Supplemental information:

VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research

Zhongwu Lai[1,]*, Aleksandra Markovets[1], Miika Ahdesmaki[2], Brad Chapman[3], Oliver Hofmann[3,4], Robert McEwen[2], Justin Johnson[1], Brian Dougherty[1], J. Carl Barrett[1], and Jonathan Dry[1]

[1]Oncology iMed, AstraZeneca, Waltham, MA 02451, USA

[2]Oncology iMed, AstraZeneca, Cambridge, United Kingdom

[3]Bioinformatics Core, Harvard T.H. Chan School of Public Health, Boston, MA 02115

[4]Wolfson Wohl Cancer Research Centre, Institute of Cancer Sciences, University of Glasgow, Bearsden Glasgow, G61 1QH, UK

* To whom correspondence should be addressed. Tel: 781-839-4495; Fax: 781-839-4200; Email: Zhongwu.Lai@astrazeneca.com

METHODS

Detecting strand bias

Strand bias is a common source of artifact in NGS, resulting in false positive calls. To detect strand bias, VarDict uses Fisher Exact test. Forward and reverse oriented reads supporting reference and variant are counted to construct the 2x2 contingency table. If the resulting p-value is less than 0.01, it's considered strand bias and filtered out.

Detecting somatic mutations and LOH variants

VarDict also uses Fisher Exact test to call somatic mutations LOH (Loss of Heterozygosity) variants. Reads supporting reference and variant in the paired samples are counted to construct the 2x2 contingency table. If the resulting p-value is less than 0.05, it's considered significant change. The variant will be classified as somatic if it's not detectable in parental sample, or as LOH otherwise.

Features calculated by VarDict

VarDict calculates many features for the variants called and provides great flexibility for user to adjust different parameters through command line options, accommodating different sequencing situations. Suppl. Table 4 list features VarDict calculates and the corresponding command line to control them, if available.

Synthesize of complex variants

To synthesize the complex variants to test VarDict's capability, we synthesized 1,122 complex variants for coding exons of common cancer genes (highlighted in bold in Suppl. Table 2). We first extract coding exon sequences with 300bp flanking at each side. We then randomly deleted 1-50bp, followed by insertion of 1-50bp of different sequences. ART was then used to simulate 2x100bp Illumina HiSeq2500 pair end reads, with mean insert sizes of 350bp and standard deviation of 75bp. The reads were aligned to hg19 using BWA MEM and variants were called from the resulting BAM files. Suppl. Table 6 list such 1,122 variants.

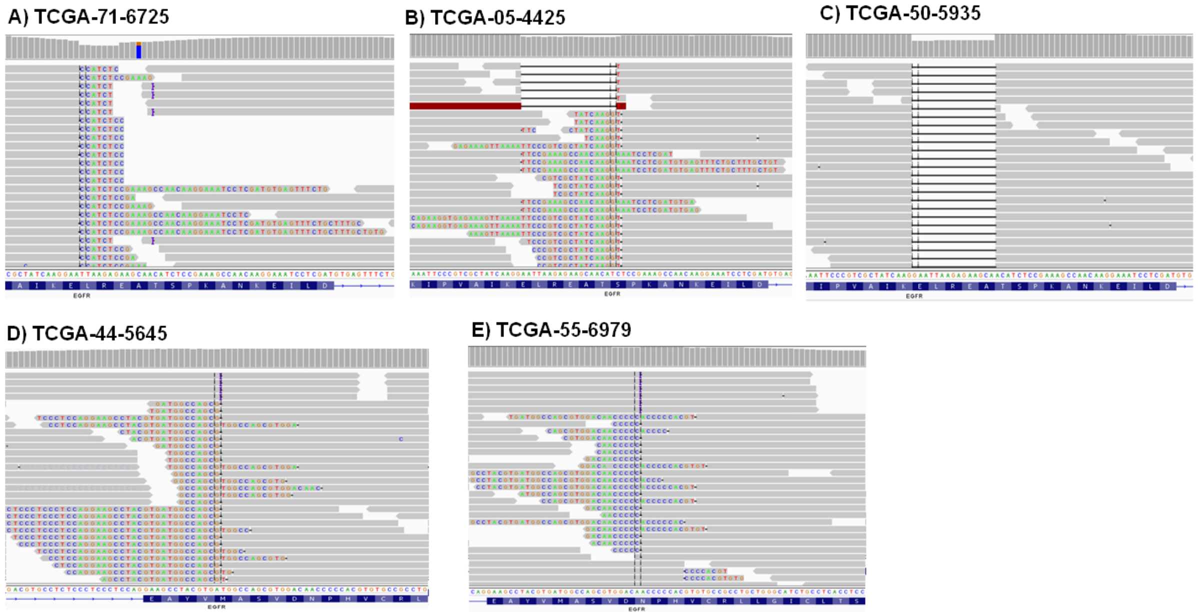
RESULTS

Timing and resource usage comparisons

We compared the timing and resource usage among VarDict, MuTect, FreeBayes, and VarScan, using DREAM challenge synthetic dataset #4, which is a single tumor/normal WGS pair (60x coverage). We used a server of 64 cores, with 3Gb memory/core and NFS file system. Supp. Table 5 showed the run time. VarDict runs as fast as MuTect, and FreeBayes, and much faster than VarScan.

SUPPLEMENTARY FIGURES

Suppl. Figure 1. IGV screenshots for five EGFR InDel mutations (3 exon 19 deletions, 2 exon 20 insertions) missed by Firehose in TCGA LUAD cohort of 230 patients. A) TCGA-71-6725 is a complex mutation in exon 19 (c.2239_2251delTTAAGAGAAGCAAinsC), resulting in in-frame deletion of 4 aa and insertion of 1aa (L747_T751delinsP); **B)** TCGA-05-4425 is a complex mutation in exon 19 (c.2237_2255delAATTAAGAGAAGCAACATCinsT), resulting in in-frame deletion of 5 aa and insertion of 1 aa (E746_S752delinsV); **C)** TCGA-50-5935 is a deletion (c.2236_2250delGAATTAAGAGAAGCA), resulting in in-frame deletion of 5 aa (E746_A750del); **D)** TCGA-44-5645 is an insertion in exon 20 (c.2300_2308dupCCAGCGTGG), resulting in in-frame insertion of 3 aa (A767_V769dup); **E)** TCGA-55-6979 is an insertion in exon 20 (c.2314_2319dupCCCCAC), resulting in in-frame insertion of 2 aa (P772_H773dup). Only portion of representative reads were shown.



A) TCGA-71-6725

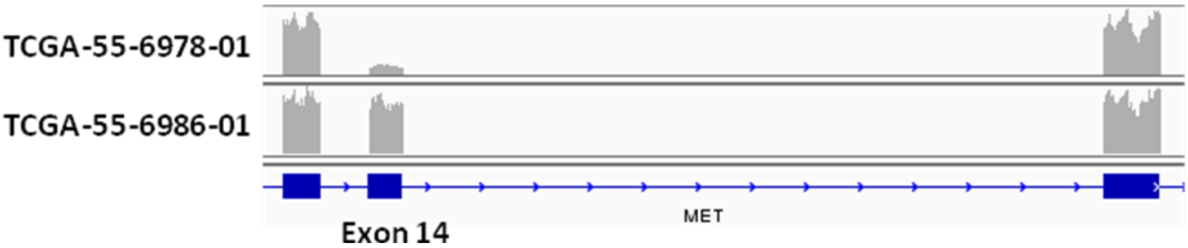
B) TCGA-05-4425

C) TCGA-50-5935

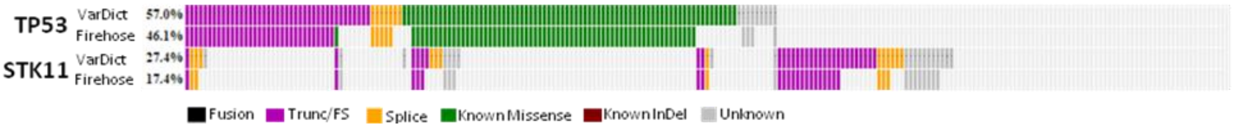
D) TCGA-44-5645

E) TCGA-55-6979

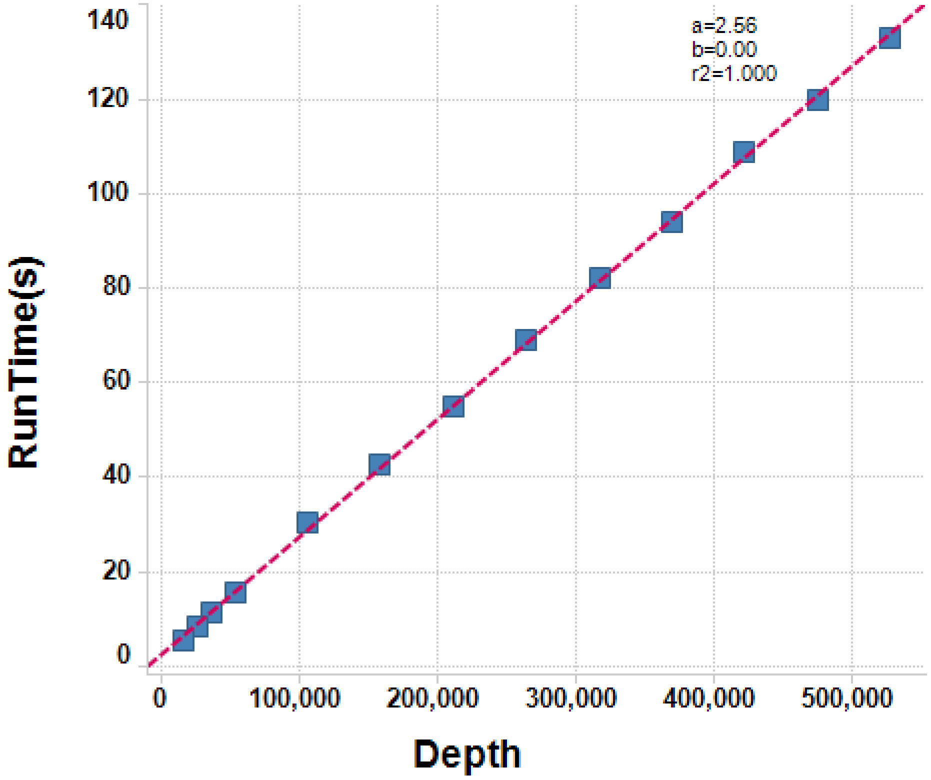
Suppl. Figure 2. No evidence of MET exon 14 skipping in sample TCGA-55-6986-01 in TCGA LUAD cohort. The figure shows the exon coverage from RNA-Seq for two samples. The middle one is the exon 14 of MET. The top sample (TCGA-55-6978-01) is a positive sample having exon 14 skipping due to splice site mutation (c.3082+1G>C), showing much lower coverage of exon 14 comparing to exon 13 and 15 due to skipping. The bottom sample TCGA-55-6986-01 showed no difference in coverage for exon 13, 14, and 15. No DNA mutation in MET was detected in this patient.



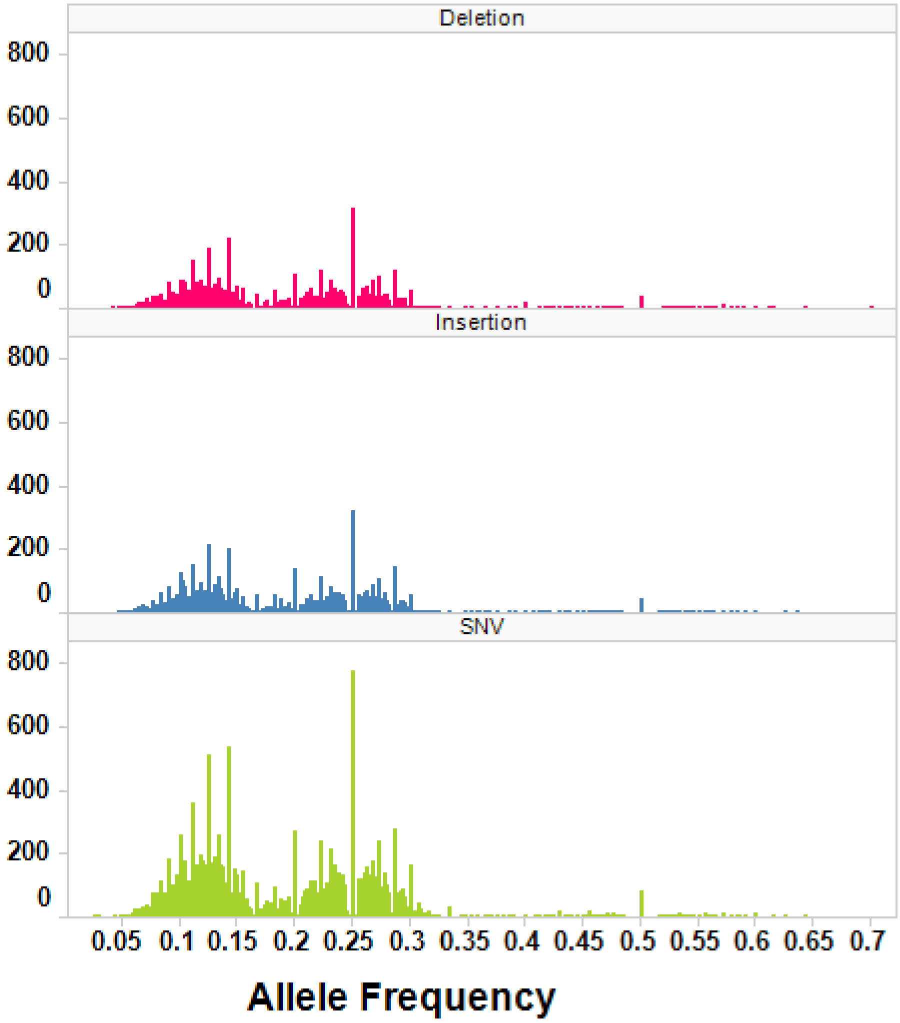
Suppl. Figure 3. The comparison of VarDict and Firehose calls for TP53 and STK11 in 230 TCGA LUAD patients. Each column represents a patient. Each gene has two rows, with top one showing calls from VarDict and bottom one calls from Firehose. Different colours indicate different mutation types. VarDict called TP53 mutations in 57% patients, compared to 46.1% in Firehose; and VarDict called STK11 mutations in 27.4% patients, compared to 17.4% in Firehose. Most of additional STK11 mutations called are truncations and splice sites, consistent with STK11's function as a tumor suppressor. VarDict called all mutations called in Firehose, except for TP53 P77L mutation in TCGA-49-4487-01, which was called by VarDict but filtered out because the allele frequency was < 7.5% and the function is unknown. However, VarDict called on frameshift InDel TP53 mutation H178fs in TCGA-49-4487-01 that was not called in Firehose, suggesting TP53 P77L was likely just a passenger mutation. The sample is marked green for TP53 by Firehose but purple by VarDict. Again, VarDict calls more mutations in both SNV and InDels. Trunc: Truncation; FS: Frameshift.



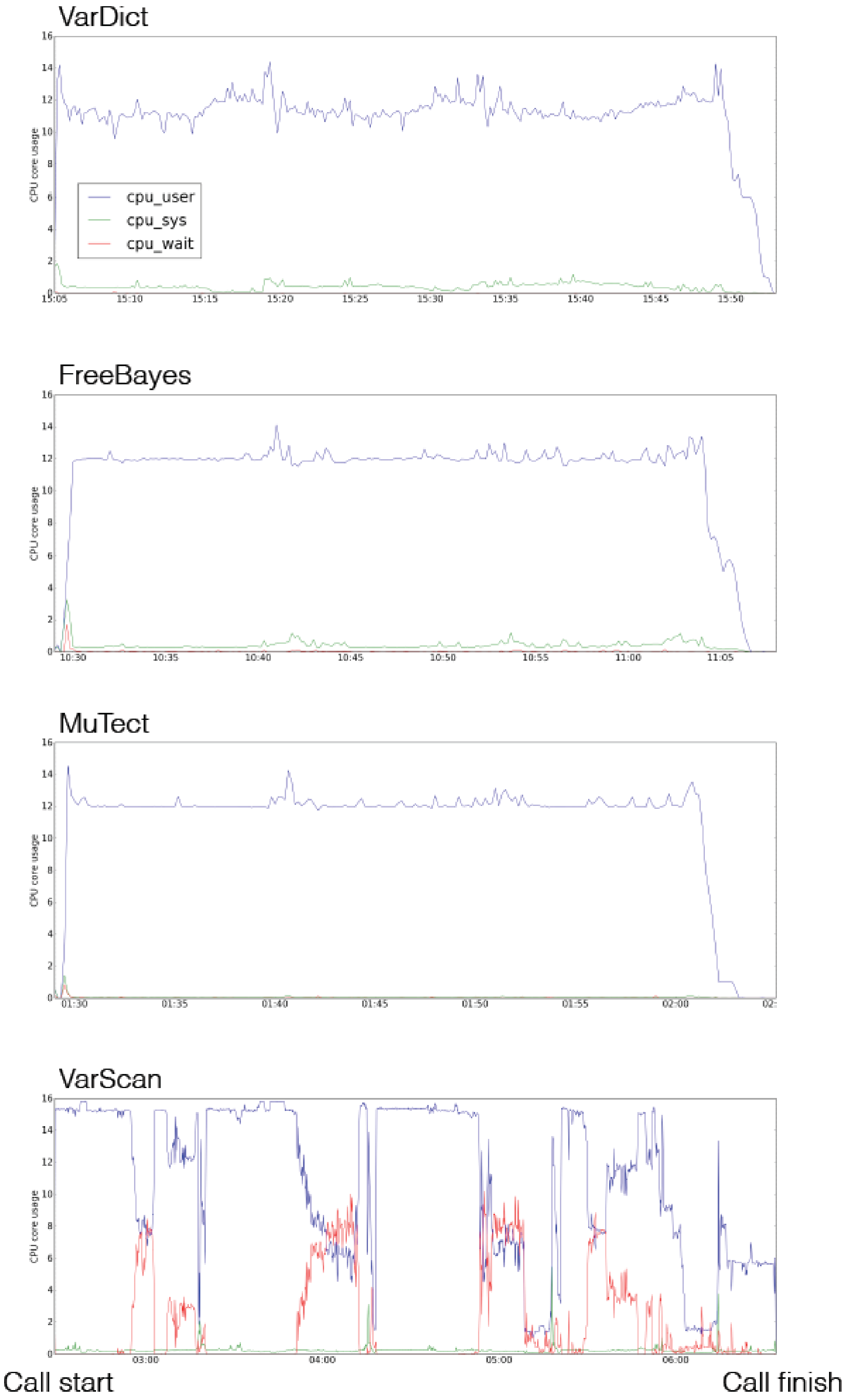
Suppl. Figure 4. Linear performance of VarDict relative to depth. The graph showed the run time of VarDict against the depth of coverage. X-axis is the depth of coverage with highest over 500k, and y-axis the running time for VarDict in seconds. The data was simulated using VarDict's –Z option, which controls the amount of downsampling in a random fashion.



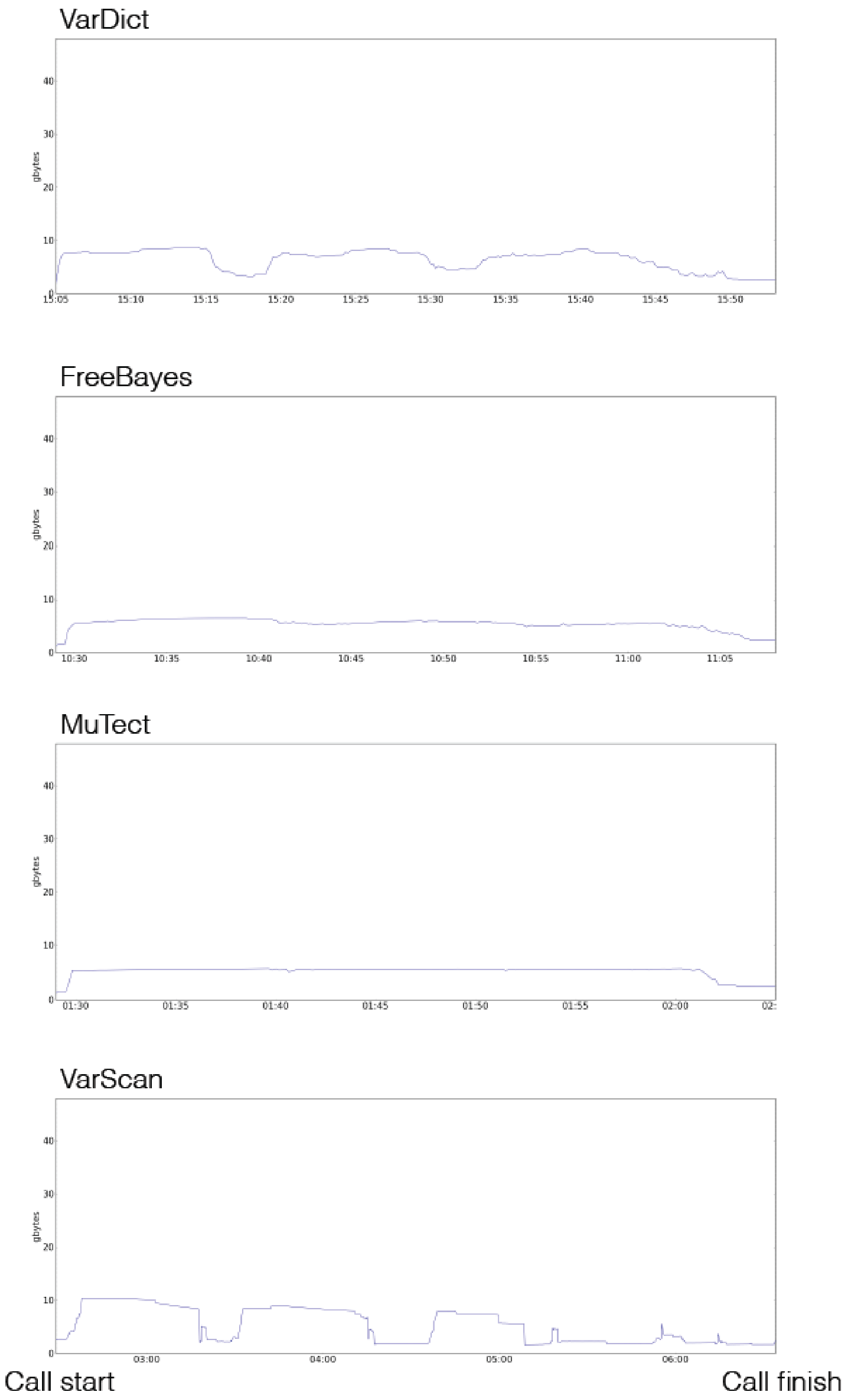
Suppl. Figure 5. Histogram of allele frequencies of somatic mutations called by VarDict. The histogram showed near identical distribution of allele frequencies for SNV, Insertion and Deletion, suggesting VarDict has accurate estimation of allele frequencies for Insertions and Deletions, as SNV estimation is relatively accurate.



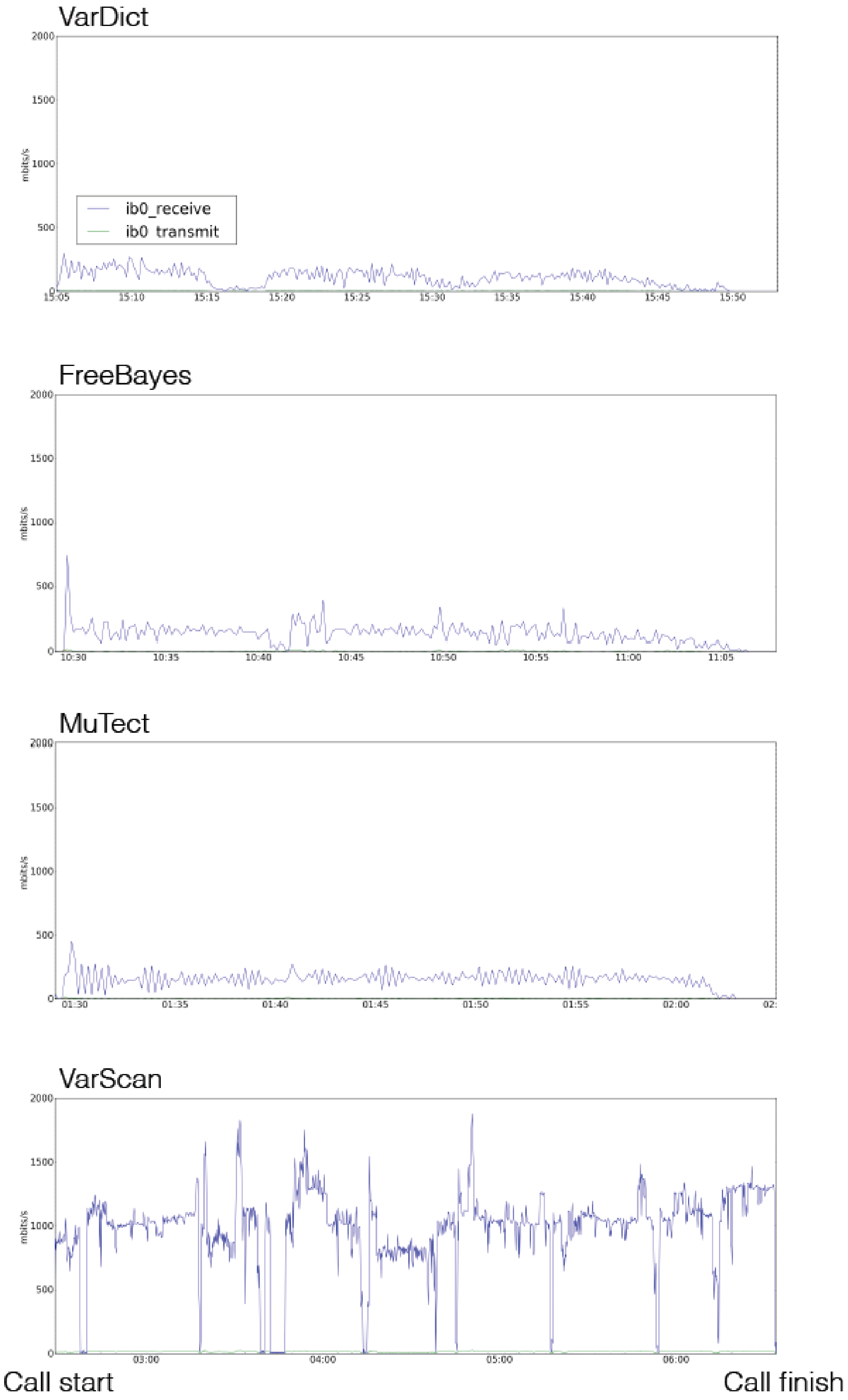
Suppl. Figure 6. CPU usage comparison of VarDict, FreeBayes, MuTect, and VarScan for DREAM Challenge synthetic dataset #4.



Suppl. Figure 7. Memory usage comparison of VarDict, FreeBayes, MuTect, and VarScan for DREAM Challenge synthetic dataset #4.



Suppl. Figure 8. Network (shared filesystem read/writes) usage comparison of VarDict, FreeBayes, MuTect, and VarScan for DREAM Challenge synthetic dataset #4.



VarDict

FreeBayes

MuTect

VarScan

Call start

Call finish

Suppl. Table 1. Comparison of complex variant calling for VarDict, Pindel, and Scalpel. The synthetic dataset contains 1,122 complex variants. Each complex variant is random deletion of 1-50bp within or near every coding exons of common cancer genes (highlighted in bold in Suppl. Table 2), inserted randomly with 1-50bp of different sequences. Pair end Illumina reads were simulated using ART (17) at 50x coverage and aligned to hg19 using BWA MEM. TP overlap: true positives that overlap with key; FP: false positives; TP exact: true positives match exactly with keys.

Caller	TP Overlap	FP	TP exact
VarDict	1,113 (99%)	16	1,073
Pindel	882 (78%)	186	63
Scalpel	1,052 (94%)	454	0

Suppl. Table 2: List of 208 genes analyzed by VarDict for 230 TCGA lung adenocarcinoma patients.

ABL1
AGTR2
AKT1
AKT2
AKT3
ALK
APC
AR
ARAF
ARID1A
ARID2
ASXL1
ATM
ATR
ATRX
AXIN2
BACH1
BAP1
BARD1
BCL2L1
BLM
BRAF
BRCA1
BRCA2
BRD4
BRIP1
C11orf30
C19orf40
CCNB1
CCND1
CCND2
CCND3
CCNE1
CD79A
CD79B
CDH1
CDK12
CDK4
CDK6
CDK9
CDKN2A
CHEK1
CHEK2
CTNNB1
CUL4A
DDR2
EGFR
EML4
EP300
ERAS
ERBB2
ERBB3
ERBB4
ERCC1
ERCC2
ERG
ESR1
EZH2
FAM175A
FANCA
FANCB
FANCC
FANCD2
FANCE
FANCF
FANCG
FANCI
FANCL
FANCM
FBXW7
FGF1
FGF10
FGF12
FGF14
FGF19
FGF2
FGF23
FGF3
FGF4
FGF5
FGF6
FGF7
FGF8
FGF9
FGFR1
FGFR2
FGFR3
FGFR4
FLT3
GATA3
GNA11
GNAQ
GNAS
HGF
HRAS
IDH1
IDH2
IGF1R
IL6
IRAK4
JAK1
JAK2
JAK3
KDM5C
KDM6A
KDR
KEAP1
KIT
KRAS
MAP2K1
MAP2K2
MAP2K4
MAP3K1
MAP3K13
MAP3K8
MAPK1
MAPK3
MCL1
MCPH1
MDM2
MDM4
MECOM
MED12
MET
MLH1
MLL2
MRAS
MRE11A
MSH2
MSH3
MSH6
MTOR
MUTYH
MYC
MYCL1
MYCN
MYD88
MYT1
NBN
NF1
NF2
NFE2L2
NFE2L3
NPM1
NRAS
NSD1
PAK1
PALB2
PARP1
PARP2
PARP3
PARP4
PBRM1
PDGFRA
PIK3C2B
PIK3C2G
PIK3C3
PIK3CA
PIK3CB
PIK3CD
PIK3CG
PIK3R1
PIK3R2
PIM1
PIM2
PIM3
PMS1
PMS2
POLE
PPP2R1A
PPP2R2A
PRKDC
PTEN
PTENP1
RAD50
RAD51
RAD51B
RAD51C
RAD51D
RAD52
RAD54L
RAF1
RB1
RET
ROS1
RPA1
RPTOR
RUNX1
SMARCA4
SMARCB1
SOS1
SOX2
SPOP
STK11
TERT
TET2
TGFBR2
TIPARP
TMPRSS2
TP53
TP53BP1
TSC1
TSC2
VHL
WEE1
XRCC1
XRCC2
XRCC3

Suppl. Table 3: List of mutations called by VarDict in 230 TCGA lung adenocarcinoma patients for 208 genes. Only mutations affecting coding regions are listed. Synonymous mutations are filtered, unless they are known in literature to be functionally impactful, such as TP53 T125T. The last part of the Sample name indicates the sequencing platform: WGS (whole genome sequencing), WXS (exome), or VALIDATION. The last column indicates the status of variants: “known” means it’s known to have a functional impact; “likely” means the variant is likely to have a functional impact on the gene; and “unknown” means the functional impact is unknown.

Suppl. Table 4: List of features VarDict calculates and corresponding command line option to control them, if available. The values in the 3rd column are default. AF: Allele fraction; DUP: Large duplication; DEL: Large deletion; INV: Large inversion; INS: Large insertion; BND: Fusion; LOH: Loss of Heterozygosity; NA: not available.

Feature	Description	Command Line (Default)
TYPE	Variant type. Possible values are: SNV, Insertion, Deletion, Complex, DUP, DEL, INV, INS, BND	NA
END	The end position for the variant	NA
DP	The total depth of coverage	NA
VD	Minimum number of reads supporting alternative	-r 2
RD	The reference forward and reverse read counts	NA
ALD	The variant forward and reverse read counts	NA
AF	Minimum allele fraction	-f 0.05
PMEAN	Mean base position in the reads	-p 8
PSTD	Indicate whether base position changes in different reads. 0 means no change and indicate of potential duplicates.	NA
QUAL	Mean base quality for the variant	-q 25
BIAS	Strand bias information	-B 2
REFBIAS	Reference depth by strand	NA
VARBIAS	Variant depth by strand	NA
SBF	The p-value for strand bias from Fisher Exact	NA
ODDRATIO	Strand bias odd ratio	NA
MQ	Mean mapping quality	-O 0
SN	Signal to noise. The ratio of high quality bases/low quality bases	-o 1.5
HIAF	AF using only high quality bases	NA
ADJAF	Additional AF for InDels from local realignment	NA
SHIFT3	No. of bases for InDels that can be shifted to 3' but still produce equivalent alignment	NA
NM	Mean mismatches in reads supporting variant	-m 4.25
MSI	No. of microsatellite repeats (>1 indicate MSI)	NA
MSILEN	The unit length of MSI	NA
DUPRATE	The duplication rate surrounding variant	NA
SPLITREAD	No. of split reads supporting structural variant	NA
SPANPAIR	No. of discordant pairs supporting structural variant	NA
LSEQ	20 bp flanking sequence at 5'	NA
RSEQ	20 bp flanking sequence at 3'	NA
SSF	Somatic p-value from Fisher Exact (paired mode only)	-p 0.05 (in VCF conversion step)
SOR	Odd ratio from somatic testing (paired mode only)	NA
STATUS	Paired status. Values are: Germline, StrongSomatic, LikelySomatic, StrongLOH, LikelyLOH, Deletion, SampleSpecific, and AFDiff	NA
GDAMP	No. of PCR amplicons supporting variant (Amplicon mode only)	NA
TLAMP	Total PCR amplicons covering variant (Amplicon mode only)	NA
NCAMP	No. of amplicons don't work (Amplicon mode only)	NA
AMPFLAG	Amplicon bias flag (Amplicon mode only)	NA

Suppl. Table 5: Run time comparison of various callers using DREAM challenge synthetic dataset #4. We used a server of 64 cores, with 3Gb memory/core and NFS file system. MuTect run time is only for SNP and doesn't include InDel calling as MuTect doesn't call InDels.

Caller	Time
MuTect	36m
FreeBayes	39m
VarDict	48m
VarScan	4h18m

Suppl. Table 6: 1,122 synthetic complex variants used in testing VarDict's complex variant calling capability.