HHS Public Access

Author manuscript

Nat Med. Author manuscript; available in PMC 2016 August 30.

Published in final edited form as:

Nat Med. 2016 January ; 22(1): 97–104. doi:10.1038/nm.4002.

Systematic Discovery of Complex Indels in Human Cancers

Kai Ye[1,2], **Jiayin Wang**[1], **Reyka Jayasinghe**[1,3], **Eric-Wubbo Lameijer**[4], **Joshua F. McMichael**[1], **Jie Ning**[1], **Michael D. McLellan**[1], **Mingchao Xie**[1,3], **Song Cao**[1], **Venkata Yellapantula**[1,3], **Kuan-lin Huang**[1,3], **Adam Scott**[1,3], **Steven Foltz**[1,3], **Beifang Niu**[1], **Kimberly J. Johnson**[5], **Matthijs Moed**[4], **P. Eline Slagboom**[4], **Feng Chen**[3,6], **Michael C. Wendl**[1,2,7], and **Li Ding**[1,2,3,6,#]

[1]McDonnell Genome Institute, Washington University in St. Louis, St. Louis, MO, USA [2]Department of Genetics, Washington University in St. Louis, St. Louis, MO, USA [3]Department of Medicine, Washington University in St. Louis, St. Louis, MO, USA [4]Leiden University Medical Center, Leiden, the Netherlands [5]Brown School Master of Public Health Program, Washington University in St. Louis, St. Louis, MO, USA [6]Siteman Cancer Center, Washington University in St. Louis, St. Louis, MO, USA [7]Department of Mathematics, Washington University in St. Louis, St. Louis, MO, USA

Abstract

Complex indels are formed by simultaneously deleting and inserting DNA fragments of different sizes at a common genomic location. Here, we present a systematic analysis of somatic complex indels in the coding sequences of over 8,000 cancer cases using Pindel-C. We discovered 285 complex indels in cancer genes (e.g., *PIK3R1*, *TP53*, *ARID1A*, *GATA3*, and *KMT2D*) in approximately 3.5% of cases analyzed; nearly all instances of complex indels were overlooked (81.1%) or mis-annotated (17.6%) in 2,199 samples previously reported. In-frame complex indels are enriched in *PIK3R1* and *EGFR* while frameshifts are prevalent in *VHL*, *GATA3*, *TP53*, *ARID1A*, *PTEN*, and *ATRX*. Further, complex indels display strong tissue specificity (e.g., *VHL* from kidney cancer and *GATA3* from breast cancer). Finally, structural analyses support findings of previously missed, but potentially druggable mutations in *EGFR*, *MET*, and *KIT* oncogenes. This study indicates the critical importance of improving complex indel discovery and interpretation in medical research.

[#]Corresponding Author: Li Ding, Ph.D, lding@genome.wustl.edu.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

CONTRIBUTIONS

L.D. designed and supervised research. L.D. and K.Y. led data analysis and K.Y., J.W., M.D.M., M.X., R.J., S.C., A.S., V.Y., K-L.H., K.J.J., and M.C.W. performed data analysis. K.Y. led methods development, with E.W.L., M.M., B.N., and E.S. contributed code to Pindel-C. and K.Y., R.J., S.C., and S.F. developing the QC code. J.F.M., K.Y., and L.D. prepared figures and tables. F.C. and J.N performed experimental validation. K.Y., M.C.W., and L.D. wrote the manuscript.

INTRODUCTION

Next-generation sequence technologies have fueled genetic research and provided unprecedented means of building an increasingly comprehensive catalog of single nucleotide variants (SNVs), small insertions and deletions (indels), and structural variants. Although cataloging these kinds of common events has continued at a brisk pace, complex indel discovery has progressed very little since the transition from Sanger sequencing to next-generation sequencing technologies.

In 2007, the first diploid genome was sequenced using Sanger dideoxy technology, revealing thousands of complex germline indels [1]. The 1000 Genomes Project recently reported 664 germline complex germline indels in NA12878 [2]. In the Genome of the Netherlands project, 291 *de novo* indels were discovered, of which 14 (4.8%) were complex indels[3]. Roerink *et al*.[4] reported complex indels in *C. elegans* strains lacking translesion synthesis polymerases and more recently, described a G-quadruplex structure induced mutagenesis characterized by occasional presence of template insertions[5]. A synthesis-dependent microhomology-mediated end joining (SD-MMEJ) model was proposed to explain the formation mechanism [6]. Complex indels have also been detected in cancer cases using traditional technologies. For example, to detect exon 19 deletions in *EGFR*, a fragment length analysis was first performed to select potential carriers and the entire exon 19 of *EGFR* was PCR-amplified and sequenced on ABI sequencer[7–9].

Since the introduction of next-generation sequencing instruments, complex indels have largely been neglected due to the lack of effective tools for mapping and detecting in short sequence reads. We are aware of one report of three somatic complex indels within *CALR* in cancer samples using next-generation sequencing[10]. We scanned the published mutation annotation files (MAFs) from ten TCGA marker papers[11–20], finding 1 *FLT3* complex indel in Acute myeloid leukemia (AML) and 5 in Ovarian cancer (*CNGA1*, *CCDC136*, *MFAP3L*, *SLC13A1* and *TP53*). Here, we report an algorithm for systematically detecting complex indels from next generation sequence data. It reveals not only a surprising prevalence of these events in human cancer, but also the potential mechanisms underlying their formation, as well as their impact on gene function. Finally, we highlight the discovery of clinically relevant complex indels and their impact on treatment strategies.

RESULTS

Implementation of Pindel-C and performance evaluation

We developed a novel module within Pindel, called Pindel-C, to specifically search for co-occurring insertion and deletion events, namely complex indels (Fig. 1a) (Methods). We examined the sensitivity and specificity of Pindel-C using three datasets. First, we randomly generated a ten Mbp reference genome. Then we simulated 1,000 complex indels with deletion and insertion size ranging from 1 to 1,000 bp, but having different values. In addition, we generated two sets of 30× coverage with distinct read lengths and insert sizes and used BWA-aln and BWA-mem for alignment (Methods). In general, the larger insertions could be detected with longer reads, although the power to detect deletions is rather consistent (Supplementary Fig. 1). When read length increases from 100 bp to 250 bp and

BWA-aln is applied, the maximum insertion size changes from 69 bp to 218 bp. For complex indels within the detection limit of read length (Methods), we observed 87.93% sensitivity for read length 100 bp and 70.00% for 250 bp. Pindel-C overall performs better on BWA-aln produced BAM files than BWA-mem (Supplementary Fig. 1).

Second, we introduced 1,128 synthetic complex indels on chromosome 1 of Craig Venter's genome and simulated 100× Illumina paired-end data (Methods and Supplementary Table 1). Pindel-C detected 541 of them (48% sensitivity) (Supplementary Table 2), while neither GATK nor VarScan captured any correctly (Methods). The latter are standard bioinformatics tools, though neither is tuned specifically for complex indels. Pindel-C mis-called 88 events as simple indels, suggesting a false discovery rate of about 14% (Fig. 1b). As the sizes of deletion and insertion increase, the detection sensitivity drops dramatically, which is largely expected for short read data (Supplementary Fig. 2).

Third, we experimentally examined Pindel-C performance for detecting complex indels in the COLO 829 cell line data. Specifically, we applied it to 40× data of COLO 829 melanoma cells (CRL-1974) and 32× coverage data of the Epstein-Barr virus- transformed control B lymphoblast cells from the same individual (CRL-1980) reported by Pleasance at al. [21]; after automated filtering, we obtained 17 somatic and 2,213 germline complex indels. A total of 75 events (all 17 somatic and 58 randomly selected germline) were selected for experimental validation (Methods). We successfully obtained PCR products and Sanger sequencing data for 51 of them (12 somatic and 39 germline). Our subsequent analysis demonstrated validation rates for somatic and germline events of 75% (9 of 12) and 100% (39 of 39), respectively. It is worth noting that the CRL-1974 batch we purchased from ATCC is not the same one sequenced by Pleasance et al.[21]; therefore the three somatic complex indels detected in the original sequencing data may not be present in our validation cell line (Supplementary Table 3 and Supplementary Information). This exercise indicates the need for purpose-designed software, as complex indels present unique challenges for detection.

Exome-wide landscape of complex indels

To obtain a more global view on complex indels across the entire coding sequence, we processed 8,060 tumor and matched normal pairs across 22 cancer types. Our initial analysis showed that excessive numbers of apparently somatic complex indels in some samples were actually attributable to sequence artifacts, such as 1 bp indels at fixed read positions, regardless of genomic location. This observation prompted a quality-control (QC) examination of the BAM files to compute the percentage of reads carrying such sequence artifacts per sample (Methods). A histogram of these percentages (Supplementary Fig. 3) suggested a 20% threshold for removing BAM files enriched with read artifacts (Methods). The remaining 4,742 cases were then deemed suitable for exome-wide analysis.

Among these 4,742 samples, Pindel-C predicted 2,948 raw somatic complex indels with variant allele frequency higher than 10%. All were manually reviewed using IGV[22], which identified 1,680 predictions having read support from both strands (Supplementary Table 4). It is not surprising that the number of complex indels is generally low in coding regions (Fig. 2a), although there are a few samples carrying significantly more instances. We investigated genes possibly contributing to these elevated numbers using MuSiC[23] (Methods), but did not

detect any substantial correlation (Supplementary Table 5). Whole genome sequencing data will be required to obtain accurate complex indel mutation frequencies across sample sets.

We annotated translational effects of 1,680 putative somatic complex indels and examined which genes are frequently affected by somatic complex indels (Methods). We noticed that 895 samples harbored one or more complex indels from 1,493 genes. Notably, the most frequently affected genes are largely well-known cancer genes. For example, 15 somatic complex indels were detected in *PIK3R1*. Other top genes were, *TP53*, *ARID1A*, *GATA3*, and *KMT2D* (Fig. 2b). This result suggests that somatic complex indels in cancer genes are likely under positive selection during tumorigenesis

We also evaluated the lengths of these 1,680 somatic complex indels, finding that deleted sequences are generally longer, but that there is no obvious correlation between deletion and insertion lengths. Insertion frequency decreases as insertion size increases. The proportions of insertion for 1 bp, 2 bp and 3 bp are 58.7%, 20.4%, 8.9%, respectively. The majority of the deletions are 2 bp in length (41.3%) while 1 bp and 3 bp are 8.8% and 18.8% respectively (Fig. 2c).

Frequency and mechanism of complex indels in cancer genes

To overcome the effects of sequencing artifacts discussed above, we used a multi-step strategy of initial discovery, manual review, re-genotyping, and DNA and RNA-seq based validation to curate a high confidence, comprehensive set of complex indels across the entire 8,060 sample set.

We compiled a list of 624 cancer genes based on the literature [24–31] (Methods and Supplementary Table 6) and found that 140 of these harbor 285 somatic complex indels in the samples analyzed (Supplementary Table 7). We examined whole genome and RNA-seq data generated within TCGA for the above sites. We found that they are largely supported if coverage is reasonably high (Supplementary Table 8 and Supplementary Information). In examining local alignments around the breakpoints and cross-checking against TCGA reports[11–20], we found that 13 of them were previously reported as substitutions, despite adjacent gapped alignments in the primary alignment result. Interestingly, 83 of them are within 100 bp of another complex event, indicating non-random distribution (geometric probability test, Methods). We argue that the local sequence context might be prone to double strand DNA breaks or under selection for cancer phenotype, elevating likelihood of complex indels. Alternatively, it is possible that the critical spots to disturb or activate key cancer genes are rather limited and these events are under selection for enrichment.

From the well-curated complex indels in cancer genes, we attempted to further search for the origin of the inserted sequences. Because the inserted sequences are relatively short, we searched the local flanking 50 bp sequences for similarities with insertions greater than 4 bp. This explained 32 of the complex indels. We propose a classification scheme (Fig. 3) with 12 classes detected in the 32 sites. Direct and reverse copies of the 5′ or 3′ flanking sequences were most common classes, representing 37.5% and 31.3% of the cases, respectively. Those single direct or inverted copies of short fragments of flanking sequences were considered to be instances of loop-out and snap-back SD-MMEJ, a model originated in

a *C. elegans* study[6]. In addition, we also discovered that one third of the template insertions originated from various combinations of two origins (Fig. 3). In the mechanism illustrated by Ref R & 5R (Fig. 3), part of the deleted sequence is inserted as a reverse complement plus a reverse complementary copy of the 5′ flanking sequence. All 12 formation mechanisms were observed in these 32 somatic complex indels. Other mechanisms might be discovered with additional samples or whole genome sequences.

Tissue specificity and functional features of complex indels

We observed 21 genes (*ALK, APC, ARHGAP35, ARID1A, ATRX, EGFR, EPHA2, FAT1, GATA3, KEAP1, LRP1B, MAP3K1, MET, NF1, PBRM1, PIK3R1, PTEN, RB1, SETD2, TP53* and *VHL*) with complex indels in at least three cancer cases. The majority (17 out of 21) are tumor suppressors. *PIK3R1* ranks first, with a remarkable 20 mutations, 16 of which are in UCEC. More than half of these result in in-frame mutations, which is consistent with the in-frame simple indels more typically found in this gene. The next four most frequent genes, *TP53, ARID1A, PTEN* and *ATRX* are not specific to one cancer type. Among the most frequent 21 genes, there are only three oncogenes (*EGFR, ALK, MET*) harboring somatic complex indels. Functionally, 8 of the 21 genes (*PIK3R1, TP53, PTEN, FAT1, RB1, APC, ALK, MAP3K1*) are related to cell growth, differentiation, proliferation, and movement. There are also five genes (*EGFR, NF1, GATA3, MET* and *EPHA2*) in either transcription factor or signal transduction pathways, four (*ATRX, ARID1A, PBRM1, SETD2*) related to chromatin structure, and three related to either energy or oxidant response.

Some of the genes have cross-cancer relevance, for example somatic complex indels in *TP53* appear in nine cancer types (Fig. 4). Others show more specificity, two examples being the eight *EGFR* in-frame somatic complex indels (likely activation mutations in oncogene) in lung adenocarcinoma (LUAD) and seven frameshift loss-of-function somatic complex indels of tumor suppressor *VHL* in kidney renal clear cell carcinoma (KIRC). Given the appearance of in-frame clusters within some cancer-gene combinations, we sought to determine whether any of these were more prevalent than what could be explained by chance. We calculated a background in-frame rate of about 0.103 from 1,680 exome-wide complex indels and used this as Bernoulli estimator for hypothesis testing of in-frame vs frame-shift under a binomial probability model (Methods). We found four groupings that were significant: *EGFR* in LUAD (FDR = 10^{-7}), *PIK3R1* in multiple cancer types and in uterine corpus endometrial carcinoma (UCEC) specifically (respective FDRs of 3×10^{-7} and 2×10^{-5}), and *TP53* in multiple cancers (FDR = 0.07). These observations are consistent with previous discoveries and underscore the importance of these three genes in tumorigenesis. The seven events in *VHL* in KIRC were all frameshifts as expected, but this was not significant (P-value = 0.47) in light of the high frameshift background rate.

Timing of the emergence of complex indels

Variant allele fractions (VAFs) of some somatic complex indels appeared higher than those of other simple forms of variant in given samples (Supplementary Table 9). We sought to determine whether there were any cancer-gene combinations in which these differences were statistically significant (Methods). However, because the indel census is typically lower than

for SNVs, statistical power is a concern. In particular, the data show an average of five simple variants per complex indel over the samples we examined. In general, this rules out the testing of singletons in favor of combinations of samples from a given cancer type having complex indels in the same gene. The exclusion process ultimately identified six cancer-gene combinations for testing (Table 1).

After correcting for multiple tests, we found *EGFR* in LUAD to show significantly higher VAFs for complex indels vs simple variants than other genes (FDR ≈ 4%), with the average VAF values differ here by almost 40%. BRCA-*TP53* and KIRC-*VHL* have VAF differences of 23% and ~10%, respectively, suggesting higher complex indel VAFs, but these did not reach significance. However, it seems likely that more data would confirm BRCA-*TP53* and KIRC-*VHL* significance in light of comparing these combinations to KIRC-*PBRM1*. Specifically, the two KIRC cases show comparable VAF differences, but the greater amount of data in KIRC-*VHL* increases statistical power with a *P*-value about half of that for KIRC-*PBRM1*, a trend which would likely continue for all three combinations with more data. Conversely, the larger amount of data, coupled with much smaller VAF difference for UCEC-*PIK3R1*, firmly indicates no considerable difference for complex versus simple indels.

Druggable complex indels supported by structure analysis

We also compared the numbers of newly discovered somatic complex indels with somatic simple indels in ten genes reported in TCGA marker papers[11–20]. These new indels increase the somatic indel census by 10%. For *EGFR*, census increased by a remarkable 25% (Fig. 4b). We also noticed that somatic complex indels are spatially distributed in tumor suppressors, but concentrated within local regions in oncogenes. This phenomenon is not surprising because it is easier to disrupt protein function given multiple locations. Conversely, activating a protein or boosting its intrinsic activity often requires a specific disturbance of the protein structure by adding or removing a few residues with in-frame variations. For example, in *EGFR*, we detected four distinct somatic complex indels affecting residues from 746 to 751 and two of them are recurrent. When visualizing the variants in the *EGFR* 3D structure 1M17 (Protein Data Bank, PDB), all of them are on the flexible loop, which is part of the ATP binding pocket. The *EGFR* inhibitor Erlotinib is co-crystalized with *EGFR* in PDB 1M17 and we noticed that Erlotinib contacts the loop directly with multiple somatic complex indels detected in our study. The four distinct somatic complex indels are coded by exon 19 of *EGFR* and removed six, four, six and eight residues, respectively. Consequently, we hypothesize that the functional impact of those mutations might be similar to the frequent exon 19 deletion. If so, our newly discovered somatic *EGFR* complex indel mutants may also exhibit increased and sustained phosphorylation of *EGFR* and other *ERBB* family proteins constitutively, selectively activating *AKT* and *STAT* signaling pathways that promote cell survival[32]. In addition, it has been reported that exon 19 deletion mutants respond well to Erlotinib and Gefitinib, with response rates >62% in various clinical trials (http://www.mycancergenome.org).

We detected two in-frame complex indels (Q556_E561delinsR and I563_E572delinsRF) in *KIT*, from TCGA SARC and SKCM cancer cases. Similar to *EGFR*, the two *KIT* mutations

are also in the kinase domain and at the same loop region compared to *EGFR*, interacting directly with the inhibitor PLX647 in the 3D structure, 4HVS. Thus we hypothesize that these two in-frame complex indels might cause the constitutive activity of the kinase. It has been reported that most *KIT* mutations in melanoma respond well to Imatinib, Sunitnib, and Sorafenib. Careful examination of *KIT* drug response results shows that our complex indels largely overlap with the in-frame mutations Del554–559 and Del556–572 and thus might also be sensitive to all three drugs.

DISCUSSION

Development and application of Pindel-C has led to the discovery of a substantial number of somatic complex indels overlooked by earlier studies[11–20] and many of these are likely to be driver mutations in the cancer samples from which they originated. Although the absolute numbers of such indels in an individual's genome might be smaller than those of somatic SNVs or simple indels, some complex indels are present at high VAFs in key cancer genes and originate in founding clones. These findings collectively suggest that the complex indel is an important factor in diseases like cancer than perhaps previously appreciated.

This study used exome sequence data from various cancer types, which essentially precludes the discovery outside of coding regions. We did not observe any large complex indels, even though Pindel-C can identify such events, in principle. Germline complex indels in cancer data are also largely unexplored but worthy of further investigation.

We designed and implemented Pindel-C for detecting somatic complex indels in cancer data. Our systematic QC identified a fraction of samples with sequence artifacts described in the Methods section. To obtain accurate complex indels in an automated fashion, we omitted samples with an excessive number of sequence artifacts. The QC script and the automated variant filtering procedure have been deposited at GitHub. Our analysis of TCGA data identified several druggable mutations in *EGFR* and *KIT* with in-frame complex indels. In the era of precision medicine, it will be critical to capture all druggable mutations including previously overlooked somatic druggable complex indels in cancer patients.

ONLINE METHODS

Data analyzed

We procured 8,060 samples from 22 distinct cancer types for our analysis. The largest cohort, BRCA, contains 990 samples and the smallest, KICH, has 66 samples, with an average of 366 samples per cancer type. All 8,060 samples have exome sequence data from tumor and matched normal, with average coverage above 100X. We curated the reported somatic indels from all previously published TCGA marker papers [11–20] and identified one complex indel in AML and five in OV. Those six somatic complex indels were initially discovered as simple variants but revealed as complex from Sanger sequencing result.

BAM QC for exome-wide analysis

In our preliminary runs of Pindel-C, we noticed excessive numbers (more than 10k) of somatic complex indel calls in a subset of samples. Using IGV [22] we manually reviewed a

random subset of those calls in 10 offending samples and found specific sequence artifacts, including extensive soft clipping and alignment gaps at fixed positions of the reads, regardless of genomic position. We subsequently implemented a BAM QC script to identify such sequence artifacts. All reads were scanned individually to count the total number, as well as the number of reads carrying non-M characters in the CIGAR string. We discarded BAM files if excessive numbers of indel carrying reads (≥20%) were detected (Supplementary Fig. 1). The BAM QC script has been deposited at https://github.com/ding-lab/VariantQC and will be merged with Pindel-C later.

Simulation and sensitivity comparison

We simulated Illumina sequence data containing complex indels of Craig Venter's genome to test Pindel-C, GATK, and VarScan for detection sensitivity. First, we examined the complex indel variants of the Venter genome (characterized with ~800 bp Sanger reads in the original discovery paper) on chromosome 1 and removed any that could be classified as simple indels or that resided within low complexity regions. This step furnished 1,128 complex indels, which were then introduced into the chromosome 1 sequence of human build hg18 (Supplementary Table 1). Then we used wgsim (github.com/lh3/wgsim v0.3.1-r13) to generate 100× Illumina paired-end synthetic sequence data, with 500 bp insert size and 100 bp read lengths. These sequence data were aligned with BWA (0.6.1-r104) using its paired-end module. This setting allows us to test whether a tool is able to capture complex indels at all, given sufficient coverage of data. Pindel-C (v0.2.5a7) was run with the same settings as those used for the exome sequence data. We also ran VarScan (v2.3.9, 06/2015) and GATK (version 2.4-46-gbc02625, 07/2015) UnifiedGenotypor on this same simulated problem. The command line we used for GATK is "java -jar GenomeAnalysisTK.jar -T HaplotypeCaller -R hg18.fa -I aln.bam --genotyping_mode DISCOVERY -stand_emit_conf 10 -stand_call_conf 30 -o aln.bam.vcf". The VarScan command line is "samtools mpileup -f hg18.fa aln.bam | java -jar VarScan.v2.3.9.jar pileup2indel > aln.bam.varscan". Neither VarScan nor GATK captured any of the Venter complex indels that were inserted.

Sanger Sequencing confirmation of complex indels

COLO 829 melanoma cells (CRL-1974) and the Epstein-Barr virus- transformed control B lymphoblast cells from the same individual (CRL-1980) were acquired from ATCC (Manassas, VA, USA) and cultured in RPMI-1640 medium supplemented with 10% FBS, 100 units/ml penicillin and 100 ug/ml streptomycin, at 37°C in 5% CO_2/95% air. DNA was purified from these cells using a mammalian genomic DNA extraction kit (Sigma-Aldrich, St. Louis, MO, USA. Catalog number: G1N10-1KT). About 10 ngs of genomic DNA were used for amplifying the genomic region containing the complex indel. PCR (50ul) was done with Taq polymerase (Promega, Madison, WI, USA. Catalog number: M8298) following these cycling conditions: 95 °C-2 min; 35 cycles of 95 °C-30 min, 45–52 °C-30 min (Annealing temperature depends on Tm of primers used), 72 °C-1 min; 1 cycle of 72 °C-5 min. 10ul PCR products were separated on 2% agarose gel to check for the quality of the amplification. Reactions with robust and specific amplifications were purified by using a PCR product cleanup kit (Qiagen, Valencia, CA, USA. Catalog number: 28104). 8–12 ng of the amplicons with a size range of 175–300 bp were bi-directionally sequenced by the Sanger method using the PCR primers. The presence or absence of the complex indels were

determined by aligning the sequencing traces to the reference sequence and sequence contained the predicted indels.

MuSiC-based correlation analysis

We took the number of complex indels in a sample as the trait and the published somatic variants in the TCGA maf files as the source for our MuSiC correlation runs.

Compilation of cancer-associated gene list

A total of 624 candidate cancer-associated genes was compiled using eleven sources, including recently published large-scale cancer studies, publicly available screening panels, and analysis of publicly available data sources (Supplementary Table 6). The 204 genes shared across at least two of the nine sources were retained and a literature search was conducted to identify evidence supporting inclusion of any remaining unique genes. A subset of 518 genes originated from recent publications, including 294 genes from Frampton *et al.* [24], 125 genes from Kandoth *et al.* [26], 212 genes from Lawrence *et al.* [27], 194 genes from Pritchard *et al.* [28], 124 genes from Vogelstein *et al.* [31], 48 genes from Rahman *et al.*, [29] and 48 from Kanchi et al. [25]. Thirty-nine additional genes were included based on the analysis of driver mutations in 20 TCGA cancer types (Supplementary Table 6), recommendations in accordance with the standards and guidelines of the American College of Genetics and Genomics [30] and 18 novel cancer driver genes identified in recently published large-scale studies.

Complex indel discovery and filtering procedure

Our analysis of variants from the Craig Venter genome indicates that a substantial number are complex (having both insertions and deletions) and are routinely missed by NGS data indel callers. A survey of several databases, such as COSMIC and dbSNP, further suggests that complex indels are vastly under-represented. To address this issue, we developed Pindel-C to specifically search for co-occurring insertion and deletion events, i.e. “complex indels” (Fig. 1). The key elements and procedures of Pindel-C are: 1) **Read Extraction.** All read pairs with one end spanning potential variant breakpoints are detected and extracted from a single alignment BAM file or multiple files. The alignment signals for read selection include soft-clip, gap alignment, unmapped, and other non-M characters in the CIGAR string. For mates of these reads, we require mapping quality to exceed a user specified cutoff (30) and use their 3′ mapping positions as anchors for local mapping. 2) **Pattern Growth-Based Alignment.** We align one base at a time from both terminals of the reads to both DNA strands around the 3′ end of the anchor read within 2 insert size distances. Pattern growth [33–35] is used for string matching to search for the maximum unique substring between the read sequence and the local reference genome. 3) **Distinguishing Complex Indels from Simple Indels.** A “simple” event is inferred if the maximum unique substrings from two terminals of the read are able to cover the entire read or reference. Otherwise, if these substrings do not cover a segment inside the read and the reference, we have likely detected a “complex” event. To characterize potential complex indels, we left-shift the mapping position and then sort reads accordingly. If a set of reads has the same left and right mapping positions and the identical middle unmapped fragment, we combine them and

report them as a potential complex indel. 4) **Remove False Predictions and Variant Allele Frequency Analysis.** The strands of the supporting reads are examined to make sure that each strand is represented. Based on the predicted complex variant, we create a reference contig, including 10kb flanks both upstream and downstream of the variant, as well as a complex indel containing contig with the same setting. We then extract all reads within a 2kb window of the variant position and remap them using BWA to the two contigs generated. Mapped reads with mapping quality of at least 30 are used for read count analysis. The calculated coverage values are noted as the numbers of reads supporting reference or variant alleles. Since part of the read is not aligned to the reference genome, we expect higher false positive rate in the raw calls because of various sequence artifacts and may perform additional manual inspection using IGV [22]. For example, we anticipate situations such as extensive soft-clipped reads without consistent breakpoints, reads with 1 bp indel at a fixed read position unrelated to genomic position, and sequence artifacts in nearby sequences. We discard the calls if any of the situations above are detected. The entire procedure has been automated and the scripts are available at https://github.com/ding-lab/VariantQC.

Identification of complex indels in cancer genes

The initial set of somatic complex indels in cancer genes contain 1,367 predictions if we require at least one supporting read from either strand and a variant allele frequency of at least 5%. Then we manually examined the supporting evidence using Integrated Genomic Viewer (IGV) [22] and also re-examined the numbers of reads supporting either allele using indel-containing contig based mapping. For each detected complex indel, we first construct a reference allele contig by including both the upstream and downstream 10k as well as the reference allele as the reference allele. We then substitute the reference allele with the detected alternative allele. This gives us two contigs representing the alleles. We next extract all reads mapped within 2kb distance to the allele position. We re-mapped those extracted reads using BWA paired-end mapping mode with parameter of −q 5. Finally we count the numbers of reads with more than a given mapping quality (30 by default) and mapped to each contig right at 10k position. Based on the new read counts, we computed variant allele fraction and discarded any calls with VAF smaller than 0.05. We took the candidate sites as input and re-ran Pindel-C to identify tumor samples carrying the same somatic variation but missed in our discovery phase due to low support reads.

Geometric probability test of proximity of 83 complex indels

We found that 83 out of 285 complex indels were within 100 bp of another, suggesting non-random distribution. This observation was tested against a null hypothesis that these 285 instances are randomly distributed across the genome using a simple geometric probability model. Consider the *a priori* placement of one of these events at an arbitrary position and the random chance of another event being placed within 100 on either side of the first event. Assuming a conservative 30Mb exome, the Bernoulli probability of any one of the remaining 284 events is 200/3e7, implying the probability that one of them from the set will be within 100 bp of the trial event is $284 \cdot 200 \cdot \exp(-200 \cdot 283/3e7)/3e7 \approx 0.00189$. The probability that 83 of these events participate in such proximity arrangements is appreciably smaller, suggesting we reject the null hypothesis of random distribution.

Statistical test on complex indel VAF vs other simple variant VAF

We assessed whether complex indel VAFs in specific gene-cancer combinations were statistically higher than their corresponding simple indels in the same samples using permutation testing. This type of test is "data driven" in that the null distribution is constructed directly from the case-control data. An important aspect of such tests is that the size of the sample space determines the lowest attainable *P*-value for any test. In fact, the low bound is the inverse of the number of relevant combinations of the pooled case-control observations. We excluded from testing those cancer-gene combinations that could not, in principle, attain a minimal *P*-value significance of 1%. This exclusion criterion eliminated essentially all single-sample combinations, as there is only an average of five simple indels per complex indel in each sample, and left six cancer-gene combinations that were found to occur in between 3 and 11 individual samples. Five of these combinations had computational sample spaces small enough to permit full permutation testing. The sixth, *EGFR* in LUAD, had much more data with a consequent sample space size on the order of 10^{12}. Here, we performed a sampling-based permutation test rather than a full test using 10^8 points of data selected randomly with replacement. The final list of *P*-values was corrected for multiple-test effects by computing the standard Benjamini-Hochberg False Discovery Rate (FDR).

Hypothesis testing of in-frame complex indels

We first estimated the Bernoulli probability, P_F, of any single event being in-frame by examining the size distribution of T=1,680 exome-wide complex indels. These are taken as the "background" information. Defining $t(k)$ as the tally of indels of length k, the Bernoulli value is the conditional

$$P_F = \sum_k P(k)P(F|k) = \sum_k \frac{t(k)}{T} P(F|k) = \frac{t(3)+t(6)+t(9)+\cdots}{T}$$

where $P(F|K)$ is 1 when k is any multiple of 3, otherwise it is 0. We found $P_F \approx 0.103$, meaning any single event is somewhat unlikely to be in-frame. Under the null hypothesis, the chances that k of a group of N complex indels that occur independent of one another are in-frame can then be described by the binomial distribution $B(N, k, P_F)$. We parsed our complex indel data set, applying the binomial test to any grouping of at least seven events, of which at least two were in-frame. These minimal cutoffs excluded the numerous low-information cases having only a few events, almost all of which were frame-shifts. Once the tailed *P*-values were computed, we applied the standard Benjamini-Hochberg FDR correction for multiple tests.

We did not perform any testing for the converse phenomenon, i.e. where numbers of frame-shift mutations are higher than that explainable by chance. Because P_F is so one-sided, our dataset lacks the power to discern any groupings where this might be true. This is illustrated by a hypothetical group, all of whose events are frameshifts. The size of this group necessary to realize a *P*-value of even 1% is the solution of $(1 - P_F)^N = 0.01$, or about $N = 43$, which is substantially larger than any of the actual groupings in our data set.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Cancer Institute grant R01CA180006 and National Human Genome Research Institute grant U01HG006517 to L.D. F.C. is supported by National Institute of Diabetes and Digestive and Kidney Diseases grant R01DK087960. R.J. is partly supported by CMB training grant (GM 007067). NHGRI Genome Analysis Training Program (T32 HG000045) to M.X. We acknowledge The Cancer Genome Atlas (cancergenome.nih.gov) as the source of primary data, thank Matthew Wyczalkowski for technical assistance, and members of the TCGA Research Network for helpful discussions.

References

1. Levy S, et al. The diploid genome sequence of an individual human. PLoS Biol. 2007; 5:e254. [PubMed: 17803354]
2. Sudmant PH, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015; 526:75–81. [PubMed: 26432246]
3. Kloosterman WP, et al. Characteristics of de novo structural changes in the human genome. Genome Res. 2015
4. Roerink SF, van Schendel R, Tijsterman M. Polymerase theta-mediated end joining of replication-associated DNA breaks in C. elegans. Genome Res. 2014; 24:954–62. [PubMed: 24614976]
5. Koole W, et al. A Polymerase Theta-dependent repair pathway suppresses extensive genomic instability at endogenous G4 DNA sites. Nat Commun. 2014; 5:3216. [PubMed: 24496117]
6. Yu AM, McVey M. Synthesis-dependent microhomology-mediated end joining accounts for multiple types of repair junctions. Nucleic Acids Res. 2010; 38:5706–17. [PubMed: 20460465]
7. Han SW, et al. Predictive and prognostic impact of epidermal growth factor receptor mutation in non-small-cell lung cancer patients treated with gefitinib. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2005; 23:2493–501. [PubMed: 15710947]
8. Lara-Guerra H, et al. Phase II study of preoperative gefitinib in clinical stage I non-small-cell lung cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2009; 27:6229–36. [PubMed: 19884551]
9. Ruppert AM, et al. EGFR-TKI and lung adenocarcinoma with CNS relapse: interest of molecular follow-up. Eur Respir J. 2009; 33:436–40. [PubMed: 19181917]
10. Nangalia J, et al. Somatic CALR mutations in myeloproliferative neoplasms with nonmutated JAK2. The New England journal of medicine. 2013; 369:2391–405. [PubMed: 24325359]
11. Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487:330–7. [PubMed: 22810696]
12. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490:61–70. [PubMed: 23000897]
13. Cancer Genome Atlas Research N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455:1061–8. [PubMed: 18772890]
14. Cancer Genome Atlas Research N. Integrated genomic analyses of ovarian carcinoma. Nature. 2011; 474:609–15. [PubMed: 21720365]
15. Cancer Genome Atlas Research N. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012; 489:519–25. [PubMed: 22960745]
16. Cancer Genome Atlas Research N. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med. 2013; 368:2059–74. [PubMed: 23634996]
17. Cancer Genome Atlas Research N. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 2013; 499:43–9. [PubMed: 23792563]
18. Cancer Genome Atlas Research N. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature. 2014; 507:315–22. [PubMed: 24476821]

19. Cancer Genome Atlas Research N. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014; 511:543–50. [PubMed: 25079552]
20. Cancer Genome Atlas Research N et al. Integrated genomic characterization of endometrial carcinoma. Nature. 2013; 497:67–73. [PubMed: 23636398]
21. Pleasance ED, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 2010; 463:191–6. [PubMed: 20016485]
22. Robinson JT, et al. Integrative genomics viewer. Nat Biotechnol. 2011; 29:24–6. [PubMed: 21221095]
23. Dees ND, et al. MuSiC: identifying mutational significance in cancer genomes. Genome Res. 2012; 22:1589–98. [PubMed: 22759861]
24. Frampton GM, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol. 2013; 31:1023–31. [PubMed: 24142049]
25. Kanchi KL, et al. Integrated analysis of germline and somatic variants in ovarian cancer. Nat Commun. 2014; 5:3156. [PubMed: 24448499]
26. Kandoth C, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013; 502:333–9. [PubMed: 24132290]
27. Lawrence MS, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014; 505:495–501. [PubMed: 24390350]
28. Pritchard CC, et al. Validation and implementation of targeted capture and sequencing for the detection of actionable mutation, copy number variation, and gene rearrangement in clinical cancer specimens. J Mol Diagn. 2014; 16:56–67. [PubMed: 24189654]
29. Rahman N. Realizing the promise of cancer predisposition genes. Nature. 2014; 505:302–8. [PubMed: 24429628]
30. Rehm HL, et al. ACMG clinical laboratory standards for next-generation sequencing. Genet Med. 2013; 15:733–47. [PubMed: 23887774]
31. Vogelstein B, et al. Cancer genome landscapes. Science. 2013; 339:1546–58. [PubMed: 23539594]
32. Sordella R, Bell DW, Haber DA, Settleman J. Gefitinib-sensitizing EGFR mutations in lung cancer activate anti-apoptotic pathways. Science. 2004; 305:1163–7. [PubMed: 15284455]
33. Ye K, Kosters WA, Ijzerman AP. An efficient, versatile and scalable pattern growth approach to mine frequent patterns in unaligned protein sequences. Bioinformatics. 2007; 23:687–93. [PubMed: 17237070]
34. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009; 25:2865–71. [PubMed: 19561018]
35. Zhang Y, et al. PASSion: a pattern growth algorithm-based pipeline for splice junction detection in paired-end RNA-Seq data. Bioinformatics. 2012; 28:479–86. [PubMed: 22219203]

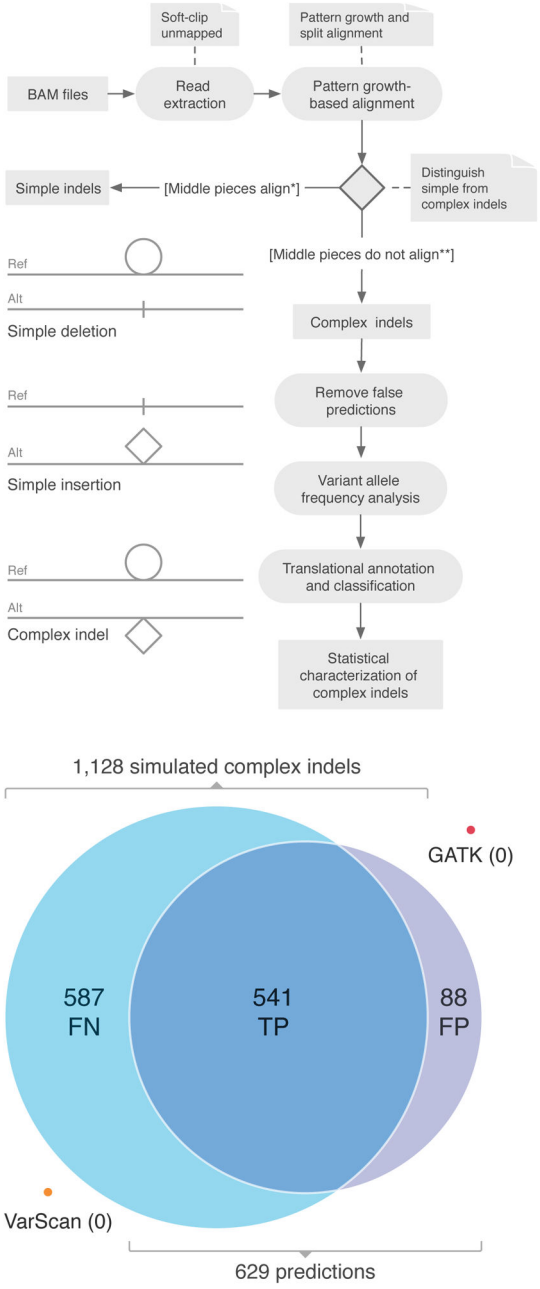

Figure 1. The somatic complex indel detection and filtering workflow and algorithm testing

(a) Soft-clipped and unmapped reads are extracted from BAM files and then split aligned with pattern growth. The alignment result is examined to determine whether certain reads support complex variants. Various filtering, annotation, and statistical analysis steps follow to maintain quality of the complex indel call list. Inset shows three 3 basic configurations as pseudo de-Bruijn graphs (where circular or square loops represent sequences removed to obtain alignment): a simple deletion (top), a complex indel with template sequence from the 5′ sense strand (middle), and a complex indel with template sequence of reverse complement to the deleted fragment (bottom). Ref is reference allele while alt is alternative

allele. **(b)** Results of simulation testing on chromosome 1 of the Venter genome for Pindel-C versus GATK and VarScan. Of 1128 simulated complex indels, Pindel-C found 541 (48% sensitivity), but neither GATK nor VarScan were able to identify any. Pindel-C also mistakenly called 88 additional events as simple indels, implying a false-discovery rate of 14%.

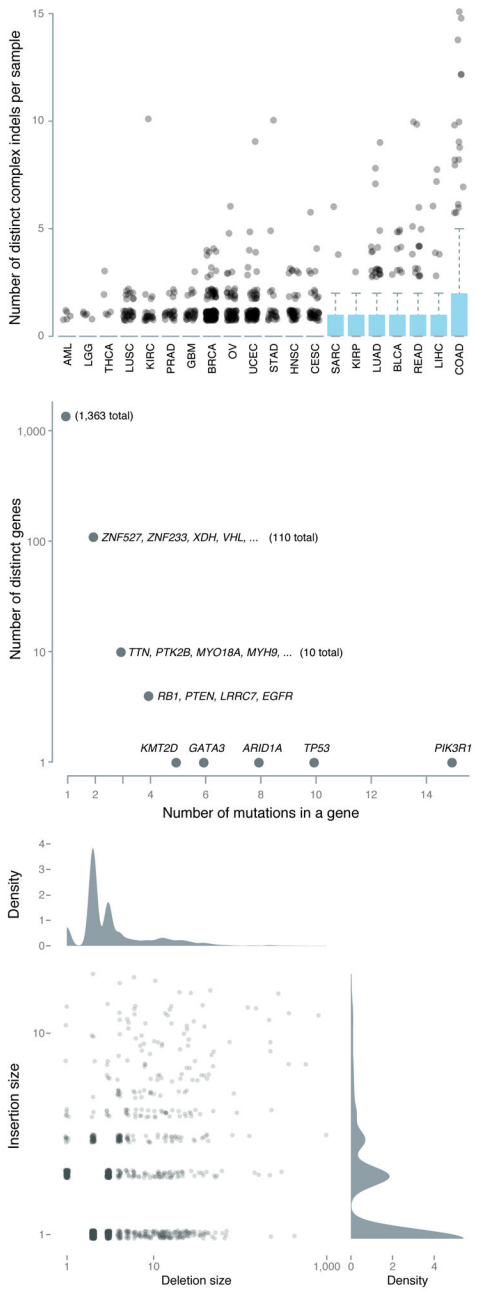

Figure 2. The exome-wide landscape and characteristics of somatic complex indels across 19 cancer types

a) Box plot of the number of somatic complex indels in 19 cancer types. In total, 5 samples having more than 15 indels are not shown. They are listed according to the following in the format of cancer type, sample name, and the number of somatic complex indels: KIRC, TCGA-AK-3430, 16; KIRC, TCGA-AK-3451, 18; COAD, TCGA-AY-5543, 18; COAD, TCGA-CM-5860, 20; COAD, TCGA-G4-6299, 22; LIHC, TCGA-G3-A25T, 69. b) Genes most frequently affected by somatic complex indels. The x-axis is the number of somatic complex indels in a given gene while the y-axis is the number of distinct genes. c) Complex

indels dissected as deletion and insertion at the same breakpoint, with sizes of each plotted per variant. Density plots of deletion and insertion sizes are depicted accordingly.

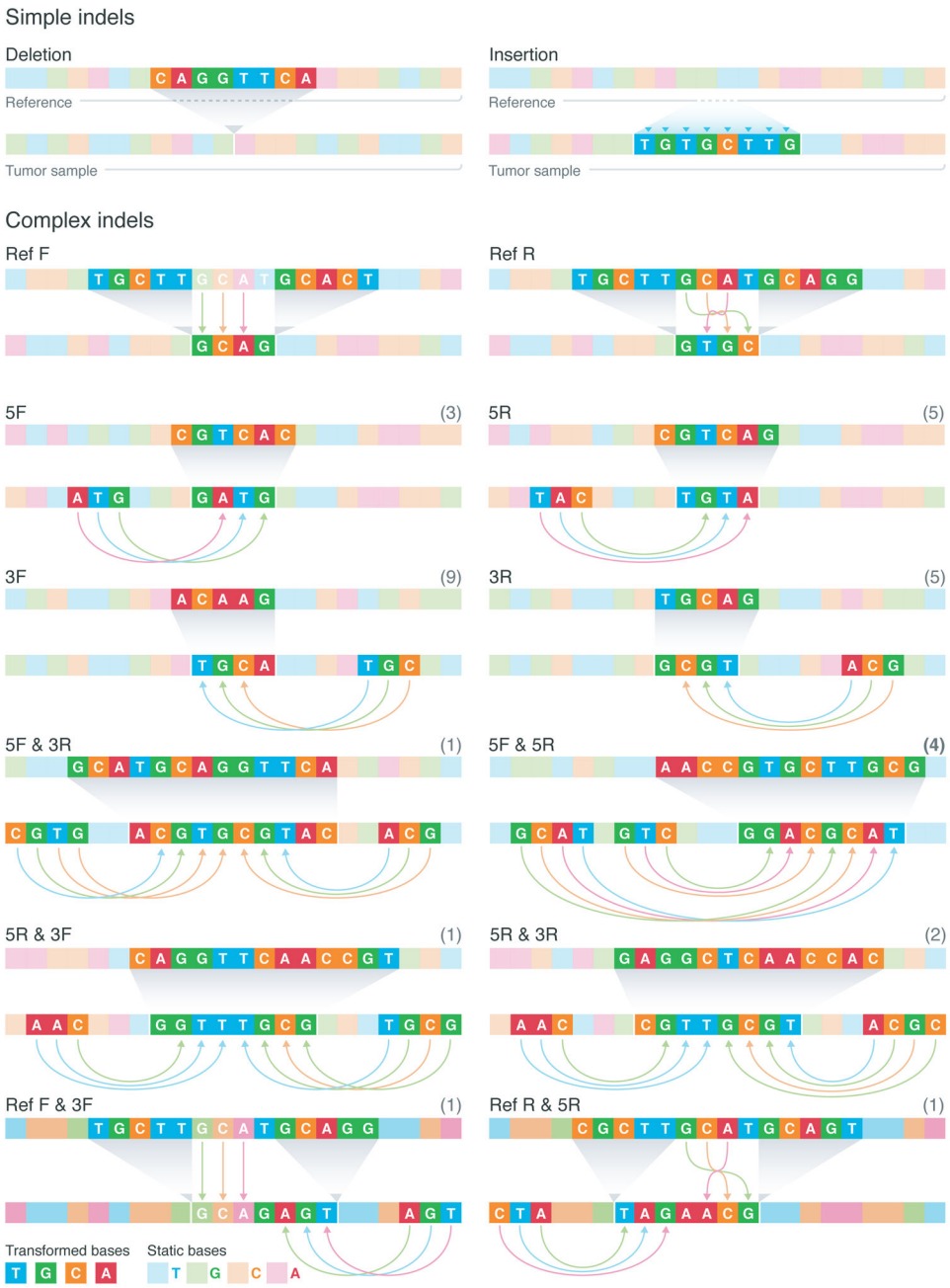

Figure 3. Schematics of simple and complex indel configurations

The first two diagrams depict simple deletion and simple insertion. A total of 12 distinct scenarios were observed. In Ref F, part of the deleted sequence is inserted right at the breakpoint, but in Ref R the reverse complementary sequence of the deleted fragment is inserted. The definitions of terms in the figures are the following: Ref 5 and 3 mean the origin of the inserted sequence is from part of the deleted 5′ flanking and 3′ flanking sequence, respectively. F and R indicate whether the inserted sequence is a direct copy or a copy of the reverse complement. Among the 12 scenarios, 6 are single source and the rest

are combinations of various single sources. The coloring scheme of unchanged (static) and mutated (transformed bases) is illustrated.

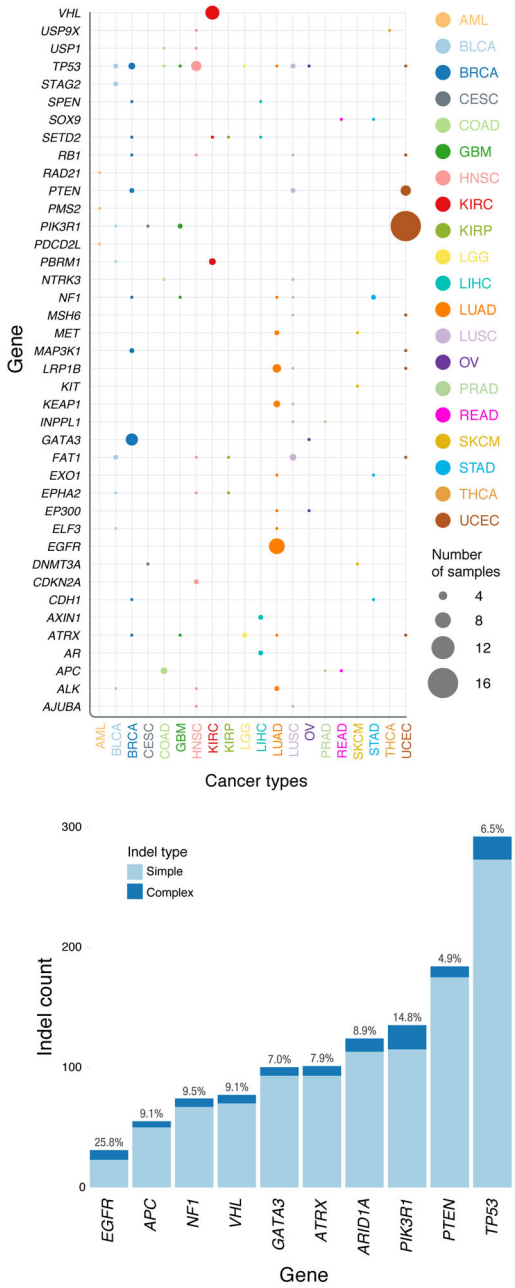

Figure 4. Abundance of somatic complex indels in key cancer genes per cancer type and the contribution of somatic complex indels to the total numbers of indels for 10 cancer genes

a) Plot of the number of samples carrying somatic complex indels in 37 cancer genes across 20 cancer types. Dot size indicates the number of samples. b) Histogram of simple and complex indel counts in 12 key cancer genes with the largest percent gain.

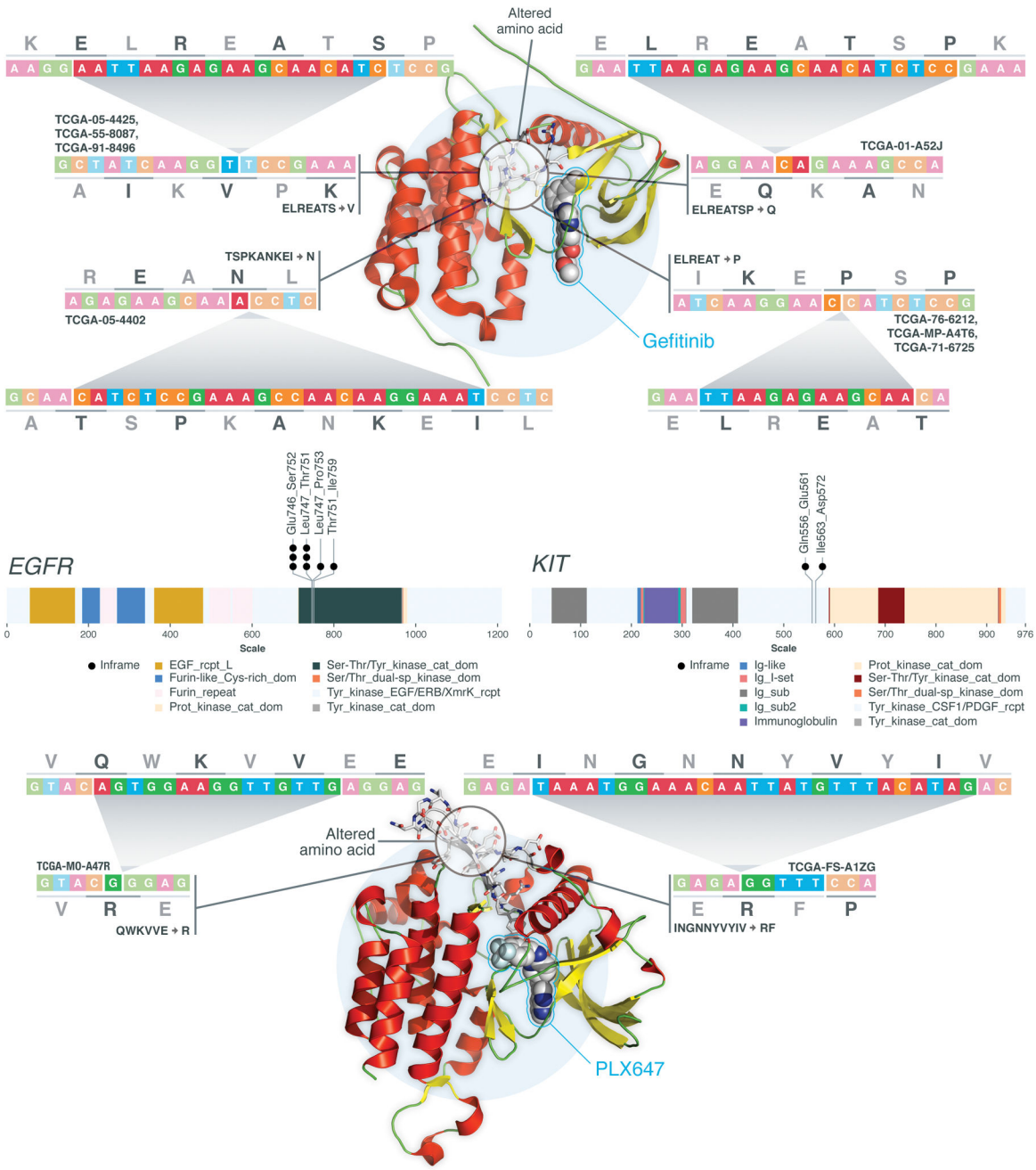

Figure 5. Druggability of somatic complex indels in *EGFR* and *KIT*

Lolliplot of somatic complex indels plotted in panels b and c. All somatic complex indel sites are in the kinase domains of *EGFR* and *KIT*. The nucleoid base and amino acid changes for distinct somatic complex indel sites are illustrated in panels a and d, respectively. The amino acids affected by the complex indels are displayed as sticks and colored, according to the element. The drug molecule co-crystalized with the protein is displayed as sphere and colored by element.

Table 1

Statistical test on whether variant allele fraction (VAF) of complex indels is higher than VAF of simple variants.

Rank	Cancer	Gene	Case VAF average	Control VAF average	Case VAFs	Control VAFs	P-value	FDR
1	LUAD	*EGFR*	53.4%	14.9%	3	12	0.00659	0.03956
2	BRCA	*TP53*	53.6%	30.6%	3	14	0.09118	0.27353
3	KIRC	*VHL*	43.9%	33.0%	6	12	0.09820	0.27353
4	KIRC	*PBRM1*	43.8%	31.8%	3	10	0.16434	0.27353
5	UCEC	*PIK3R1*	42.3%	39.4%	11	50	0.26250	0.31500
6	UCEC	*PTEN*	40.6%	39.1%	4	21	0.39976	0.39976