

## BGI Publishes Largest Ever Genomic Study of Chinese Population Discoveries in more than 140,000 Genomes throughout China

**October 4, 2018** -- BGI has analyzed the world's largest set of genome data from pregnant women, a total 141,431 expectant mothers from across China, in a groundbreaking study published today in the leading academic journal *Cell*.



Figure1. Publication of the study on *Cell*, Oct 4<sup>th</sup>, 2018

The genomic analyses from non-invasive prenatal testing (NIPT) revealed new genetic associations, patterns of viral infections, and history in Chinese populations. The study is the largest genomic study of the Chinese population to date and represents the first phase of BGI's Million Chinese Genome Project.

“For the first time, we proved that NIPT data can be used in genome-wide association studies to understand the genetic architecture of complex traits and disease,” said Siyang Liu, senior research scientist in BGI and first author of the study. “This brings up a new and promising model of investigating genetic mechanisms underlying traits related to maternal and child’s health.”

The research has revealed novel genetic associations between genes and pregnancy-related traits, including the birth of twins and the woman’s age at pregnancy, one of the well-known measurements of fertility. The analysis also allowed researchers to reconstruct the recent movement and intermarriage of different ethnic groups in China, and promises to help identify genes that make people susceptible to viral infection.

The data used in the study were from 141,431 anonymized and consented pregnant women participants, who took non-invasive prenatal genetic tests. The mothers-to-be had provided blood samples to be tested for fetal chromosomal abnormalities, primarily Down syndrome, which is possible because mothers have DNA from their unborn child

in their bloodstream. The participants came from all the 31 provincial administrative units in mainland China and included not only Han Chinese but also 36 ethnic minorities. The large sampling, about 1/10,000 of the Chinese population, is a good representation of the entire population.

The researchers sequenced, on average, only about 10 percent of each mother's genome, relying on large numbers of genomes so as to leverage individual sequencing data to discover new genetic links.

Non-invasive prenatal testing is gaining popularity and is common in China. It has been administered to approximately 6-7 million Chinese women and an estimated 10 million women worldwide. Sampling the mother's blood can be done as early as her tenth week of pregnancy and is risk-free, whereas standard prenatal testing in the U.S. involves amniocentesis and chorionic villus sampling, both of which require obtaining fetal cells from inside the uterus and risk harming the unborn child.

"This study is a first attempt at applying NIPT sequencing data from participants with written informed consent for population-scale genetic and medical inference," said Xun Xu, President of BGI-research and the study's lead author, "We developed series of practical and novel analytical methods and obtained numerous interesting results."

"It's amazing that this is even possible; that you can take these massive samples and do association mapping to see what the genetic variants are that explain human traits," said co-author Rasmus Nielsen, a professor of integrative biology at the University of California, Berkeley, who jointly oversaw the study at BGI in Shenzhen, China.

The data also has the potential to be used for viral epidemiology and infection monitoring, said Liu. "The fact that we are able to study DNA viral infection in maternal plasma also suggests the potential for infection monitoring using NIPT." She added that this will require further investigation with experimental validations.

### **The Chinese population history**

China is the world's most populous country and the second largest economy. The genetic resources are rich and unique but largely under-studied. The largest Chinese population study until now numbered only in the thousands of participants, mainly Han Chinese from the eastern coast. The current study has sampled 141,431 individuals across the entire population and provides an opportunity to learn about the genetic structure, the gene flow and the genetic adaptation in the Chinese population.

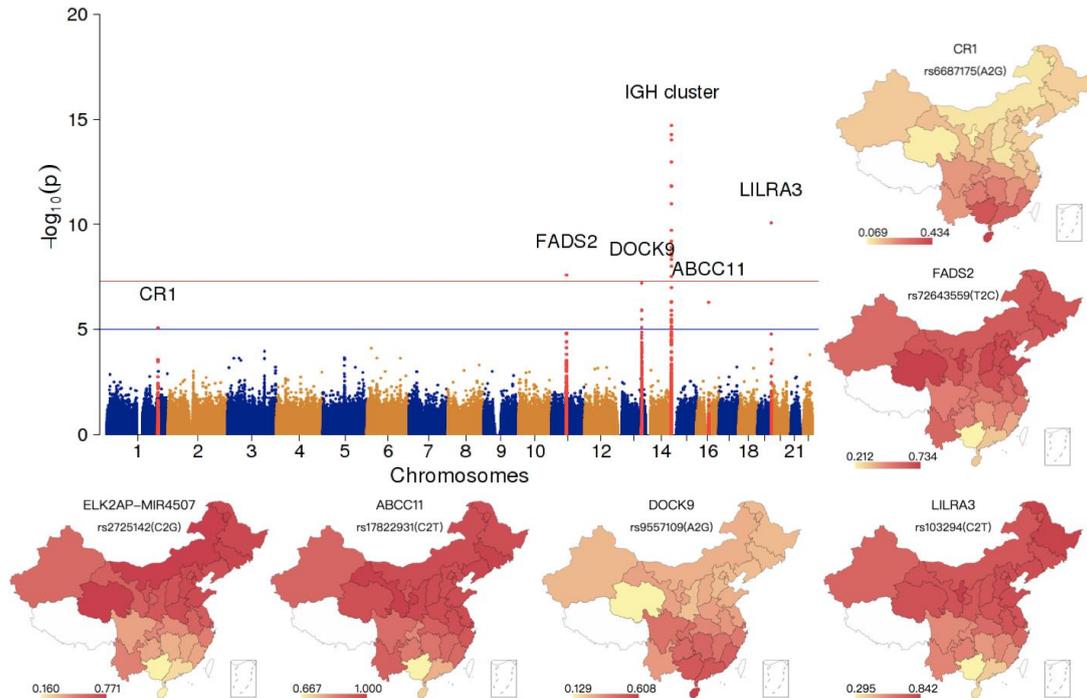


Figure 2. Unravelling six genes that have experienced significant natural selection in Chinese population across varying latitudes

By constructing novel statistics, the research team has developed analytical strategies and algorithms suitable for the unique NIPT data. They revealed the genetic structure of the various ethnic groups, the gene flow patterns between the Europeans, South Asians, East Asians and the 31 provincial administrative units in China, and six genes that display significant differences in gene frequency across latitudes. Those genes are related to immune responses, fatty acid metabolism and climate, unraveling how environmental factors have influenced the evolution of the Chinese population. For example, the well-known *FADS2* gene has a much higher allele frequency for the derived allele in most of the northern provinces that consume a meat-oriented diet due to lack of vegetables in those places in ancient China (Figure 2). For the first time, the study provides such information clearly across the Chinese map.

### Potentiality of NIPT data in genome-wide association studies



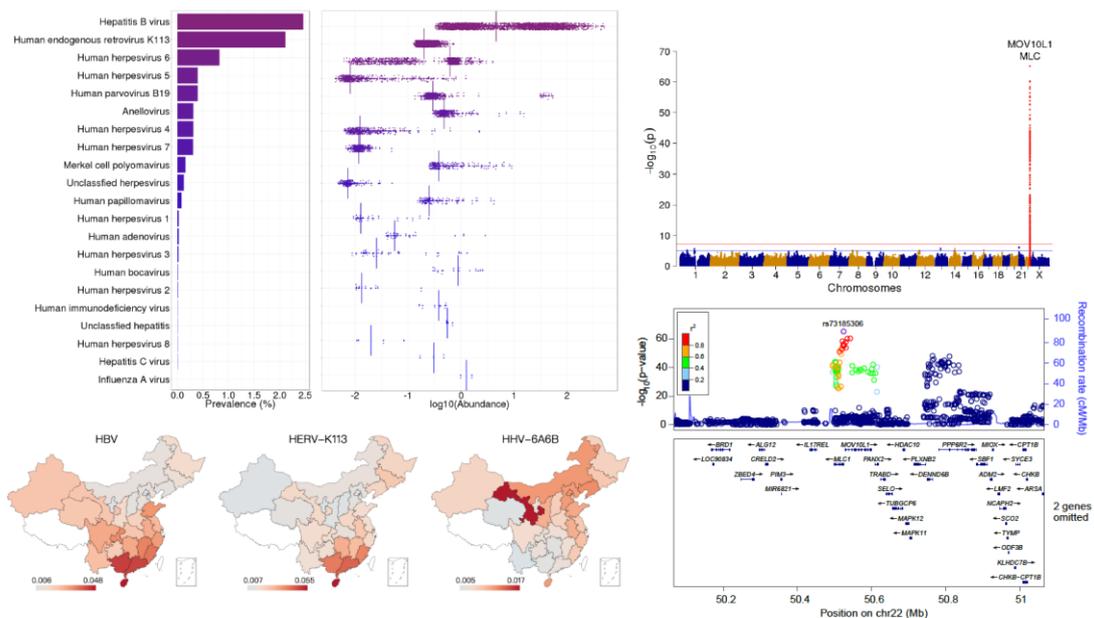


Figure 4. Viral profile in maternal plasma and the loci associated with HHV6 infection

Another interesting discovery is the significant association between a genetic variant in *MOV10L1-MLC* loci and the Herpesvirus 6 infection.

The research team identified a different pattern in terms of viral DNA burden in the Chinese pregnant women population compared to that of the European population in a previous study. The Hepatitis B DNA displays the highest prevalence in the Chinese population, followed by a human endogenous retrotransposon virus HERV-K113 and the Herpesvirus 6 that is related to skin and neural disorders. On the other hand, the top viral DNA in baseline European populations are Herpesvirus 7, related to skin disease, and Herpesvirus 4, related to nasopharyngeal cancer.

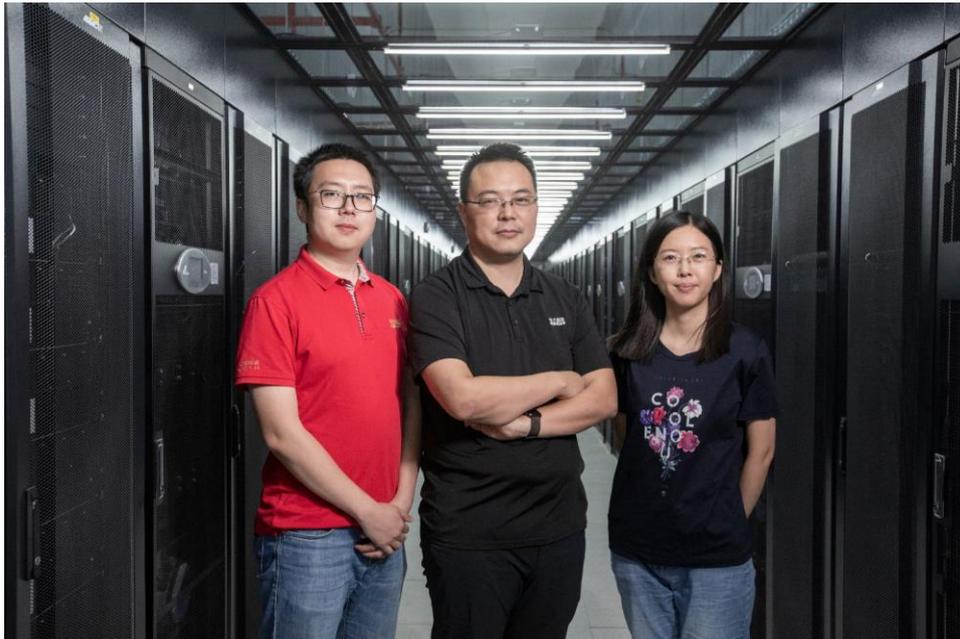
When studying the correlation between the genes and the presence of the viruses, the team has found a variant that significantly increases a person's susceptibility to Human Herpesvirus 6 infection, which is important since HHV6 infection is known to be related to both the childhood illness roseola and Alzheimer.

### The CMDB database

Strictly following the ethical and data privacy regulations in China, the study has built up the largest scale allele frequency database of the Chinese population, called Chinese Millionome Database (CMDDB) <https://db.cng.org/cmdb>. Allele frequency information of 9.05 million single nucleotide polymorphisms identified in the 141,431 Chinese genomes, as well as the summary statistics of GWAS conducted for the four phenotypes, were released on the website.

The study effectively demonstrated the utility of NIPT data for population and medical genetic studies. "It opens up new avenues for investigating hypothesis related to the

association between DNA genetic variability and important traits,” says Xu. “For me, this is a very exciting new model for biology research. It provides powerful tools and a platform for future study. Here, we show the proof of concept of the data and the structure and the methods, and that it could be used to study a lot of things. It’s just the beginning.”



Dr. Xin Jin (co-corresponding author), Dr. Xun Xu (lead author) and Dr. Siyang Liu (first author) at BGI’s computational center.

#### Related information

Liu et al., Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Association, Patterns of Viral Infections, and Chinese Population History, *Cell* (2018),

<https://doi.org/10.1016/j.cell.2018.08.016>

The CMDDB database: <https://db.cngb.org/cmdb/>

Contact: [bigdata@genomics.cn](mailto:bigdata@genomics.cn)