Systematic analysis of genetic alterations in tumors using Cancer Genome WorkBench (CGWB)

Jinghui Zhang, Richard P. Finney, William Rowe, et al.

Genome Res. 2007 17: 1111-1117 originally published online May 24, 2007
Access the most recent version at doi:10.1101/gr.5963407

Resource

Systematic analysis of genetic alterations in tumors using Cancer Genome WorkBench (CGWB)

Jinghui Zhang,[1,3] Richard P. Finney,[1] William Rowe,[1] Michael Edmonson,[1] Sei Hoon Yang,[1,2] Tatiana Dracheva,[1] Jin Jen,[1] Jeffery P. Struewing,[1] and Kenneth H. Buetow[1]

[1]*Laboratory of Population Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA;* [2]*Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Wonkwang University Hospital, Cheonbuk 570-749, Korea*

Systematic investigations of genetic changes in tumors are expected to lead to greatly improved understanding of cancer etiology. To meet the analytical challenges presented by such studies, we developed the Cancer Genome WorkBench (http://cgwb.nci.nih.gov), the first computational platform to integrate clinical tumor mutation profiles with the reference human genome. A novel heuristic algorithm, IndelDetector, was developed to automatically identify insertion/deletion (indel) polymorphisms as well as indel somatic mutations with high sensitivity and accuracy. It was incorporated into an automated pipeline that detects genetic alterations and annotates their effects on protein coding and 3D structure. The ability of the system to facilitate identifying genetic alterations is illustrated in three projects with publicly accessible data. Mutagenesis in tumor DNA replication leading to complex genetic changes in the *EGFR* kinase domain is suggested by a novel deletion–insertion combination observed in paired tumor–normal lung cancer resequencing data. Automated analysis of 152 genes resequenced by the SeattleSNPs group was able to identify 91% of the 1251 indel polymorphisms discovered by SeattleSNPs. In addition, our system discovered 518 novel indels in this data set, 451 of which were found to be valid by manual inspection of sequence traces. Our experience demonstrates that CGWB not only greatly improves the productivity and the accuracy of mutation identification, but also, through its data integration and visualization capabilities, facilitates identification of underlying genetic etiology.

[Supplemental material is available online at www.genome.org.]

Tumor mutation analysis has led to important insights into the molecular basis of cancer (Knudson 1971; Shih et al. 1981; Vogelstein and Kinzler 2004; Sjoblom et al. 2006). For example, studies of tumor mutations have led to the development of therapeutic drugs targeting specific genetic changes (e.g., Gleevec) and the discovery of a correlation between mutation profile and drug response (Lynch et al. 2004; Paez et al. 2004). The Cancer Genome Atlas Project (TCGA) (http://cancergenome.nih.gov/about/message.asp) has now been launched, a project that aims to develop a comprehensive catalog of genetic alterations in tumors. This project is expected to lead to great improvement in cancer diagnosis and treatment.

In a systematic investigation of tumor genetic alterations, informatics is a critical component because manual compilation of mutation profiles for hundreds or thousands of patients in hundreds or thousands of genes is time-consuming and error-prone. There is a critical need for an integrated system to support data analysis in mutation discovery, which includes mutation detection, mutation annotation, and data integration. Moreover, user-friendly visual displays would be very useful for validation and interpretation of the findings.

To meet these challenges, we developed Cancer Genome WorkBench (CGWB), a comprehensive informatics package designed to facilitate systematic investigation of genetic changes in tumors. As outlined in Figure 1, CGWB consists of the following three components: (1) an automated analysis pipeline that detects and annotates substitution and insertion/deletion (indel) changes using our novel algorithm's SNPdetector (Zhang et al. 2005) and IndelDetector (described here); (2) a database management system for managing project, sample, sequence trace archive, and genetic variation data (somatic mutations and polymorphisms); and (3) an application tool, Cancer Genome Browser, an enhancement of the UCSC genome browser (Kent et al. 2002) tailored for integrated tumor variation analysis.

The mutation analysis pipeline of CGWB originated from SNPdetector, an automated method developed for identifying substitution variations in fluorescence-based resequencing. SNPdetector is able to find homozygous indels (Zhang et al. 2005) but does not support the decoding of heterozygous indels that generate many overlapping fluorescence signals. Automated analysis of heterozygous indels is extremely useful for cancer genetic research involving a large number of clinical samples because indel mutations are important genetic changes in many familial cancer genes (Couch and Weber 1996; Neuhausen et al. 1996), and somatic deletions have been found to be correlated with patient response to cancer treatment (Lynch et al. 2004; Paez et al. 2004). However, existing computational methods for indel detection either lack the high accuracy required for automated data analysis (Manaster et al. 2005) or are designed primarily for finding polymorphic indels that are expected to constitute ~50% of the overall fluorescence signal in resequencing analysis (Manaster et al. 2005; Bhangale et al. 2006). Somatic indels in tumors often

[3]Corresponding author.
E-mail jinghuiz@mail.nih.gov; fax (301) 402-9325.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.5963407.

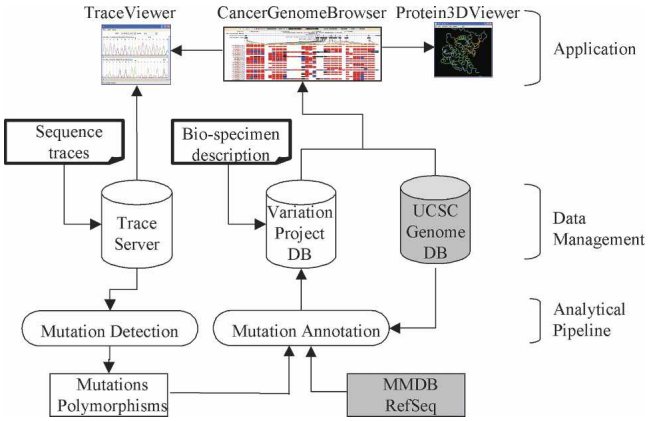

Figure 1. System overview of Cancer Genome WorkBench (CGWB). CGWB takes two sets of input data, sequence traces and biospecimen descriptions, which are loaded into the trace server and the variation database, respectively. The mutation analysis pipeline, which consists of SNPdetector and IndelDetector, processes the traces to detect mutations and polymorphisms. These variations are integrated into the reference human genome and their annotations on protein coding and 3D structure are computed using our HapScope pipeline (Zhang et al. 2002). The mRNA and protein annotation are based on the latest version of NCBI RefSeq database (Wheeler et al. 2006). The 3D structure mapping is based on NCBI's MMDB database (Wheeler et al. 2006). The variation data are then loaded into the variation database. Application programs then access the databases to generate visual displays.

have a much lower fluorescence signal than germline indels because of tumor heterogeneity and contamination of the tumor sample by nontumor tissue. We have developed a new algorithm, IndelDetector, to support the identification of indels in both germline samples and tumor tissues. In this report we show that IndelDetector is highly accurate in finding low-abundance somatic indel mutations as well as inherited indel polymorphisms in large-scale resequencing data.

CGWB provides access control security. Data for a project may be kept private or made accessible to the public. The majority (74%) of the genes analyzed to date are from confidential projects. In this report we present results from three public data sets. The first involves resequencing of PCR-amplified exonic regions of candidate genes in paired tumor–normal lung cancer samples. The second is a reanalysis of 152 genes resequenced by the SeattleSNPs group in normal individuals (Carlson et al. 2003). The third aims to discover germline mutations in candidate genes in familial ovarian cancer probands. The results demonstrate that CGWB substantially improves the accuracy of mutation identification and, through its data integration and visualization capabilities, facilitates understanding of underlying genetic etiology.

Results

Algorithm for heterozygous indel detection

A heterozygous indel allele results in two DNA molecules with different sequence lengths. In fluorescence sequencing, this gives rise to a shift of one sequence trace relative to the other, resulting in many overlapping peaks downstream of the indel (examples in Figs. 2, 3). Decoding an indel allele requires first identifying the two corresponding DNA sequences in such a region and then aligning them to derive the gap size. In a peak-overlapping region, the base calls often have very low-quality scores, and the two alleles representing the two DNA sequences may not be the two highest peaks observed because sequencing artifacts such as bubbles or spills can overshadow the peaks representing the bona fide alleles. In addition, artifacts such as polymerase slippage (i.e., "stutter") can also generate a peak-overlapping region almost indistinguishable from that caused by a heterozygous indel. For these reasons, a heuristic algorithm was developed to address the ambiguities and the artifacts in the input data.

Indel discovery was carried out in the following five steps: (1) identify peak-overlapping regions in a chromatogram; (2) decode each of the two DNA sequences in these regions; (3) compute an optimal alignment between the two decoded sequences and determine whether the secondary peaks are caused by background noise or by a potential insertion/deletion change; (4) determine whether the observed indel allele is an artifact of "stutter" or a real genetic change; and (5) determine whether indels in a region of simple tandem repeats (STR) represent multiallele simple tandem repeat polymorphisms (STRP).

In fluorescence sequencing, the peak profile of a somatic indel can be quite different from that of a polymorphic indel because a somatic indel in a tumor can have very low abundance while a polymorphic indel usually has equimolar representation

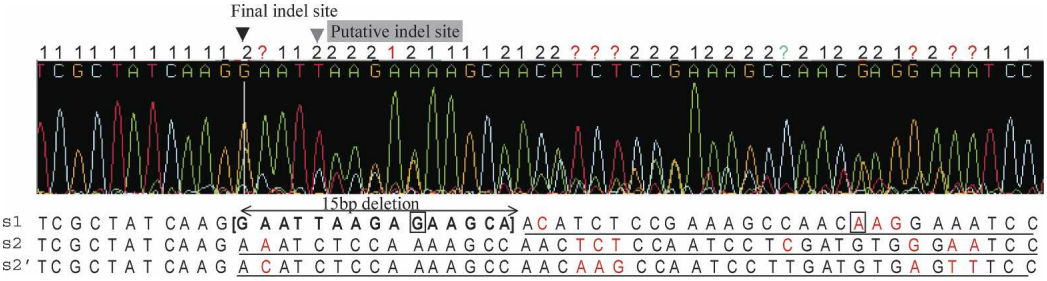

Figure 2. An example of decoding the two sequences in the peak-overlapping region using ambiguous sites. The read covers exon 19 of *EGFR*. (*Top*) The double-peak sites (2), single-peak sites (1), and ambiguous sites (?) calculated using $P_{high} = 30$ and $P_{low} = 15$ are labeled. The two decoded sequences, s1 and s2, are displayed *below*. s2′ is an alternative representation of s2 that incorporates the bases at the secondary peaks at the ambiguous sites. The alignment between s1 and s2 has a 15-bp deletion in s2 (bold). The aligned bases after the deletion in s1, s2, and s2′ are underlined. The ambiguous sites are labeled in red in s1, s2, and s2′. Seven out of the eight ambiguous sites show that bases in s2′ match those in s1 in the alignment after the 15-bp deletion (red "?" at *top*). At the remaining ambiguous site (green "?" at *top*), the base in s2 matches s1 but not that of s2′. One single-base site is labeled (red) because neither s2 nor s2′ matches s1 at this site. Two sites in s1 are labeled (boxes) because the bases represented by their secondary peaks are selected for s1. In this instance, s2′ best represents the sequence of the deletion allele, while s1 represents the sequence of the wild-type allele. The final indel position computed in step 3 is 4 bp upstream of the putative indel site found by step 1.

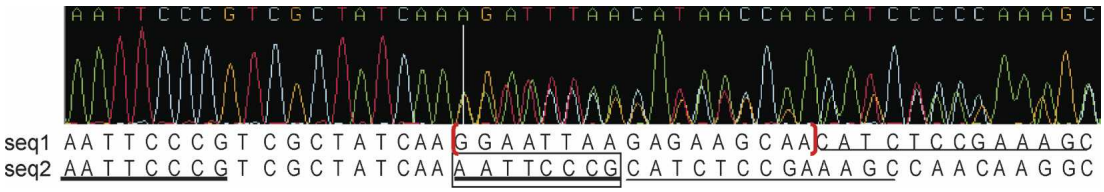

Figure 3. A complex somatic indel on *EGFR* exon 19. Seq1 and seq2 represents the two decoded alleles. Seq1 matches the reference human sequence, while seq2 represents the mutant indel allele that has a 17-bp deletion (red parentheses) coupled with a 8-bp insertion caused by replication of the upstream AATTCCCG sequence (box).

compared with the wild-type allele. IndelDetector employed the following strategy to analyze these two different types of peak profiles. The value of *P*, defined as the percent of primary peak area occupied by the secondary peak area (i.e., *P* = [secondary_peak_area/primary_peak_area] × 100), was used to determine whether a site in the chromatogram represents one or two bases. $P_{high,}$ and P_{low} are constants. Sites with $P \leq P_{low}$ were classified as single-peak (e.g., the sites marked "1" in Fig. 2). Sites with $P \geq P_{high}$ were considered to represent double-peak (e.g., the sites marked "2" in Fig. 2). Those with $P_{low} < P < P_{high}$ were classified as ambiguous (e.g., the sites marked "?" in Fig. 2). We used $P_{high} = 30$ and $P_{low} = 15$ to decode two sequences expected to have equimolar representation in the PCR product (e.g., the majority of germline genetic changes). This threshold was based on our previous analysis of chromatogram peak fluctuation of germline tissues; the heterozygous sites in such samples usually have $P > 30$ and rarely have $P < 15$ (Zhang et al. 2005). We used $P_{high} = 10$ and $P_{low} = 0$ for PCR products anticipated to have a strong bias of one sequence (e.g., some of the somatic mutations in tumor samples). This threshold permitted detection of a low-abundance sequence in a region of low background noise.

The details of each of the five steps in the indel discovery process are described in Supplemental material.

In the analysis presented in this report, the low-abundance indel detection was applied only to the tumor samples but not to the germline samples. The remaining parameters are the same for all the data sets.

Analysis of three public data sets using CGWB

1. Analysis of somatic mutations in paired tumor–normal lung cancer samples

The lung cancer study is a typical mutation discovery project conducted in an individual research laboratory. Five candidate genes have been analyzed for this study by CGWB to date, two of which (*EGFR* and *STK11* [also known as *LKB1*]) were previously analyzed by a genotyping assay and manual review of sequence traces (Yang et al. 2005). In *EGFR*, exons 18–21 were resequenced in 219 paired tumor–normal tissues. The tumor tissues contain at least 50% tumor content. Our mutation analysis pipeline was able to find all 11 substitution variations (six somatic and five germline) as well as all seven indel mutations (six somatic and one germline, summarized in Table 1) that were previously discovered either by genotype assay or manual data review. One of the somatic indels initially found to represent a 9-bp deletion by genotype assay had an unusual alignment between its wild-type allele and the deletion allele. A follow-up analysis revealed that it was a complex mutation involving a 17-bp deletion coupled with

Table 1. Comparison of indel mutations discovered in tumor samples

Indel and flanking sequence[a]	Protein change[b]	Samples[c]	Indel signal[d]	PP6[e]
Tumor sequencing data of *EGFR* from Yang et al. 2005				
tcgctatcaa[ggaattaagagaagc/-]aacatctccg	746–750[ELREA/−]	It0006*, M119*	15%–24%	No
cgctatcaag[gaattaagagaagca/-]acatctccga	746–750[ELREA/−]	H14753*, It1064*, M46125*, M114*, H1650*, H1535	39%–43%	Yes
atcaaggaat[taagagaagcaacatctc/-]cgaaagccaa	747–753[LREATSP/S−]	H1501*	100%	No
gtttctgctt[tgctgtgtgggggtccatggct/-]ctgaacctca	Intron 19	M217*	18%–20%	No
tcgctatcaa[ggaattaagagaagcaa-/-catctccg]catctccgaa$	746–750[ELREA−/−IPA]	H11318*, M49	43%–45%	Yes
cagcgtggac[-/ggc]aacccccacg	770D[−/G]771N	M96*	13%–14%	No
cgtggacaac[-/aac]ccccacgtgt	770D[−/N]771N	M23*	22%	Yes
WIBR tumor sequencing data of *EGFR*				
tcgctatcaa[ggaattaagagaagc/-]aacatctccg	746–750[ELREA/−]	BIGRTSP0000001863	19%–25%	No
cagcgtggac[-/cagcta]aacccccacg	770D[−/QL]771N	BIGRTSP0000105936	33%–38%	Yes
ggccacagcc[a/-]gggggcggc^	Intron 10	BIGRTSP0000118086	34%	No
WIBR tumor sequencing data of *STK11*				
ggtgtcaggtgggggg[-/g]ctattggccc^	Intron 6	BIGRTSP0000104956	15%–19%	No
gcgactgtggcccccc[-/c]gctctctgac	281PPALX	BIGRTSP0000106586	29%	No

[a]The mutation is displayed with the following format [wild type/mutant]. The mutation noted with $ is a 17-bp deletion coupled with an 8-bp insertion. The two mutations noted with a carat (^) are the only ones that have not been reported in published literature.
[b]An in-frame protein deletion is displayed as [wild type/−], an in-frame protein insertion is displayed as [−/insertion], and a frame-shift mutation is displayed using the first amino acid that is altered with respect to the wild-type sequence followed by the translations in the altered reading frame. The stop codon is displayed as X.
[c]Samples noted with an asterisk (*) were originally identified by a genotype assay based on fragment size changes.
[d]The range of percentage of fluorescence signals representing the indel allele observed in sequencing traces.
[e]Whether the indel is found by polyphred version 6.0.2 (Bhangale et al. 2006).

an 8-bp insertion that replicates an 8-bp upstream sequence (Fig. 3, details in Supplemental material). An additional 12 variations were found in *EGFR*, 11 of which were verified by manual data review. The performance of the CGWB pipeline in the mutation analysis of *STK11* is similar to that in *EGFR*. The mutation profiles assembled using all variations identified by the CGWB pipeline are presented in Figure 4.

2. Reanalysis of indel polymorphisms in 152 genes

The second data set we analyzed is the resequencing data of candidate genes by SeattleSNPs group (Carlson et al. 2003). In this data set, indel polymorphisms were detected in an earlier study by extensive manual review (Bhangale et al. 2005, 2006), making it useful for measuring the accuracy of IndelDetector as well as for developing and testing the high-throughput data analysis capability of CGWB. We reanalyzed a total of 574,217 sequence traces of 152 genes and found 1659 indels and 42 STRPs. 1141 of the 1669 indels were found to match those that were manually identified by SeattleSNPs (Bhangale et al. 2005). Ninety-five percent of the matching indels were found by our method to have the same indel size as those reported by SeattleSNPs, the largest of which is a 155-bp deletion in *SFTPD* (dbSNP rs17882135). Of the 518 novel indels discovered exclusively by IndelDetector (Supplemental Table S1), 451 were found valid by manual review of sequence traces. To summarize the accuracy of IndelDetector, we calculated the true positive rate (TPR), which is the proportion of indels that our method successfully detects among the 1251 known indels manually identified by SeattleSNPs, as well as the false discovery rate (FDR), which is the proportion of invalid indels among all indels identified. In this 152-gene data set, IndelDetector has a FDR of 0.04 at a TPR of 0.91.

3. Analysis of germline mutations in familial ovarian cancer probands

This project aims to identify germline mutations in the coding region of candidate genes for ovarian cancer. The subjects used in mutation screening include probands with ovarian cancer who are *BRCA1*/*BRCA2* mutation-negative and have a family history of ovarian cancer. It is an ongoing study in its early stages. Here, we present early results for one gene, *BRD4*, which has been sequenced in 50 samples that include subjects from NCI's Family Cancer Registry and from the Gilda Radner Familial Ovarian Cancer Registry.

The automated CGWB analysis identified 18 variations, two of which, based on manual review, appear to be false positives. These were removed using our Cancer Genome Browser's genotype editing function (Fig. 5). Of the 16 legitimate variations summarized in Supplemental Table S2, 10 are novel SNPs not found in dbSNP.

Comparison of IndelDetector with polyphred in the analysis of indel polymorphism and somatic mutation

The accuracy of IndelDetector in the identification of indel polymorphisms and indel somatic mutations was compared with that of the recently published polyphred version 6 software (Bhangale et al. 2006). The test data sets for the polymorphism analysis include: (1) the four-gene test data used by Bhangale et al. to compare the performance of polyphred with that of InSNP, MutationSurveyor, and novoSNP; and (2) the 16-gene test data used by Bhangale et al. to evaluate the performance of polyphred. Known indels identified by Bhangale et al. were used to calculate TPR, and all the novel indels identified by IndelDetector were manually reviewed. IndelDetector has an FDR of 0.03 at a TPR of

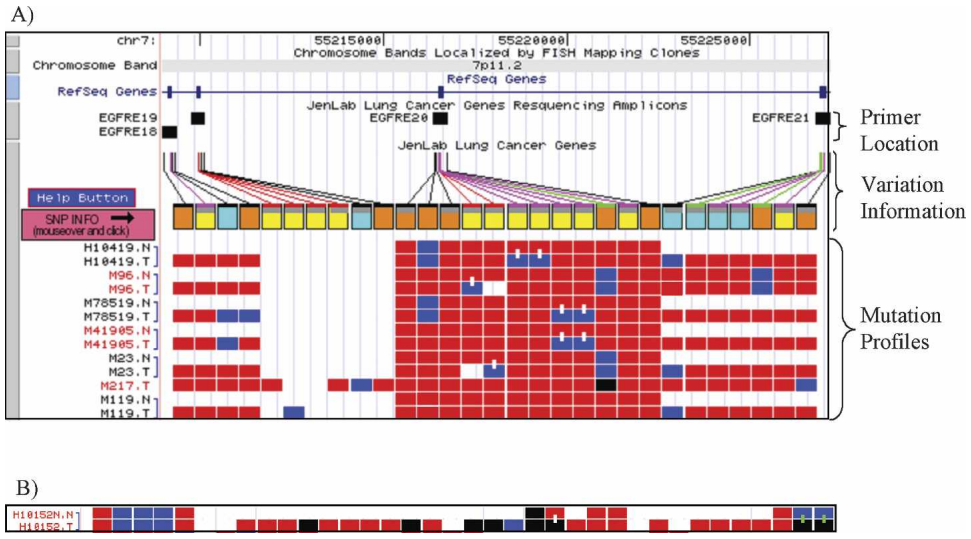

Figure 4. (*A*) Mutation profiles of *EGFR* exon 18–21 presented in the Cancer Genome Browser. The complete legend can be found by clicking "Help Button." In this view, the genomic locations of the primers for each exon are displayed at the *top*. The variation information is shown in the *middle*. The color of the line connecting the genomic location of a variation to the genotypes indicates whether a genetic change is silent (green), missense (magenta), in-frame indel (red), or intronic (black). The same color was applied on the first layer of the information block. A variation with a gray bar in the second layer represents a novel genetic change not found in dbSNP. The color in the remaining portion of the block indicates whether a variation is somatic (yellow), germline (orange), or found only in tumor samples with no matching normal samples (cyan). The genotypes of paired tumor–normal samples (labeled with extension .T and .N, respectively) are displayed *below*, representing homozygous major alleles (red boxes), homozygous minor alleles (blue boxes), heterozygous alleles (black boxes), or missing data (white boxes). Genotype difference in a paired tumor–normal sample is highlighted by a vertical line across the paired samples, representing somatic mutation (white) or allele loss (green). (*B*) The mutation profile of the paired normal–tumor sample of H10152 in *STK11*. There are two SNPs indicating allele loss (two green lines) and one somatic mutation (one white line) in this sample.

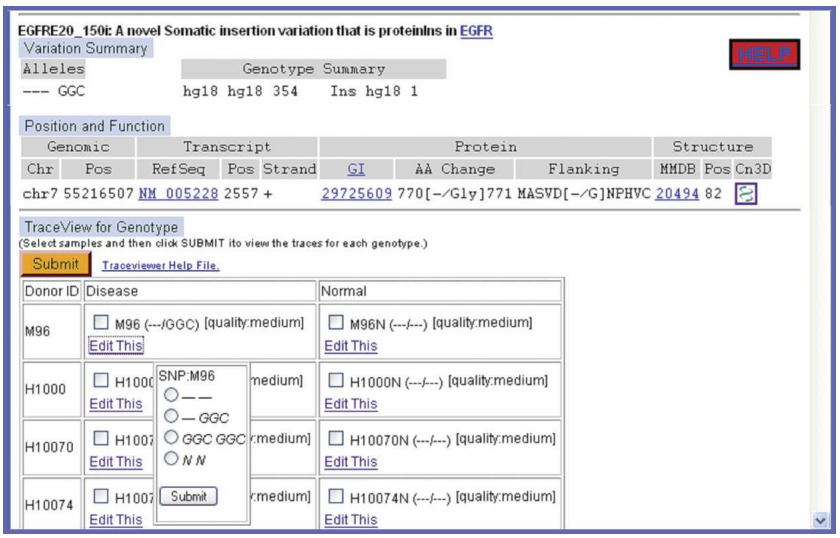

Figure 5. A variation report page for a project owner who logs on to the CGWB. The genotype editor is activated to allow manual edit of the computationally determined genotype.

0.97 for test data 1 (the reported accuracy of polyphred 6 is an FDR of 0.09 at a TPR of 0.95) and an FDR of 0.05 at a TPR of 0.92 for test data 2 (the reported accuracy of polyphred 6 is an FDR of 0.03 at a TPR of 0.90). Details are presented in Supplemental Tables S3 and S4 and Supplemental Figures S6–S8.

For the tumor somatic mutation analysis, we used genes *EGFR* and *STK11* since somatic indels in these two genes have been well characterized in previous studies (Sanchez-Cespedes et al. 2002; Lynch et al. 2004; Paez et al. 2004; Yang et al. 2005; Thomas et al. 2006), some of which employed technologies other than the fluorescence-based sequencing (Thomas et al. 2006; Yang et al. 2005). Sequence traces from our own laboratory (Yang et al. 2005) as well as those in the NCBI trace archive deposited by the WIBR/Broad Institute were analyzed. IndelDetector identified 12 unique indel mutations, 10 of which are previously published mutations that cause in-frame protein insertion/deletion or frame-shift truncation (Table 1). The two novel indels are intronic and were verified by manual review. Using the same data set, polyphred found five mutations, one of which was determined to be false-positive by follow-up manual review. For this data set, polyphred has a FDR of 0.2 at a TPR of 0.33.

Graphical user interface of CGWB

The Cancer Genome Browser (http://cgwb.nci.nih.gov) is an enhanced version of the UCSC Genome Browser (Kent et al. 2002). It provides an integrated view of somatic mutations, germline mutations, and copy number changes in paired tumor–normal samples. Figure 4 shows the graphical view of *EGFR* and *STK11* in the Cancer Genome Browser, which presents clinical mutation profiles in parallel with the reference human genome. The samples are sorted by the number of somatic alternations, which include somatic mutations and copy number changes (highlighted with white and green vertical bars, respectively), so that tumor samples with important mutagenesis events can be evaluated easily. The tumor sample shown in Figure 4B has both allele loss and somatic mutation in *STK11*, which fits the "two-hit" model proposed for a tumor suppressor gene. Clicking on a sample of interest launches a multiple-alignment trace viewer that displays all traces of the normal and the disease tissue.

Clicking on a variation of interest launches the variation report page, which shows its population frequency as well as descriptions of its effect on mRNA transcription and the protein coding (Fig. 5). For genes that have 3D structure records in NCBI's MMDB database, their variations' structural implication can be assessed using the Cn3D viewer (Hogue 1997). Currently, a total of 530 variations in CGWB have been mapped to MMDB. The variation report page also allows a user to review genotypes of multiple samples using our trace viewer and a project owner to manually edit the genotype data (Fig. 5).

CGWB variation database

The CGWB variation database supports variation definition, variation annotation, primer data, sample description, population frequency, manual edit, and project access control (Supplemental Fig. S9). The database loading process verifies data integrity, identifies novel variations (i.e., those variations that have not been found in dbSNP), and determines whether a variation is somatic or germline. A variation is considered "somatic" if its minor allele is found in tumor tissues but not in paired normal tissue. This determination is carried out independently for each project so that a somatic mutation in a tumor resequencing project may turn out to be a germline variation in an epidemiology study. Currently, the variation database stores 24,646 germline variations, 1046 somatic mutations, and 11,856 disease-only variations. The latter are found in tumor tissues for which there is no corresponding genotype data for paired normal samples. Forty percent of all variations in the database are novel.

Performance and availability

IndelDetector was written in C and currently runs on a UNIX/LINUX platform. Its performance is dependent on the number of the samples being sequenced as well as the indel frequency, but not the indel size. Processing 26 amplicons resequenced in 100 tumor samples on an HP DL585 with 28 Gb of memory took an average of 7 sec per amplicon for IndelDetector, while the average run time of polyphred was 21 sec per amplicon for the same data set.

IndelDetector is a component of the SNPdetector3 pipeline and can be obtained by anonymous ftp at ftp://ftp1.nci.nih.gov/pub/SNPdetector3.

Discussion

A comprehensive and integrated bioinformatics platform can dramatically improve the accuracy and the productivity of mutation analysis. The examples described here demonstrate that CGWB is useful both for in-depth analysis of mutation profiles in an individual research laboratory and for large-scale analysis of high-throughput data.

In the analysis of somatic indels in tumor samples, the sensitivity of our new algorithm, IndelDetector, was much higher than that of the recently published polyphred algorithm (Table 1). Polyphred was designed to identify indel polymorphisms of

which the wild-type and the indel sequences are expected to have equimolar representation in the PCR product. Our analysis shows that a somatic indel can have much lower abundance (Table 1), which explains the lack of sensitivity of polyphred in the somatic mutation analysis presented here. The current implementation of polyphred can identify indel polymorphisms only up to 30 bp long; IndelDetector does not have this limitation. The longest validated indel reported in this study is a 155-bp polymorphic deletion previously identified by manual data review.

In *EGFR*, IndelDetector was able to find all indel sites but missed two tumor samples (98M and 0059T) originally found to have a 15-bp deletion based on fragment size reduction using a genotype assay. These two somatic deletions were estimated to have 2%–5% frequency by the genotype assay, and their sequence traces are indistinguishable from those of the tumor samples without the indel mutations. This suggests that the failure to identify somatic indels in these two samples is attributable to limitation of sequencing technology rather than the computational algorithm. The lowest frequency of somatic indels identified by IndelDetector is 13% while the highest is 100%, found in an 18-bp homozygous deletion of *EGFR* (Table 1) formerly described as the Del-4 mutation (Paez et al. 2004). The Del-4 mutation was found to be heterozygous in different tumor samples analyzed in two previous studies (Paez et al. 2004; Thomas et al. 2006), one of which showed that it had a very low frequency (3%) and was detectable only by microreactor-based pyrosequencing (Thomas et al. 2006) but not by fluorescent sequencing. One possible interpretation for the homozygous genotype observed in our sample is that additional somatic alteration events, such as the copy number change or loss of heterozygosity, resulted in the great reduction of the wild-type allele in this tumor.

In addition to the public data sets presented in this report, CGWB is hosting several ongoing projects with unpublished data that are accessible only to the project owners. Currently, 60% of the novel variations detected by CGWB belong to confidential projects. These are expected to be made publicly accessible upon completion of the projects. One such example is the resequencing of candidate regions identified by the Breast Cancer Association Consortium (Breast Cancer Association Consortium 2006). A data submission portal is under development to facilitate submission of confidential projects for individual research laboratories.

Methods

Mutation detection pipeline

Prior to running IndelDetector, base calls of each read, quality scores, and primary and secondary peak information are computed using the program *phred* (Ewing et al. 1998). Each read sequence is aligned to the reference sequence using the program *sim* (Huang et al. 1990), and repetitive regions are identified using the program *ptrfinder* (Collins et al. 2003). This pipeline was previously developed for SNP analysis; details are presented in Figure 1 of Zhang et al. (2005).

Analysis of SeattleSNPs data

To create the gene list for our reanalysis of SeattleSNPs data, three sets of the accessions were obtained from NCBI and SeattleSNPs: (1) 185 reference accessions in the trace information of NCBI's trace archive, obtained by querying reads with the technology defined as "resequencing;" (2) 259 GenBank accessions submitted by SeattleSNPs; and (3) 218 accessions in the GenBank flatfiles downloaded from the SeattleSNPs Web site (http://pga.gs.washington.edu/finished_genes.html). A total of 152 accessions were common to all three lists, and the genes encoded in these genomic sequences were reanalyzed in this study.

We discovered that there were three disjoint sample sets in the SeattleSNPs data, so we created a project for each subset. The three projects were named SeattleSNPs1, SeattleSNPs2, and EGP. The first two corresponded to SeattleSNPs donor panels 1 and 2 (http://pga.gs.washington.edu/platemaps.html#table1) and the third was based on the donors from the EGP panel 1 (http://egp.gs.washington.edu/samplepop.html); the corresponding genes for each project can be viewed on the CGWB Web site.

Traces were batch-downloaded from NCBI's trace archive and converted into standard SCF format using their "query_tracedb" and "rcf2scf" programs. Our mutation/polymorphism detection pipeline is designed to analyze one amplicon at a time to ensure computing efficiency and accuracy. However, we found that ~20% of primer pair annotations obtained from the trace archive did not match the corresponding traces. Therefore, we remapped each trace to an amplicon using the process described in Supplemental material.

To compare indels found by IndelDetector with those reported in SeattleSNPs data, indels in SeattleSNPs data were first mapped to our template genomic sequence by aligning their 50-bp flanking sequence (stored in the *.snpcontext.fasta file in the SeattleSNP data download) to the template genomic sequence. SeattleSNPs indels that failed to map to any amplicon were excluded. We added a 10-bp buffer to compare the overlap in the locations of these two data sets because the location of an indel in a repetitive region (polynucleotide repeat or STR) is arbitrary. IndelDetector always reports the rightmost location as the indel position, which can differ from the location reported by SeattleSNPs even if the two represent the same variant.

In our manual review of indel changes, indel locations were manually verified so that a discrepancy in representing an indel in a repetitive region was not counted as a difference in indel analysis. Redundant indels that arise from a subtle alignment discrepancy in overlapping amplicons were consolidated to ensure uniqueness. The sequence alignments were reviewed using the program HapScope (Zhang et al. 2002), while the sequence quality and trace chromatogram were reviewed using the program consed (Gordon et al. 1998).

Test data sets for comparison with polyphred

Of the four genes used for assessing the accuracy of polymorphic indel detection of multiple algorithms in the previous study (Bhangale et al. 2006), *ACTB* and *ALAD* have traces from both the NCBI's trace archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi) and the polyphred Web site (http://droog.mbt.washington.edu/poly_data_v6.html). The trace files of *F2RL2* and *SERPINE1* were not found in NCBI's trace archive by direct query or by blast search using their genomic sequences, so their traces from the polyphred Web site were used instead for our analysis. All traces for the 16-gene test data set were obtained from the polyphred Web site. Traces were mapped to the corresponding amplicons following the same protocol as that described in the section "Analysis of SeattleSNPs data."

The known indels used for assessing the accuracy of IndelDetector were obtained from the genotype type files (*.prettybase.txt) downloaded from the polyphred web page. The numbers of known indels for the four-gene test data and the 16-gene test data are 33 and 170, respectively.

For somatic mutation analysis we obtained traces from NCBI's trace archive by querying the center name WIBR and the project name TSP in October 2006.

Mapping of a variation to a 3D structure

We mapped human RefSeq proteins to the PDB records by running a blast search and retained the best protein hit with ≥95% identity. The start and the stop positions of RefSeq proteins that matched PDB records were loaded into the variation database. A protein variation that resides in such an interval is then automatically mapped to the 3D structure.

Acknowledgments

We thank Ms. Mary Johnson for her support in deploying the production software. We thank Drs. Robert Clifford and Peisen Zhang for critical review of the manuscript.

References

Bhangale, T.R., Rieder, M.J., Livingston, R.J., and Nickerson, D.A. 2005. Comprehensive identification and characterization of diallelic insertion–deletion polymorphisms in 330 human candidate genes. *Hum. Mol. Genet.* **14:** 59–69.

Bhangale, T.R., Stephens, M., and Nickerson, D.A. 2006. Automating resequencing-based detection of insertion–deletion polymorphisms. *Nat. Genet.* **38:** 1457–1462.

Breast Cancer Association Consortium. 2006. Commonly studied single-nucleotide polymorphisms and breast cancer: Results from the Breast Cancer Association Consortium. *J. Natl. Cancer Inst.* **98:** 1382–1396.

Carlson, C.S., Eberle, M.A., Rieder, M.J., Smith, J.D., Kruglyak, L., and Nickerson, D.A. 2003. Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat. Genet.* **33:** 518–521.

Collins, J.R., Stephens, R.M., Gold, B., Long, B., Dean, M., and Burt, S.K. 2003. An exhaustive DNA micro-satellite map of the human genome using high performance computing. *Genomics* **82:** 10–19.

Couch, F.J. and Weber, B.L. 1996. Mutations and polymorphisms in the familial early-onset breast cancer (BRCA1) gene. Breast Cancer Information Core. *Hum. Mutat.* **8:** 8–18.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8:** 175–185.

Gordon, D., Abajian, C., and Green, P. 1998. Consed: A graphical tool for sequence finishing. *Genome Res.* **8:** 195–202.

Hogue, C.W. 1997. Cn3D: A new generation of three-dimensional molecular structure viewer. *Trends Biochem. Sci.* **22:** 314–316.

Huang, X.Q., Hardison, R.C., and Miller, W. 1990. A space-efficient algorithm for local similarities. *Comput. Appl. Biosci.* **6:** 373–381.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res.* **12:** 996–1006.

Knudson Jr., A.G. 1971. Mutation and cancer: Statistical study of retinoblastoma. *Proc. Natl. Acad. Sci.* **68:** 820–823.

Lynch, T.J., Bell, D.W., Sordella, R., Gurubhagavatula, S., Okimoto, R.A., Brannigan, B.W., Harris, P.L., Haserlat, S.M., Supko, J.G., Haluska, F.G., et al. 2004. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.* **350:** 2129–2139.

Manaster, C., Zheng, W., Teuber, M., Wachter, S., Doring, F., Schreiber, S., and Hampe, J. 2005. InSNP: A tool for automated detection and visualization of SNPs and InDels. *Hum. Mutat.* **26:** 11–19.

Neuhausen, S., Gilewski, T., Norton, L., Tran, T., McGuire, P., Swensen, J., Hampel, H., Borgen, P., Brown, K., Skolnick, M., et al. 1996. Recurrent BRCA2 6174delT mutations in Ashkenazi Jewish women affected by breast cancer. *Nat. Genet.* **13:** 126–128.

Paez, J.G., Janne, P.A., Lee, J.C., Tracy, S., Greulich, H., Gabriel, S., Herman, P., Kaye, F.J., Lindeman, N., Boggon, T.J., et al. 2004. *EGFR* mutations in lung cancer: Correlation with clinical response to gefitinib therapy. *Science* **304:** 1497–1500.

Sanchez-Cespedes, M., Parrella, P., Esteller, M., Nomoto, S., Trink, B., Engles, J.M., Westra, W.H., Herman, J.G., and Sidransky, D. 2002. Inactivation of *LKB1/STK11* is a common event in adenocarcinomas of the lung. *Cancer Res.* **62:** 3659–3662.

Shih, C., Padhy, L.C., Murray, M., and Weinberg, R.A. 1981. Transforming genes of carcinomas and neuroblastomas introduced into mouse fibroblasts. *Nature* **290:** 261–264.

Sjoblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N., et al. 2006. The consensus coding sequences of human breast and colorectal cancers. *Science* **314:** 268–274.

Thomas, R.K., Nickerson, E., Simons, J.F., Janne, P.A., Tengs, T., Yuza, Y., Garraway, L.A., LaFramboise, T., Lee, J.C., Shah, K., et al. 2006. Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nat. Med.* **12:** 852–855.

Vogelstein, B. and Kinzler, K.W. 2004. Cancer genes and the pathways they control. *Nat. Med.* **10:** 789–799.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., et al. 2006. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **34:** D173–D180.

Yang, S.H., Mechanic, L.E., Yang, P., Landi, M.T., Bowman, E.D., Wampfler, J., Meerzaman, D., Hong, K.M., Mann, F., Dracheva, T., et al. 2005. Mutations in the tyrosine kinase domain of the epidermal growth factor receptor in non-small cell lung cancer. *Clin. Cancer Res.* **11:** 2106–2110.

Zhang, J., Rowe, W.L., Struewing, J.P., and Buetow, K.H. 2002. HapScope: A software system for automated and visual analysis of functionally annotated haplotypes. *Nucleic Acids Res.* **30:** 5213–5221.

Zhang, J., Wheeler, D.A., Yakub, I., Wei, S., Sood, R., Rowe, W., Liu, P.P., Gibbs, R.A., and Buetow, K.H. 2005. SNPdetector: A software tool for sensitive and accurate SNP detection. *PLoS Comput. Biol.* **1:** e53.

Received September 18, 2006; accepted in revised form March 21, 2007.