



Promoting international collaboration through highly accurate genome analysis and the genome bank project

NAKAMURA, Yasukazu Professor

Genome Informatics Laboratory
Department of Informatics

Professor Nakamura graduated from the Faculty of Agriculture, Shinshu University and completed his master's course at the Graduate School of Agriculture, Kyoto University. After serving as an engineer at the Yamanashi Agricultural Research Center, he entered the doctoral course in Genetics at the School of Life Science, The Graduate University for Advanced Studies (NIG). After graduating, he assumed the position of research scientist at the Kazusa DNA Research Institute in 1996. He was promoted to be head of the laboratory at the same institution in 2006. He became a Professor at the National Institute of Genetics in January 2009 in the Genome Informatics Laboratory at the then Bioinformation Research Center. Currently, he is a Professor and Project Research Scientist at the Genome Informatics Laboratory at NIG. He also serves as Head of the International Affairs Division, Bioinformation and DDBJ Center.

"The Genetic code consists of four nucleotides, which are common to both liverworts and humans." — Having carefully analyzed the entire genome of diverse organisms, Professor Yasukazu Nakamura has been fascinated by genome analysis, which can be performed with any biological species. In the course of conducting research at various levels within industry, government, and academia, Professor Nakamura is now in charge of an international collaboration to expand the nucleotide sequence data bank project as the representative of the National Institute of Genetics.

The twists and turns of becoming a specialist in bioinformatics

While I am currently conducting research and development as a bioinformatician, I was an experimental researcher when I was in graduate school. I was a member of Professor Kanji Oyama's laboratory at the Faculty of Agriculture, Kyoto University, and participated in research such as figuring out the entire mitochondrial genome DNA sequence of a liverwort —the first to be determined in a plant. The method of genome analysis we were using at the time was very primitive and something you could hardly imagine nowadays. It involved creating a genomic library of liverwort using *E. coli*, reading the sequence using a radioisotope, and then manually connecting the fragments of the genome that had been read separately. I grew tired of the manual work involved and started to use a 16-bit computer in the lab instead. I wrote a program that made it possible to connect

DNA fragments automatically. Looking back, this may have been the starting point for me to pursue bioinformatics.

However, concepts such as "bioinformatics" didn't actually exist at the time. My seniors and colleagues around me at the time really excelled in their fields, and it seemed like it would be difficult for me to survive in the academic world even if I continued my doctoral course. After obtaining my master's degree, I decided to work at the Yamanashi Agricultural Research Center which was in my hometown. I was in charge of soil research and development together with rice and vegetable cultivation. It wasn't a bad life, but I was also asking myself "Do you really want to do this for the rest of your life?" I got married around that time, and my wife was financially independent. So, I planned on getting my PhD in a field where I could combine both computer analysis and biology. That was what I was really interested in.

Fortunately, I heard that Professor Toshimichi Ikemura of NIG was

looking for a student. When looking through the "Application Guidelines for the 3-year Course Program" of the newly established Graduate University for Advanced Studies, the word "Computer biology" came up. When I saw this, I thought "this is what I'm looking for!" This was in 1992. At that time, the Human Genome Project was in its golden age, and the advent of the sequencer made it possible to read nucleotide sequences at a very high speed. I went straight to Professor Ikemura to speak with him. In our talk, he said, "We are planning to analyze the large-scale structure of the human genome," and this is what made me decide to apply for the entrance exam.

I entered The Graduate University for Advanced Studies as a 3-year course student and started work. The target of my research changed drastically from liverworts to humans, but no matter the species, the genetic information is written with four nucleotides. In that sense, bioinformatics is a field that has much freedom. Professor Ikemura conducted his research through both experiments and analysis. When chromosomes are stained, both dark and light bands are visible. It is known that the dark areas are rich in GC content (GC-rich) and there are many genes present. Professor Ikemura was working on a theory that there was some kind of signal present in the boundary between the bands.

I was one of the few people lucky enough to have access to a high-performance computer at the time, which cost around 30 million yen. While researching the human genome, which was my assigned work, I also started building a codon database, creating a program that could automatically convert codons for amino acids, which differ slightly depending on the species. Professor Ikemura was the person who started to develop the codon database, but he became too busy with his work on human genome analysis, so I picked up where he left off.

After finishing the database, I launched a webserver and made it public through the internet. This was when the internet was just starting to gain traction, and there were only one or two hundred web servers in Japan including the one that I had made.

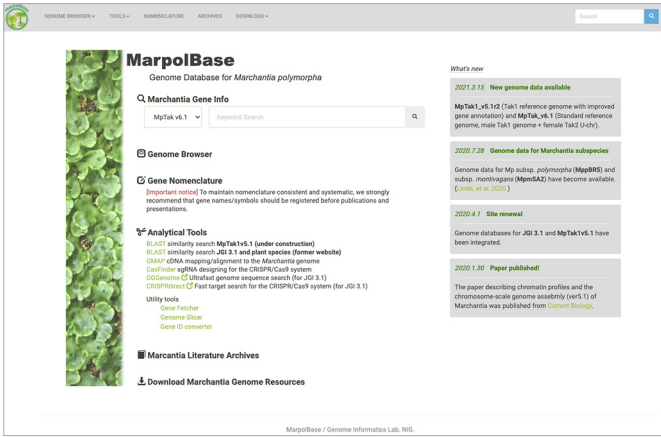
I received my PhD in 1996 and was employed as a bioinformatics researcher at the Kazusa DNA Research Institute in Chiba Prefecture. I joined the group that successfully sequenced the entire genome of cyanobacteria, with the sequence being completed in the same year. It was the world's first completely sequenced genome of a photosynthetic organism, and also the fourth completely sequenced bacterial genome. It was around this time that I started working on data analysis, and at the same time, I became a member of the steering committee of the data bank of NIG (DDBJ: DNA Data Bank of Japan). In 2001, we completed the genome sequence analysis of *Arabidopsis*, a generic model of higher plants, and in 2008, NIG offered me a position as a professor. I decided to accept the offer and return to NIG, my former home, in 2009. DDBJ was the main work to which I was initially assigned, but I also became responsible for international collaboration later down the line, and we're now working in collaboration with NCBI (United States), EBI (EU), and Genome Bank (China). Of course, we are also continuing to conduct genome analyses of various biological species. The main targets of analysis so far include liverwort, citrus, and the domestic cat.

Liverwort (*Marchantia polymorpha*) as the basis of terrestrial plants

Liverwort is an organism that I have spent an enormous amount of

time and effort on for genome analysis. Liverwort retains the most primitive mechanisms of terrestrial plants and is the ideal material for studying their morphology and evolution. It has a long history as a model plant, and the whole chloroplast and mitochondrial genomes of plants were first sequenced in liverwort. Despite the entire genome also being analyzed, and an outline of the chromosome level results being reported in 2007, a thorough analysis of the complete chromosome did not make significant progress at that time.

Under these circumstances, 39 international and domestic universities including the National Institute of Genetics, Kinki University, Kobe University, the National Institute for Basic Biology, Tohoku University, the US Department of Energy, and Monash University in Australia, created a consortium to shed light on the overall liverwort genome structure, which was a significant project. I believe that this consortium was assisted by the full-scale introduction of the next-generation sequencer that can read DNA sequences at a very high speed and with high precision at a relatively low cost, and the development of the "long-read method" that made it possible to read DNA nucleotide sequences longer than 10,000 bases in one go. My laboratory group was responsible for using the NIG supercomputer to link fragmented genome sequences accurately, and for constructing a database by predicting genes.



The results of the analysis clarified that there are around 20,000 genes and the number of transcription factors that work as switches for genes is smaller than in other terrestrial plant species, but there is a set of genes that is the basis for all land organisms. Accordingly, there is now information on the use of liverwort as the basis for research into a broad range of terrestrial plants on evolution, gene structure, morphology, information transmission, and metabolic systems. The paper was published in Cell, 2017 (*1).

I think that the image most people have of liverwort is just "a moss growing in damp shady places." For plant researchers, though, the completion of the entire genome database was a long-awaited and priceless accomplishment.

Catalogue of a wide variety of citrus genomes, including mandarin orange (*Citrus unshiu*)

Shizuoka Prefecture, where NIG is located, is famous for its oranges due to its mild climate. For this reason, the Institute of Fruit Tree and Tea

Science of The National Agriculture and Food Research Organization (NARO) is located there. We started genome analysis of citrus species after I was invited to work with Dr. Tokuro Shimizu of the Genome Research Unit in the same institution. The aim of the research was not to carry out a detailed analysis of the entire linear genome sequence as in liverwort but to roughly analyze as many species of the citrus genus as possible and make a catalogue of their genomes. While there are more than 150 varieties available on the market throughout the world, the genetic diversity was too broad and there were many cases where the parental relationships were unknown. It would be extremely useful for breeding citrus species if we could figure out their phylogenetic relationships.

Accordingly, Dr. Shimizu and his team extracted nuclear and mitochondrial DNA targeting 269 varieties planted in their institute and sequenced each fragment. After receiving the sequence data, we analyzed and compared the gene structure data and the sequence polymorphisms in detail. As a result, the parents of 22 species, including *Citrus unshiu*, were identified and the combination of seed parent line and pollen parent line was also clarified. In addition, we were able to obtain phylogenetic information on 45 native varieties, such as identification of single parents, place of origin, and parent-offspring relationships. The paper was published in 2017 (*2).

In some phenotypes such as those that are associated with taste or disease resistance, multiple genes are involved. Accordingly, a database that can provide a comprehensive view of all genes is likely to help produce varieties with both improved taste and better disease resistance. We have already registered the raw and analyzed data in DDBJ.

The sequence of the whole genome of the American Shorthair —a popular housecat breed both in Japan and overseas

We analyzed the genome of the domestic cat because of my affection for them, and because it is known that there are genetic diseases that tend to only arise in certain breeds. However, research on the prevention and treatment of these diseases hasn't made significant progress because important genomic information hasn't been available. The domestic cat that we know today came about when humans domesticated wild cats and bred them as companion animals. Crossbreeding with wild cats can be seen in some breeds. The only breed with an analyzed genome is the Abyssinian, which was done in the US in 2007. However, the data is not comprehensive. Since Abyssinians are a breed in which a high degree of inbreeding has occurred, there were few genetic markers (marker genes) that could be used to compare with other breeds. In addition, the entire genome was obtained by reading the sequence of some 4,500 separate fragments and simply connecting them. There were also many unclear parts in the genome structure at the chromosome level.

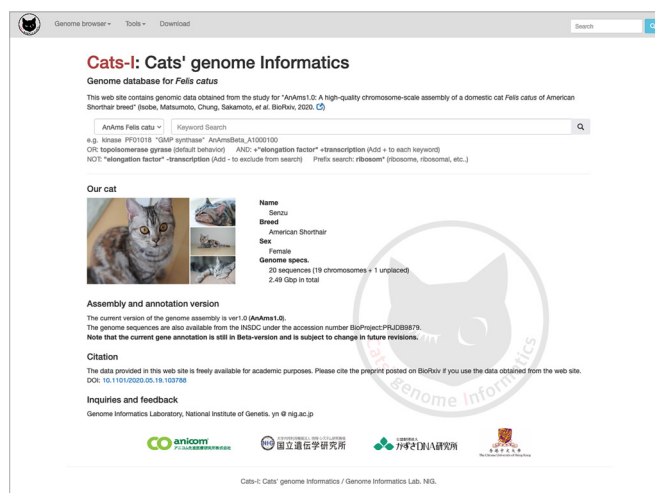
Fortunately, we could undertake our domestic cat genome project in collaboration with three organizations that have been at the forefront of research in this field: a well-known pet insurance company, Anicom Insurance, Inc. (Anicom Specialty Medicinal Institute), the Kazusa DNA Research Institute, and The Chinese University of Hong Kong. The breed we selected to analyze was the American Shorthair (female).

The American Shorthair is well known throughout the world and is frequently used in crossbreeding with other breeds such as the Scottish Fold. In other words, there are many breeds of cats that are genetically related to the American Shorthair, so the genome information gathered can be used for a broad range of purposes, and it is expected to contribute to research in veterinary medicine. Furthermore, the ancestors of the American Shorthair are the cats that first landed on the American continent, so the genome information is also important for research into the evolution of the domestic cat.

The DNA was sequenced at the Kazusa DNA Research Institute. We used the NIG supercomputer to conduct gene prediction and gene structure analysis based on the sequence data. The key to our success was making effective use of the long read method—the same method used for the analysis of liverwort, and we were able to analyze 19 chromosomes very accurately, at almost full length. The entire genome of the American Shorthair comprises 2,493,141,643 base pairs, and 23,119 genes were found (*3). Making full use of our groups' experience and intuition, we were constantly posing questions to ourselves such as "Isn't this gene too large?" and "Is there any possibility that we have used a fragmented gene to make the prediction?" I think making use of AI in these proofreading steps could be effective in the future.

By comparing the sequences of the American Shorthair and the Abyssinian, we also found that they have different genomic structures in some parts of their chromosomes. In the future, we can expect to gain insight into the genes associated with hypertensive cardiomyopathy common in the American Shorthair by comparing the genes of the two breeds in detail.

It is often not possible to publicly disclose the results obtained when we collaborate with private companies, but Anicom was willing to release the results to the public. The results were published in DDBJ in May last year. While only the sequence data is available at the moment, we intend to continue to gather and organize medical-related and other information and link it to pathogenesis-related genes to make it easier for veterinarians to use.



Collaborating with researchers, whether from industry, government, or academia

We are conducting collaborative research on genome analysis for a wide variety of biological species, at a broad range of scales and

levels, with researchers regardless of whether they are in industry, government, or academia. This is because there is not a sufficient national budget for scientific research and there are not many informaticians who can conduct large-scale genome analysis. My research attitude of "emphasize sequence information regardless of the species to be analyzed" may also have something to do with it.

I will continue working in collaboration with various industries, governments, and academic institutions. At present, COVID-19 mutations are an emerging problem. My research group is contributing to society by carrying out genome analyses of mutant strains of the virus originating from those who have tested positive in Shizuoka Prefecture.

Although it may not seem useful at first, I think genome analysis is also a very important form of basic research. For example, genome information from liverwort may not be directly linked to industrial applications, but it is essential to understand basic mechanisms in terrestrial plants. I often tell students in my lectures that when Heinrich Hertz discovered the radio wave, he said, "This is totally useless." However, without the radio wave, there would be no smartphones that we use today. Even if it doesn't seem to be applicable, pursue what you are interested in until you really understand it. This kind of attitude is important for researchers.



Currently, we are also working on "genetic diseases caused by genome instability" as a collaborative research project. It's a fairly large project, and many genetic diseases are sporadic with very few patients. Many people say that there is little information that can be obtained by reading the entire genome in one patient, so it lacks meaning. However, I don't believe this to be the case. If the entire genome is sequenced in detail and the data is gathered, I think important relationships regarding pathology will appear in the future. If you want to read the sequence of a genome, I think the most important thing to do is to sequence everything properly and gather the data.

Bioinformatics at The Graduate University for Advanced Studies

I believe today's students at The Graduate University for Advanced Studies are exceptional compared to the times when I was a doctoral student. Communication in English is considered a matter of course, and I get the impression that many students are motivated to leave their mark in the world of academia. Some graduates even get employed by private companies after they obtain their doctorate. Everyone has a strong sense of purpose, and they are coming to the university not just

because of its prestige, but because they have professors in their mind whom they want to work with.

In fact, the genome analysis of the domestic cat was a joint research project organized by Dr. Matsumoto, who was employed at the Anicom Specialty Medical Institute. I was a member of the screening committee for his doctoral dissertation.

In the past, there were only a limited number of post-doctoral careers, but now there is much more freedom. In particular, in the field of information analysis, private companies are willing to employ data scientists, so the work environment and salary are excellent. I think given the chance, young researchers should experience a variety of different environments, both in academia and in the private sector. However, because informaticians are more likely to get employed by private companies that have the appeal of better conditions, they tend to leave academia, which is an issue. If there continue to be only a limited number of laboratory head (or similar) posts in academia, and no improvements are made, it is difficult for me to ask students to remain in academia considering their future.

In Japan, the University of Tokyo and Tokyo Institute of Technology are very well known for their bioinformatics research, but I would like to take this opportunity to emphasize that it is possible to engage in a variety of research at the Graduate University for Advanced Studies as well. The pandemic continues to take its course, but we can conduct research even from home if we have a computer at hand, meaning there is little to no impact on our work. For those who think "This is for me" and would like to join us, please don't hesitate to knock on the door of the National Institute of Genetics. People like me who are already working members of society but would like to study again are also very welcome.



Interviewer:
Naoko Nishimura, Science writer

Photographer:
Mitsuhiko Kuruu, NIG ORD

July 2021

■ Links to databases

■ [MarpolBase: Genome Database for *Marchantia polymorpha*](#)



■ [Cats-I: Cats' genome Informatics](#)



■ References

*1-----

■ [Insights into Land Plant Evolution Garnered from the *Marchantia polymorpha* Genome.](#)

Cell. 2017 Oct 5;171(2):287-304.e15.



*2-----

■ [Draft sequencing of the heterozygous diploid genome of Satsuma \(*Citrus unshiu* Marc.\) using a hybrid assembly approach.](#)

Front Genet. 2017 Dec 5;8:180.



*3-----

■ [AnAms1.0: A high-quality chromosome-scale assembly of a domestic cat *Felis catus* of American Shorthair breed.](#)

bioRxiv

doi: <https://doi.org/10.1101/2020.05.19.103788>

