华大基因 BGI | Premier Scientific Partner

Human Disease Research in the Era of **Next-Generation Sequencing**

Introduction

In the current post-human-genome-project era, next-generation sequencing (NGS) technologies promise to accelerate human genetic studies tremendously. The International HapMap and 1,000 Genomes Projects have identified numerous valuable genetic variants—many of which have been annotated extensively at single-base resolution. This information in turn has fueled researches in population genetics, human disease association studies, comparative genomics, pharmacogenomic studies, and more.

However, much discovery yet remains: we seek to identify important rare variants, understand disease inheritance, and realize the goal of therapies based on personal genomics. NGS technologies make these goals possible and promise a way to exploit the complexity of the human genome. With NGS technologies fueling the diverse palette of analytical and bioinformatics technologies available today, we can now take a systems biology approach to human disease research that includes genetic, epigenetic, and pathway perspectives.

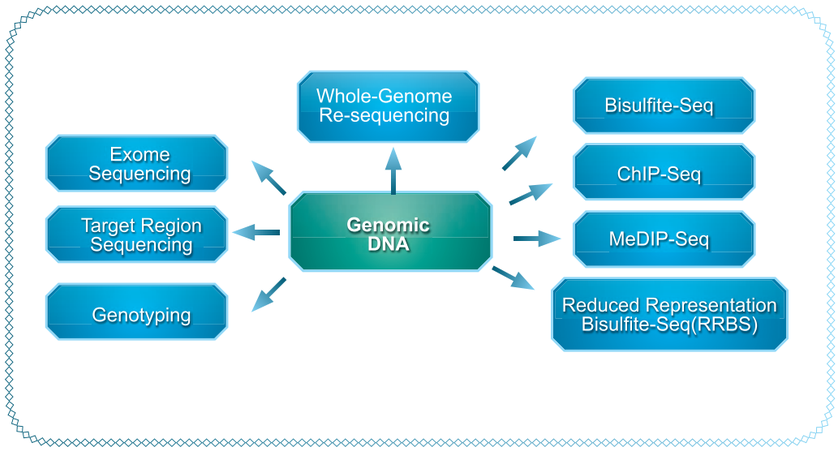
At BGI, we leverage the full spectrum of these multi-omics technologies to find answers to the important research questions that impact human health. Our collaborations include projects in cancer, Mendelian disease, and complex disease research, which rely on a wide range of NGS-based methods and bioinformatics approaches. From whole exome, whole genome, and GWAS studies to leading-edge metagenomics and single cell sequencing methods, BGI is delivering available technologies to enable our customers collaborators to meet their research goals.

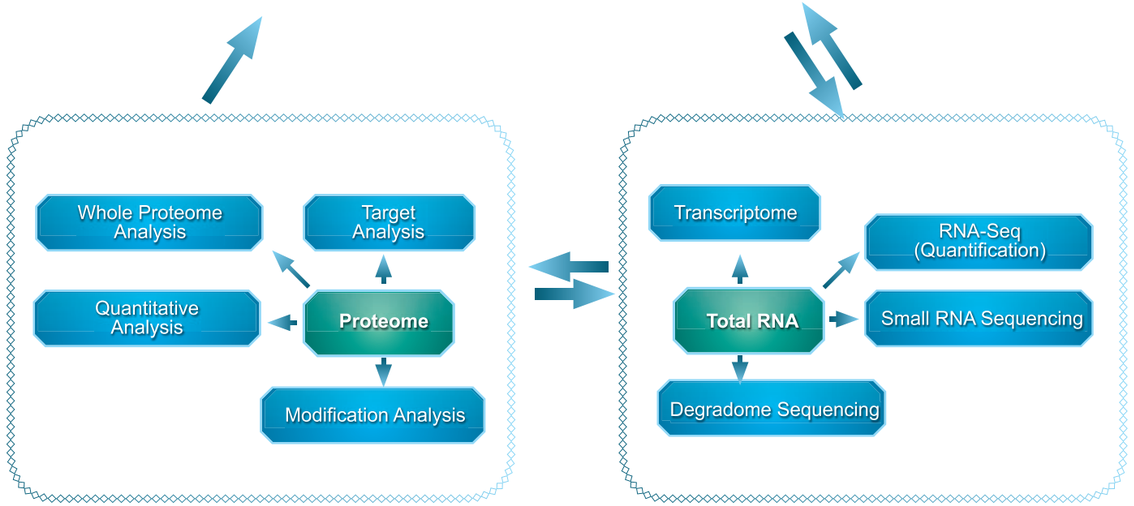
Selected Publications of BGI

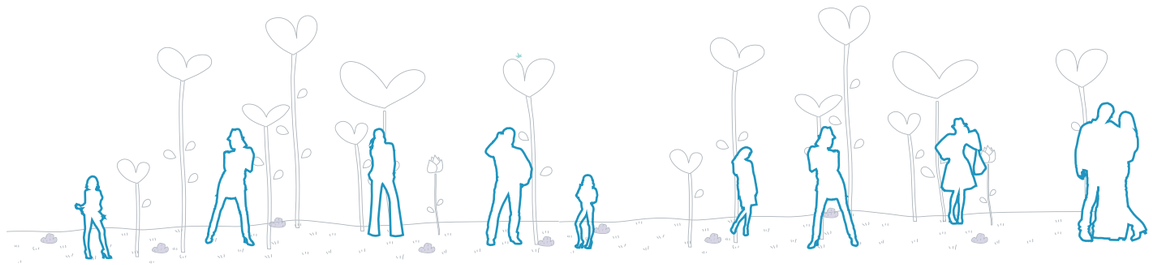
- The Chicken Genome
  Nature
  Dec 2004
- Domestication of Silkworm by Large Scale Resequencing
  Science
  Oct 2009
- Human Pan-Genome by Sequence Assembly
  Nature Biotechnology
  Jan 2010
- Epigenomics Map of Silkworm by Methylation Study
  Nature Biotechnology
  May 2010
- Genes for High Altitudes by Exome Sequencing
  Science
  Jul 2010
- Resequencing of 200 Human Exomes Identifies An Excess of Low-Frequency Non-Synonymous Coding Variants
  Nature Genetics
  Oct 2010

Contents

Multi-Omics Technologies for Human Genomics Research







Human Evolution and Population Genetics Research by NGS

Introduction

The development of higher-throughput, cost-effective NGS technology has taken population genetics from the HapMap project to sequencing individual whole genomes, enabling us to gain more detailed sequence information useful in medical research.

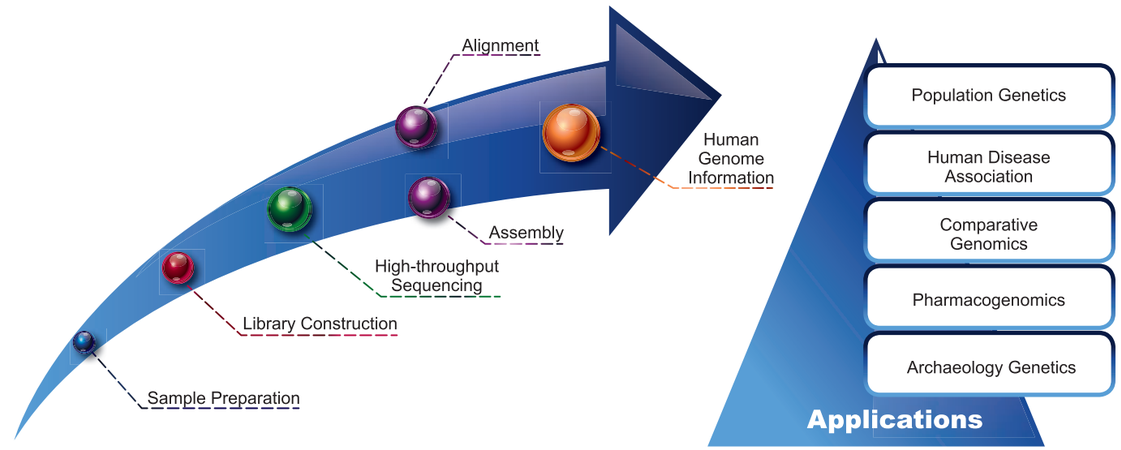


Figure1. Human population genetics research workflow and applications

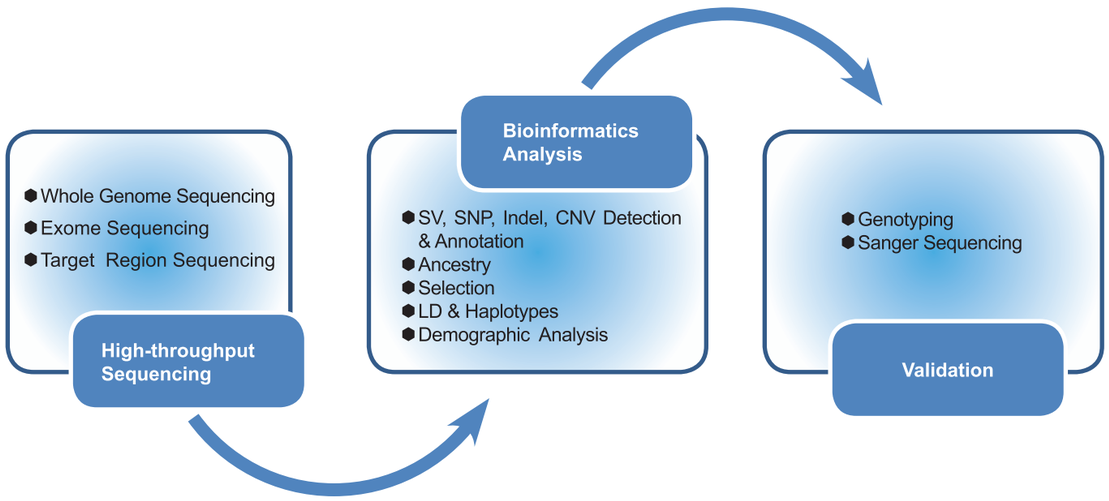


Figure 2. Strategies for high-throughput sequencing in human population genetics research

BGI Research Cases

(1) Ancient Human Genome Sequence of an Extinct Palaeo-Eskimo. ***Nature*** 2010; 463: 757-762.

In 2010, BGI and Copenhagen University obtained DNA from a 4,000-year-old permafrost-preserved human hair, and isolated and sequenced its genomic DNA. The genome is from a male individual who was a member of the first known population to settle in Greenland. SNP analysis of the sequence data enabled the researchers to assign possible phenotypic characteristics of this individual and identify the population to which he is most closely related. This analysis provided evidence that there was human migration from Siberia into the New World about 5,500 years ago, and this migration was independent of the one that gave rise to the modern Native Americans and Inuits.

(2) The Diploid Genome Sequence of an Asian Individual. ***Nature*** 2008; 456 (7218): 60-65.

The first Asian individual genome was sequenced for 36X by massively parallel sequencing technology. The NCBI human reference genome was covered at 99.97%. And the resulting high-quality consensus sequence represents 92% of the individual genome. The data quality and analyses demonstrated the potential usefulness of next-generation sequencing technologies for personal genomics.

(3) Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. ***Science*** 2010; 329 (5987): 75-78.

50 exomes of ethnic Tibetans were sequenced for 18X per individual. Genes showing population-specific allele frequency changes, which represent strong candidates for altitude adaptation, were identified. One single-nucleotide polymorphism (SNP) at EPAS1 shows a 78% frequency difference between Tibetan and Han samples, representing the fastest allele frequency change observed at any human gene to date. This research can help us to prevent and cure the disease of plateau-anoxia.

Mendelian Disorders Research

Mendelian disorders, which result from mutations of a single gene, are an important area of active research. However, many important modern human diseases are examples of complex disease. Examples include diabetes, obesity, hypertension, and many cancers—all complex diseases that result from a web of interactions among multiple genes and environmental factors. Careful analysis across multiple data sources (for example, clinical data, transcriptome, and genetic sequencing data sets) can reveal much about a complex disorder. Researchers can point to causal genes and identify distinct disease subtypes, which may present as one phenotype. Understanding these can be critical to designing effective therapies.

To identify the causal factors for disease—both Mendelian and Complex—researchers are moving from traditional hypothesis-driven research to data-driven research. The rapid development of sequencing technologies has enabled this shift, allowing researchers to broaden their focus from examination a few suspect-genes to simultaneous interrogation of most or all genes across the human genome.

BGI 1000 Mendelian Disorders Project

In May 2010, BGI launched the “1000 Mendelian Disorders Project.” With this initiative, we seek to understand the molecular basis of key Mendelian disorders to facilitate their early prediction and diagnosis and to identify potential interventions. Currently, we are collaborating on more than 100 Mendelian disorder projects with diverse and talented groups of collaborators across the globe. More than 40 of these projects are in the validation stage, and we continue to seek additional collaborators. Within the first six months of the launch of this initiative, one of our collaborators identified a gene, *TGM6*, previously unassociated with any disease state, as responsible for a rare Mendelian disorder—spinocerebellar ataxia (SCA). The research summary is presented below.

BGI Research Cases

TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. Oxford Journals Medicine Brain Volume 133, 2010. Issue12 Pp. 3510-3518.

This study was a collaboration among BGI Shenzhen, Xiang Ya Hospital, Central South University, and several other organizations. Candidate causal genes were identified using whole-exome sequencing from disease sufferers across multiple generations of one Chinese family.

A summary of study findings:

- Exome sequencing and analysis identify *TGM6* as a candidate causative gene for spinocerebellar ataxia.
- In confirming analyses, found a different mutation in *TGM6* and this novel variant was predicted to be damaging.
- Identified a novel SCA causative gene, *TGM6*.

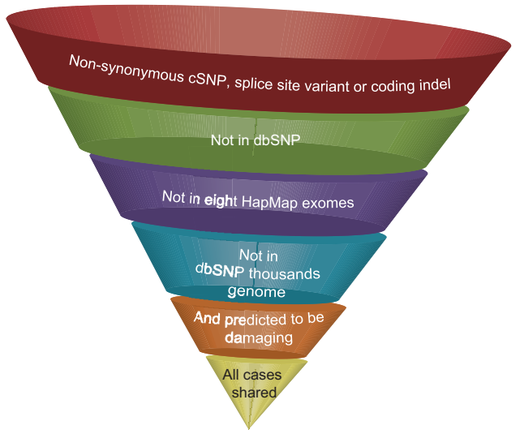
Mendelian Disorders Research Strategies

Using NGS technologies, BGI provides whole exome and whole genome sequencing workflows suited to the study of Mendelian disorders. Project workflows from data acquisition to identification and validation of candidate genes are shown in Figures 3.

A. Research Strategy

Sample selection	≥4
Timeline	60 working days
Method	Exome Sequencing : 50 X Whole Genome Sequencing : 30X

B. Screening of Candidate Genes (Analysis Pipeline)



C. Validation of Candidate Genes

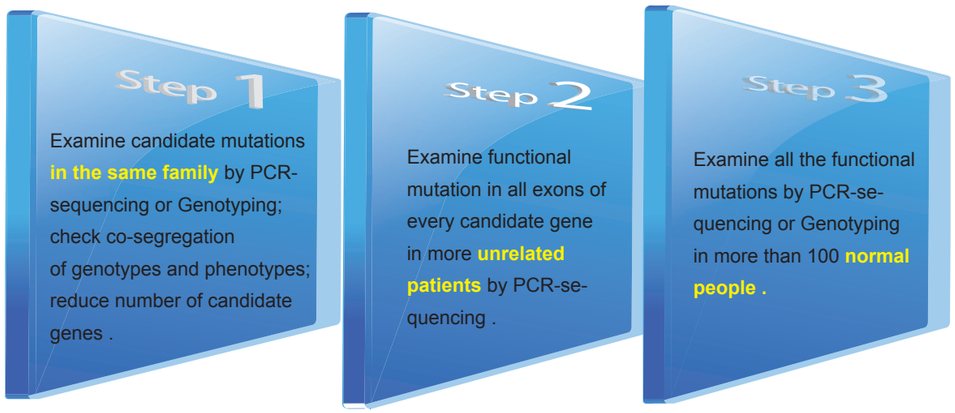


Figure 3. BGI workflows for Mendelian disorder research

Complex Disease Research

1. Next-Generation GWAS Strategy

Next-generation GWAS combines high-throughput sequencing and genotyping to uncover novel causative genetic mutations of complex human disease. Newer GWAS technologies, like human genome sequencing and CNV arrays, can provide detailed information, from common SNPs, to millions of less-common and rare variants, to completely novel variants. These results can complement those of traditional GWAS methods.

To accurately identify causative genetic mutations in complex disease, a two-stage strategy is proposed:

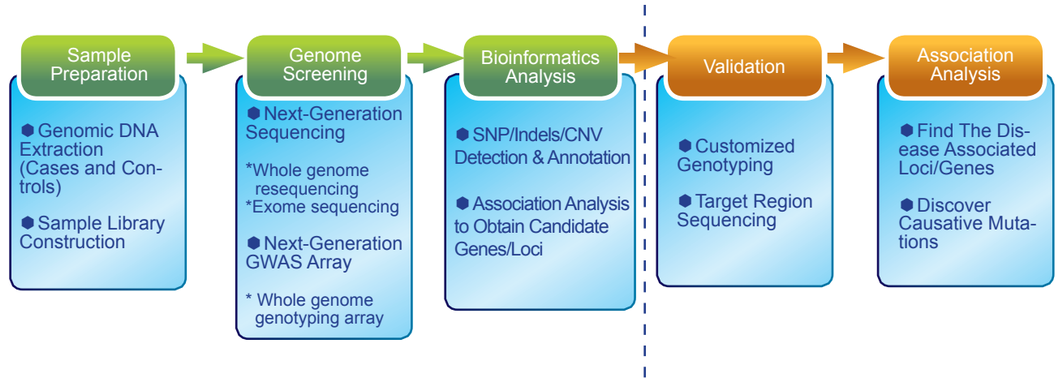
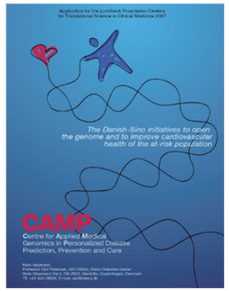


Figure 4. Workflow of the two-stage strategy

BGI Research Cases

Diabetes-Associated Genes and Variations Study



The Sino-Danish diabetes-associated genes and variations study was launched in 2008 at BGI-Shenzhen. The LuCAMP consortium comprising ten research organizations from Europe and China was formed to carry out this project.

Design

- Select 1,000 patients with gender-defined visceral obesity, type 2 diabetes and essential hypertension
- Select 1,000 age-matched and gender-matched controls, who are glucose-tolerant, lean and normotensive
- Collect samples for exome sequencing

Objectives

- Identify novel genetic variations, both common and rare associated with the disease state.
- Identify differences in the allele frequency of genetic variations between disease-affected and control groups

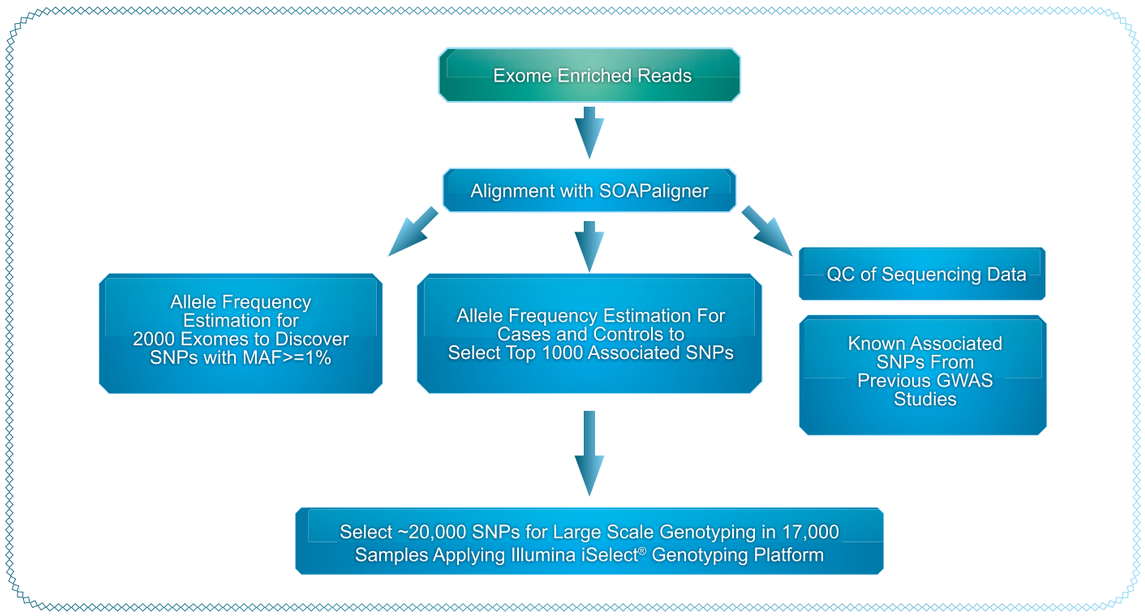


Figure 5. Bioinformatics pipeline of exome sequencing & genotyping

Partial Results From This Project

Summary of SNP Calling

Exon			Intron/Intergenic	Total	Total SNPs Classified by Allele Frequency		
CDS		UTR					
Synonymous	Non-Synonymous				<0.01	<0.02	<0.05
27,718	30,871	10,523	83,717	152,829	41%	57%	66%

Figure 6. The classification of SNPs discovered in this project

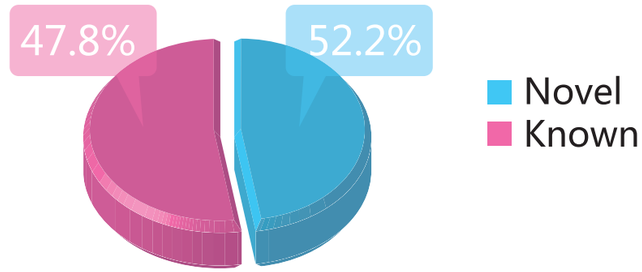


Figure 7. The percentage of detected SNPs

Association Study

We first surveyed the allele frequency difference between the case and control groups. We then assessed the likelihood ratio for each association. We detected some genes previously associated with metabolic disorders, such as ADRB3 and GSK3A.

Publication

Li Y, Vinckenbosch N, Tian G, Huerta-Sanchez E, Jiang T, et al. Resequencing of 200 human exomes identifies an excess of low-frequency, non-synonymous coding variants. Nature Genetics. 2010; 42 (11): 969-972.

Work in Progress

The Illumina iSelect HD Custom Genotyping Array was constructed based on the selected SNPs. The large-scale genotyping experiment is complete. The association analysis is in process to further validate the disease-associated genes and loci.

Detect Variants of Drug Targets

We collaborated with GlaxoSmithKline on a project designed to better inform selection of clinical trial subjects. The project uses target region sequencing to examine common variants in known drug targets of existing therapeutics.

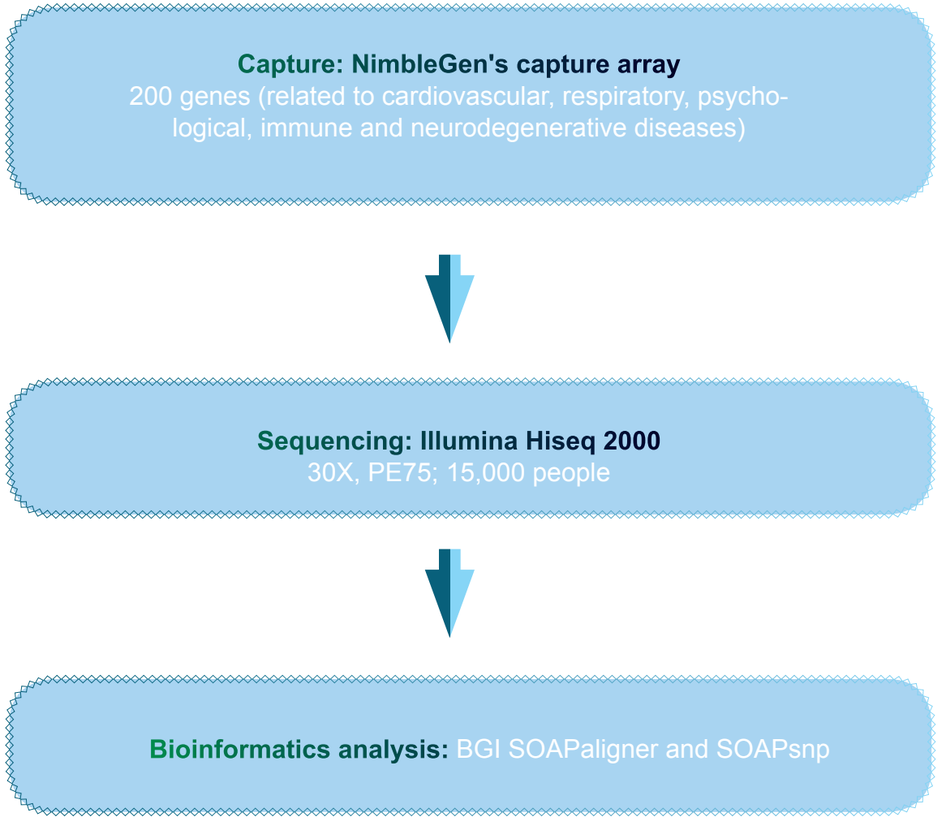


Figure 8. Workflow of target region sequencing

Currently, the researchers have analyzed results from 5,000 patients and have found approximately 50,000 SNPs, 7,000 of which are unique and non-synonymous. The association study is a work in progress, but already the amount of small associations is more than expected.

2.RNA Sequencing Strategy

Compared with DNA, RNA is more dynamic, as it reflects the functional status of specific cells in a particular time and space. Using transcriptome sequencing, RNA-Seq and small RNA sequencing technologies, one can monitor gene expression from a specific tissue in a particular functional state. These technologies have the multiple advantages of digital signals, high resolution and no requirement for previous information, making them applicable in many research fields including fundamental disease research, pharmacogenomics and disease typing.

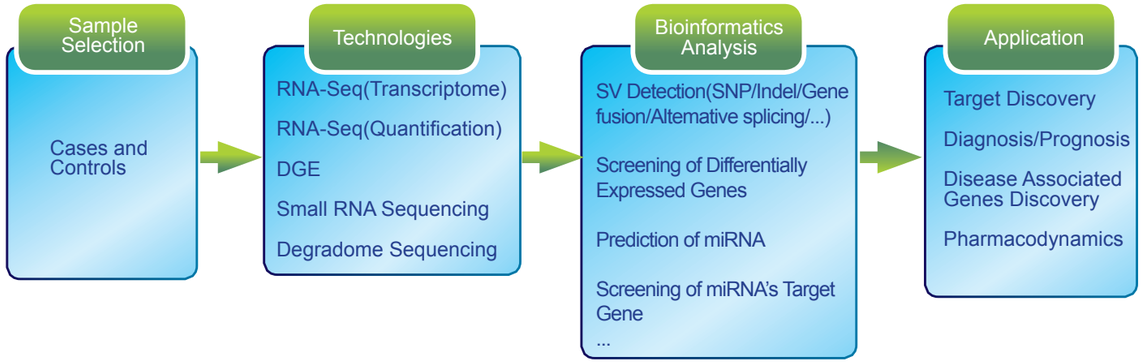


Figure 9. BGI workflow of RNA sequencing strategy

3.Epigenomics Strategy

Sequencing technologies

1. Bisulfite-Sequencing

Genome-wide DNA methylome analysis at single base-resolution. Research examples include the study of methylation-regulated genes potentially involved in cancers.

2. Reduced Representation Bisulfite Sequencing (RRBS)

A cost-effective, high-throughput method for studying DNA methylation, focused on CpG sites.

3. MeDIP Sequencing

Difference analysis of the genome-wide DNA methylation patterns among multiple samples, based on an enrichment method.

4. ChIP Sequencing

Genome-wide profiling of DNA-binding proteins (histones). Research examples include study of the mechanism of regulation of cancer-related proteins.

Epigenomics for Study of complex Disease

Epigenetics refers to the regulation of various genomic functions, including gene expression, brought about by heritable, but potentially reversible mechanisms. Epigenetic regulatory mechanisms include DNA methylation and histone modification. Recent research points to epigenetic mechanisms as important factors in the development of complex disease. By understanding these mechanisms, we may find targets for therapeutic interventions that interfere with the primary epigenetic cause of a disease.

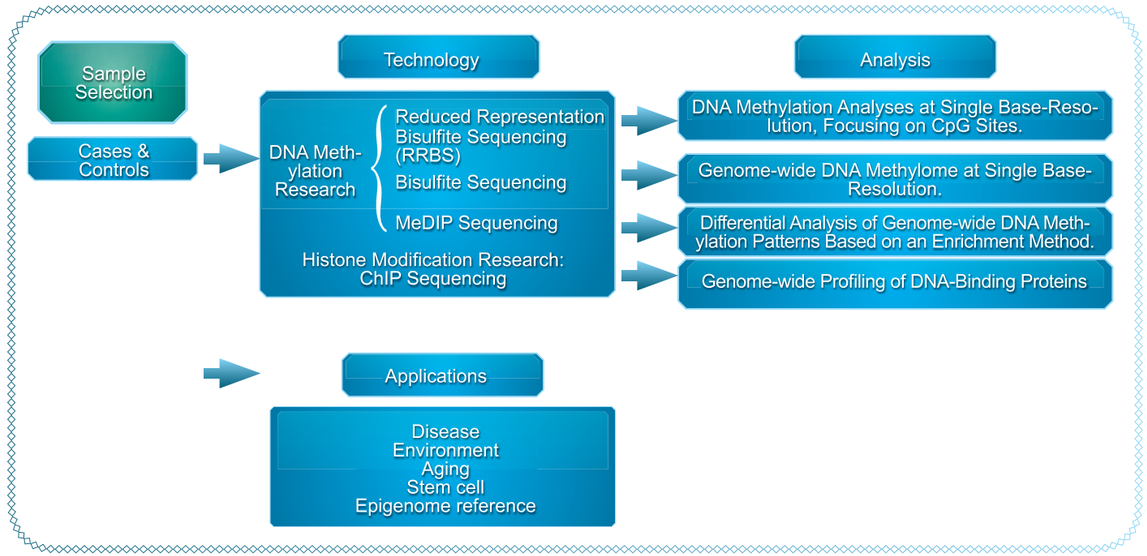


Figure 10. Epigenomics research workflow

Example research

⬢ MeDIP sequencing of breast cancer

Ruike Y, Imanaka Y, Sato F, Shimizu K, Tsujimoto G. Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-DNA immunoprecipitation combined with high-throughput sequencing. BMC Genomics. 2010; 11(137)

⬢ This study provides the first comprehensive, detailed map of DNA methylation patterns in human mammary cell lines, by covering almost the entire genome at sufficient depth and resolution.

⬢ Results: Massively reduced methylation level, particularly in CpG-poor regions, was found in human breast cancer cell lines (BBC) compared to normal human mammary epithelial cells (HMEC).

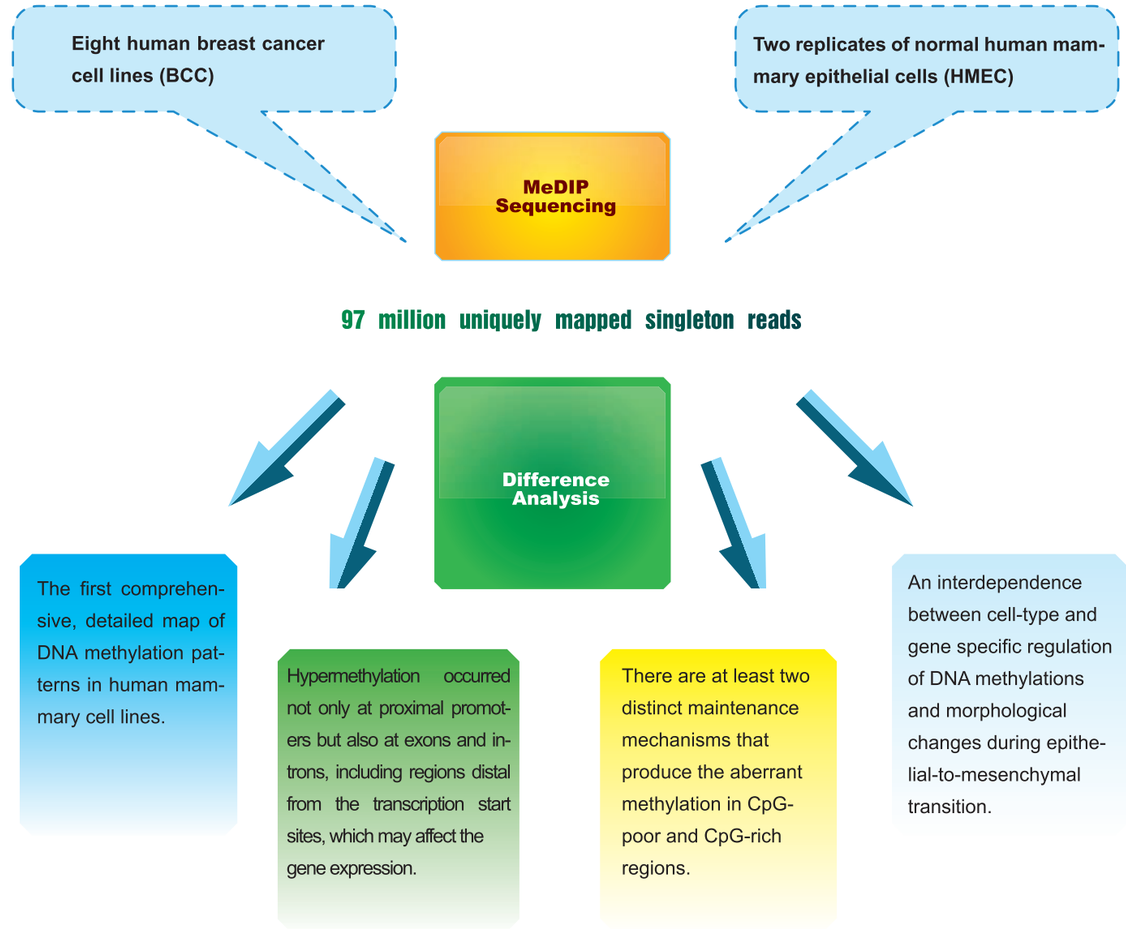


Figure 11. Workflow of MeDIP sequencing of breast cancer

4. Proteomics Strategy

The goal of proteomics analysis is not only to determine the expression level of proteins, but also to map their localization and understand their dynamics, post-translational modifications and interaction partners. We are pushing the existing proteomics technologies to provide answers useful to medicine (for example, diagnostic markers for cancer), the development of bioenergy sources, and other disciplines.

Example Research

Chen YT, Chen CL, Chen HW, Chung T, et al. Discovery of novel bladder cancer biomarkers by comparative urine proteomics using iTRAQ technology. J Proteome Res. 2010; 9 (11):5803-5815.

An iTRAQ-based quantitative proteomics approach for biomarker discovery from urine

- Using the strategy shown in Fig12, the global differences in the urine proteome between non-tumor controls and three bladder cancer patient subgroups with different grades/stages were analyzed.
- Combining the results of two independent clinical sample sets, a total of 638 urine proteins were identified and 507 were quantified. Furthermore, 55 proteins consistently showed greater than 2-fold differences in both sample sets. After validation using Western blot analysis and ELISA, several urine proteins were identified as novel candidates for bladder cancer diagnosis.
- Collectively, these results provide the first iTRAQ-based quantitative profile of bladder cancer urine proteins. These findings represent a valuable resource for the discovery of bladder cancer biomarkers.

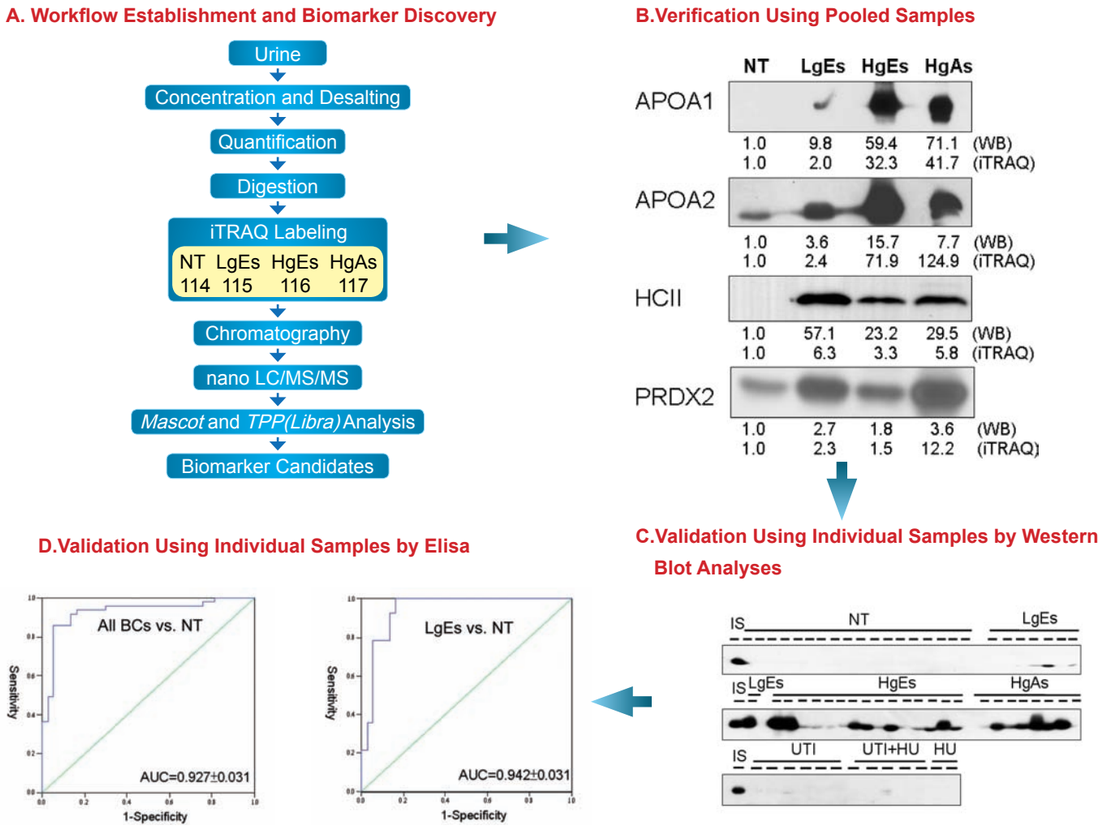


Fig12. Experiment pipeline and validation

5. Multi-omics Strategy

The flow of information from DNA to RNA to protein makes for a complex network of information. A systematic multi-omics research method can provide a detailed picture of complex disease processes. With this approach, researchers relate diverse data sets, such as genomics, transcriptomics, epigenomics, and proteomics. The full picture of a disease state may include cancer germline mutations, somatic mutations (including SNV, InDel, SV, CNV, etc.), gene expression information, gene structure changes, and epigenetic regulation of disease-causing genes. The multi-omics strategy provides an information-intensive approach to understand the many layers involved in producing the phenotype of a complex disease, and ultimately may provide a holistic understanding of the mechanism underlying the disease phenotype.

Example Research

- Whole genome sequencing and transcriptome sequencing of breast cancer
- Shah SP, Morin RD, Khattra J et al. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature. 2009; 461 (7265): 809-813.
- Method: Sequence the genomes (43-fold coverage) and transcriptomes of an estrogen-receptor-alpha-positive metastatic lobular breast cancer.
- Results: The combined analysis of genome and transcriptome data revealed two new RNA-editing events that recode the amino acid sequence of SRP9 and COG3.

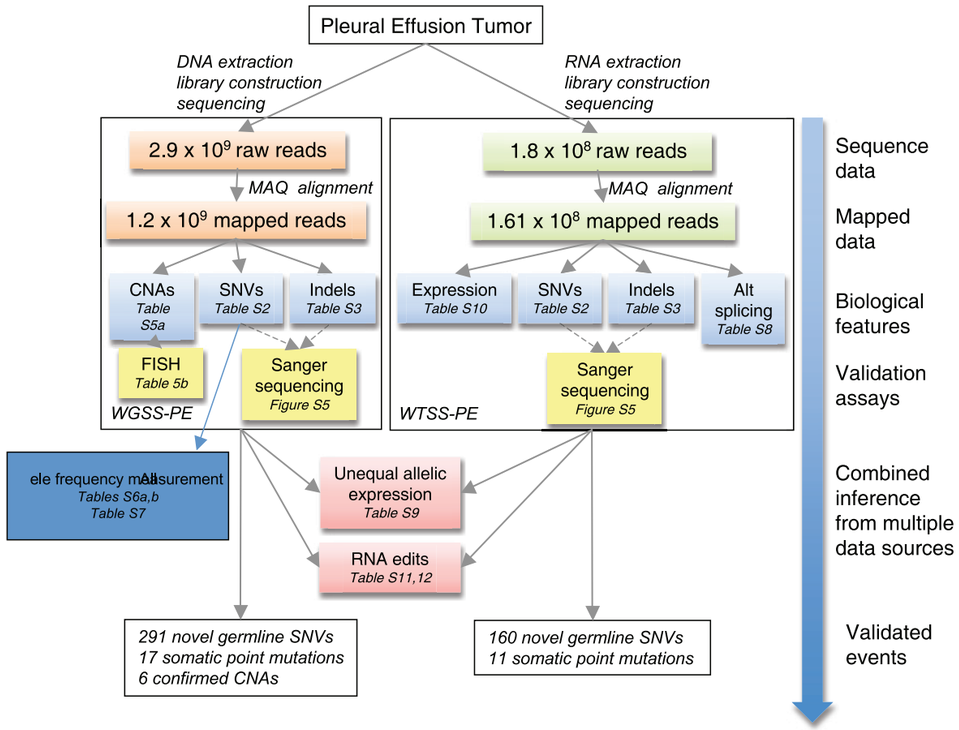


Figure 13. Schematic diagram of analysis workflow depicting (from top to bottom) the analytical steps and resulting inferences. Tables and Figures referred are from the original article. Taken from Shah SP, 2009 (full reference above).

Cancer Research Highlight

Single Cell Sequencing

Single cell sequencing is an innovative technology that utilizes whole genome amplification and next generation sequencing to obtain sequence data at the level of single-cell genomes. It can be applied to discover genetic information in single cells, detecting SNV at high accuracy with low false positives. This technique is useful in the study of many cancers, in which there is often high cell heterogeneity across a single cancerous tissue. Single cell sequencing allows for the differentiation of those mutations that coincide with the development of cancerous cells and those that spur the cancer's progression.

Applications

- Analysis of cancer-cell evolution during tumor progression
- Large-scale epidemiological tumor research
- Early diagnosis and prognosis prediction of cancers

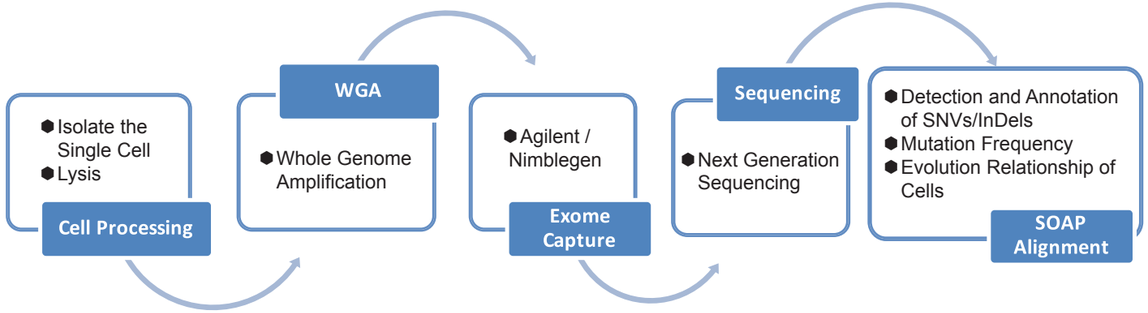


Figure 14. Workflow of single cell sequencing

BGI Research Cases

Single cell sequencing of cancerous tumors (ongoing):

Genetic comparisons between cancer cells, normal cells and leukocytes are in progress

Single cell sequencing reveals subpopulations of cancer cells [*] (CSHL)

- Five major subpopulations of cells were detected, three of which were cancerous
- Each subpopulation was clustered within a specific region of the tumor
- Mutations in the cancer genes EFNA5 and COL4A5 were identified

Note: [*]Single cell sequencing reveals subpopulations of cancer cells, Reported by Monica Heger, 2010 Cell Line Sequencing

Cell Line Sequencing

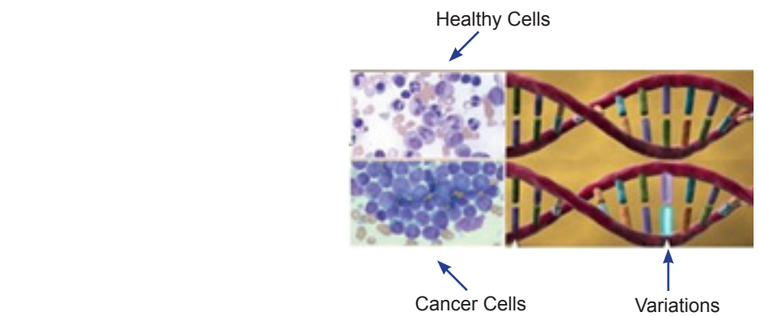
Human immortal cancer cell lines serve as an accessible and widely adopted biological model for investigating basic cancer biology and the efficacy of anticancer drug candidates. With knowledge of the genetic abnormalities inherent to cell lines, investigators may better match their choice of cell lines to the goals of their study. Doing so, they increase both the relevance of their results and their ability to interpret the validity of their results in the context of their test system. Despite the widespread use of immortal cell lines, it remains necessary to systematically characterize the genetics and genomics of large numbers of the cancer cell lines in use today.

Applications

- Describe the genome of the immortal cell lines used for cancer research
- Highlight suitable models for anticancer drug investigations
- Facilitate the development of personalized medicine for cancer

Technology Advantages

- Detect high-accuracy SNV through alignment
- Identify SV/CNV at the single-base resolution ratio
- Get a clear genetic picture of a cell line of interest
- Characterize both cancer cell lines and primary culture cells



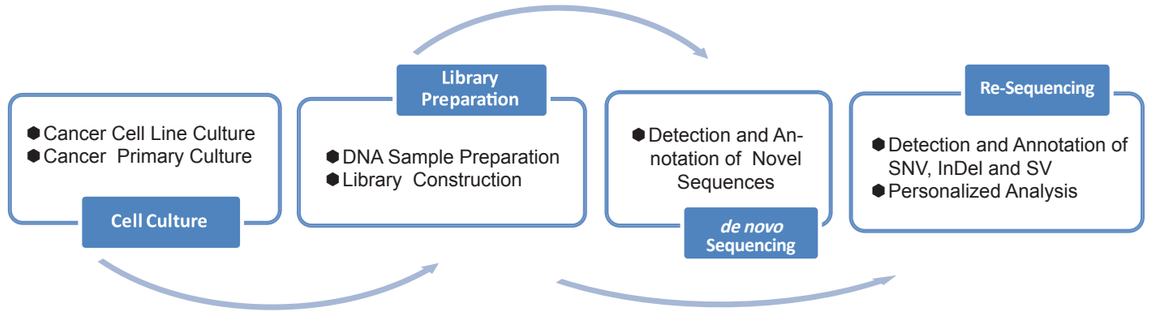


Figure 15. Workflow of cell line sequencing

Example Research

1.Comparison of breast cancer cell lines with primary tumor cells by NGS

Researchers used next-generation DNA sequencing to produce maps of genome rearrangements in 24 breast cancer samples (9 immortal cancer cell lines and 15 primary tumors)[1].

2.Small-cell lung cancer genome revealed by cell line sequencing

Using massively parallel sequencing technology, researchers sequenced a small-cell lung cancer cell line (NCI-H209) to explore the

[1] Pleasance ED, Stephens PJ, et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature. 2009; 462 (7276): 1005-1010.
[2] Pleasance ED, Stephens PJ, et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature. 2010; 463 (7278): 184-190

Metagenomics —A New Way to Research into Human Microbiome

There is an increasing awareness of the relationship between human health and the microbe populations living in the human body. Research has shown diverse diseases, such as obesity, enteritis, diabetes, and colon cancer as influenced by the microbe populations resident in the body. To understand the impact of resident microbes on human health, it is crucial to assess their genetic potential—to understand our “other genome.” Metagenomics based on NGS technologies allows for sequencing of the genomic DNA of microbial communities directly, bypassing microbe isolation and cloning steps. Based on a BGI in-house profiling algorithm and association analysis method, we were able to identify the mixes of microbial populations in each study group and, by comparison of groups, explore the correlation between microbial populations, microbial genes, and disease states.

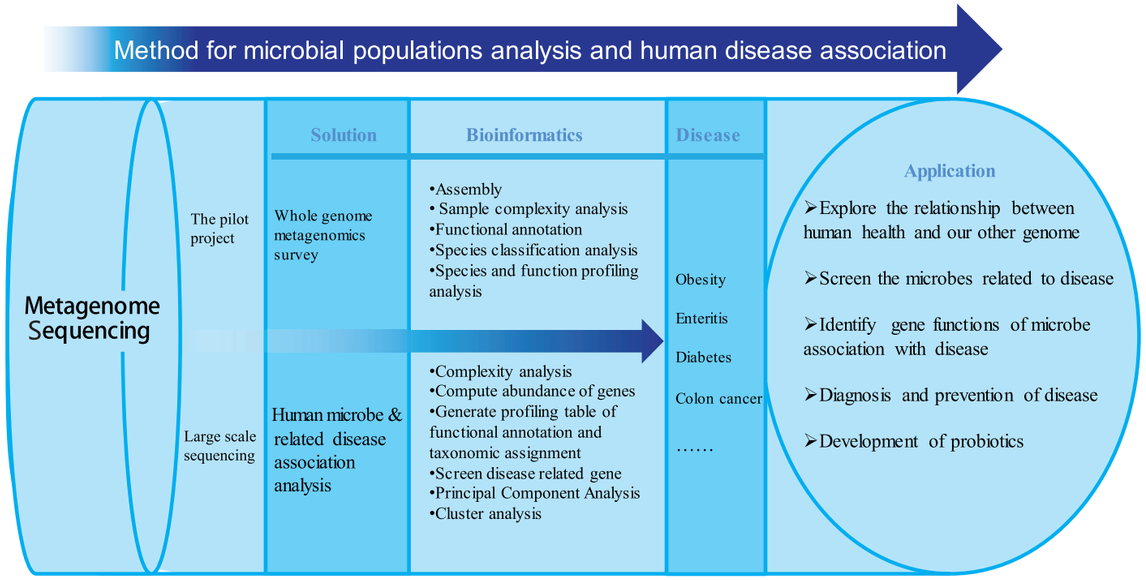


Figure 16. Method for microbial populations analysis and human disease association

BGI research case

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010; 464: 59-65.

Sequenced 124 European Faecal Samples:

1) Found 1,000 to 1,150 prevalent bacterial species and 3.3 million non-redundant genes.
2) Established the first human gut microbial gene catalogue.
3) Confirmed that bacterial species abundance and bacterial genes differentiated between healthy individuals and disease-affected patients. These findings offer an important theoretical basis for further exploring the relationship between human gut microbes and obesity, enteritis and other diseases.

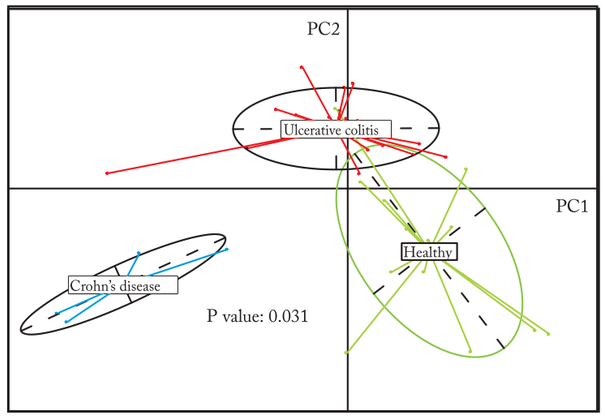


Fig17. Bacteria species abundance differentiates IBD patients and healthy individuals.

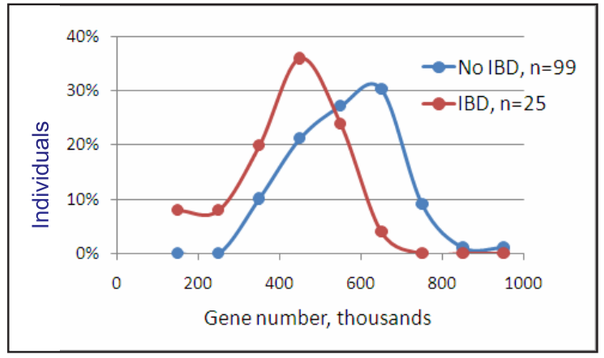


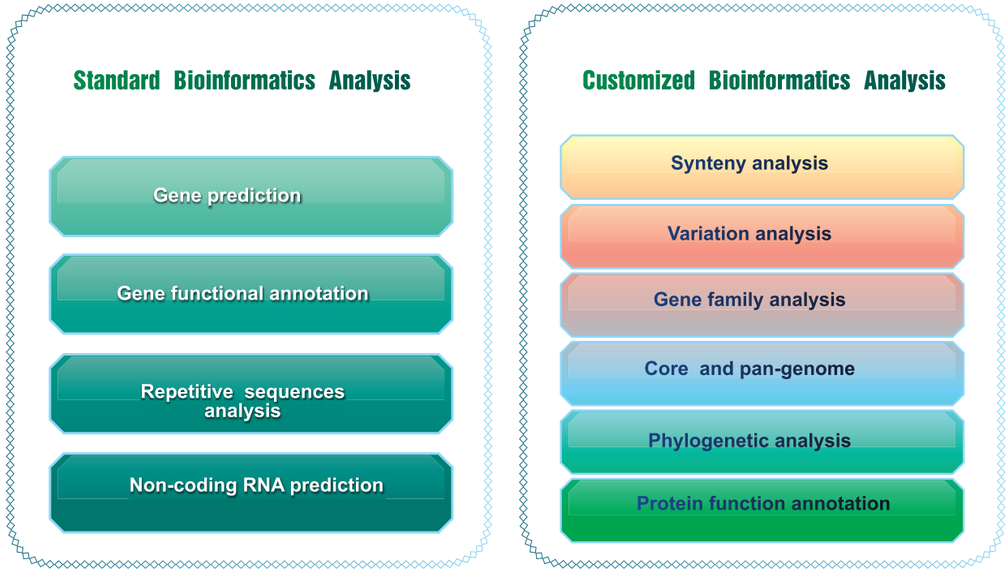
Fig18. Distribution of non-redundant bacterial genes in IBD (inflammatory bowel disease) patients and healthy controls.

BGI has both the expertise and leading platforms for pathogen genome sequencing and bioinformatics analysis. Our history includes sequencing studies of pathogenic microbes, including species of Staphylococcus, Streptococcus, Tuberculosis and others, from single strains to population levels.

Application

- The genetic relationships of novel pathogens
- Pathogenesis
- The origin and evolution of a pathogen
- Clues for pharmaceutical Innovation
- Potential methods for pathogen detection

Bioinformatics Analysis



Example Research

Morelli, G., Song, Y., Mazzoni, et al. (2010). *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. Nat Genet 42 (12), 1140-1143.

1: Plague is a pandemic human invasive disease caused by the bacterial agent *Yersinia pestis*. By understanding the population structure and evolutionary forces responsible for the emergence and spread of this pathogen, we can better understand the origin of relevant emerging pathogens.

2: We compared non-repetitive core genomes of 17 isolates of *Y. pestis*, and identified 1,364 non-synonymous SNPs in coding regions (Figure 19). 933 SNPs of 286 isolates from diverse sources were used to calculate a minimal spanning tree and to assign isolates to populations and inference historical transmission routes (Figure 20).

3: Our phylogenetic analysis suggests that *Y. pestis* evolved in or near China and spread through multiple radiations to Europe, South America, Africa and Southeast Asia, leading to country-specific lineages that can be traced by lineage-specific SNPs.

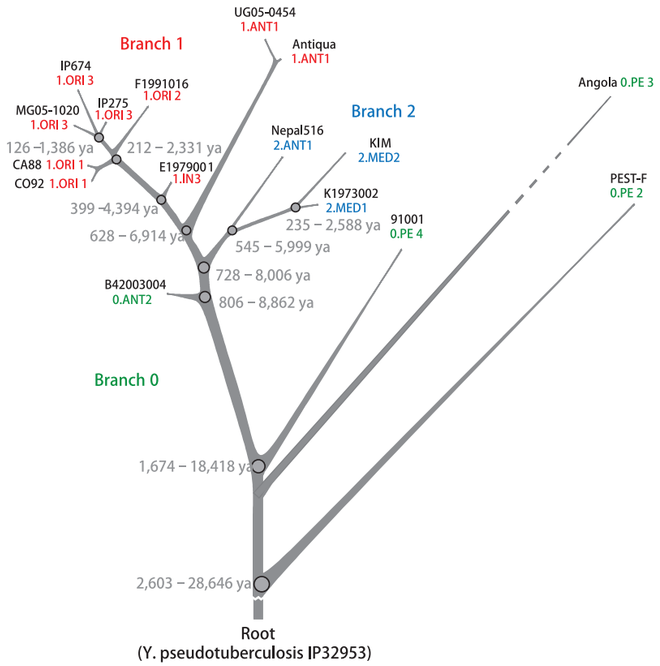


Figure 19. Genomic maximum parsimony tree and divergence dates based on 1,364 non-repetitive, non-homoplastic SNPs from 3,349 coding sequences in 16 *Y. pestis* genomes (excluding FV-1).

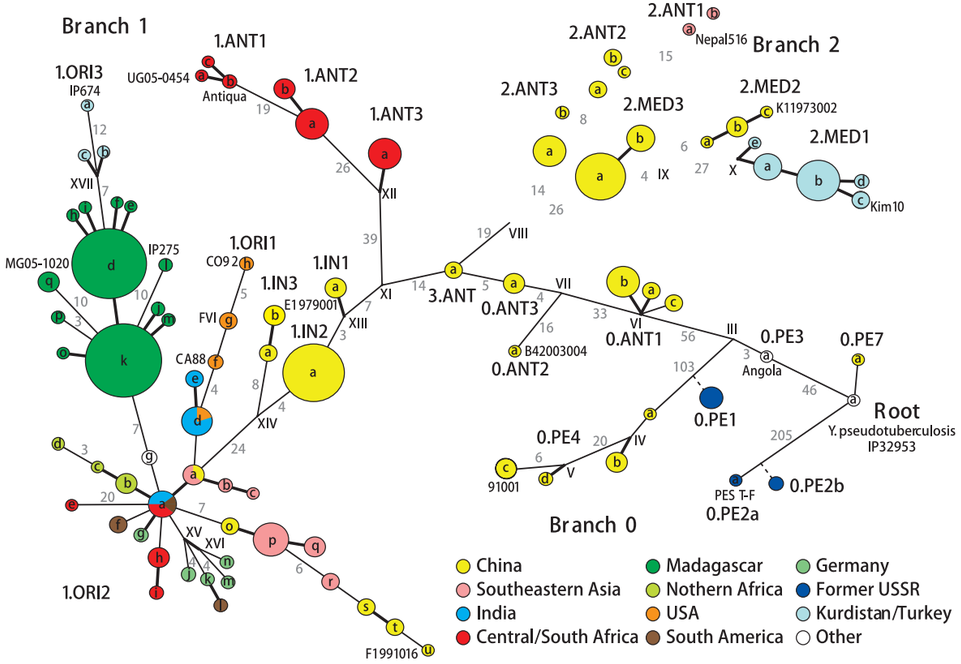


Figure 20. Fully parsimonious minimal spanning tree of 933 SNPs for 282 isolates of *Y. pestis* colored by location

NGS and related technologies provide great promise for a sophisticated understanding of the cellular mechanisms key to human health and disease. This diverse palette of analytical technologies allows us to understand the path from genes to phenotype with richer detail than ever before. BGI offers the deep experience and sequencing capacity needed to generate the multi-omics data sets needed for this approach. But accurate and rapid analysis is only one step on the path to useful and valid results. In our collaborations and publications, we have shown our bioinformatics expertise in our ability to mine these complex data sets to answer relevant biological research questions. By leveraging our combined expertise in data generation and interpretation, we strive to take the analysis burden from our collaborators and to work with them to accelerate their research. Please contact us to discuss your research interests. We would look forward to the chance to work with you to answer your most pressing research questions.

Contact Us Offices & Locations

China (Mainland)

BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen, 518083, China
Tel: 400-706-6615
Fax: +86-755-25273620
Email: tech@genomics.cn
www.genomics.cn

North America (Boston)

BGI-Americas Corporation, One Broadway, 3rd Floor, Cambridge, MA 02142, USA
Tel: 617-500-2741
Email: info@bgiamericas.com
www.bgisequence.com/us
www.bgiamericas.com

Asia Pacific

BGI-Hong Kong Co. Limited,16th Dai Fu Street, Tai Po Industrial Estate, Tai Po, Hong Kong
Tel: +852-9812 2524
www.bgisequence.com

Europe (Copenhagen)

BGI-Europe A/S, Bülowsvej 15 DK-1870 Frederiksberg C
Tel: +45-5030-1666
www.bgisequence.com/eu

华大基因 BGI | Premier Scientific Partner