DATABASE IN BRIEF

HUMAN MUTATION

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

NGS Catalog: A Database of Next Generation Sequencing Studies in Humans

Junfeng Xia[1,†], Qingguo Wang[1,†], Peilin Jia[1], Bing Wang[1], William Pao[2,3], and Zhongming Zhao[1,3,4,*]

[1]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN, USA,[2]Department of Medicine/Division of Hematology-Oncology, Vanderbilt University School of Medicine, Nashville, TN, USA,[3]Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN, USA,[4]Department of Psychiatry, Vanderbilt University School of Medicine, Nashville, TN, USA

†These authors contributed equally to this work.

*Correspondence to Zhongming Zhao, Ph.D., Department of Biomedical Informatics, Vanderbilt University School of Medicine, 2525 West End Avenue, Suite 600, Nashville, TN, 37203, USA; Phone: +1-615-343-9158; FAX: +1-615-936-8545; E-mail: zhongming.zhao@vanderbilt.edu.

Contract grant sponsor: This study was partially supported by the Stand Up To Cancer-American Association for Cancer Research Innovative Research Grant (SU2C-AACR-IRG0109) and the VICC Cancer Center Core grant P30CA68485 from National Institutes of Health.

Communicated by Johan T. den Dunnen

ABSTRACT: Next generation sequencing (NGS) technologies have been rapidly applied in biomedical and biological research since its advent only a few years ago, and they are expected to advance at an unprecedented pace in the following years. To provide the research community with a comprehensive NGS resource, we have developed the database Next Generation Sequencing Catalog (NGS Catalog, http://bioinfo.mc.vanderbilt.edu/NGS/index.html), a continually updated database that collects, curates and manages available human NGS data obtained from published literature. NGS Catalog deposits publication information of NGS studies and their mutation characteristics (SNVs, small insertions/deletions, copy number variations, and structural variants), as well as mutated genes and gene fusions detected by NGS. Other functions include user data upload, NGS general analysis pipelines, and NGS software. NGS Catalog is particularly useful for investigators who are new to NGS but would like to take advantage of these powerful technologies for their own research. Finally, based on the data deposited in NGS Catalog, we summarized features and findings from whole exome sequencing, whole genome sequencing, and transcriptome sequencing studies for human diseases or traits. ©2012 Wiley Periodicals, Inc.

KEY WORDS: next generation sequencing (NGS), exome sequencing, whole genome sequencing, RNA sequencing, disease genome, gene fusion, database

INTRODUCTION

Next generation sequencing (NGS), also known as massively parallel sequencing, is rapidly transforming biomedical and biological research from single gene to genome scale [Koboldt et al., 2010; Metzker, 2009]. Compared with gene expression microarray developed in the late 1990s and early 2000s, the NGS technologies

Received 15 November 2011; accepted revised manuscript 9 March 2012.

DOI: 10.1002/humu.22096

have a much higher impact on diverse biological applications, especially the clinical diagnostic applications. Within only a few years of the advent of NGS technologies, it is now possible to allow researchers to apply whole exome sequencing (Exome-Seq), whole genome sequencing (WGS), whole transcriptome sequencing (RNA-Seq), or a combination of them to investigate individual genome(s), especially those related to disease. Due to the massive amount of genetic information generated, these sequencing strategies are quickly shifting the paradigm of basic and translational research. Consequently, the number of applications of NGS to various genomic studies has increased at an exceptional speed over the past several years.

As of December 2011, the genomes of at least 13 healthy individuals have been sequenced and published, each of which had high coverage (average number of reads of each genomic site) using the NGS technologies (Supp. Table S1). Analysis of the first reported personal genome (the Watson Genome), which was sequenced to approximately 7× coverage on the 454 GS FLX platform, revealed 3.3 million single nucleotide variants (SNVs) [Wheeler et al., 2008b]. Among these SNVs, 0.6 million were novel; that is, they were not found in the dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/) [Sherry et al., 2001]. Analyses of other human individual genomes by WGS reported a similar number of SNVs (~3-4 millions per genome) (Supp. Table S1). For those SNVs, approximately 80-90% was included in the dbSNP database. The genotype concordance of the identified SNVs is usually in the range of 98-99% in the reported personal genome sequencing studies, indicating a high quality of the NGS data. Of importance, the ratio of the number of transitions to the number of transversions (Ti/Tv) is consistently found to be ~2.1. This ratio of 2.1 has been well reported in previous studies of specific genomic regions, and it has become the gold standard for SNV detection in WGS studies [Koboldt et al., 2010].

NGS technologies have demonstrated their power in detecting disease causing or causative genetic variants of human diseases [Gilissen et al., 2011; Shendure, 2011], especially in cancer [Ding et al., 2010b; Robison, 2010]. The genetic variants identified by NGS technologies typically include single nucleotide variants (SNVs), small insertions and deletions (indels), copy number variations (CNVs) and large structural variants (SVs). In 2009, Ng et al. [2009a; 2009b] first demonstrated that Exome-Seq was able to uncover candidate genes for a human Mendelian disease (the Miller syndrome). In another study, Hoischen et al. [2010] first successfully applied Exome-Seq to identify a dominant disease gene *SETBP1* that caused Schinzel-Giedion syndrome. These studies illustrated the great potential of Exome-Seq technology to uncover Mendelian disease genes in a high-throughput fashion. Not surprisingly, as the cost of NGS has decreased dramatically since the advent of NGS technologies, many researchers have quickly followed and applied Exome-Seq for disease gene discovery. However, because of its targeted genomic regions (exons and flanking regions), limitations of Exome-Seq to date include its incapability and inefficiency to identify SVs, CNVs, or non-coding variants, which is within the capacity of WGS. For WGS applications, Campbell et al. [2008] were the first to apply low-coverage WGS to detect CNVs and SVs in cancer genomes. They applied the Illumina sequencing technology platform to generate sequence reads from the genomes of two lung cancer cell lines and identified two fusion transcripts in addition to other genomic rearrangements. Subsequently, the Ley et al. [2008] study was the first to sequence the entire cancer genome of a patient. They used high-coverage WGS to sequence a typical acute myeloid leukemia genome (32.7 × coverage) and its matched normal counterpart (13.9 × coverage) obtained from the same patient, and this design had the goal of unbiased identification of tumor-specific mutations that altered the protein-coding genes. The pioneering work above has stimulated the application of NGS to disease studies and is shifting the paradigm of biomedical research. Accordingly, NGS studies have led to a substantial expansion of the realm of disease/trait associated genetic studies.

Rapidly emerging NGS studies provide us with an exceptional opportunity to examine the potential impact of genetic variants on diseases by systematically cataloging and summarizing key characteristics of the observed SNVs, indels, CNVs, and SVs. Although there is a review of some features such as identification of causative mutations using Exome-Seq for Mendelian disease [Gilissen et al., 2011], a comprehensive survey across all NGS publications, to date, has not been conducted yet. Knowledge synthesis of these analyses can translate the rapidly emerging data from disease genetic association research into potential applications for clinical practice, which is important for researchers to expand the utilities of NGS technologies based on previous work, to generate new hypotheses for follow up functional validations, and to explore new discoveries from existing and newly generated NGS datasets.

To provide the research community with a comprehensive NGS resource and facilitate translational research on disease genetics, we have developed the Next Generation Sequencing Catalog (NGS Catalog, http://bioinfo.mc.vanderbilt.edu/NGS/index.html), a continually updated database that collects, curates, and

manages available human NGS data from published literature. To the best of our knowledge, this is the first online resource for the published NGS studies that focus on human diseases/traits. The NGS Catalog deposits all available genomic characteristics of disease/trait associated SNVs, small indels, CNVs, and SVs. It also collects mutated genes, the variants in the mutated genes, and gene fusions detected by NGS technologies if such information can be obtained from the literature. Additionally, NGS Catalog provides a comprehensive and updated list of software widely used in the NGS community, as well as general pipelines for the analysis of Exome-Seq, WGS and RNA-Seq data. These features are particularly useful for numerous investigators who are new to NGS but would like to take the advantage of these powerful technologies for their own research. Finally, based on the data deposited in NGS Catalog, we summarized features and findings from germline mutations and somatic mutations studies for human diseases/traits.

MATERIALS AND METHODS

The design of the NGS Catalog, including data collection, curation, database design and implementation, and website development, follows the guidelines proposed by the GWAS Catalog [Hindorff et al., 2009] and our two recent databases: Ethanol Related Gene Resource (ERGR) [Guo et al., 2009] and Schizophrenia Gene Resource (SZGR) [Jia et al., 2010]. Details are provided below.

Data Collection

Studies are eligible to be included in our database if they meet the following criteria. First, the publications should be online or in peer reviewed journals and published in English. Second, the studies are disease- or trait-oriented in humans. Third, the studies provide sufficient information and include technical details such as bioinformatics analysis procedures and number of variants detected.

Published literature related to or employing NGS techniques were collected primarily through weekly PubMed searches (http://www.ncbi.nlm.nih.gov/pubmed). We also periodically performed literature searches using Google Scholar (http://scholar.google.com/). We applied several strategies in the PubMed searches. First, a literature search was performed with general terms: “high throughput sequencing” OR “next generation sequencing.” Experiment-specific terms were then applied to filter the search results. For RNA-Seq, we used “RNA sequences” OR “transcriptome sequencing”. Of note, RNA-Seq in our database is primarily addressed as a method to sequence the transcribed portions of genome, rather than to measure expression level. For Exome-Seq, we used “exome sequencing”; for WGS, we used “whole genome sequencing.” Among the searched results, review articles and publications primarily focusing on NGS tool development were removed. As a specific example, to identify Exome-Seq related publications in disease studies, we searched the PubMed database using the key term “Exome-sequencing AND disease” to retrieve all studies and used the limits setting “Humans.” This specific search yielded 39 publications to be further manually checked. Of these publications, 37 were included in our database (as of August, 2011).

We retrieved all the publications that met the criteria above, along with their supplementary materials, which often contained more related information, for the NGS Catalog database.

Data Extraction and Preparation

The information extracted from each eligible publication can be distinguished by three categories: publication information, study information and result summary. Specifically, publication information includes first author (last name), publication date (online/electronic publication date if available), full journal name, and the title of the paper. For study information, we extracted biological information, including the diseases/traits examined, mutation type, population or ethnic background of the samples, sample composition, as well as technical information, including NGS methods and platforms, maximum read length, average coverage, computational tools, reference genome and public SNPs database used in the study. For result summary information, the numbers of total and novel variants, including SNVs, indels, CNVs and SVs, if available, were collected. Moreover, we collected the disease/trait associated genes, the identified variants in these genes and gene fusions reported by the authors. In our extraction process, if some information was not available, it was labeled as missing.

Database Design and Implementation

The data extracted from NGS publications was managed through MySQL (http://www.mysql.com/), an open source relational database management system that has been widely used in biomedical and biological research. We created a main table to store a description of each publication and a summary of the corresponding study. Other information was stored in separate tables and linked by keys. For example, one table was designed specifically to map each gene name to its Entrez Gene identifier, by which the system could retrieve gene information from the National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov/gene/). This design allows us to easily update the database frequently and keep public information updated through dynamic links.

RESULTS AND DISCUSSION

Web Interface

The NGS Catalog database can be browsed or accessed in three ways: (1) the study-oriented web function, (2) the technique-oriented function, and (3) the user upload function. We describe each of them below.

The studies deposited in the NGS Catalog database can be retrieved through its web interface. For ease of use, we provided a search function in multiple ways. For example, users may use any of the keywords related to disease/trait, gene, mutation type, journal, publication date, sequencing platform, author, or their combination to search for studies deposited in the database. Figure 1 shows an example using two keywords, *TP53* and *Link*, to retrieve the study by Link et al., which was published in the *Journal of the American Medical Association* in April 2011 [Link et al., 2011]. By default, the records retrieved from the database are ordered by publication date. If the query result is not empty, a web link will be provided to allow the users to save the query result, as shown in Figure 1.

Figure 1. The NGS Catalog main search page, one example of a search result is shown at the bottom.

To present the search results, a summary table is generated in the first page with each row for each study. In each row, the first column shows the authors, title, and publication date, followed by other columns listing the mutation type (somatic mutation or germline mutation), the number of SNVs, the number of short indels, the number of CNVs, the number of large SVs, and the reported genes and gene fusions whenever available. For more information, users may click the publication title link, and then a detailed web page describing the study pops up, including NGS platforms, average coverage of the NGS experiment, mapping tools, and variant calling software, among others. Moreover, NGS Catalog collects variants from the studies if available. A detailed web page for these variants, including chromosomal position and specific changes in the cDNA and protein sequences, will pop up when the mutation type link is clicked. Currently, we mainly provide SNVs and short indels of the reported gene(s). For the studies that contain more than 50 variants, we do not list these variants due to workload constraints and beyond our main aim; rather, a brief summary ">50 mutations" is displayed. In addition to the study information, NGS Catalog has collected software widely used in the NGS community. The software is grouped into the following eight categories: (i) Mapping tools, (ii) SNV detection; (iii) Indel detection; (iv) CNV detection; (v) SV detection; (vi) Annotation; (vii) Data visualization, and (viii) Fusion gene detection. The web page presenting this information can be opened by clicking "Software" on the functional menu. Moreover, three pipelines for WGS, RNA-Seq and Exome-Seq, respectively, are provided on the NGS Catalog website, allowing new investigators to have a general view of the data process and analysis.

Aside from data retrieval from NGS Catalog, users are encouraged to upload additional publication information to the website. Users may first search the NGS Catalog database to check if their publication has already been deposited into the database. If not, users may upload the related publication information, which will be stored in NGS Catalog. The new record will be forwarded to the NGS Catalog developer via email and will become available after a manual check and confirmation.

Finally, users can provide feedback to us through the webpage, accessible via the menu function "Feedback." This important feature will help us gain feedback from users, including information regarding data quality or errors, new features, new publications, and so on, so that NGS Catalog can be improved through users' experiences.

All the descriptions above are available in our user-friendly online manual, which is accessible through the menu item "Documents" on the website.

Data Overview

We restricted our analyses to 240 NGS studies deposited in NGS Catalog by December 31, 2011. As summarized in Figure 2A, they included 157 Exome-Seq, 47 WGS, and 39 RNA-Seq studies, respectively (note that in some studies these sequencing technologies were combined to sequence one sample). We expect this number to increase dramatically considering the increasingly widespread applications of NGS technologies to the acquisition of genomic data, as indicated in Figure 2B. Most of these NGS studies were based on the Illumina platform (Supp. Table S2). So far, The American Journal of Human Genetics, Nature and Nature Genetics are the top journals that published these studies (Supp. Table S3).

Germline Mutation Detection

NGS can be applied to identify germline mutations. For example, family-based Exome-Seq has been applied to identify causative germline mutations that underlie Mendelian diseases. Considering most variations are mediated by non-synonymous, frame-shifting and canonical splice variation in such diseases, Exome-Seq is ideal for researchers to understand high-penetrance allelic mutations and their relationships to diseases/traits [Gilissen et al., 2011].

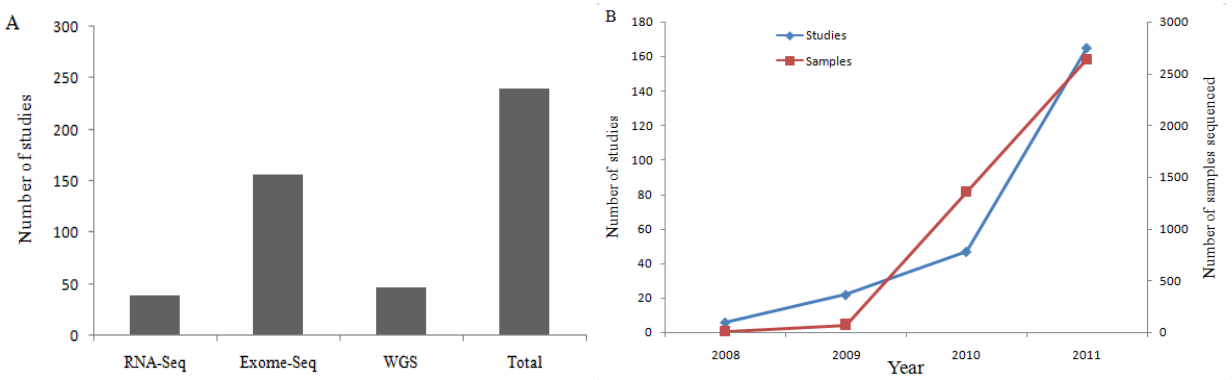

Figure 2. Distribution of publications using next generation sequencing technologies. (A) Number of studiess by genomic category (transcriptome, exome, and whole genome). (B) Number of studies and samples sequenced by publication year.

During the past two years we have seen numerous proof-of-concept studies using Exome-Seq technologies to identify new Mendelian disease genes related to recessive and dominant disorders, such as Miller syndrome [Ng et al., 2009a], Kabuki syndrome [Ng et al., 2010], and several others. These studies have led to the identification of over 40 new disease genes (Supp. Table S4). These publications paint a mixed picture of phenotypes, genes and mutations underlying Mendelian diseases. As shown in Supp. Table S4, there is a trend toward recessive disorders, in which the genetic causes are easier to detect than those of dominant disorders. In a recent review of the Mendelian diseases that have been studied by Exome-Seq to date [Gilissen et al., 2011], the authors pointed out that an Exome-Seq study could not always identify a new disease gene. For example, Xiao et al. used Exome-Seq to identify pathogenic mutations in a large Chinese family with congenital motor nystagmus (CMN); however, no causative gene was identified in that family [Xiao et al., 2011]. Several reasons such as failure in capturing the genomic region where causative gene resides may explain this failure of identification of CMN gene. Furthermore, *NOTCH2* was identified as the disease gene for Hajdu-Cheney syndrome by two different groups at almost the same time [Isidor et al., 2011; Simpson et al., 2011]. This clearly indicates that Exome-Seq is rapidly expanding all over the world and there is fierce competition among NGS Studies.

In addition to Exome-Seq, WGS can also be applied to study Mendelian diseases [Swami, 2010]. Lupski et al. [2010] applied the WGS method to identify variants and genes involved in Charcot-Marie-Tooth disease. This is the first study using WGS to discover a recessive disease gene (*SH3TC2*). In that study, the authors sequenced the whole genome of one individual with Charcot-Marie-Tooth and called approximately 3.4 million SNVs. In addition, WGS has been used to detect germline mutations in cancer as well. For example, Yokoyama et al identified a germline mutation in MITF that was associated with sporadic melanoma via sequencing the genome of an affected individual from a number of melanoma families [Yokoyama et al., 2011].

Somatic Mutation Detection

Besides germline mutations, NGS has been widely used to detect somatic mutations in complex diseases and traits, especially cancer. Exome-Seq has been applied to study somatic SNVs and short indels in many types of cancers, such as acute monocytic leukemia [Yan et al., 2011], melanoma [Nikolaev et al., 2011; Stark et al., 2011; Wei et al., 2011], and gastric cancer [Wang et al., 2011]. However, one drawback of Exome-Seq is its lack of capacity in detecting SVs [Karakoc et al., 2011], which occur frequently in cancer. Compared with Exome-Seq, WGS is particularly appealing because it can detect a full spectrum of genetic variants (SNVs, indels, CNVs and SVs) that may contribute to human cancer. Obviously, however, much more cost and effort is needed in a WGS project than an Exome-Seq project.

The complete genome sequencing of several types of human cancer (Table 1), such as melanoma [Pleasance et al., 2009a], lung cancer [Lee et al., 2010], and hepatocellular carcinoma [Totoki et al., 2011], has dramatically expanded the catalog of somatic mutations that may contribute to cancer development and growth.

It has been observed that transitions (A↔G and C↔T) occur at a higher frequency than transversions (A↔C, A↔T, C↔G, and G↔T), which is known to be a general property of DNA sequence change and evolution [Jiang and Zhao, 2006]. Recent human studies have shown that for a whole human genome, the ratio of the number of transitions to the number of transversions (Ti/Tv) is estimated to be around 2.1 when the genome is assessed as a whole or large scale [Moore et al., 2011; Pelak et al., 2010]. With a large number of WGS studies of cancer and the corresponding control samples collected in our database (see Table 1), it is worthwhile to examine this feature specifically in cancer genomes. We calculated the Ti/Tv ratio of somatic mutations and then compared this feature. The results are shown in Figure 3 and Supp. Table S5. It is important to note that the number of somatic mutations and the values of Ti/Tv ratio varied greatly both within and between classes of cancer, indicating that the somatic mutational spectrum is complicated and highly different from the germline mutational spectrum.

As indicated in Figure 3, the Ti/Tv ratios of somatic mutations in most cancers were below 2, whereas a melanoma (COLO-829 cell line) showed an exceptionally large ratio value (3.26). This high Ti/Tv ratio was similarly observed in our recent WGS study of a chemotherapy-naive metastatic melanoma (5.25, Dahlman et al., manuscript submitted). T>C/A>G transition is the second most frequent type of mutations in our case, while in COLO-829 C>A/G>T transversion is the second most frequent type of mutations; this difference might result in even a larger Ti/Tv ratio in our melanoma dataset. The lowest Ti/Tv ratio was observed in non-small-cell lung cancer (0.36), followed by small-cell lung cancer (0.67). While we know that somatic mutations in melanoma and lung cancer studies may reflect previous exposure to mutagens such as ultraviolet radiation light and tobacco smoke carcinogens, respectively, the detailed pathogenesis is not quite understood. Figure 3 also shows the differences within the same types of cancer. For example, in colorectal cancer, the highest Ti/Tv value is 1.23 for the sample CRC-3 while the lowest value is 0.79 for the sample CRC-5. One reason for this difference is that the signatures derived in the past from known cancer genes or functionally important genes, such as *TP53*, are inevitably influenced by biological selection, which distorts the patterns generated by the underlying mutational processes [Greenman et al., 2007]. This is supported by the fact that the Ti/Tv ratio of therapy-related AML (t-AML) with *TP53* mutations is 1.06, whereas two *de novo* AML (without chemo/radiotherapy) genomes without *TP53* mutations have much higher ratios. A similar phenomenon exists in chronic lymphocytic leukemia, where the samples with mutations in immunoglobulin genes (CLL3 and CLL4) have a relatively lower value of Ti/Tv when compared with the samples without mutations in these genes (CLL1 and CLL2).

Besides Exome-Seq and WGS, another prominent development in the field of NGS is to apply RNA-Seq to examine the whole transcriptome in high resolution. RNA-Seq only sequences the regions of the genome that are transcribed and spliced into mature mRNA, which is approximately 2% of the entire genome [Sboner et al., 2010]. While the major application of RNA-Seq is to identify differentially expressed transcripts or genes, here, we focus on its advantages in detecting variants/mutations, especially somatic mutations, including gene fusions in cancers. The advantage that makes RNA-Seq ideal for discovery of expressed fusion genes in cancer is that it allows the detection of multiple alternative splice variants resulting from a fusion event. The feasibility of applying RNA-Seq to detect fusion genes in cancer genomes was first evaluated by Maher et al. [2009a], who not only pinpointed known chimeras such as *TMPRSS2-ERG* and *BCR-ABL1* from RNA-Seq data of tumor and cancer cell lines but also found novel fusions that were subsequently validated by experiments. This pioneering work demonstrated the full power of RNA-Seq for hybrid gene detection and stimulated RNA-Seq applications in cancer research. Table 2 summarizes recent RNA-Seq studies that have resulted in the findings of novel hybrid oncogenes, which are available through the NGS Catalog. However, one limitation in the RNA-Seq application is that it cannot detect gene fusion events involving non-transcribed regions [Kim and Salzberg, 2011]. Other factors that complicate RNA-Seq data analysis are the tissue-specificity and the broad dynamic range of expression in the human transcriptome.

Table 1. Whole genome sequencing (WGS) of disease genomes

Disease[a]	Samples	Read length (bp)[b]	Coverage	SNVs[c]	Indels[c]	Somatic SVs	Reference
CMT	1 case	2×50	29.9×	561,719	-	-	[Lupski et al., 2010]
MS, PCD	2 cases and their parents	-	~61× (average)	323,255 (within family)	-	-	[Roach et al., 2010]
HA	10 cases/10 controls	2×75	31.1× (average)	~441,846 (average)	-	-	[Pelak et al., 2010]
GBM	1 case	2×50	>30×	243,622	116,964	-	[Clark et al., 2010]
M1-AML	1 case/1 control	-	32.7× (case), 13.9× (control)	31,632	333	-	[Ley et al., 2008]
M1-AML	1 case/1 control	-	23.3× (case), 21.3× (control)	20,256	-	-	[Mardis et al., 2009]
Melanoma	1 case/1 control	2×75	40× (case), 32× (control)	33,345	66 validated	37 validated	[Pleasance et al., 2009a]
SCLC	1 case/1 control	2×25	39× (case), 31× (control)	22,910	65	58	[Pleasance et al., 2009b]
BC	3 cases (tumor, metastasis, xenograft)/1 control	-	~28× (case), 38.8× (control)	27,173 (tumor), 51,730 (metastasis), 109, 078 (xenograft)	45,946 (tumor), 61,857 (metastasis), 19,844 (xenograft)	-	[Ding et al., 2010a]
NSCLC	1 case/1 control	-	60× (case), 46× (control)	83,054	54,921	43 validated	[Lee et al., 2010]
PC	13 cases	2×37	-	-	-	381	[Campbell et al., 2010]
PR	7 cases/7 controls	2×101	30× (average)	3,866 (average)	-	90 (average)	[Berger et al., 2011]
CC	9 cases/9 controls	2×101	30.7× (case), 31.9× (control)	15,330 (average)	-	75 (average)	[Bass et al., 2011]
T-AML	1 case/1 control	2×75	28.7× (case), 29.9× (control)	26 validated	2 validated	8 translocations	[Link et al., 2011]
CLL	4 cases/4 controls	-	~39.5× (case), ~40.5× (control)	~1,000 (average)	5 validated	-	[Puente et al., 2011]
HCC	1 case/1 control	2×50	35.9× (case), 28.1× (control)	11,731	670	22 validated	[Totoki et al., 2011]
MM	23 cases/23 controls	2×101	32.8× (case), 32.2× (control)	~7,450 (average)	-	-	[Chapman et al., 2011]

[a]CMT, Charcot-Marie-Tooth disease; MS, Miller syndrome; PCD, primary ciliary dyskinesia; HA, hemophilia A; GBM, glioblastoma multiforme; M1-AML, French-American-British subtype M1 acute myeloid leukemia; SCLC, small-cell lung cancer; BC, breast cancer; NSCLC, non-small-cell lung cancer; PC, pancreatic cancer; PR, prostate cancer; CC, colorectal cancer; t-AML, therapy-related acute myeloid leukemia; CLL, chronic lymphocytic leukemia; HCC, hepatitis C virus–positive hepatocellular carcinoma; MM, multiple myeloma. [b]Read length: maximum read length (bp). [c]For CMT, MS and PCD, HA, and GBM, SNV and Indels indicate the number of SNVs and indels that were not present in the dbSNP or other public databases. For the remaining diseases, SNVs and Indels indicate the number of somatic SNVs and indels, respectively.

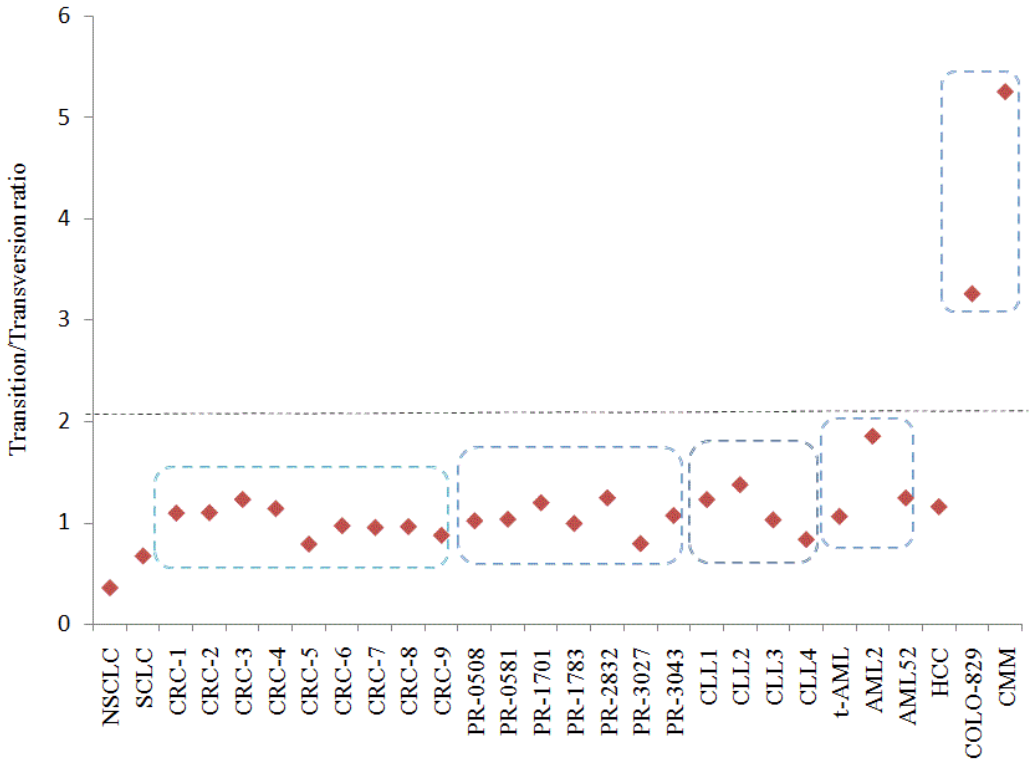

Figure 3. Transition to transversion (Ti/Tv) ratios of somatic mutations reported in different types of cancer. The Ti/Tv ratio of AML was derived by the mutational spectrum figure in Link et al. [2011]; the Ti/Tv ratio of HCC was calculated based on the mutation data provided by Totoki et al. [2011]; the Ti/Tv ratio of CMM was based on our recent melanoma WGS study (Dahlman et al., manuscript submitted); the remaining Ti/Tv ratios were calculated based on the supplementary data of original papers. NSCLC, non-small-cell lung cancer; SCLC, small-cell lung cancer; CRC, colorectal cancer; PR, prostate cancer; CLL, chronic lymphocytic leukemia; AML, acute myeloid leukemia; HCC, hepatitis C virus–positive hepatocellular carcinoma; CLL1 and CLL2, with no mutations in the immunoglobulin genes (IGHV-unmutated) and CLL3 and CLL4, with mutations in these genes (IGHV-mutated); t-AML, therapy-related AML with TP53 mutations; AML2 and AML52 are two *de novo* AML (without chemo/readiotherapy) genomes without TP53 mutations; CMM, chemotherapy-naive metastatic melanoma.

With increasingly widespread applications of RNA-Seq, major advancements have been made in the discovery of gene fusions. As of November 21, 2011, as many as 706 gene fusions have been documented in the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer (http://cgap.nci.nih.gov/Chromosomes/Mitelman) [Mitelman et al., 2007]. These gene fusions enhance our understanding of the origins of cancer. More significantly, a number of fusions have been recognized as important prognostic tools or therapeutic targets in anti-cancer treatments. In a study reported recently [Tomlins et al., 2011], a newly invented urine test has been developed to detect the presence of *TMPRSS2-ERG*, thereby assessing the risk of prostate cancer.

Table 2. Summary of novel gene fusions discovered by RNA-Seq

Cancer type	Sample	# validated gene fusions	Example fusion event(s)	Reference
Breast cancer	1 breast cancer cell line MCF-7	5	*AHCYL1-RAD51C*	[Maher et al., 2009b]
Breast cancer	4 breast cancer cell lines BT-474, MCF-7, KPL-4 and SK-BR-3	24	*ACACA-STAC2*	[Edgren et al., 2011]
Chronic myeloid leukemia	1 chronic myeloid leukemia cell line K562	1	*NUP214-XKR3*	[Maher et al., 2009b]
Gastric cancer	1 gastric cancer GCT-15	1	*AGTRAP-BRAF*	[Palanisamy et al., 2010]
Hodgkin lymphoma	2 Hodgkin lymphoma cell lines KM-H2 and L428 C	3	*CIITA-BX648577*	[Steidl et al., 2011]
Lung cancer	12 lung cancer cell lines	1	*R3HDM2-NFE2*	[Wang et al., 2009]
Melanoma	10 melanoma samples (2 cell lines, 8 patient-derived short-term cultures)	11	*CCT3-C1orf61*	[Berger et al., 2010]
Prostate cancer	8 (3 human primary prostate tumors and adjacent matched normal tissue samples, 1 human brain reference RNA sample, 1 universal human reference sample)	8	*SEC31A-C6orf62*	[Nacu et al., 2011]
Prostate cancer	25 prostate cancer samples enriched for ETS fusion negative samples and 3 benign prostate tissues	7	*ALG5-PIGU*	[Pflueger et al., 2011]
Prostate cancer	5 *ETS* gene fusion positive and 10 *ETS* gene fusion negative prostate cancers	3	*SLC45A3-BRAF, ESRP1-RAF1*	[Palanisamy et al., 2010]
Prostate cancer	2 prostate cancer cell lines (VCaP, LNCaP), 1 benign immortalized prostate cell line RWPE, 2 prostate cancer tissues (VCaP-Met, Met 3), 1 metastatic prostate tissue (Met 4)	13	*STRN4-GPSN2, SLC45A3–ELK4*	[Maher et al., 2009a]
Prostate cancer	2 prostate cancer cell lines (VCaP, LNCaP), 2 prostate tumors (aT52, aT64)	8	*ZDHHC7-ABCB9*	[Maher et al., 2009b]
Serous ovarian cancer	12 late-stage serous ovarian tumors	1	*ESRRA-C11orf20*	[Salzman et al., 2011]

CONCLUSION

The application of NGS technologies to biomedical and biological research has advanced rapidly over the past few years and is expected to advance at an unprecedented pace in the following years. A simple example illustrates this point: Ley and colleagues identified a somatic mutation in *DNMT3A* from a woman with acute myeloid leukemia [Ley et al., 2010], which was not found in their previous sequencing study based on the Illumina sequencing technology platform using the same case [Ley et al., 2008]. The main reason behind this fact is the improved sequencing technology that has become more and more sensitive, efficient, and sophisticated.

NGS studies have added significant genetic information and deepened our knowledge and understanding of the genetic variants in the human genome. To facilitate NGS studies in human diseases/traits, we have developed the Next Generation Sequencing Catalog (NGS Catalog, http://bioinfo.mc.vanderbilt.edu/NGS/index.html), a continually updated database that collects, curates and manages available human NGS data published in the literatures. NGS Catalog is a unique database for NGS studies in human diseases and is freely available to the public. This valuable database could significantly reduce investigators' efforts to develop NGS studies in human diseases/traits, and NGS Catalog makes important information in the field easily accessible and available to the research community. We will continue to collect NGS publications and make the NGS Catalog database more comprehensive in the near future. Moreover, we will develop more online tools that allow users to custom browse and search the website, as well as to communicate with each other through the NGS Catalog.

ACKNOWLEDGMENTS

We thank Satishkumar R. Ganakammal and Rebecca Hiller Posey for their assistance and Jingchun Sun for helpful discussion. We also thank Kimberly Brown Dahlman for sharing with us melanoma whole genome sequencing data.

REFERENCES

Ahn SM, Kim TH, Lee S, Kim D, Ghang H, Kim DS, Kim BC, Kim SY, Kim WY, Kim C and others. 2009. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res 19:1622-1629.

Bass AJ, Lawrence MS, Brace LE, Ramos AH, Drier Y, Cibulskis K, Sougnez C, Voet D, Saksena G, Sivachenko A. 2011. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. Nat Genet 43:964-968.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53-59.

Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C and others. 2011. The genomic complexity of primary human prostate cancer. Nature 470:214-220.

Berger MF, Levin JZ, Vijayendran K, Sivachenko A, Adiconis X, Maguire J, Johnson LA, Robinson J, Verhaak RG, Sougnez C and others. 2010. Integrative analysis of the melanoma transcriptome. Genome Res 20:413-427.

Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C and others. 2008. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 40:722-729.

Campbell PJ, Yachida S, Mudie LJ, Stephens PJ, Pleasance ED, Stebbings LA, Morsberger LA, Latimer C, McLaren S, Lin ML and others. 2010. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature 467:1109-1113.

Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, Schinzel AC, Harview CL, Brunet JP, Ahmann GJ, Adli M and others. 2011. Initial genome sequencing and analysis of multiple myeloma. Nature 471:467-472.

Clark MJ, Homer N, O'Connor BD, Chen Z, Eskin A, Lee H, Merriman B, Nelson SF. 2010. U87MG decoded: the genomic sequence of a cytogenetically aberrant human cancer cell line. PLoS Genet 6:e1000832.

Ding L, Ellis MJ, Li S, Larson DE, Chen K, Wallis JW, Harris CC, McLellan MD, Fulton RS, Fulton LL and others. 2010a. Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature 464:999-1005.

Ding L, Wendl MC, Koboldt DC, Mardis ER. 2010b. Analysis of next-generation genomic data in cancer: accomplishments and challenges. Hum Mol Genet 19:R188-R196.

Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 327:78-81.

Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, Rye IH, Nyberg S, Wolf M, Borresen-Dale AL and others. 2011. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biol 12:R6.

Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe T, Boroevich KA, Nagasaki M, Yamaguchi R, Shibuya T, Kubo M and others. 2010. Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. Nat Genet 42:931-936.

Gilissen C, Hoischen A, Brunner HG, Veltman JA. 2011. Unlocking Mendelian disease using exome sequencing. Genome Biol 12:228.

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C and others. 2007. Patterns of somatic mutation in human cancer genomes. Nature 446:153-158.

Guo AY, Webb BT, Miles MF, Zimmerman MP, Kendler KS, Zhao Z. 2009. ERGR: An ethanol-related gene resource. Nucleic Acids Res 37:D840-D845.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362-9367.

Hoischen A, van Bon BWM, Gilissen C, Arts P, van Lier B, Steehouwer M, de Vries P, de Reuver R, Wieskamp N, Mortier G and others. 2010. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat Genet 42:483-485.

Isidor B, Lindenbaum P, Pichon O, Bézieau S, Dina C, Jacquemont S, Martin-Coignard D, Thauvin-Robinet C, Le Merrer M, Mandel JL and others. 2011. Truncating mutations in the last exon of NOTCH2 cause a rare skeletal disorder with osteoporosis. Nat Genet 43:306-308.

Jia P, Sun J, Guo A, Zhao Z. 2010. SZGR: a comprehensive schizophrenia gene resource. Mol Psychiatry 15:453-462.

Jiang C, Zhao Z. 2006. Mutational spectrum in the recent human genome inferred by single nucleotide polymorphisms. Genomics 88:527-534.

Karakoc E, Alkan C, O'Roak BJ, Dennis MY, Vives L, Mark K, Rieder MJ, Nickerson DA, Eichler EE. 2011. Detection of structural variants and indels within exome data. Nat Methods doi:10.1038/nmeth.1810.

Kim D, Salzberg SL. 2011. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol 12:R72.

Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, Mudge J, Miller NA, Hong D, Bell CJ. 2009. A highly annotated whole-genome sequence of a Korean individual. Nature 460:1011-1015.

Koboldt DC, Ding L, Mardis ER, Wilson RK. 2010. Challenges of sequencing human genomes. Brief Bioinform 11:484-498.

Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, Yue P, Zhang Y, Pant KP, Bhatt D and others. 2010. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature 465:473-477.

Ley TJ, Ding L, Walter MJ, McLellan MD, Lamprecht T, Larson DE, Kandoth C, Payton JE, Baty J, Welch J and others. 2010. DNMT3A mutations in acute myeloid leukemia. N Engl J Med 363:2424-2433.

Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M and others. 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 456:66-72.

Link DC, Schuettpelz LG, Shen D, Wang J, Walter MJ, Kulkarni S, Payton JE, Ivanovich J, Goodfellow PJ, Le Beau M and others. 2011. Identification of a novel TP53 cancer susceptibility mutation through whole-genome sequencing of a patient with therapy-related AML. JAMA 305:1568-1576.

Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DCY, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA and others. 2010. Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. N Engl J Med 362:1181-1191.

Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM. 2009a. Transcriptome sequencing to detect gene fusions in cancer. Nature 458:97-101.

Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, Luo S, Khrebtukova I, Barrette TR, Grasso C, Yu J and others. 2009b. Chimeric transcript discovery by paired-end transcriptome sequencing. Proc Natl Acad Sci USA 106:12353-12358.

Mardis ER, Ding L, Dooling DJ, Larson DE, McLellan MD, Chen K, Koboldt DC, Fulton RS, Delehaunty KD, McGrath SD and others. 2009. Recurring mutations found by sequencing an acute myeloid leukemia genome. N Engl J Med 361:1058-1066.

McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC and others. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res 19:1527-1541.

Metzker ML. 2009. Sequencing technologies—the next generation. Nat Rev Genet 11:31-46.

Mitelman F, Johansson B, Mertens F. 2007. The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer 7:233-245.

Moore B, Hu H, Singleton M, De La Vega FM, Reese MG, Yandell M. 2011. Global analysis of disease-related DNA sequence variation in 10 healthy individuals: Implications for whole genome-based clinical diagnostics. Genet Med 13:210-217.

Nacu S, Yuan W, Kan Z, Bhatt D, Rivers C, Stinson J, Peters B, Modrusan Z, Jung K, Seshagiri S and others. 2011. Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. BMC Med Genomics 4:11.

Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC and others. 2010. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet 42:790-793.

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA and others. 2009a. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42:30-35.

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE and others. 2009b. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461:272-276.

Nikolaev SI, Rimoldi D, Iseli C, Valsesia A, Robyr D, Gehrig C, Harshman K, Guipponi M, Bukach O, Zoete V and others. 2011. Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. Nat Genet

Palanisamy N, Ateeq B, Kalyana-Sundaram S, Pflueger D, Ramnarayanan K, Shankar S, Han B, Cao Q, Cao X, Suleman K and others. 2010. Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. Nat Med 16:793-798.

Pelak K, Shianna KV, Ge D, Maia JM, Zhu M, Smith JP, Cirulli ET, Fellay J, Dickson SP, Gumbs CE and others. 2010. The characterization of twenty sequenced human genomes. PLoS Genet 6:e1001111.

Pflueger D, Terry S, Sboner A, Habegger L, Esgueva R, Lin PC, Svensson MA, Kitabayashi N, Moss BJ, MacDonald TY and others. 2011. Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Genome Res 21:56-67.

Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin ML, Ordóñez GR, Bignell GR and others. 2009a. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463:191-196.

Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, Lin ML, Beare D, Lau KW, Greenman C and others. 2009b. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature 463:184-190.

Puente XS, Pinyol M, Quesada V, Conde L, Ordóñez GR, Villamor N, Escaramis G, Jares P, Beà S, González-Díaz M and others. 2011. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature 475:101-105.

Pushkarev D, Neff NF, Quake SR. 2009. Single-molecule sequencing of an individual human genome. Nat Biotech 27:847-850.

Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M and others. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328:636-639.

Robison K. 2010. Application of second-generation sequencing to cancer genomics. Brief Bioinform 11:524-534.

Salzman J, Marinelli RJ, Wang PL, Green AE, Nielsen JS, Nelson BH, Drescher CW, Brown PO. 2011. ESRRA-C11orf20 is a recurrent gene fusion in serous ovarian carcinoma. PLoS Biol 9:e1001156.

Sboner A, Habegger L, Pflueger D, Terry S, Chen DZ, Rozowsky JS, Tewari AK, Kitabayashi N, Moss BJ, Chee MS and others. 2010. FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. Genome Biol 11:R104.

Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, Harris RS, Petersen DC, Zhao F, Qi J. 2010. Complete Khoisan and Bantu genomes from southern Africa. Nature 463:943-947.

Shendure J. 2011. Next-generation human genetics. Genome Biol 12:408.

Sherry S, Ward MH, Kholodov M, Baker J, Phan L, Smigielski E, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308-311.

Simpson MA, Irving MD, Asilmaz E, Gray MJ, Dafou D, Elmslie FV, Mansour S, Holder SE, Brain CE, Burton BK and others. 2011. Mutations in NOTCH2 cause Hajdu-Cheney syndrome, a disorder of severe and progressive bone loss. Nat Genet 43:303-305.

Stark MS, Woods SL, Gartside MG, Bonazzi VF, Dutton-Regester K, Aoude LG, Chow D, Sereduk C, Niemi NM, Tang N and others. 2011. Frequent somatic mutations in MAP3K5 and MAP3K9 in metastatic melanoma identified by exome sequencing. Nat Genet 44:165-169.

Steidl C, Shah SP, Woolcock BW, Rui L, Kawahara M, Farinha P, Johnson NA, Zhao Y, Telenius A, Neriah SB and others. 2011. MHC class II transactivator CIITA is a recurrent gene fusion partner in lymphoid cancers. Nature 471:377-381.

Swami M. 2010. Disease genetics: Whole-genome sequencing identifies Mendelian mutations. Nat Rev Genet 11:313-313.

Tomlins SA, Aubin SMJ, Siddiqui J, Lonigro RJ, Sefton-Miller L, Miick S, Williamsen S, Hodge P, Meinke J, Blase A and others. 2011. Urine TMPRSS2: ERG fusion transcript stratifies prostate cancer risk in men with elevated serum PSA. Sci Transl Med 3:94ra72.

Totoki Y, Tatsuno K, Yamamoto S, Arai Y, Hosoda F, Ishikawa S, Tsutsumi S, Sonoda K, Totsuka H, Shirakihara T and others. 2011. High-resolution characterization of a hepatocellular carcinoma genome. Nat Genet 43:464-469.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J. 2008. The diploid genome sequence of an Asian individual. Nature 456:60-65.

Wang K, Kan J, Yuen ST, Shi ST, Chu KM, Law S, Chan TL, Kan Z, Chan ASY, Tsui WY and others. 2011. Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. Nat Genet 43:1219-1223.

Wang XS, Prensner JR, Chen G, Cao Q, Han B, Dhanasekaran SM, Ponnala R, Cao X, Varambally S, Thomas DG and others. 2009. An integrative approach to reveal driver gene fusions from paired-end sequencing data in cancer. Nat Biotech 27:1005-1011.

Wei X, Walia V, Lin JC, Teer JK, Prickett TD, Gartner J, Davis S, NISC Comparative Sequencing Program, Stemke-Hale K, Davies MA and others. 2011. Exome sequencing identifies GRIN2A as frequently mutated in melanoma. Nat Genet 43:442-446.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT. 2008a. The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872-876.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT and others. 2008b. The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872-876.

Xiao X, Li S, Guo X, Zhang Q. 2011. A novel locus for autosomal dominant congenital motor nystagmus mapped to 1q31-q32. 2 between D1S2816 and D1S2692. Human Genet doi:10.1007/s00439-011-1113-7.

Yan XJ, Xu J, Gu ZH, Pan CM, Lu G, Shen Y, Shi JY, Zhu YM, Tang L, Zhang XW and others. 2011. Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. Nat Genet 43:309-315.

Yokoyama S, Woods SL, Boyle GM, Aoude LG, MacGregor S, Zismann V, Gartside M, Cust AE, Haq R, Harland M and others. 2011. A novel recurrent mutation in MITF predisposes to familial and sporadic melanoma. Nature 480:99-103.