Oncogene (2015) **34,** 5943–5950

www.nature.com/onc

REVIEW

Cancer whole-genome sequencing: present and future

H Nakagawa, CP Wardell, M Furuta, H Taniguchi and A Fujimoto

Recent explosive advances in next-generation sequencing technology and computational approaches to massive data enable us to analyze a number of cancer genome profiles by whole-genome sequencing (WGS). To explore cancer genomic alterations and their diversity comprehensively, global and local cancer genome-sequencing projects, including ICGC and TCGA, have been analyzing many types of cancer genomes mainly by exome sequencing. However, there is limited information on somatic mutations in non-coding regions including untranslated regions, introns, regulatory elements and non-coding RNAs, and rearrangements, sometimes producing fusion genes, and pathogen detection in cancer genomes remain widely unexplored. WGS approaches can detect these unexplored mutations, as well as coding mutations and somatic copy number alterations, and help us to better understand the whole landscape of cancer genomes and elucidate functions of these unexplored genomic regions. Analysis of cancer genomes using the present WGS platforms is still primitive and there are substantial improvements to be made in sequencing technologies, informatics and computer resources. Taking account of the extreme diversity of cancer genomes and phenotype, it is also required to analyze much more WGS data and integrate these with multi-omics data, functional data and clinical-pathological data in a large number of sample sets to interpret them more fully and efficiently.

Oncogene (2015) **34,** 5943–5950; doi:10.1038/onc.2015.90; published online 30 March 2015

INTRODUCTION

Cancer is essentially a 'disease of the genome' with an evolutional process, which develops and evolves with accumulations of diverse types of somatic mutations, copy number alterations (SCNAs) and structural variants (SVs), with and without a background of heritable factors (germline variants) and epigenetic factors.[1–3] The germline study on familial cancer cases and gross copy number analysis on cancer tissues identified[4] germline and somatic mutations of many classical tumor suppressor genes (*RB1*, *TP53* and *APC*),[2,3] and copy number analysis found some oncogenes and underlying oncogenic activators such as *HER2/ERBB2*.[2,5] Some of these recurrent driver mutations have since been targeted for cancer therapy and genomic mutational profiling on single or multiple genes is being used for clinical diagnosis, drug sensitivity, residual disease detection and prognosis prediction.[2,3]

Recent explosive advances of next-generation sequencing (NGS) technology and bioinformatics/computational approaches to massive data enable us to comprehensively analyze a number of cancer genome profiles by targeted sequencing, whole-exome sequencing (WES), RNA sequencing (RNA-Seq) and whole-genome sequencing (WGS).[2,3,6–8] So far, many types of cancer genomes have been sequenced and analyzed worldwide, including large projects such as TCGA (The Cancer Genome Atlas)[3,9] and ICGC (The International Cancer Genome Consortium),[10] to explore cancer genomic alterations and their diversity comprehensively and intensively. In these global and local projects, WES is the main platform for cancer genome sequencing and vast amounts of mutational data in protein-coding genes or regions for many types of both common and rare tumors have been accumulated. These systematic studies of the cancer genome have reveled scores of new cancer genes, pathways and mehanisms,[3] and saturation analysis suggests that most driver genes that are frequently mutated in cancer have been almost elucidated.[11,12] Now, it is important to focus on the 'long tail' of rare mutated driver candidates,[3,13] in addition to validating these using functional studies.[11]

On the other hand, there is limited information on somatic mutations in non-coding regions,[3] including UTRs (untranslated regions), introns, regulatory elements and diverse non-coding RNAs (ncRNAs). Genomic SVs including large deletion/insertion, inversion, duplication, translocation and pathogen integration in cancer genomes remain widely unexplored[3] (Figure 1). Different from WES, WGS approaches can detect these unexplored mutations and help us better understand the whole landscape of cancer genome and elucidate the functions of these unexplored human genomic regions, and it can clarify the underlying carcinogenesis and achieve molecular sub-classification of cancer, which facilitates discovery of genomic biomarkers and personalized cancer medicine. This review describes the recent approaches in cancer genome analysis based on WGS and the future direction of cancer WGS and its use as an analysis platform for cancer genomics.

TARGET RE-SEQUENCING AND COPY NUMBER ANALYSIS

Sanger sequencing following PCR was an initial approach to discover mutations in cancer genomes. Although this method can analyze limited numbers of genes, it has been applied to characterize hundreds of samples and most protein-coding exons. Vogelstein and his group have extended these approaches to discover somatic mutations by screening thousands of protein-coding exons in the human genome and succeeded in identifying many important driver genes and their mutations. In 2003, they sequenced 819 exons of tyrosine kinase genes and all exons of 16 phosphoinositide 3-kinases genes, and identified recurrent

Laboratory for Genome Sequencing Analysis, RIKEN Center for Integrative Medical Sciences, Tokyo, Japan. Correspondence: Dr H Nakagawa, RIKEN Center for Integrative Medical Sciences, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan.
E-mail: hidewaki@ims.u-tokyo.ac.jp
Received 18 January 2015; revised 27 February 2015; accepted 27 February 2015; published online 30 March 2015

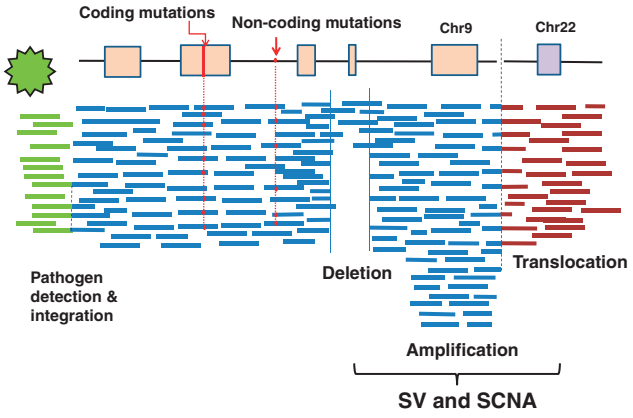

Figure 1. WGS by NGS can detect non-coding mutations, SVs (structural variants) including SCNAs (somatic copy number alterations) and translocations, and pathogen detection, as well as protein-coding mutations.

somatic mutations of *PIK3CA* in many types of cancer.[14] In 2008, they also sequenced 20 661 protein-coding genes in glioblastoma and found *IDH1* hotspot mutations,[15] which is now targeted for drug development. These achievements proved the concept for high-throughput sequencing study for cancer genome by Sanger sequencing, which has been replaced by NGS.[11] Targeted sequencing is still an important method for verification and validation of mutations detected by NGS analysis. Targeted sequencing for some dozens or hundreds of genes by using NGS is an important approach at the time when most of driver genes or 'actionable' genes are listed for validation and clinical application.

SCNAs are a hallmark of genomic abnormality in cancer cells. Copy number analysis using DNA arrays is a powerful method to characterize SCNAs in cancer genomes. In the 1980s, intensive copy number analyses found many of loss of heterozygosity and amplified regions, which was important information to identify classical tumor suppressor genes and oncogenes.[2,3,11,16] High-density DNA arrays or single-nucleotide polymorphism arrays measure copy number in genomic regions using hybridization signals or allele intensity ratios of single-nucleotide polymorphisms, whereas high-throughput sequencing analysis on WES and WGS can detect SCNAs using sequence read numbers aligned to each of genomic regions. The resolution of sequencing analysis is expected to be higher than DNA chip approach by combining them with SVs information detected by WGS.

EXOME SEQUENCING (WES) AND MUTATIONS IN PROTEIN-CODING REGIONS

WES is taking center stage for cancer genome sequencing and now huge amounts of mutational data in exonic protein-coding regions for common and rare tumors have been accumulated, with most driver genes that are frequently mutated in common cancers identified.[11] Focus is shifting to the rarer mutated driver candidate among the 'long tail' of mutated gene lists[3,12] and to validate these biological functions in cancer. Pan-cancer analysis on exome data demonstrated that carcinogen-exposed cancer such as ultraviolet-related melanoma and smoking-related lung cancer have far higher numbers of somatic mutations in coding regions among common types of cancer.[17] On the other hand, pediatric tumors and leukemia have fewer mutations and only several non-silent mutations are present in their whole-coding regions, which is likely to be solely driver mutations in their carcinogenesis.[17] TCGA and ICGC provide comprehensive mutational data of coding regions in $>12\,000$ cancers and the COSMIC database is extensively curating mutations from targeted sequencing and WES, summarizing coding mutations for $>1\,000\,000$ cancer samples.[18]

WES usually captures protein-coding exons spanning 37–50 Mb (1–2%) of the human genome by DNA–RNA or DNA–DNA hybridization insolution.[19] Its performance is 30–70% on-target rate, and WES sequences 70–100× and more 'depth' (coverage for target regions) for each sample, which is more accurate than 30× WGS, because accuracy of mutation detection by NGS is primarily dependent on the sequencing depth. WES is much cheaper than WGS and a reasonable stop-gap solution until WGS becomes cheap enough in the future. However, the present WES captures miss 3% of coding regions due to probe failure for extremely GC-rich coding regions and complicated regions.[20] This method cannot analyze off-target regions such as non-coding regions and shows some bias specific to each capture method. For example, reads with indels of a dozen nucleotides or more,[20] or those with multiple variants like human leukocyte antigen (HLA) regions can be missed or show lower depth because of lower efficiency of hybridization.

'WET' PLATFORM FOR WGS

WGS by NGS is technically straightforward. DNA is randomly fragmented and 30× or more depth (90–100 Gb) of each human whole genome is usually sequenced for both cancer and normal genomes.[8,21] It can cover 99% of the entire human genome,[21] but considering heterogeneity of both cancer tissue and cancer genome, more sequencing depth is preferable for detecting somatic mutations comprehensively. In several types of cancer such as pancreatic cancer tissues, cancer cells may comprise only 5–30% of the sample. These stroma-abundant tumors require more sequence depth to detect their somatic mutations (100–200×).[22] Common NGS technology reads are 100–150-bp sequences of both ends of a 300–600-bp DNA fragment, which is randomly sheered.[21] However, this technology is still dependent on PCR and shows some PCR bias. GC-rich or AT-rich regions are difficult to sequence and WGS produces lower depth in these regions. PCR-free protocols have been presented and show less GC-bias and may be more comprehensive. The other limitation of the present WGS technology is the short read length. The 3-Gb human reference sequence reveals many repetitive regions and pseudogenes and when short sequence reads are aligned to this redundant reference sequence, a lot of alignment errors occur and lead to mutation calling errors.[23] The third-generation single-molecule sequencing[24] or nanopore sequencing technologies can yield some kb and longer sequences without PCR process and are expected to be applied to WGS, but they currently suffer from a high error rate (5–10%) in each read and are prohibitively expensive for WGS at present.

'DRY' PLATFORM FOR WGS

One of the greatest difficulties for cancer WGS is informatics and computational analysis. Cancer WGS is required to produce >90–100 Gb × 2 (cancer and normal) of sequence data, corresponding to about one T(tera)-bytes for one sample. Hence, a large computer facility or equivalent resources are required to handle these large data sets and to perform alignment and variant calling promptly for some hundreds or thousands of cancer WGS. World-class genome centers are increasing their computer resources for WGS, but it would be not sufficient in these academia resources for analysis on tens of thousands of WGS data set. Informatics teams are interested in high-performance commercial-based cloud computing systems that can solve these problems, although there are technical problems of data transfer and ethical and legal issues because cancer genome analysis handles potentially sensitive genomic information of each patient.[25]

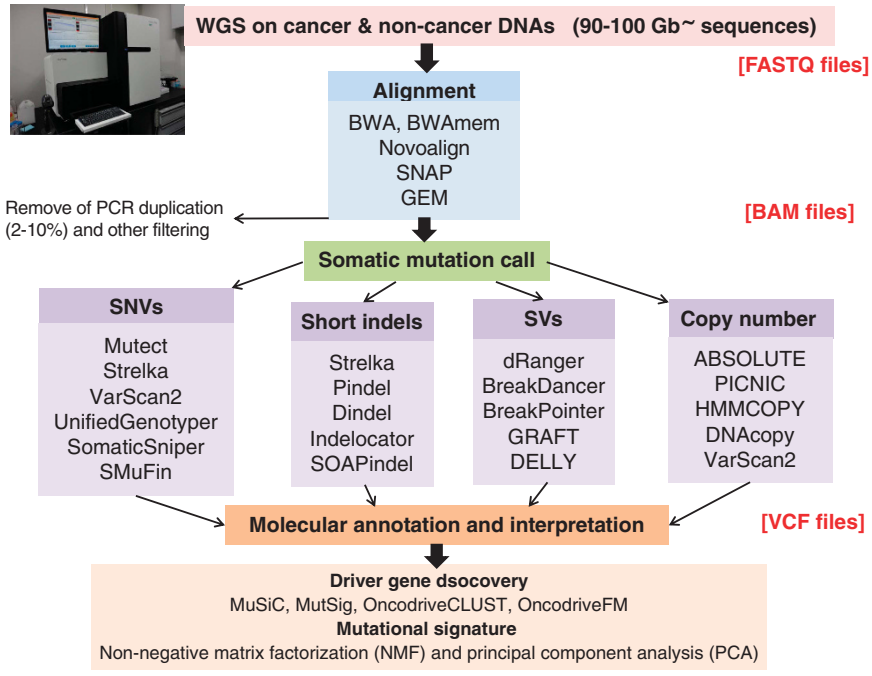

Figure 2. A representative set of analysis tools for cancer WGS data. As an initial step, raw sequence data (90–100 Gb × 2: FASTQ files) from NGS of cancer genome and normal genome are aligned to the 3-Gb human reference sequence (3 Gb), producing BAM files, and PCR duplication is removed from the BAM file (usually 2–10%). Somatic mutations are called by several types of algorithms specific to types of mutations (SNVs, short indels, SCNAs, SVs and others), comparing variant allele number in cancer genome with those in normal genome by statistical analysis and creates a list of somatic mutations (VCF files).

A representative set of computational analysis tools for cancer WGS data is shown in Figure 2. As an initial step, raw sequence data from NGS (FASTQ files) are aligned to the 3-Gb human reference sequence, producing BAM files, and PCR duplicates are removed from the BAM file (usually 2–10%). Somatic mutations are called by several types of algorithms or specific to types of mutations (single-nucleotide variants (SNVs), short indels, SCNAs and SVs), comparing variant allele fraction in the cancer genome with those in the normal genome. Accuracy is dependent on read depth in each genomic region, with WGS typically having lower depth than WES, although WGS can cover variants in coding and non-coding regions. The other issue to determine the accuracy of WGS is alignment error. Taking account of the complexity and redundancy of the human whole genome, alignment error can occur with some frequency when short read sequences are aligned to repetitive and redundant genome sequences.[23] The most serious problem of WGS is that its result is dependent on these mutation call algorithm and each pipeline calls different results,[26] especially in low-depth regions or for mutations with low frequency of the mutant allele. To evaluate the consistency of mutation calling for cancer WGS, there are some ongoing efforts for comparing calling results from several algorism and optimizing mutation calling through a community-based challenge such as DREAM Challenge[27] and ICGC validation/verification working group.

As a next step, several informatics/statistics analyses are performed to interpret these mutational profiles, such as non-negative matrix factorization[17] and principal component analysis of mutational signatures for speculation of carcinogenic process and background, and identification of driver genes, which is described below (Figure 2).

SNVs AND SHORT INDELS IN NON-CODING REGIONS

The human genome contains genes encoding ~20 000 ncRNAs,[28,29] including transfer RNA, ribosomal RNA, microRNA and other long non-coding RNAs (Figure 3). These functional ncRNAs are expected to contribute to epigenetics, transcription regulation, RNA splicing and the translational machinery.[28,29] In addition to these ncRNAs, the initial transcripts of protein-coding genes usually contain extensive non-coding sequences, in the form of introns, 5′ UTRs and 3′ UTRs, which are also involved with the regulation of RNA transcription, RNA splicing and protein translational processes.[30,31] Mutations in 3′ UTR may tend to occur in cancer driver genes[32] and are likely to control RNA stability and protein translation through microRNA binding.[30,31] Furthermore, intergenic regions are likely to have many different regulatory element sequences, which are crucial to regulating

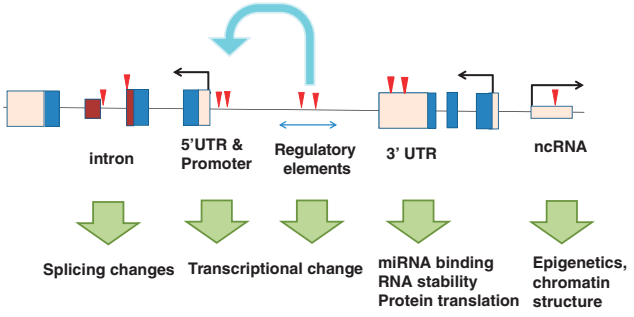

Figure 3. Non-coding mutations and gene expression. Intronic mutations can affect splicing forms. Mutations in 5′ UTR and promoter regions can alter transcriptional activity and regulatory elements such as enhancers and silencers in intergenic regions also change transcriptional activity. Mutations in 3′ UTR can alter RNA stability and protein translation through change of microRNA (miRNA) binding and other mechanism. Mutations in non-coding RNA, especially miRNAs and lincRNAs, may change the interaction of coding RNAs/proteins and regulatory elements, and alter epigenetic status and chromatin structure. lincRNA, long non-coding RNA.

gene expression around these elements. Genome-wide association studies (GWAS) on common types of cancers have identified some hundred cancer-predisposing loci, many of which are located in intergenic non-coding regions, and they are expected to be involved with regulatory elements controlling gene expression around these loci.[33] These findings from genome-wide association study indicate that non-coding regulatory elements could have some important roles in carcinogenesis. Extrapolations from the ENCODE[34] project and FAMTOM[35] suggest that 20–40% of the human genome could be regulatory elements. Efforts have shifted towards finding interactions between DNA and regulatory proteins by ChIP-Seq, or open chromatin elements where the DNA is not packaged by histones (DNase hypersensitive sites), both of which tell where there are active regulatory sequences in a cell type-specific or cancer-specific manner.[34,35]

One of the biggest advantages of WGS over WES is that WGS can detect non-coding mutations in cancer and mutational analysis focusing on non-coding regions in a number of WGS data is now underway. Initially, WGS analysis in melanoma samples discovered hotspot mutations in the *TERT* promoter, which are located − 124-bp and − 146-bp upstream from its translation start ATG site, and conferred enhanced *TERT* promoter activity by putatively generating a consensus binding site for the ETS transcription factor.[36] These promoter mutations were frequently detected in brain tumors, bladder cancer, thyroid cancer, melanoma and liver cancer, although the strength of association between these *TERT* promoter mutations and *TERT* expression is highly variable across cancer types.[37] In a subset of T-cell ALL, somatic mutations in non-coding regions were found, which introduced binding motifs for the *MYB* transcriptional factor and created a super-enhancer upstream of the *TAL1* oncogene.[38] Recent systematic or statistical analysis for non-coding somatic mutations using WGS data sets from TCGA and ICGC indicated that some non-coding regions are frequently mutated, such as the promoters or regulatory elements of *PLEKHS1* and *WDR74* in addition to *TERT*,[39,40] although it is still unclear whether they are related with gene expression around these loci as regulatory regions. More systematic approaches targeting non-coding regions are required. Many data sets targeting non-coding regions such as ENCODE[34] and FANTOM5[35] are annotating regulatory regions and ncRNAs and by combining these ChIP-Seq and gene expression data sets, an amount of non-coding mutations can be annotated and interpreted.

Communities and projects involved with cancer genome are now producing much more WGS for many types of cancer and analyzing them, thanks to the markedly decreasing cost of WGS and expanding computer resources. It will become more important to interpret these non-coding mutations and their functions in the human genome by integration with other omics data sets such as ChIP-Seq, RNA-Seq and DNA methylation (bisulfite sequencing) data.

MUTATIONAL ANALYSIS IN REPEAT OR REPETITIVE REGIONS

Repetitive DNA sequences comprise ~ 50% of the human genome, and $<$ 10% of the human genome consists of tandem repeats or low complexity repeat sequences with variable lengths of two nucleotides to tens of nucleotides.[23] These sequences are highly variable and are used for forensic DNA analysis and diagnosis for cancer with DNA mismatch-repair deficiency using a microsatellite (MS) instability test. Among the repetitive sequences, trinucleotide repeats are important as they sometimes occur within protein-coding regions and may lead to change in protein function, such as the tandem repeat regions of *FLT3*[41] and *AR*.[42] It is still difficult to analyze mutations and variants in such repetitive regions by WGS and NGS approaches, because of alignment issues for short read sequences.[23] Park and his group[43] analyzed several whole genomes of MS instability-positive colorectal and endometrial cancers and found 11 000–332 000 MS mutations in whole genomes of these tumors in addition to many more exonic mutations, whereas MS-stable cancer revealed 5–7000 MS mutations.

Transposable genetic elements, which can replicate and insert copies of themselves at other locations, are an abundant component in the human genome.[23] The most abundant transposon lineage, *Alu*, has about 50 000 active copies and LINE-1 has about 100 active copies per genome. These transposons had a major role and were a driving force for genomic evolution and diversity.[23,42] Several studies analyzed transposon-mediated mutations or SVs using cancer WGS data and identified 4–5 somatic retrotransposon insertion per tumor.[44,45] These somatic retrotransposon insertions tend to occur in genes that are commonly mutated in cancer, and change their expressions.

GENOMIC REARRANGEMENTS (SVs) AND CHROMOTHRIPSIS

Distinct rearrangements or translocations in leukemia and rare sarcoma lead to the activation of proto-oncogene products or creation of cancer-specific fusion genes, some of which are diagnostic tools for sarcomas such as *STY-SSX1* fusion in synovial sarcoma and *EWS-FLI1* fusion in Ewing sarcoma.[46] Philadelphia chromosome in CML,[47] translocation between chromosome *9q34* and *22q11* gives rise to the *BCR-ABL* fusion gene with the *ABL* kinase inhibitor imatinib, the first successful drug targeting cancer mutations. A small inversion on chromosome *2p* creates the *EML4-ALK* fusion gene, which is found in 1–2% lung adenocarcinomas and kinase inhibitors are targeting such a kinase fusion gene as *ALK*.[48] Rearrangement involving *ROS1* at chromosome *6q22* (mainly translocation) and *RET1* at chromosome *10q11.2* (mainly inversion) were also identified in a few percent of lung cancers with unique clinical and pathological features, which produce fusion kinases as driver genes and are now molecular targets for lung cancer.[49,50] Not only in rare tumors, 40–70% of prostate cancers was found to have rearrangements involving *ERG* at chromosome *21q22* and multiple ETS family genes, producing *TMPRSS2-ERG* fusion and ETS family gene fusion.[51] Recent analysis for medulloblastoma found recurrent rearrangement activated *GFIB* proto-oncogene by enhancer hijacking (swapping).[52]

These SVs or chromosomal rearrangements are pervasive but our ability to characterize and interpret their impacts has been limited,[3] NGS can detect SVs comprehensively and systematically[52] and WGS can detect these SVs by several approaches (Figures 4a and b). First, paired-end information and apparent fragment length and orientation by NGS are used to detect deletion, insertion, duplication, inversion and translocation[53] (Figure 4a). For deletion and duplication, integration with read number distribution or copy number information is likely to more effective. Split read or chimera read analysis can detect these SVs and also determine their breakpoints directly.[53,54] Ultimately, for more complicated SVs and pathogen detection, a *de novo* assembly approach may be more useful and attractive[55] (Figure 4b). The *de novo* assembly using short read sequence with long read sequences produced by third-generation NGS is now attempted to clarify complicated genomic structures in human genome, cancer genome and other species.[24]

WGS analysis on cancer revealed a distinct phenomenon, termed chromothripsis (from the Greek for 'chromosome' (chromo) and 'shattering into pieces' (thripsis)).[55] In chromothripsis, one or a few chromosomes in one cell produce dozens to hundreds of clustered rearrangements[56,57] (Figure 4c). The mechanism for such complicated rearrangements is that at one or more carcinogenic stages distinct chromosomes or genomic regions become fragmented into many segments, which are then pieced together inaccurately by DNA repair mechanisms.[57] Recent

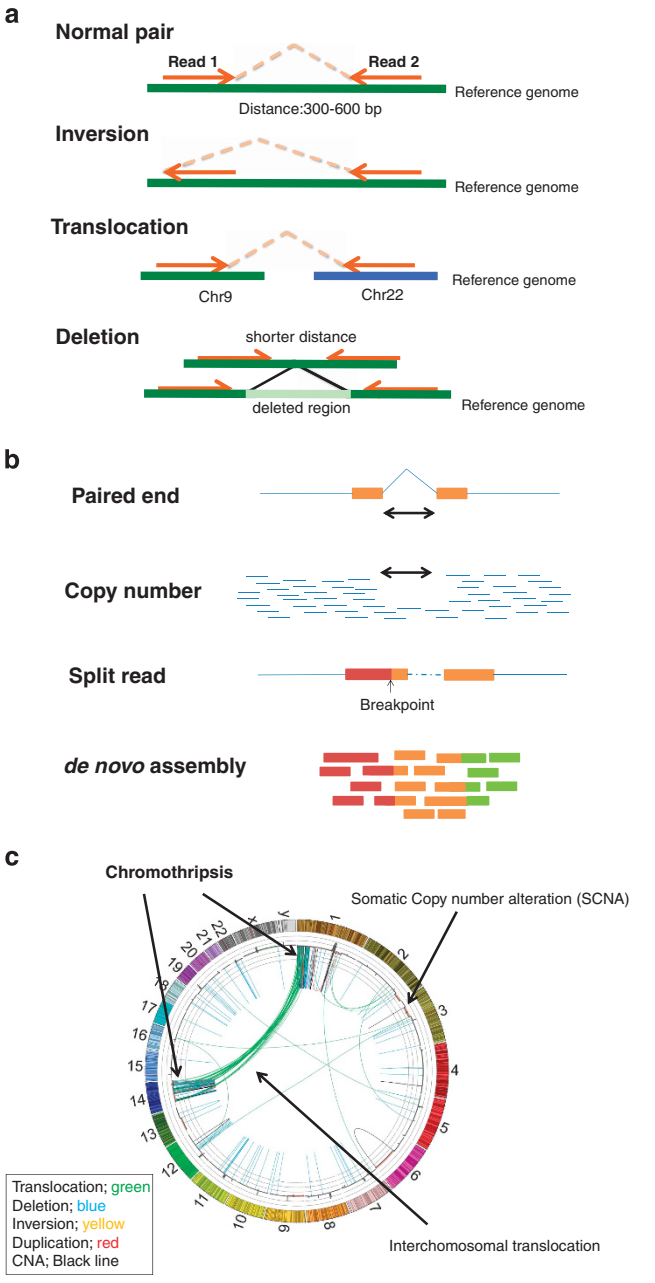

Figure 4. Genomic rearrangement and chromothripsis in cancer genome. (**a**) Paired-end sequencing by NGS detects genomic rearrangements using the information on read direction and read distance that are aligned to the reference human genome. (**b**) In addition to paired-end information, copy number information from WGS, split read alignment, breakpoint spanning reads and *de novo* assembly of reads that are unaligned to the reference genome are used to detect SVs and pathogen integration. (**c**) A representative Circos plot of cancer genome structure and chromothripsis. A Circos plot indicates SVs and SCNAs in all of human chromosomes (1–22+XY).

study suggested chromothriptic signatures in cancers arising from patients with inherited p53 mutations,[58] as well as strong associations with somatic p53 mutations, this event is associated with the functions of p53 and other genomic stability in various DNA damage response signaling. Furthermore, WGS analysis for pancreatic cancers indicated that such a genomic instability 'phenotype' was co-segregated with inactivation of DNA maintenance genes (*BRCA1*, *BRCA2* and *PALB2*), and also related with high response to DNA-damaging agents.[59]

PATHOGEN DETECTION AND INTEGRATION TO THE HOST GENOME

Cancer development is strongly associated with pathogen infection and chronic inflammation.[60] For example, hepatitis B virus or hepatitis C virus infection is a strong carcinogenic factor for liver cancer development. Human papillomavirus infection initiates and promotes carcinogenesis of cervix and *Helicobacter pylori* infection and its related chronic gastritis is also a strong carcinogenic factor for gastric cancer. Hence, it is important in cancer genomes to detect DNA or RNA sequences derived from known and unknown pathogens (virus and bacteria) leading to chronic inflammation and to identify genomic integrations of pathogens to the host human cancer genome. Technically, unaligned reads to the human genome sequences are extracted from WGS data and they can be checked whether they match known pathogen genome sequences with or without pre-assembling.[8] Especially for tumors in digestive organs, bacteria detection and metagenome analysis of gut flora are important to understand the genome–environmental interaction in tumor development.[61]

Using pair-end information of NGS reads, WGS analysis can detect genomic integration sites of pathogen to human genome efficiently and accurately. Hepatitis B virus (DNA virus) infection is proved to be integrated in liver cancer and infected non-tumorous liver cells. WGS of liver cancers detected several integration sites of hepatitis B virus DNA genome (3 kb), which were reported to preferentially integrate to the genomic regions of the *TERT* and *MLL4* loci.[62,63] The Human papillomavirus DNA genome and its integrations are detected by WGS of cervical cancer,[64] and head and neck cancer.[65] Several rare cancers have a strong viral component, such as Epstein–Barr virus in Burkitt lymphoma[66] and nasopharyngeal carcinoma, and RNA retrovirus HTLV-1 in adult T-cell leukemia/lymphoma.[67] These viral integrations/interactions are likely to lead to genomic instability followed by copy number changes, overexpression around the integration sites, and human–human or human–virus gene fusion events, in addition to oncoproteins derived from viral genome.

IDENTIFICATION OF DRIVER GENES

In the above sections, we described mutation detection. Identification of driver genes from the mutation catalog and their interpretation is an important next step.[3] One of the broadest definitions of 'driver gene' is a gene that confers growth advantage to the cancer cell.[2] As driver genes are expected to be positively selected throughout cancer development, the number and pattern of mutations in the driver genes are likely to be different from the expectation under the null hypothesis. Therefore, driver genes are identified by examining the number of non-silent mutations, position of mutations and functional bias of mutations. To test the number of mutations, expected number of mutation is estimated for each gene. As the number of mutations is influenced by gene length and nucleotide composition, such as CpG, the expected number is calculated by considering these factors.[68–70] In addition, gene expression level and replication timing are also considered to standardize the effects of transcription-coupled repair and regional mutation rate.[69] Then the observed number of mutation is compared with the expected number of mutations to test the recurrence of the mutations.[12,62,70] This method is commonly used to identify driver genes. Besides coding genes, this framework was applied to non-coding regions and significantly mutated regulatory regions were reported.[40]

Analysis of location and functional bias of mutations is useful for interpretation. Driver genes are generally classified into oncogenes and tumor suppressor genes. Mutations in oncogenes render the gene constitutively active (gain of function) or confer different function (change of function).[71] Therefore, we can expect that the mutations in oncogenes are clustered at a few functionally important codons.[71] These mutation clusters can be identified by non-uniform distribution of mutations in the coding regions.[70,71] On the other hand, mutations in tumor suppressor gene reduce activity of the gene, and loss-of-function mutations, such as nonsense or frameshift indels, are likely to be enriched in tumor suppressor gene. Therefore, test for the proportion of loss-of-function mutations can identify tumor suppressor genes. In addition, the comparison of the functional impact of mutations also helps us interpret the role of mutated genes.[72] Systematic analysis on the location and functional bias of mutations requires large number of mutations, therefore, only a few studies have been reported.[12,70] However, recent accumulation of mutation data would make it possible to perform detailed analyses on the mutation patterns or combination among multiple driver and passenger candidate mutations, and incorporating other information, such as evolutional conservation of functional domains, protein–protein interactions and protein complexes, should expand our knowledge of driver genes.

In addition to point mutation and short indels, SVs and pathogen integrations also alter gene functions and gene expressions. Although statistical methods to analyze SCNAs are available,[73] testing for the significance of somatic rearrangements is still a difficult task. How to comprehensively integrate data of point mutations, short indels, SCNAs and rearrangement is an important topic for large-scale WGS study.

INTEGRATIVE WGS ANALYSIS WITH RNA-SEQ FOR INTERPRETATION

The impact of mutations in non-coding regions and SVs in cancer genomes is usually difficult to evaluate and interpret so far.[3] RNA-Seq data can evaluate the impact of deep intronic and synonymous mutations and check the transcriptional or functional consequences of these genomic alterations. Changes in splicing patterns can be driven by somatic mutations and contribute to oncogenic processes and acquisition of more aggressive and survival phenotypes of cancer cells.[74,75] In addition to mutations in exon–intron junction sites (AT–GC consensus sites), mutations in deep intronic regions can generate new splice-donor or -acceptor sites, giving rise to new splicing forms.[76] Synonymous mutations in coding regions and intronic regions can alter exonic motifs that regulate splicing and the functions of cancer-related genes.[31] There are also many fusion events detected by RNA-Seq and most of them are caused by genomic rearrangement.[48–51] Some SVs affecting non-coding regions can also change the expression of nearby genes owing to changes of regulatory elements and chromatin structures.[52] There is much potential for transcriptional or functional consequences of SVs and non-coding mutations and they should be further explored. However, at present, few studies systematically compare genomic mutations or variants and transcriptional aberrations from WGS and RNA-Seq data and evaluate the transcriptional consequences of these genomic variants comprehensively.[76,77] In the future, to interpret mutational profiles from cancer WGS, integrative analysis of RNA-Seq and multi-omics analysis with RNA profile,[76] DNA methylation profile,[78] protein expression profiles[79] and chromatin structure profiles such as ChIP-Seq[34,35] and HiC-Seq[80] will be required and should be a powerful approach to interpret mutational consequences and understand the biology of cancer genomes.

Another integrated approach to interpret mutations and identify driver genes is to combine high-throughput functional cellular assays[81] with WES or WGS and more functional data should be integrated for cancer genomics.

CONCLUSION

As sequencing costs are decreasing and computer resources are expanding, WGS analysis for cancer profiling research and clinical utilities will become more common and more sophisticated. WGS for cancer provides information to understand the biology underlying the cancer genome and the function of unexplored non-coding regions and SVs in the human genome. Current WGS analysis of the whole picture of the cancer genome is still primitive and there are improvements to be made in NGS technology, informatics and computational methods. Taking into account the diversity of cancer genomes and phenotypes, interpretation of the mutational data from cancer WGS will also require the analysis of much more WGS data and integration with multi-omics data, functional data and clinic-pathological data in a larger number of sample set.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank researchers and technical staffs in RIKEN-IMS and Professor Miyano and his fellows in Human Genome Center, Institute of Medical Science, The University of Tokyo for their great efforts to cancer genome sequencing in ICGC project. The supercomputing resource 'SHIROKANE' was provided by Human Genome Center, The University of Tokyo (http://sc.hgc.jp/shirokane.html). This work was supported partially by RIKEN President's Fund 2011, the Princess Takamatsu Cancer Research Fund, and Takeda Science Foundation.

REFERENCES

1 Dulbecco R. A turning point in cancer research: sequencing the human genome. *Science* 1986; **231**: 1055–1056.
2 Stratton M, Campbell PJ, Futreal A. The cancer genome. *Nature* 2009; **458**: 719–724.
3 Garraway LA, Lander ES. Lessons from the Cancer Genome. *Cell* 2013; **153**: 17–37.
4 Kinzler KW, Vogetstein B. Lessons from hereditary colorectal cancer. *Cell* 1996; **87**: 159–170.
5 King CR, Kraus M, Aaronson SA. Amplification of a novel v-erbB related gene in human mammary carcinoma. *Science* 1985; **229**: 974–976.
6 Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 2008; **456**: 66–72.
7 Mardis ER, Wilson RK. Cancer genome sequencing. *Hum Mol Genet* 2009; **18**: R163–R168.
8 Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* 2010; **11**: 685–696.
9 Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 2008; **455**: 1061–1068.
10 International Cancer Genome Consortium, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C *et al.* International network of cancer genome projects. *Nature* 2010; **464**: 993–998.
11 Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz Jr LA, Kinzler KW. Cancer genome landscapes. *Science* 2013; **339**: 1546–1558.
12 Lowrence M, Stojanov P, Mermel C, Robinson JT, Garraway LA, Golub T *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 2014; **505**: 495–501.
13 Leiserson MD, Vandin F, Wu H, Dobson JR, Eldridge JV, Thomas JL *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet* 2014; **47**: 106–114.
14 Samuels Y, Wang Z, Bardelli A, Silliman N, Ptak J, Szabo S *et al.* High frequency of mutations of the PIK3CA gene in human cancers. *Science* 2004; **304**: 554.
15 Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* 2008; **321**: 1807–1812.
16 Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* 2010; **463**: 899–905.

17 Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV *et al.* Signatures of mutational processes in human cancer. *Nature* 2013; **500**: 415–421.
18 Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 2014, pii: gku1075 **43**: D805–D811.
19 Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* 2009; **27**: 182–189.
20 Clark MJ, Chen R, Lam HY, Karczewski KJ, Chen R, Euskirchen G *et al.* Performance comparison of exome DNA sequencing technologies. *Nat Biotechnol* 2011; **29**: 908–914.
21 Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008; **456**: 53–59.
22 Biankin AV, Waddell N, Kassahn KS, Gingras MC, Muthuswamy LB, Johns AL *et al.* Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature* 2012; **491**: 399–405.
23 Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* 2011; **13**: 36–46.
24 Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G *et al.* Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol* 2012; **30**: 693–700.
25 Dove ES, Joly Y, Tassé AM, Public Population Project in Genomics and Society (P3G) International Steering Committee, International Cancer Genome Consortium (ICGC) Ethics and Policy Committee, Knoppers BM. Genomic cloud computing: legal and ethical points to consider. *Eur J Hum Genet* 2014; e-pub ahead of print 24 September 2014; doi:10.1038/ejhg.2014.196.
26 Wang Q, Jia P, Li F, Chen H, Ji H, Hucks D *et al.* Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Med* 2013; **5**: 91.
27 Boutros PC, Ewing AD, Ellrott K, Norman TC, Dang KK, Hu Y *et al.* Global optimization of somatic variant identification in cancer genomes with a global community challenge. *Nat Genet* 2014; **46**: 318–319.
28 Esteller M. Non-coding RNAs in human disease. *Nat Rev Genet* 2011; **12**: 861–874.
29 Prensner JR, Chinnaiyan AM. The emergence of lncRNAs in cancer biology. *Cancer Discov* 2011; **1**: 391–407.
30 Zhao W, Pollack JL, Blagev DP, Zaitlen N, McManus MT, Erle DJ. Massively parallel functional annotation of 3' untranslated regions. *Nat Biotechnol* 2014; **32**: 387–391.
31 Oikonomou P, Goodarzi H, Tavazoie S. Systematic identification of regulatory elements in conserved 3' UTRs of human transcripts. *Cell Rep* 2014; **7**: 281–292.
32 Supek F, Miñana B, Valcárcel J, Gabaldón T, Lehner B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* 2014; **156**: 1324–1335.
33 Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C *et al.* Principles for the post-GWAS functional characterization of cancer risk loci. *Nat Genet* 2011; **43**: 513–518.
34 ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; **489**: 57–74.
35 Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* 2014; **507**: 455–461.
36 Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA. Highly recurrent TERT promoter mutations in human melanoma. *Science* 2013; **339**: 957–959.
37 Vinagre J, Almeida A, Pópulo H, Batista R, Lyra J, Pinto V *et al.* Frequency of TERT promoter mutations in human cancers. *Nat Commun* 2013; **4**: 2185.
38 Mansour MR, Abraham BJ, Anders L, Berezovskaya A, Gutierrez A, Durbin AD *et al.* An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* 2014; **346**: 1373–1377.
39 Fredriksson NJ, Ny L, Nilsson JA, Larsson E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat Genet* 2014; **46**: 1258–1263.
40 Weinhold N, Jacobsen A, Schultz N, Sander C, Lee W. G enome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet* 2014; **46**: 1160–1165.
41 Nakao M, Yokota S, Iwai T, Kaneko H, Horiike S, Kashima K *et al.* Internal tandem duplication of the flt3 gene found in acute myeloid leukemia. *Leukemia* 1996; **10**: 1911–1918.
42 Gottlieb B, Beitel LK, Wu JH, Trifiro M. The androgen receptor gene mutations database (ARDB): 2004 update. *Hum Mutat* 2004; **23**: 527–533.
43 Kim TM, Laird PW, Park PJ. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. *Cell* 2013; **155**: 858–868.
44 Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ 3rd *et al.* Landscape of somatic retrotransposition in human cancers. *Science* 2012; **337**: 967–971.
45 Shukla R, Upton KR, Muñoz-Lopez M, Gerhardt DJ, Fisher ME, Nguyen T *et al.* Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell* 2013; **153**: 101–111.
46 Oda Y, Tsuneyoshi M. Recent advances in the molecular pathology of soft tissue sarcoma: implications for diagnosis, patient prognosis, and molecular target therapy in the future. *Cancer Sci* 2009; **100**: 200–208.
47 Groffen J, Stephenson JR, Heisterkamp N *et al.* Philadelphia chromosomal breakpoints are clustered within a limited region, bcr, on chromosome 22. *Cell* 1984; **36**: 93–94.
48 Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S *et al.* Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* 2007; **448**: 561–566.
49 Kohno T, Ichikawa H, Totoki Y, Yasuda K, Hiramoto M, Nammo T *et al.* RET, ROS1 and ALK fusions in lung cancer. *Nat Med* 2012; **18**: 375–377.
50 Takeuchi K, Soda M, Togashi Y, Suzuki R, Sakata S, Hatano S *et al.* KIF5B-RET fusions in lung adenocarcinoma. *Nat Med* 2012; **18**: 378–381.
51 Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 2005; **310**: 644–648.
52 Northcott PA, Lee C, Zichner T, Stütz AM, Erkek S, Kawauchi D *et al.* Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* 2014; **511**: 428–434.
53 Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* 2008; **40**: 722–729.
54 Yang L, Luquette LJ, Gehlenborg N, Xi R, Haseley P, Hsieh C *et al.* Diverse mechanisms of somatic variations in human cancer genomes. *Cell* 2013; **153**: 919–929.
55 Nagarajan N, Pop M. Sequence assembly demystified. *Nat Rev Genet* 2013; **14**: 157–167.
56 Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 2011; **144**: 27–40.
57 Korbel JO, Campbell PJ. Criteria for inference of chromothripsis in cancer genomes. *Cell* 2013; **152**: 1226–1236.
58 Rausch T, Jones DT, Zapatka M, Stütz AM, Zichner T, Weischenfeldt J *et al.* Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* 2012; **148**: 59–71.
59 Waddell N, Pajic M, Patch A, Chang DK, Kassahn KS, Bailey P *et al.* Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* 2015; **518**: 495–501.
60 Elinav E, Nowarski R, Thaiss CA, Hu B, Jin C, Flavell RA. Inflammation-induced cancer: crosstalk between tumours, immune cells and microorganisms. *Nat Rev Cancer* 2013; **13**: 759–771.
61 Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM *et al.* Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res* 2012; **22**: 292–298.
62 Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Nguyen HH *et al.* Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat Genet* 2012; **44**: 760–764.
63 Sung WK, Zheng H, Li S, Chen R, Liu X, Li Y *et al.* Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat Genet* 2012; **44**: 765–769.
64 Ojesina AI, Lichtenstein L, Freeman SS, Pedamallu CS, Imaz-Rosshandler I, Pugh TJ *et al.* Landscape of genomic alterations in cervical carcinomas. *Nature* 2014; **506**: 371–375.
65 Parfenov M, Pedamallu CS, Gehlenborg N, Freeman SS, Danilova L, Bristow CA *et al.* Characterization of HPV and host genome interactions in primary head and neck cancers. *Proc Natl Acad Sci USA* 2014; **111**: 15544–15549.
66 Cao S, Strong MJ, Wang X, Moss WN, Concha M, Lin Z *et al.* High-throughput RNA sequencing-based virome analysis of 50 lymphoma cell lines from the cancer cell line encyclopedia project. *J Virol* 2015; **89**: 713–729.
67 Cook LB, Melamed A, Niederer H, Valganon M, Laydon D, Foroni L *et al.* The role of HTLV-1 clonality, proviral structure and genomic integration site in adult T cell leukemia/lymphoma. *Blood* 2014; **123**: 3925–3931.
68 Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* 2007; **44**: 153–158.
69 Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 2013; **499**: 214–218.
70 Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J *et al.* Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci Rep* 2013; **3**: 2650.

71 Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. *Nat Med* 2014; **10**: 789–799.
72 Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res* 2012; **4**: e169.
73 Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* 2011; **1**: R41.
74 Pajares MJ, Ezponda T, Catena R, Calvo A, Pio R, Montuenga LM. Alternative splicing: an emerging topic in molecular and clinical oncology. *Lancet Oncol* 2007; **8**: 349–357.
75 Chen J, Weiss WA. Alternative splicing in cancer: implications for biology and therapy. *Oncogene* 2015; **34**: 1–14.
76 Shiraishi Y, Fujimoto A, Furuta M, Tanaka H, Chiba K, Boroevich KA *et al*. Integrated analysis of whole genome and transcriptome sequencing reveals diverse transcriptomic aberrations driven by somatic genomic changes in liver cancers. *PLoS One* 2014; **9**: e114263.
77 Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RK *et al*. The human splicing code reveals new insights into the genetic determinants of disease. *Science* 2015; **347**: 1254806.
78 Shen R, Mo Q, Schultz N, Seshan VE, Olshen AB, Huse J *et al*. Integrative subtype discovery in glioblastoma using iCluster. *PLoS One* 2012; **7**: e35236.
79 Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z *et al*. Proteogenomic characterization of human colon and rectal cancer. *Nature* 2014; **513**: 382–387.
80 Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet* 2013; **14**: 390–403.
81 Liang H, Cheung LW, Li J, Ju Z, Yu S, Stemke-Hale K *et al*. Whole-exome sequencing combined with functional genomics reveals novel candidate driver cancer genes in endometrial cancer. *Genome Res* 2012; **22**: 2120–2129.