

# Latest NGS Strategies for Cancer Research @ BGI

# Contents

<i>5</i>	<b>B</b>	Single-cell Sequencing
<i>12</i>	<b>D</b>	PDX Toolkits @ BGI
<i>19</i>	<b>F</b>	FFPE Sequencing
	<b>A</b>	Introduction <i>3</i>
	<b>C</b>	Circulating Tumor Cell Sequencing <i>9</i>
	<b>E</b>	Virus Integration Sequencing <i>15</i>

# Latest NGS Strategies for Cancer Research

## Introduction

Cancer is the most popular cause of death in the global world. Due to population growth and aging, the burden of this disease is sharply increasing in developing countries<sup>1</sup>. Human cancers usually carry several genomic rearrangements such as alteration of copy number and point mutations. Despite the prevalence of cancer, conventional strategies for cancer research are generally laborious and low-throughput<sup>2</sup>. The emergence of effective and powerful technologies for analyzing cancer genome sequences and DNA structures is necessary and urgent. Next-generation (next-gen) sequencing for genome-wide analysis of genetic alterations has recently revolutionized this field. The application of gene sequencing technologies through whole-transcriptome, whole-exome, and whole-genome approaches will allow substantial advances in cancer genomics<sup>3</sup>. These technological developments offer insights, which will lead to a greater understanding of the causes of cancer, as it is principally a disease of accumulation of genome alterations. These insights will ultimately lead to new strategies for cancer diagnosis and therapy. Next-gen sequencing methods will eventually result in the comprehensive identification of all the major abnormalities in isolated cancer genomes. The new challenge is to make computational, biological, and clinical sense of the huge amount of genomic data being generated<sup>4</sup>.

Since results of the human genome project (HGP) began to emerge in 1999, BGI has been striving to utilize genomic research for advancing disease treatment. The subsequent explosion of cancer research projects and collaborations (e.g., Internal Cancer Genome Project, ICHG) has dramatically improved our understanding of cancer diagnostics.

## Why Choose BGI Tech?

- Extensive experience in cancer sequencing (18,406 human cancer samples analyzed so far)
- Comprehensive and innovative service solutions available - single cell sequencing, patient derived xenograft (PDX) toolkits, etc.
- In-house cancer genome databases for cross-validation and auto-correction of genetic variants

Equipped with the industry's broadest array of cutting-edge technologies coupled with an experienced team of scientists and bioinformaticians, BGI now provides our innovative "One Stop, Total Solutions" service package, which includes **whole genome sequencing, whole exome sequencing, and RNA sequencing (RNA-Seq)**, to fully explore cancer research using an entire tool-kit of next-generation sequencing technologies. In this brochure, we will highlight a few of BGI Tech's unique and innovative techniques, which are designed to facilitate cancer research.

Research Goal	BGI's Solution
Investigate the mechanism of tumor heterogeneity and molecular evolution	Single-Cell Sequencing
Identify circulating blood biomarkers and analyze metastatic progression	Circulating Tumor Cell Sequencing
Investigate the genome (e.g., somatic single nucleotide variations) and the transcriptome of patient-derived xenografts	PDX Toolkits @ BGI
Analyze tumorigenesis caused by viral infection	Virus Integration Sequencing
Investigate gene variations, gene expression profiling changes, and protein biomarkers with FFPE tissues	FFPE Sequencing

## References

1. Jemal A., Bray F., *et al.* (2011) Global cancer statistics. *CA Cancer J. Clin.*, 61, 69-90.
2. Campbell P.J., Stephens P.J., *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.*, 40, 722-9.
3. Bentley D.R., Balasubramanian S., *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456, 53-9.
4. Meyerson M., Gabriel S. & Getz G. (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.*, 11, 685-96.

# Single-cell Sequencing

## Introduction

It is becoming increasingly apparent that seemingly homogeneous cell populations *in vivo* or cell cultures *in vitro* can exhibit considerable heterogeneity in expression patterns<sup>1</sup>, due to both intrinsic stochastic processes and extrinsic factors, such as the surrounding microenvironment. Cellular heterogeneity is frequently observed in many cancers as well. Previous approaches have focused on genomic differences in complex mixtures of cells, and these techniques may obscure the heterogeneity of single cells.

The next-gen sequencing based technologies are increasingly being targeted to individual cells, providing a means to address many longstanding questions. For example, single-cell genomics will facilitate elucidation of cell lineage relationships, and single-cell transcriptomics will supplant the imprecise notion of marker-based cell types<sup>2</sup>. Individual cells can easily be isolated using micromanipulation, such as with a simple mouth pipette<sup>3,4</sup>, or by serial dilution<sup>5,6</sup>. As micromanipulation methods are easy and inexpensive, they are the most commonly used single-cell isolation methodologies. In one recent study<sup>7</sup>, single-cell whole exome sequencing of a sample from a patient with myeloproliferative neoplasm was performed to reconstruct tumor ancestries and to identify candidate driver mutations. In another study, single-cell RNA sequencing revealed dynamic, random monoallelic gene expression in mammalian cells<sup>8</sup>. Furthermore, allelic expression analysis demonstrated the *de novo* inactivation of the paternal X chromosome<sup>8</sup>.

To fully explore its application in tumor heterogeneity<sup>9</sup> and microevolution<sup>10</sup>, in 2012 BGI developed an innovative method for genomics and transcriptomics analyses at the single-cell level, and to date, BGI has sequenced hundreds of cells with publications in top-tier journals<sup>7,11</sup>.

## Benefits

- BGI's rich experience in sequencing of single cells from various tumors, such as bladder cancer, essential thrombocythemia, colorectal cancer, and renal cancer
- Much lower quantity of input DNA required than traditional sequencing methods
- Uniform and unbiased whole genome amplification technology
- Longer length of amplified product (>10 kb), which is better for copy-number variation (CNV) and structural variation (SV) detection
- Cost-effective method for target region sequencing in pre-implantation genetic diagnosis

## BGI Solutions

### Single-cell DNA Sequencing

BGI provides researchers in academia and pharmaceutical companies with an innovative approach that utilizes whole-genome amplification (WGA) and next-generation sequencing to obtain sequence data at the single cell level. This novel methodology can be applied to discover genetic information in single cells and allows for the differentiation of those mutations that coincide with the development of cancerous cells and those that spur the cancer progression. BGI has sequenced hundreds of cells, and results from application of this novel method to identify the genetic characteristics of essential thrombocythemia and clear cell renal cell carcinoma have been published in *Cell*<sup>7,11</sup>.

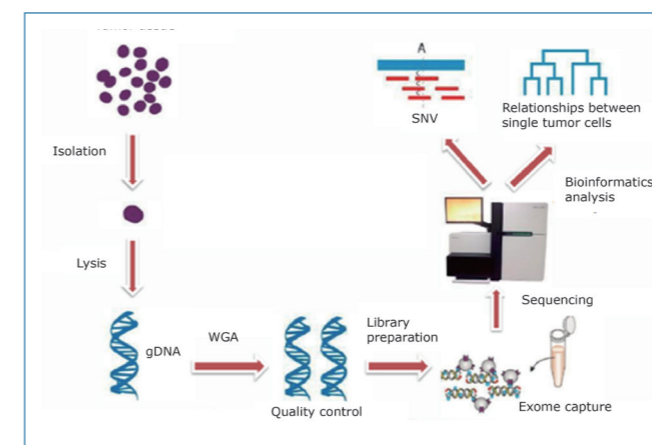


Figure 1. Workflow of single-cell exome sequencing. Single cells are isolated from tissues. Every single qualified cell is then lysed, and genomic DNA is amplified, followed by quality control using housekeeping genes as internal indicators. After library construction with exome capture and next-gen sequencing, SNPs of the coding regions and the relationships between single cells can be analyzed.

### Single-cell RNA Sequencing

The BGI Single-cell RNA sequencing system provides an end-to-end solution for strand-specific RNA-Seq library construction using as little as 10 picograms of total RNA or single-cell lysates. Customized bioinformatics analysis is available upon request. Single-cell RNA sequencing quantifies the mRNA of a single cell through single-tube reverse transcription and PCR amplification, allowing several micrograms of cDNA to be collected for traditional library construction and Hiseq2000 sequencing. This technique is useful in the study of many cancers, which often exhibit high cell heterogeneity across a single cancerous tissue.

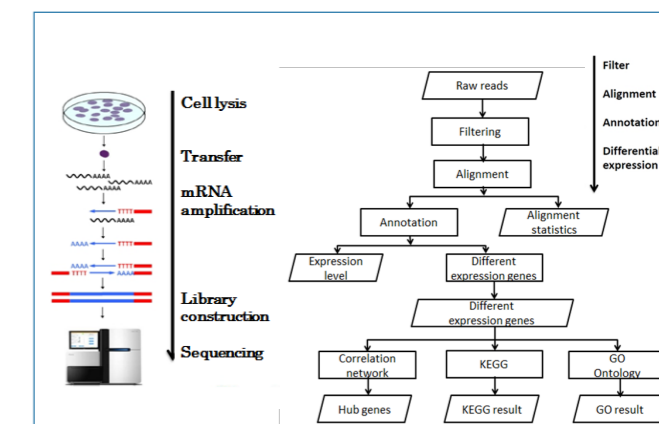


Figure 2. Workflow of single-cell RNA sequencing. Single cell isolation, cell lysis, and cDNA amplification are the key steps for library construction.

## BGI Cases

### Single-Cell Exome Sequencing and Monoclonal Evolution of a JAK2-Negative Myeloproliferative Neoplasm.

Yong Hou *et al.*, *Cell*. 148, 873-885 (2012)

Tumor heterogeneity presents a challenge for inferring clonal evolution and driver gene identification. Here, we described a method for analyzing the cancer genome at a single-cell nucleotide level.

- High-throughput whole-genome single-cell sequencing was conducted on single cells from two lymphoblastoid cell lines.
- Whole-exome single-cell sequencing was performed on 90 cells from a JAK2-negative myeloproliferative neoplasm patient.
- Essential thrombocythemia (ET)-related candidate mutations, including *SESN2* and *NTRK1*, were identified, suggesting that these factors may be involved in neoplasm progression.
- We established a single-cell sequencing method that paves the way for detailed analyses of a variety of tumor types, including those with high genetic variations between patients.

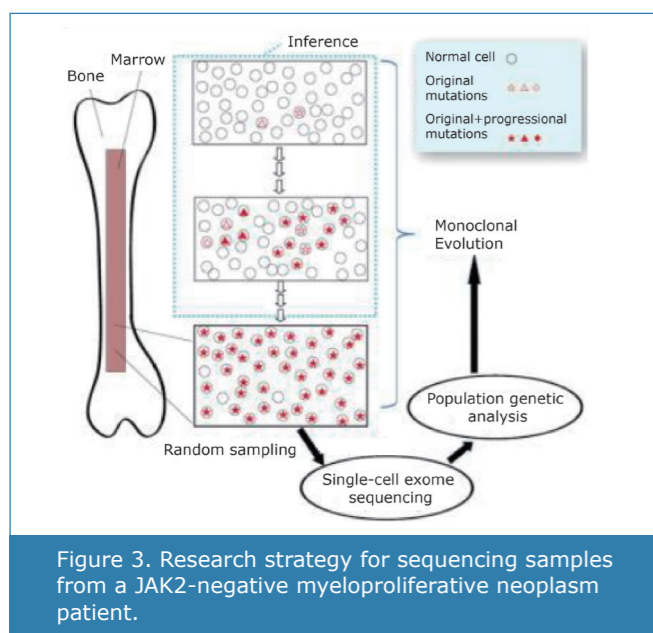


Figure 3. Research strategy for sequencing samples from a JAK2-negative myeloproliferative neoplasm patient.

Samples:  
- one *JAK2*-negative ET patient  
Strategy:

- 30X whole exome sequencing of 82 cells from fresh bone marrow and 8 cells from normal oral mucosal epithelium
- 91.12X whole exome sequencing of matched tissue samples from fresh bone marrow
- 57.62X whole exome sequencing of matched tissue samples from normal oral mucosal epithelium

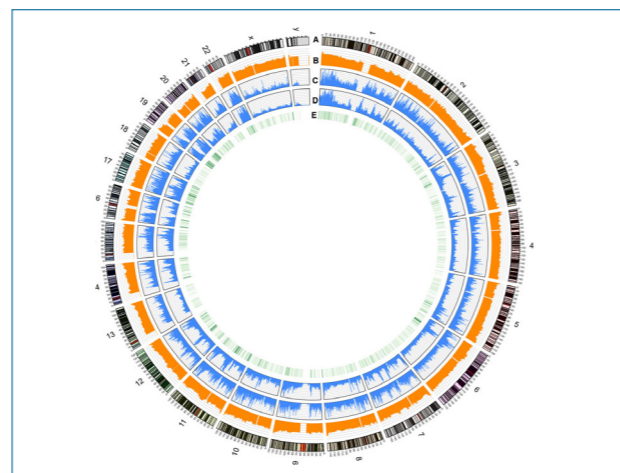


Figure 4 Graphic representation of the whole genome of two single YH cells  
(A) Karyotype of the human reference genome (Hg18). (B) GC content distribution of the reference genome (height of orange rectangles ranges from 0%–70%, bin = 1 Mb). (C) Whole-genome coverage of YH-2 (height of blue rectangles ranges from 03–403, bin = 1 Mb). (D) Whole-genome coverage of YH-1 (height of blue rectangles ranges from 03–403, bin = 1 Mb). (E) Gene density across the reference genome (Hg18). Gradually changing green represents 0 to 30 genes per 100 kb.

### Single-Cell Exome Sequencing Reveals Single-Nucleotide Mutation Characteristics of a Kidney Tumor

Xun Xu *et al.*, *Cell*. 148, 886–895 (2012)

Clear cell renal cell carcinoma (ccRCC) is the most common kidney cancer, and these tumors exhibit very few mutations that are shared between different patients. We performed an investigation of the intratumoral heterogeneity of ccRCC using single-cell sequencing.

- Single-cell exome sequencing was conducted on a ccRCC tumor and its adjacent kidney tissue, allowing us to delineate a detailed intratumoral genetic landscape at the single-cell level.
- Quantitative population genetic analysis indicated that the tumor did not contain any significant clonal subpopulations and that mutations that had different allele frequencies within the population also had different mutation spectrums.
- Our pilot study provides information that may lead to new ways to investigate individual tumors (e.g., exome sequencing) with the aim of developing more effective cellular targeted therapies.

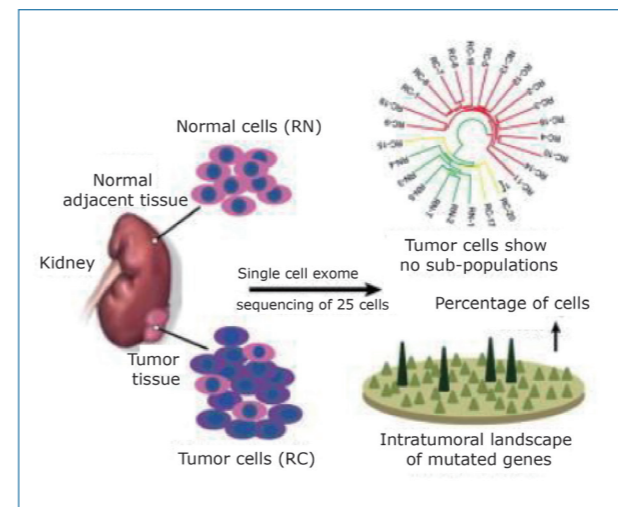


Figure 5. Research strategy for single-cell exome sequencing of a ccRCC tumor and its adjacent kidney tissue

Samples:

- One patient with ccRcc

Strategy:

- 30X whole-exome sequencing on 20 single cancer cells and 5 single normal cells
- 100X whole-exome sequencing on a mixture of cancer cells
- 30X whole-exome sequencing on a mixture of normal cells

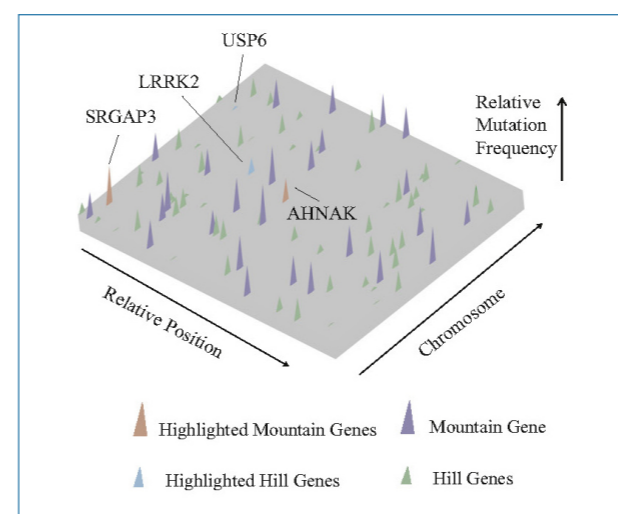


Figure 6. The intratumoral mutational landscape of the single ccRCC patient  
The landscape display shows a small number of mutant genes that are present in a large fraction of individual cells (which we term “mountains”) and a significantly greater number of genes mutant in only one or a few cells (“hills”). Moreover, the “mountains” and “hills” were evenly distributed across the patient genome and had no significant bias for any chromosome.

## References

1. Wilkinson D.J. *et al.* (2009) Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat Rev Genet.* 10: 122–133
2. Shapiro E., Biezuner T., Linnarsson S. (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet.* 14(9): 618–630
3. Zong C., Lu S., Chapman A. R. and Xie X.S. (2012) Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science.* 338: 1622–1626
4. Kurimoto K., Yabuta Y., Ohinata Y. and Saitou M. (2007) Global single-cell cDNA amplification to provide a template for representative high-density oligonucleotide microarray analysis. *Nature Protoc.* 2: 739–752
5. Reizel Y. *et al.* (2011) Colon stem cell and crypt dynamics exposed by cell lineage reconstruction. *PLoS Genet.* 7(7): e1002192
6. Shlush L. I. *et al.* (2012) Cell lineage analysis of acute leukemia relapse uncovers the role of replication-rate heterogeneity and microsatellite instability. *Blood.* 120: 603–612
7. Hou Y. *et al.* (2012) Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell.* 148: 873–885
8. Deng Q., Ramsköld, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science.* 343: 193-196
9. Dalerba P. *et al.* (2011) Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature Biotech.* 29: 1120–1127
10. Cristofanilli M. *et al.* (2005) Circulating tumor cells: a novel prognostic factor for newly diagnosed metastatic breast cancer. *J Clin Oncol.* 23: 1420–1430
11. Xu X., Hou Y. *et al.* (2012) Single-Cell Exome Sequencing Reveals Single-Nucleotide Mutation Characteristics of a Kidney Tumor. *Cell.* 148(5): 886-895

# Circulating Tumor Cell Sequencing

## Introduction

Circulating tumor cells (CTCs) represent rare cancer cells, which are derived from tumor tissues and found in the peripheral blood and lymphatic vessels<sup>1,2</sup>. CTCs have been proved to play a key role in cancer metastasis<sup>2,3</sup>. In addition, tumor related cell-free DNA, which circulates in the blood of cancer patients, is released by tumor cells in various forms and amounts. DNA can be shed as both single-stranded and double-stranded DNA. The release of DNA from tumor cells occurs through various physiological events such as apoptosis, necrosis and secretion<sup>3,4</sup>. Both CTCs and cell-free DNA can serve as biomarkers and are considered a real-time “liquid biopsy” in cancer patients<sup>4</sup>. These circulating blood biomarkers promise to become non-invasive real-time surrogates for tumor tissue-based biomarkers. They have been demonstrated to be effective tools for cancer diagnosis, therapeutic response monitoring, and prognostication<sup>4-7</sup>. In addition to clinical applications, analyzing the concentration, as well as genomic mutations, of CTCs and cell-free DNA provides unique insights into the biology of metastatic progression.

While promising, CTC and cell-free DNA genomics are still in its infancy, mainly due to a lack of technologies capable of isolating sufficient numbers of CTCs to analyze somatic mutations, as well as the lack of a suitable method to enrich the tumor DNA above the background of normal DNA<sup>1</sup>.

As pioneers in the technical development of single cell sequencing, BGI has extensively explored the application of this technology to cancer research. We have developed and provide a comprehensive solution to investigate genomic alterations in CTCs and cell-free DNA<sup>11</sup>.

## Benefits

- Innovative and efficient method to isolate CTCs from blood samples
- Comprehensive service from sample preparation to optimized bioinformatics analysis

## BGI Solutions

BGI provides a comprehensive solution to isolate CTCs and cell-free DNAs from blood samples, followed by next-generation sequencing (NGS; e.g., whole genome sequencing and whole exome sequencing) and optimized bioinformatics analysis (Figure 1, 2). This solution enables researchers to better identify heterogeneity in CTCs, as well as to select biomarkers for cancer diagnosis, monitoring and treatment.

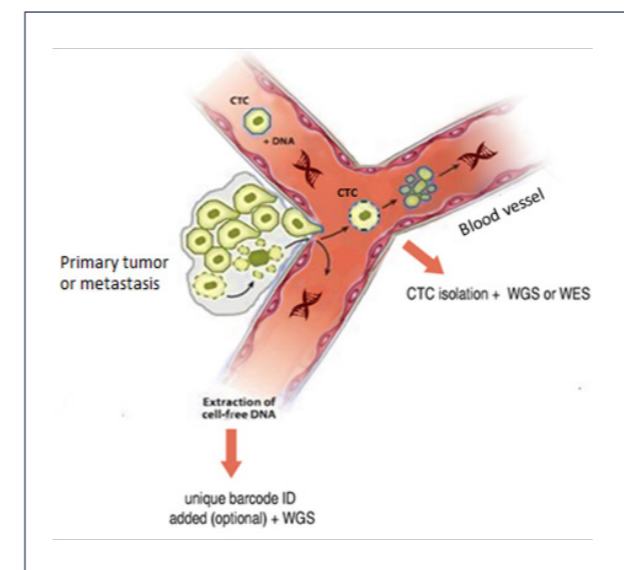


Figure 1. CTC and cell-free DNA sequencing solution at BGI

Cell-free tumor DNA circulates in the blood of cancer patients. To detect mutations in this DNA, a unique barcode ID (UID) is added to the end of cell-free DNA for more efficient identification (optional step), followed by library construction and sequencing (Figure 1).

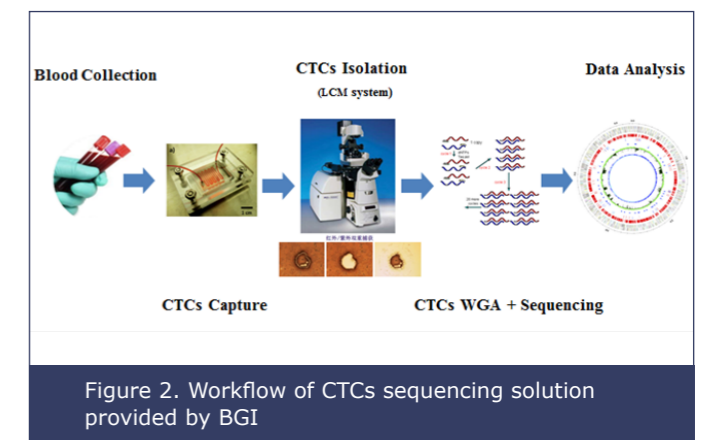


Figure 2. Workflow of CTCs sequencing solution provided by BGI

To fully explore the clinical application of CTC research, BGI provides an end-to-end solution from CTC isolation and sequencing to bioinformatics analysis\*. In particular, the CTCs are identified by immunocytochemistry (Figures 3-5) and collected via our CTC capture and isolation system, a modified NanoVelcro Chip coupled with ArcturusXT laser capture micro dissection technology which allows for isolation of CTCs well suited for NGS<sup>11</sup> (Figure 3). Next, whole genome amplification (WGA) is conducted on isolated CTCs, followed by whole genome sequencing or whole exome sequencing using the WGA material, and finally, bioinformatics analysis of the sequencing results (Figure 6).

\* Due to shipping regulation of human blood samples, we may only accept isolated CTCs from outside of China. Please do not hesitate to contact us for your specific case enquiry.

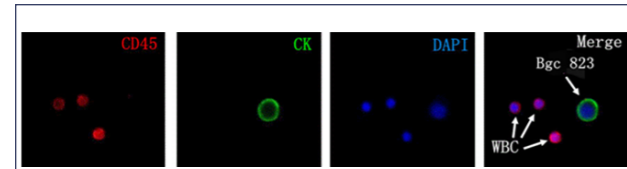


Figure 3. Fluorescent microscopy (40X) of the Bgc 823 cell line spiked in normal blood captured on polymer chip. Identification of CTCs was based on triple immunocytochemistry staining. CTCs are CK+, CD45-, and nucleated. Whole blood cells (WBCs) are CK-, CD45+, and nucleated.

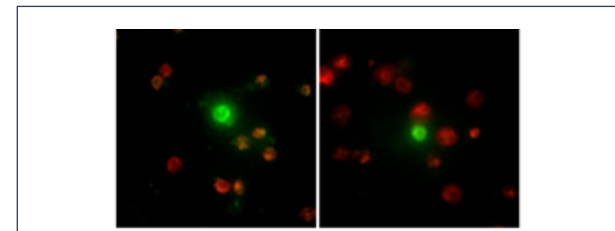


Figure 4. Identification of CTCs (green) from surrounding WBCs (red) in a prostate cancer patient sample.

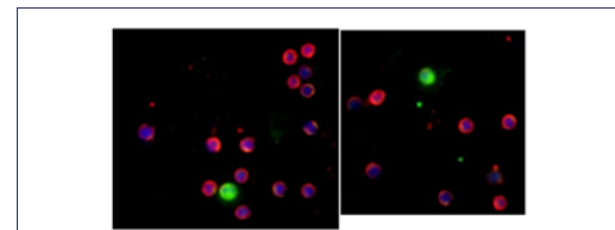


Figure 5. Identification of CTCs (green) from surrounding WBCs (red) in a breast cancer patient sample.

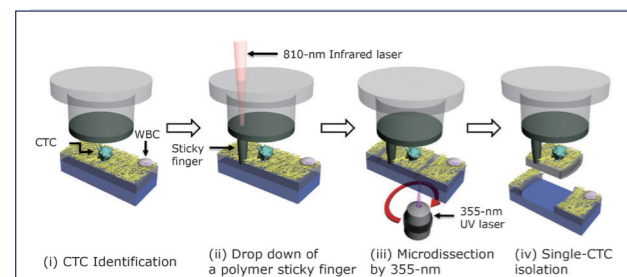


Figure 6. Isolation of CTCs for further NGS analysis from blood samples. (i) CTCs (CK+/CD45-) are first identified from surrounding WBCs (CK-/CD45+), and an laser capture microdissection (LCM) cap (with a polymer membrane on bottom) is then aligned over the CTC and substrates. (ii) An IR-laser is used to melt the polymer membrane, dropping down a "sticky finger" to facilitate adherence onto the PN-NanoVelcro substrate. (iii) A 355-nm UV laser is utilized to dissect the substrates underneath the identified CTC. (iv) By removing the LCM cap, high-purity CTCs can be obtained for subsequent molecular analysis.

## References

1. Plaks, V., C. D. Koopman, *et al.* (2013) Cancer. Circulating tumor cells. *Science*. 341(6151): 1186-1188.
2. De Mattos-Arruda, L., J. Cortes, *et al.* (2013). Circulating tumour cells and cell-free DNA as tools for managing breast cancer. *Nat. Rev. Clin. Oncol.* 10(7): 377-389.
3. Alix-Panabieres, C., H. Schwarzenbach, *et al.* (2012) Circulating tumor cells and circulating tumor DNA. *Annu. Rev. Med.* 63: 199-215.
4. Schwarzenbach, H., D. S. Hoon, *et al.* (2011) Cell-free nucleic acids as biomarkers in cancer patients. *Nat. Rev. Cancer* 11(6): 426-437.
5. Cristofanilli, M., G. T. Budd, *et al.* (2004) Circulating tumor cells, disease progression, and survival in metastatic breast cancer. *N. Engl. J. Med.* 351(8): 781-791.
6. Dawson, S. J., D. W. Tsui, *et al.* (2013) Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N. Engl. J. Med.* 368(13): 1199-1209.
7. Punnoose, E. A., S. Atwal, *et al.* (2012) Evaluation of circulating tumor cells and circulating tumor DNA in non-small cell lung cancer: association with clinical endpoints in a phase II clinical trial of pertuzumab and erlotinib. *Clin. Cancer Res.* 18(8): 2391-2401.
8. Ni, X., M. Zhuo, *et al.* (2013) Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. *Proc. Natl. Acad. Sci.* 110(52): 21083-21088.
9. Swanton, C. (2013) Plasma-derived tumor DNA analysis at whole-genome resolution. *Clin. Chem.* 59(1): 6-8.
10. Lianidou, E. S. and A. Markou (2011) Circulating tumor cells in breast cancer: detection systems, molecular characterization, and future challenges. *Clin. Chem.* 57(9): 1242-1255.
11. Zhao, L., Y. T. Lu, *et al.* (2013) High-Purity Prostate Circulating Tumor Cell Isolation by a Polymer Nanofiber-Embedded Microchip for Whole Exome Sequencing. *Adv. Mater.* DOI:10.1002

# PDX Toolkits @ BGI

## Introduction

To better reflect human disease pathology in mouse models, patient-derived xenografts (PDX) have been widely used to evaluate new anti-cancer drugs for potential development in human clinical trials<sup>1,2</sup>. Tumors obtained in the clinic can be maintained by serial xenografting in athymic (nude) or severe combined immunodeficient (SCID) mice<sup>3</sup>. One limitation of this approach is that the mouse genome is almost 90% homologous to the human genome<sup>4</sup>, leading to possible contamination and complicating downstream bioinformatics analyses. The application of xenograft technology is further impeded by several technical issues, such as a lack of qualified paired controls for accurate variant profiling and a mixture of genetic factors from the host in the xenograft<sup>5</sup>. To address these issues, BGI developed comprehensive patient-derived xenograft toolkit sets ("PDX Toolkits") with a modular design comprised of tools encompassing all functions for HiSeq data from basic mapping to variant recalibration and annotation. These modules enable the prediction of somatic mutations without requiring normal tissues, provide a more efficient method for eliminating mouse genome contamination, and enhance the validation of genetic variants using our comprehensive cancer genome databases.

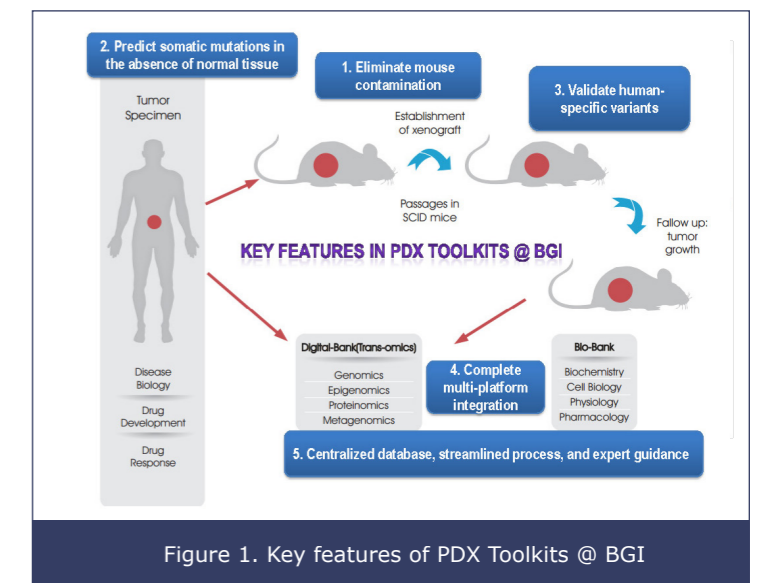


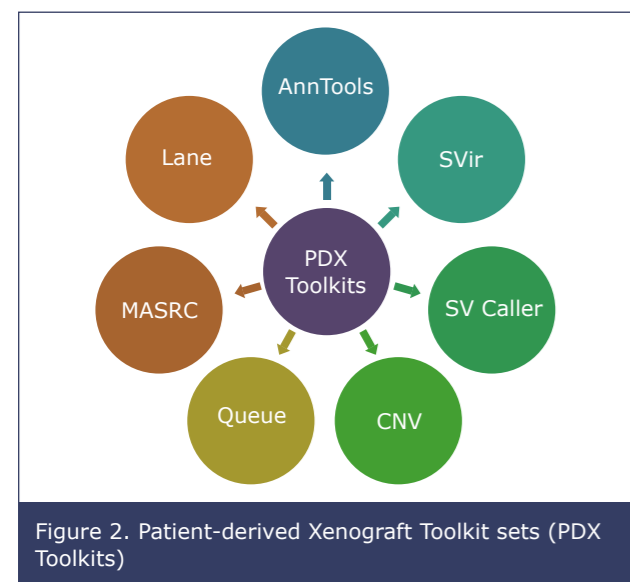
Figure 1. Key features of PDX Toolkits @ BGI

## Benefits

- **Novel and efficient** algorithm (**PDXomics**) to filter out mouse genome contaminants and acquire a highly accurate variant set of PDX models
- **First-in-kind** solution (**PDXsnv**) to identify germline mutations and predict somatic single nucleotide variations (SNVs) in the absence of normal tissues
- **Comprehensive** cancer genome database (18,406 human cancer samples sequenced at BGI) for cross-validation and auto-correction of genetic variants
- **Robust** bioinformatics pipeline to detect SNP, Indel, and CNV calling with high accuracy
- **Integrative** methods (four reliable bioinformatics tools available) to identify structural variations (SV).
- **Cost-effective and rapid** validation by incorporation of in situ and RNA-Seq validation into the pipeline

# BGI Solutions

PDX Toolkit is a software package developed at BGI to reveal intrinsic mechanisms and features of PDX models systematically and comprehensively (Figure 2), which facilitates translational research and drug discovery.



The toolkit offers a wide variety of tools (modules), including a basic mapping and removal of mouse contamination module, a statistics module at the sample level, a primary variant discovery and genotyping module, as well as powerful processing variant recalibrating and annotating modules (Table 1).

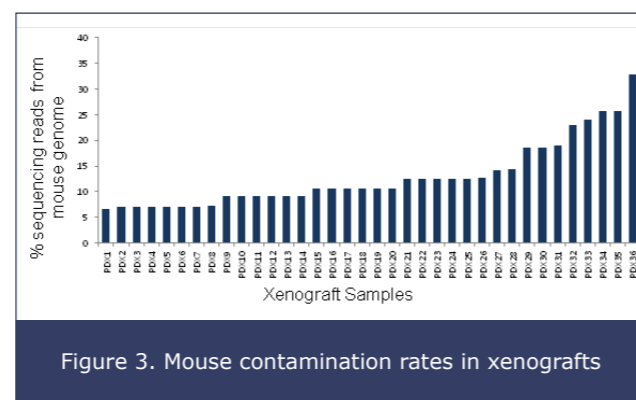
Table 1. Modules and their functions in PDX Toolkits

Module(s)	Function(s)
[Lane]	Distinguish human reads from mouse reads with very high accuracy for downstream analysis using BGI's self-developed PDXomics algorithms
[MARSC]	Merge files as defined by users Recalibrate variants with Gaussian error model
[Queue] & [CNV]	Identify SNPs, SNVs, Indels, and CNVs in PDX samples without requiring normal matched controls using BGI's self-developed PDXsnv algorithms
[SVcaller]	Identify SVs from sequencing data using a combination of six methods of analysis (e.g., BreakDancerMax, Crest, Pindel)
[SVir]	Integrate all SV calling results from SVcaller and locate accurate breakpoints
[AnnTools]	Annotate variants from above modules and infer mechanisms for the involvement of SVs

# Method Validation

## 1. Filter out mouse contaminants with PDXomics@BGI

Based on our data from PDXomics@BGI, anywhere from 5%–33% of the sequencing reads from xenograft samples are actually contaminants from the mouse genome sequence. The amount of contamination varied between different models, different vendors, and in various cancer types. We found that there is an obvious concordance between DNA and RNA data.



## 2. Somatic SNV prediction in the absence of normal control tissue

Using our PDXsnv algorithm, the number of somatic SNVs decreased greatly from more than 3,500,000 to less than 20,000. Moreover, PDXsnv predicts the somatic SNVs of major cancer types with at least 75% sensitivity in the absence of corresponding normal controls, covers the known driver and suppressor genes, and detects novel SNVs.

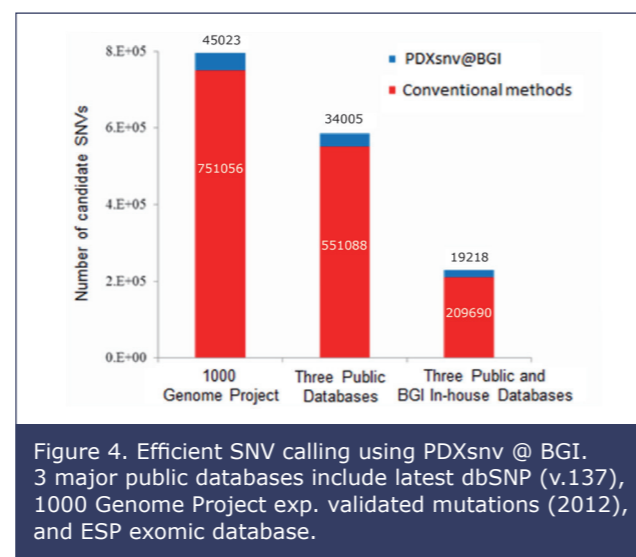


Figure 4. Efficient SNV calling using PDXsnv @ BGI. 3 major public databases include latest dbSNP (v.137), 1000 Genome Project exp. validated mutations (2012), and ESP exomic database.

Figure 4 shows that PDXsnv@BGI significantly reduces the number of candidate somatic SNVs.

- Our proprietary in-house database reduces the number to 209,690.
- Our unique machine learning algorithm further decreases the number to 19,218.

## 3. PDX genomic data is concordant with clinical samples

Somatic mutations of seven pairs of primary tumors and their corresponding xenograft samples are identified in the absence of normal control samples at an accuracy of >80% when analyzing a panel of cancer associated genes. PDXs are highly consistent with primary tumor samples in the variation patterns of cancer associated genes (Figure 5A). A more detailed analysis of one primary tumor-xenograft pair shows highly concordant gene expression (Figure 5B).

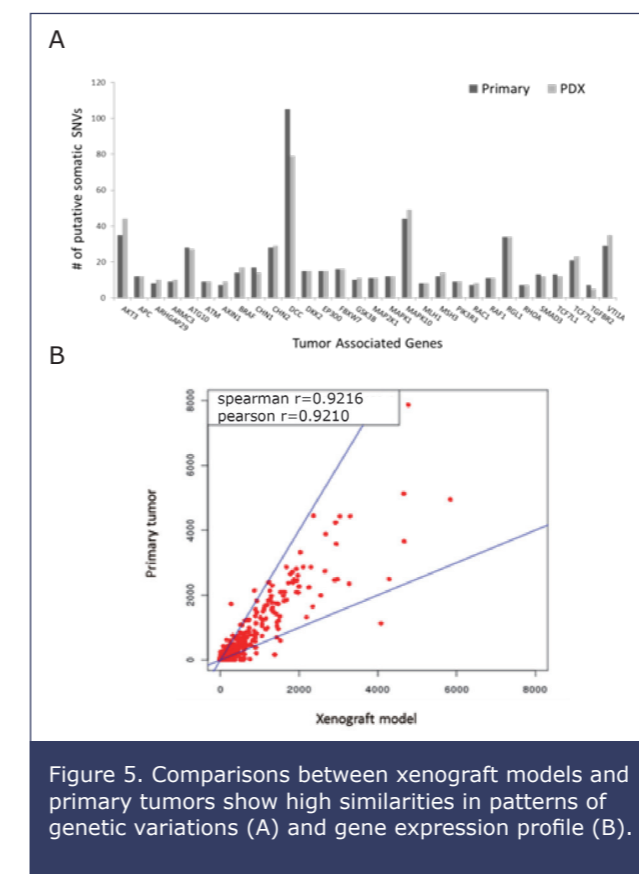


Figure 5. Comparisons between xenograft models and primary tumors show high similarities in patterns of genetic variations (A) and gene expression profile (B).

## 4. Conclusions

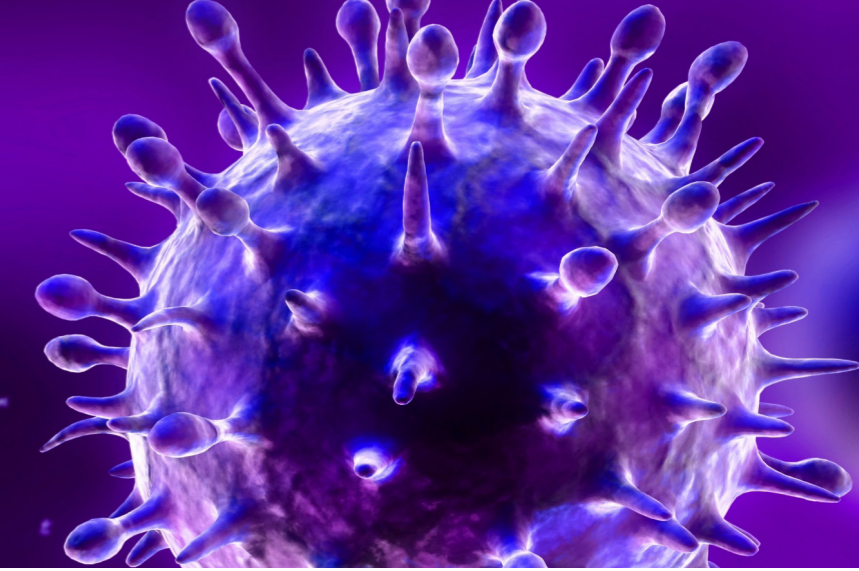
While patient-derived xenografts are suitable for assessing potential anti-cancer therapeutics, limitations of this technology have not allowed this potential to be realized. Our innovative PDX Toolkits overcome these limitations using a bioinformatics approach. This software can successfully recognize and filter out sequencing reads from the mouse genome (mouse sequence contaminates PDX sequencing data at a rate of 5%–33%). Our innovative PDXsnv algorithm and proprietary cancer database can significantly narrow down the number of candidate somatic SNVs to less than 20,000 (Fig. 4), even in the absence of corresponding normal controls.

## References

1. Tentler, J.J., *et al.*, (2012) Patient-derived tumour xenografts as models for oncology drug development. **Nat. Rev. Clin. Oncol.**, 9(6): p. 338-350.
2. Kopetz, S., R. Lemos, and G. Powis, (2012) The promise of patient-derived xenografts: the best laid plans of mice and men. **Clinical Cancer Research**, 18(19): p. 5160-5162.
3. Morton, C.L. and P.J. Houghton, (2007) Establishment of human tumor xenografts in immunodeficient mice. **Nat. Protoc.**, 2(2): p. 247-50.
4. Chinwalla, A.T., *et al.*, (2002) Initial sequencing and comparative analysis of the mouse genome. **Nature**, 420(6915): p. 520-562.
5. Greenman, C., *et al.*, (2007) Patterns of somatic mutation in human cancer genomes. **Nature**, 446(7132): p. 153-158.



# Virus Integration Sequencing



## Introduction

A century of tumor virology has revealed that seven types of viruses cause 10–15% of all human malignancies<sup>1</sup>. Viruses can cause cellular transformation by expression of viral oncogenes, genomic integration to alter the activity of cellular proto-oncogenes or tumor suppressors, or inducing inflammation that promotes oncogenesis.

Viral etiology is particularly evident in cervical carcinoma and ovarian cancer, which are almost exclusively caused by high-risk human papillomaviruses (HPV)<sup>2,3</sup>, and in hepatocellular carcinoma, where infection with hepatitis B virus (HBV) or hepatitis C virus (HCV) is the predominant cause in some countries<sup>4</sup>. In addition, several rare cancers have a strong viral component, including Epstein–Barr virus (EBV), which causes most Burkitt’s lymphomas<sup>5</sup>. Therefore, it is of high importance to clarify the mechanism of tumorigenesis and development relative to viral infection by studying the relationship between viral integration and cancer formation in the host.

Traditional research methods include chromosome walking PCR, qPCR, and FISH. However, these methods are tedious, low throughput, and imprecise

with regards to location and copy number, which greatly limit development of the field. To alleviate these problems, new techniques, such as whole genome sequencing (WGS), have been developed to study virus integration. WGS has single base resolution, thus all integration sites could be detected accurately in a single experiment. Unfortunately, this technology has been cost prohibitive thus far. To resolve this issue, BGI has developed an HBV Integration capture sequencing technique that captures probes based on viral sequence. This technique can comprehensively, accurately, and cost-effectively identify virus integration sites and virus type in virus-related tumors.

## Benefits

- Novel and more accurate method to capture probes based on virus sequence
- Comprehensive investigation of virus integration sites in a cost-effective way

## BGI Solutions

BGI provides comprehensive solutions to identify virus integration sites in viral-related tumors using whole genome sequencing and target region sequencing, which will lead to an understanding of the mechanism of virus-induced tumorigenesis and development. The workflow is as follows:

Scientific problem	• To study virus integration sites in virus-related tumor
Design a plan	• Sequence DNA from virus positive and negative tumor patients • Virus integration breakpoint analysis
Sample collection	• Virus positive and negative tumor samples • Whole genome sequencing/target region sequencing
Bioinformatics analysis	• Develop virus integration analysis pipeline • Identify integration events • Analyse integration events in tumors & adjacent tissues
Technical validation	• RNA-Seq and Sanger sequencing for validation • Identify recurrent integration events in multiple genes

Figure 1. Workflow of virus integration sequencing @ BGI

Moreover, to efficiently enrich viral DNA for sequencing, BGI has developed a virus capture array using Agilent SureSelect target enrichment. We have unique virus capture chips for four of the most relevant viruses (HBV, HPV, EBV, and HIV), and we have the most experience and the most sophisticated process for HBV capture. To date, more than 1,000 liver cancer samples have been analyzed by HBV integration research at BGI.

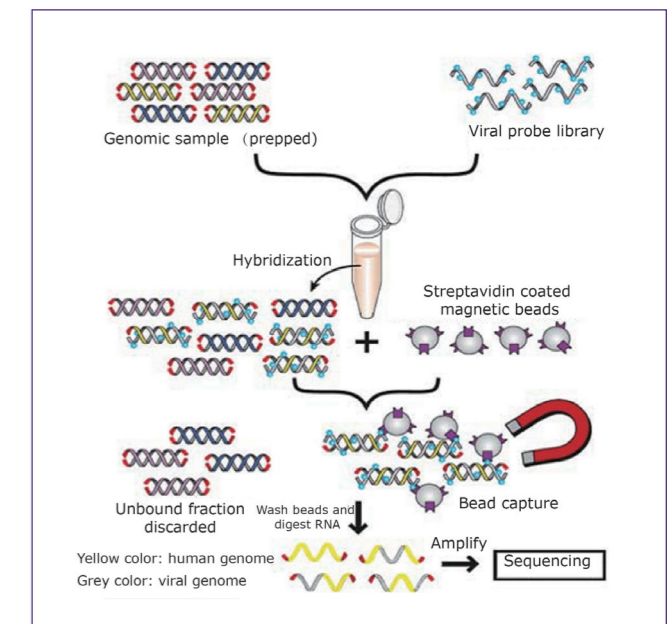


Figure 2. Virus capture array developed by BGI. (Courtesy of Agilent Technologies, Inc.)

Virus integration capture sequencing was highly concordant with whole genome sequencing (WGS) and could detect some low frequency integration, which was not detected by WGS. For example, when these two methods were conducted on the same lung cancer sample, WGS detected 63 breakpoints, while virus integration capture sequencing identified an additional 41 new breakpoints as well as those 63 ones (our internal data). This indicates that the sensitivity of virus integration capture sequencing is even higher than the sensitivity of WGS.

## BGI CASE

Genome-wide Survey of Recurrent HBV Integration in Hepatocellular Carcinoma<sup>6</sup>  
Wing-Kin Sung, Hancheng Zheng et al. *Nature Genetics*. 44, 765–769 (2012)

Hepatocellular carcinoma (HCC) is a common solid tumor and represents the third leading cause of cancer deaths worldwide. Thus far, three mechanisms have been reported to explain how hepatitis B virus (HBV) promotes carcinogenesis. In this study, we investigated the events of HBV integration and their effects on the HCC genome using whole-genome sequencing (~30X depth, coverage >99%) and integrated expression profiling analyses. Ultimately, 339 HBV integration breakpoints were discovered using the workflow outlined below (Figure 3A).

- Parallel sequencing was conducted on 81 HBV-positive and 7 HBV-negative hepatocellular carcinomas (HCCs) and adjacent normal tissues.
- HBV integration was observed more frequently in the tumors (86.4%) than in adjacent liver tissues (30.7%).
- Copy-number variations (CNVs) were significantly increased at HBV breakpoint locations where chromosomal instability was likely induced.
- With validation using RNA-Seq and Sanger sequencing, recurrent HBV integration events (i.e., in  $\geq 4$  HCCs) were identified at several genes that have been linked to cancer, including TERT, MLL4 and CCNE1 (Figure 3B and 3C), which also showed upregulated gene expression in tumors compared to normal tissue (Figure 3D).

## References

1. Moore, P.S. and Y. Chang, (2010) Why do viruses cause cancer? Highlights of the first century of human tumour virology. *Nat. Rev. Cancer*, 10, 878-889.
2. Ojesina, A.I., et al., (2013) Landscape of genomic alterations in cervical carcinomas. *Nature*, doi:10.1038/nature12881.
3. Al-Shabanah, O.A., et al., (2013) Human papillomavirus genotyping and integration in ovarian cancer Saudi patients. *Viol. J.*, 10:343.
4. Williams, R., (2006) Global challenges in liver disease. *Hepatology*, 44(3):521-6.
5. Schmitz, R., et al., (2012) Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nature*, 4; 490(7418):116-20.
6. Sung, W.K., et al., (2012) Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat. Genet.*, 44 (765-769)

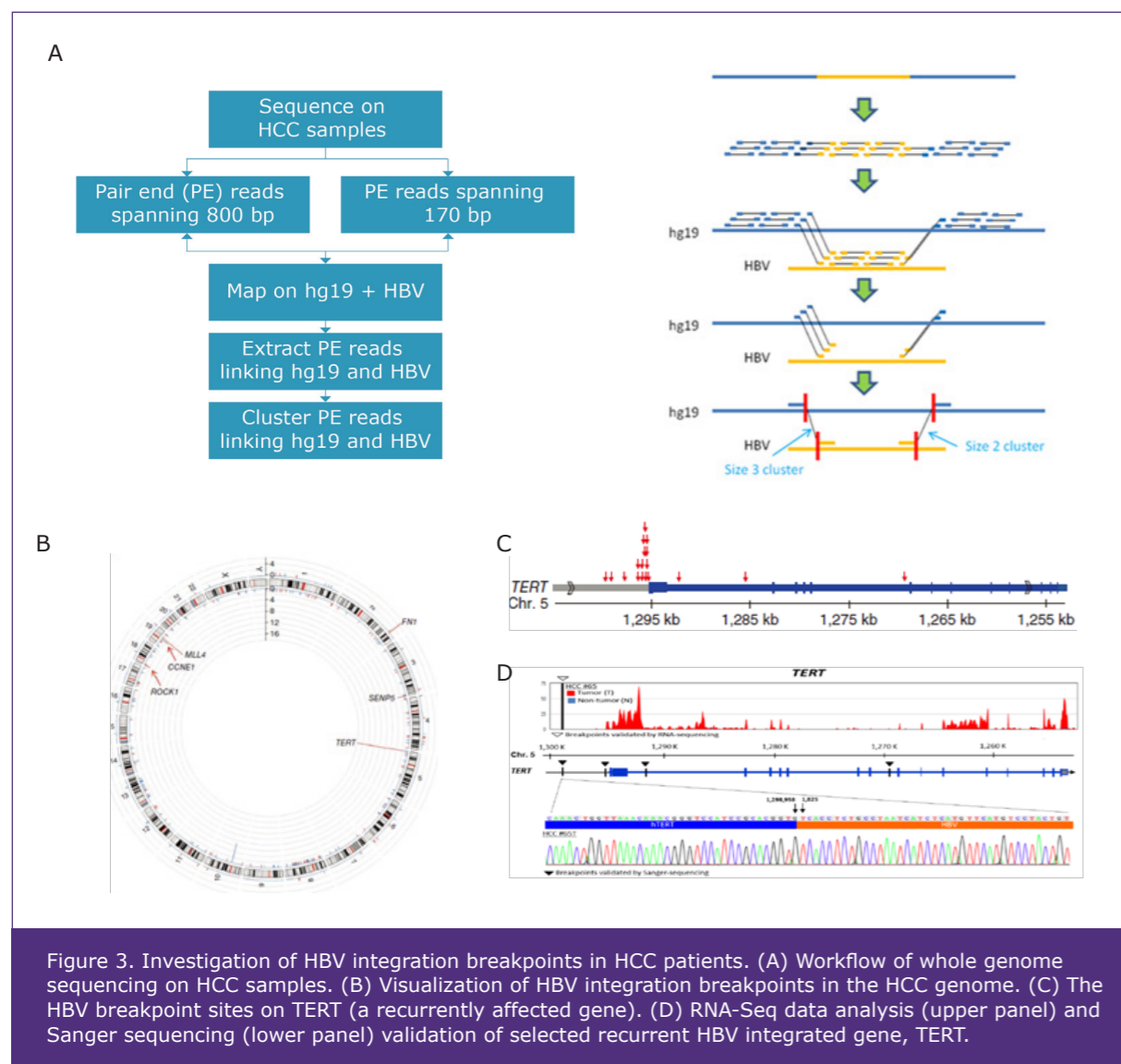


Figure 3. Investigation of HBV integration breakpoints in HCC patients. (A) Workflow of whole genome sequencing on HCC samples. (B) Visualization of HBV integration breakpoints in the HCC genome. (C) The HBV breakpoint sites on TERT (a recurrently affected gene). (D) RNA-Seq data analysis (upper panel) and Sanger sequencing (lower panel) validation of selected recurrent HBV integrated gene, TERT.

# FFPE Sequencing

## Introduction

Most of the clinical tumor samples available for molecular studies are formalin-fixed, paraffin-embedded (FFPE) specimens stored in tissue banks and biorepositories worldwide<sup>1,2</sup>. These samples are often very well-characterized with histological, pathological and follow-up clinical data. Moreover, FFPE has been the standard clinical sample preparation technique for pathology; thus, performing parallel sequencing of the large numbers of FFPE samples would connect multiple existing clinical trials and transform genomic analysis into standard clinical practice<sup>3,4</sup>. Unfortunately, challenges such as severe degradation, damage and molecular modifications make the recovery of nucleic acids from these tissue specimens difficult. In addition, formalin-induced cross-linking of DNA, RNA, and proteins, as well as the presence of fragmented and insufficient nucleic acids, restricts the ability of researchers to comprehensively sequence FFPE samples<sup>4-7</sup>.

To address these issues and take advantage of preserved tissues, we established a standard protocol for sequencing FFPE samples, including whole exome sequencing (WES), transcriptome sequencing, and small RNA sequencing.

## Benefits

- Much less DNA and RNA required: As low as 200 ng for WGS, WES, and RNA-Seq.
- High throughput: Millions of reads available for downstream analysis.
- Comprehensive: Integrated analysis of DNA and RNA data on FFPE samples.

## BGI Solutions

BGI provides a comprehensive solution to apply WES, transcriptome sequencing, and small RNA sequencing to FFPE samples, which provides an opportunity to link molecular biology research to disease, diagnosis and biomarker discovery.

In particular, for transcriptome sequencing in FFPE samples, BGI employed two efficient methods, duplex-specific nuclease (DSN) normalization technology and Ribo-Zero™ rRNA Removal Kits (Human/Mouse/Rat), to remove ribosomal RNA (more than 97% of the transcripts output in eukaryotes) without severe degradation of nucleic acids, followed by library construction and high-throughput sequencing.

## Method Validation

The following demonstrates the feasibility of applying WES to FFPE samples based on BGI internal data.

### 1. Similar sequencing coverage

50 FFPE samples and fresh frozen (FF) samples were analyzed by whole exome sequencing at 100X coverage captured by Agilent SureSelect Human All Exon 51M. The degree of coverage was roughly the same between FFPE and FF samples (Table 1).

Table 1. Coverage statistics of FF and FFPE samples

Sample	Clean Data	Map rate	Coverage of target	>=20X	>=10X	>=4X
FF	14G	0.99	0.99	0.97	0.99	0.99
FFPE	14G	0.98	0.99	0.96	0.98	0.99

Note: >=20X denotes that the proportion of target region covered with >=20 reads

### 2. Overlapping mutation profiles

We conducted whole exome sequencing on two pairs of tumor tissues and normal controls with two different storage conditions, FFPE and FF. The map rate and coverage performance are similar (data not shown). To investigate whether information obtained from FFPE samples was reliable, we compared the concordance of single nucleotide polymorphisms (SNPs) between these two pairs of FF and FFPE samples. The genome variation pattern obtained from FFPE samples is comparable to that from FF samples (Figure 1).

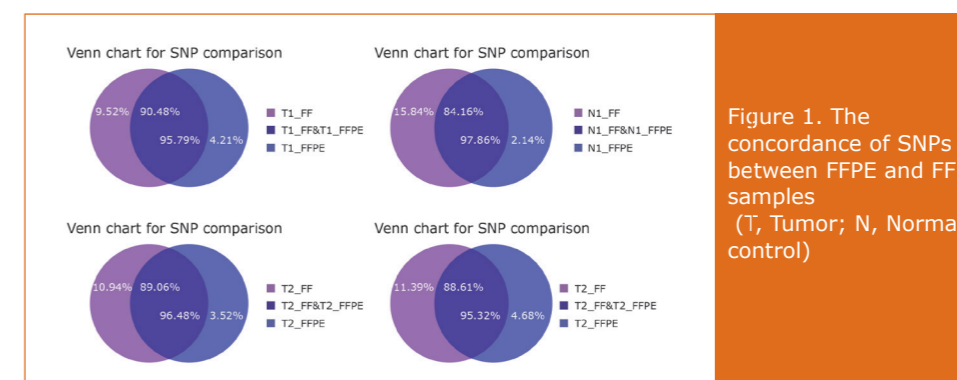


Figure 1. The concordance of SNPs between FFPE and FF samples (T, Tumor; N, Normal control)

### 3. Reliable results (SNP detection)

Most of the SNPs analyzed were detected in the same positions. To explore the accuracy of SNP analysis from FFPE samples, we compared the SNPs from FFPE and FF samples with dbSNP and found that most SNPs are consistent with dbSNP, indicating that the SNPs called from FFPE samples were as accurate as those from FF samples.


Table 2 Concordance of SNPs from FFPE and FF samples with dbSNP

Sample Name	Overlap with dbSNP (FF samples)	Overlap with dbSNP (FFPE samples)
T3_FF vs. T3_FFPE	97.84%	97.58%
N3_FF vs. N3_FFPE	97.82%	98.32%
T4_FF vs. T4_FFPE	97.33%	97.51%
N4_FF vs. N4_FFPE	97.89%	97.83%

Note: T, Tumor; N, Normal control

## References

1. Becker, K. F., Schott, *et al.*, (2007) Quantitative protein analysis from formalin-fixed tissues: implications for translational clinical research and nanoscale molecular diagnosis. *J Pathol*, 211, 370-8.
2. Berg, D., Malinowsky *et al.*, (2011) Use of formalin-fixed and paraffin-embedded tissues for diagnosis and therapy in routine clinical settings. *Methods Mol Biol*, 785, 109-22.
3. Lim, M. S. & Elenitoba-Johnson, K. S. (2004) Proteomics in pathology research. *Lab Invest*, 84, 1227-44.
4. Corless, C. L. and Spellman, P. T. (2012) Tackling formalin-fixed, paraffin-embedded tumor tissue with next-generation sequencing, *Cancer Discov*, 2 (1), 23-4.
5. Schweiger, M. R., *et al.*, (2009) Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor tissues for copy-number- and mutation-analysis, *PLoS One*, 4 (5), e5548.
6. Wagle, N., *et al.*, (2012) High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing, *Cancer Discov*, 2 (1), 82-93.
7. Yost, S. E., *et al.*, (2012) Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens, *Nucleic Acids Res*, 40 (14), e107.
8. Marie-Christine Maurel., *et al.*, (2005) The RNA world: Hypotheses, facts and experimental results. *Advances in Astrobiology and Biogeophysics*, pp 571-594



NGS and related technologies provide great promise for a sophisticated understanding of the cellular mechanisms key to human health and diseases. This diverse palette of analytical technologies allows us to understand the path from genes to phenotype with richer detail than ever before, which is crucial for cancer research because cancer is principally a disease of accumulation of genome alterations. To fully explore the application of NGS in cancer research, BGI offers our innovative "One Stop, Total Solutions" service package to help you improve the understanding of the mechanisms underlying various types of cancer (e.g., single-cell sequencing) as well as accelerate translational research and drug discovery (e.g., PDX Toolkits). With our rich experience in cancer sequencing and bioinformatics expertise, we look forward to working with you to answer your most pressing research questions.

## **China**

BGI Shenzhen (HQ)  
Beishan Industrial Zone, Yantian District, Shenzhen, 518083, China  
Tel: 400-706-6615 (within China); +86-755-25273045 (international)  
Email: info@bgitecholutions.com

## **North and South America**

BGI Americas Corporation  
One Broadway, 3rd Floor, Cambridge, MA 02142 U.S.A.  
Tel: +1-617-500-2741  
Fax: +1-617-500-2742  
Email: info@bgiamericas.com

## **Europe**

BGI Europe  
Copenhagen Bio Science Park, Ole Maaloes Vej 3, 2200 Copenhagen,  
Denmark  
Tel: +45-7026-0806  
Email: bgieurope@genomics.cn

## **Hong Kong**

BGI Hong Kong  
Dai Fu Street, Tai Po Industrial Estate, Tai Po, New Territories, Hong Kong  
Email: bgihk-marketing@genomics.cn  
Tel: +852-3610-3524

## **Japan**

BGI Japan  
Kobe KIMEC Center BLDG. 8F 1-5-2 Minatojima-minamimachi, chuo-ku,  
Kobe City, Hyogo-pref. 650-0047, JAPAN  
Tel: 078-599-6108  
Fax: 078-599-6109  
Email: bgijapan@genomics.cn

## **Asia Pacific / Oceania**

BGI Asia Pacific  
Main Building 2nd Floor, Beishan Industrial Zone, Yantian District,  
Shenzhen, 518083, China  
Tel: +86-755-25273120  
Fax: +86-755-25011756  
Email: p\_info\_asiapacific@genomics.cn