

New York Times referred to the discovery as "A Pearl Harbor in schizophrenia research."

Back to the early 1990s: Hoehe soon reached the conclusion that the entire approach to genetic research would have to be reversed. Research should not start with the complex disease and look for the associated genes; instead, it should search the genome for inter-individual DNA sequence differences and observe their effect on genes *in vitro*. She compiled a corresponding project application and sent it to George Church, one of the leading genetic researchers. He immediately liked the idea, and a short time later, Hoehe was accepted as a postdoctoral researcher at Harvard Medical School. She developed the Multiplex PCR Sequencing technology there, which enabled the simultaneous reading of 20, and later 50, DNA snippets. That Kary B. Mullis, Nobel laureate and inventor of the polymerase chain reaction, had said that it could "never" work was something she didn't learn until later.

Hoehe returned to Germany in 1995 and established a research group at the Max Delbrück Center for Molecular Medicine in Berlin – yet another adventure. She had the machines for the research flown in from the Machine Shop at Harvard University. She resequenced the μ-opiate receptor gene in 250 patients and healthy subjects, and found first haplotypes that were associated with addictive disorders. This was one of the two initial studies that showed that it is important to consider the haplotypes of a gene and not just randomly selected single mutations. In 2002, Margret Hoehe moved to the Max Planck Society. One of her main projects there was the sequencing of the haplotypes of the most complex region of the human genome: the histocompatiblity complex that serves as the blueprint for the immune system's antibodies, and that is also rich in disease genes. This eventually gave rise to the sequencing of the entire haploid genomes of human individuals.

All of this work took a great deal of energy and dedication. "It wasn't easy to swim against the tide for such a long time." But that's changing now. And Margret Hoehe is already focusing on individuality again, this time from the perspective of a personalized genome project. Illness can best be studied where it occurs – in the patient. The project will initially focus on a breast cancer patient whose haploid genome will be studied in the blood and in the tumor tissue. The two women have known each other for a long time and are writing a joint diary on the project, which may one day become a book.

An unusual aspect of this project is that the patient, a journalist with medical training, is involved in the study and is being given access to her genetic information. Anonymized research is generally the rule in medicine. "People make their samples available to science voluntarily. We should give them something in return for this," says Margret Hoehe. "This topic is already being hotly debated in the US." Patients who want to know and have the necessary background knowledge should also be given the information about their genomes, following the motto: "My genome is mine!"

Rumor has it that Craig Venter is taking prophylactic drugs since he found out about his genome. The German subject whose genome was sequenced can't do this. Because his blood sample came from a biobank, it was anonymized for ethical reasons. All that is known about him is that, when he gave the blood sample, he was 51 years old and in good health. The scientists would have both good and bad news for him: he also has two mutations in the breast cancer gene BRCA1. Fortunately, they are both on the same copy of the gene. This means he also has a completely healthy version. ◂

With her method of sequencing a genome separately based on the two haplotypes, Margret Hoehe has laid the foundations for personalized medicine.

GLOSSARY

Chromosome
The genetic material of living organisms with a cell nucleus is distributed on varying numbers of chromosomes. A chromosome consists of the thread-like DNA molecule and various proteins. The biggest chromosome in humans is around 250 million base pairs long, while the smallest contains only around 50 million base pairs. The chromosome with the most genes has around 3,000 genes, while the male Y chromosome has only around 100.

Diploidy
Diploid organisms have a double set of chromosomes. For example, humans have 46 chromosomes, which occur in the form of 23 chromosome pairs. Each homologous chromosome pair contains one chromosome from the mother and one from the father. The 23rd chromosome pair contains either two XX sex chromosomes (women) or one XX and one XY sex chromosome (men).

Haplotype
The two parental sequence versions of each chromosome. Because humans have two (variant) forms of each chromosome, the chromosome exists in two haplotypes.

Photo: MPI for Molecular Genetics