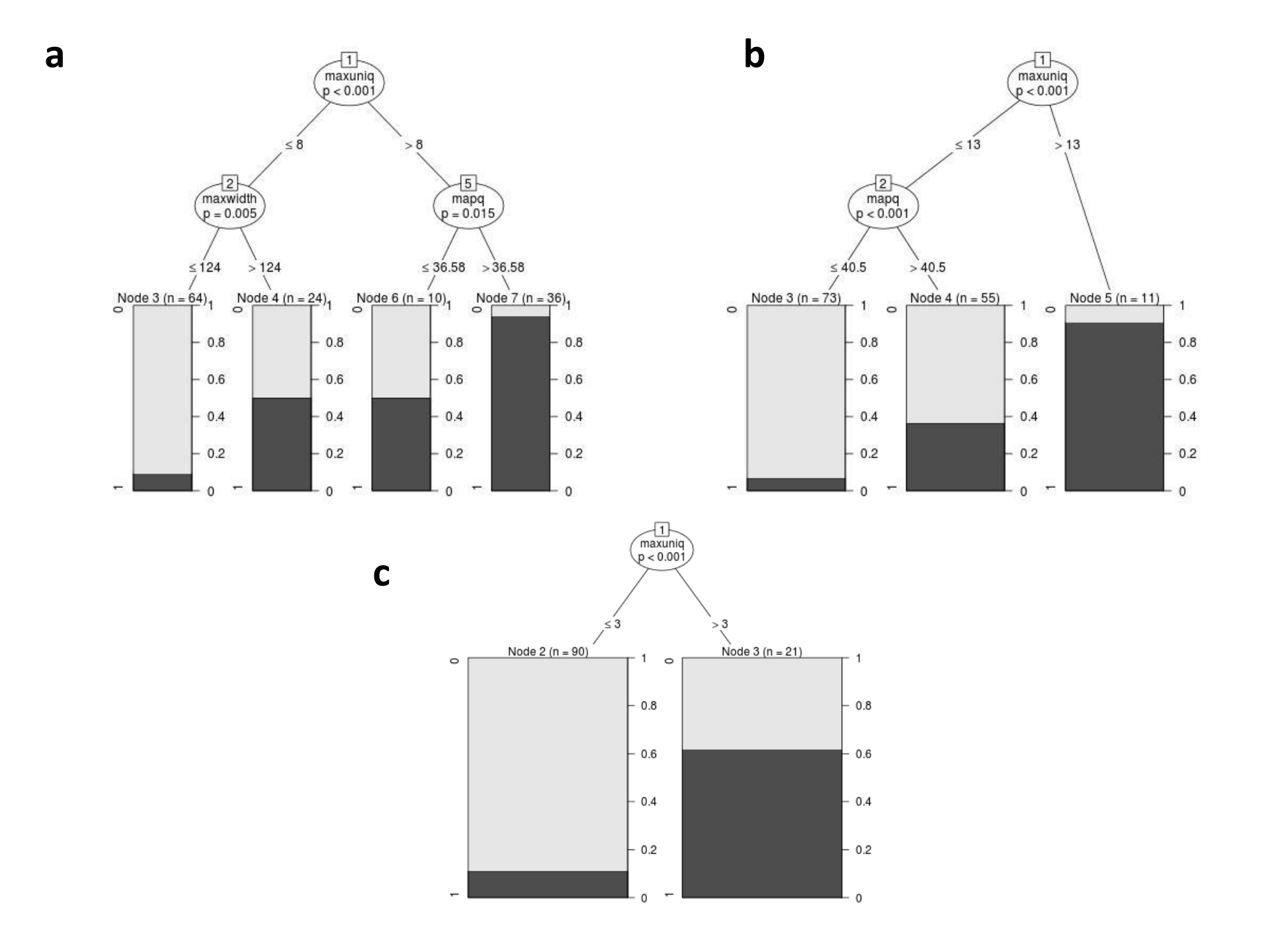
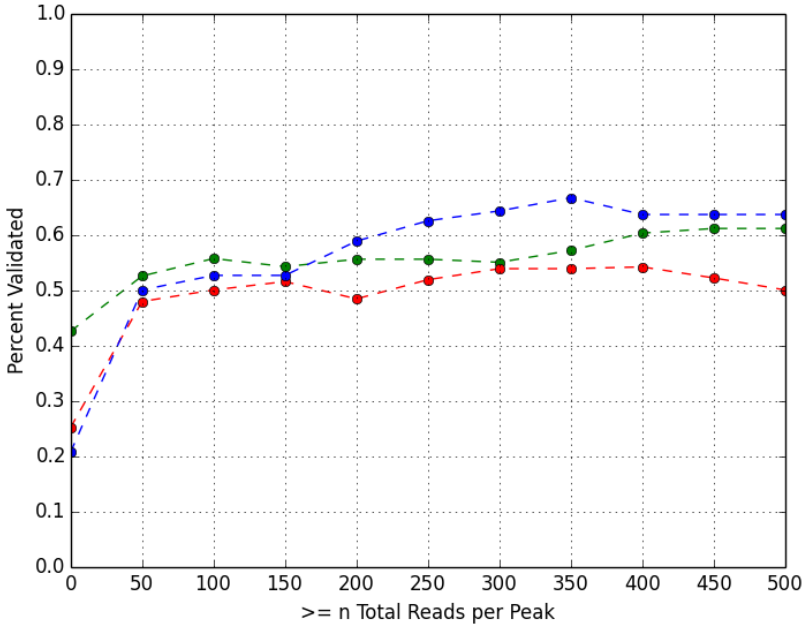
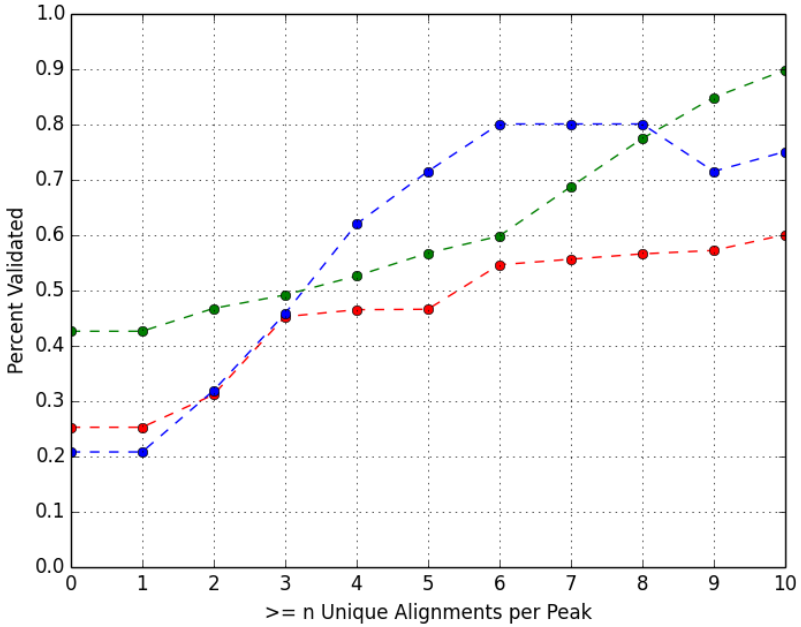

a
1
maxuniq
p < 0.001
≤ 8
> 8
2
maxwidth
p = 0.005
5
mapq
p = 0.015
≤ 124
> 124
≤ 36.58
> 36.58
Node 3 (n = 64)
Node 4 (n = 24)
Node 6 (n = 10)
Node 7 (n = 36)
b
1
maxuniq
p < 0.001
≤ 13
> 13
2
mapq
p < 0.001
≤ 40.5
> 40.5
Node 3 (n = 73)
Node 4 (n = 55)
Node 5 (n = 11)
c
1
maxuniq
p < 0.001
≤ 3
> 3
Node 2 (n = 90)
Node 3 (n = 21)

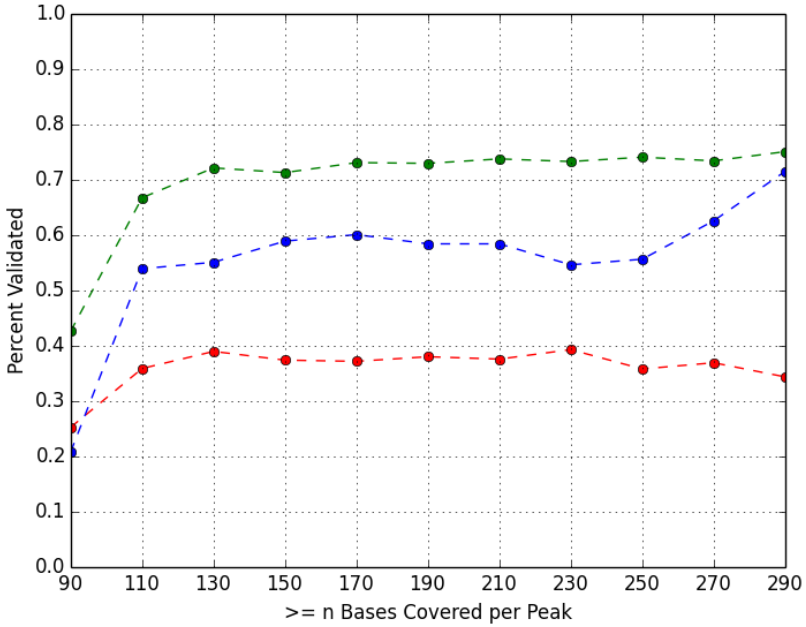
d


Percent Validated
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
50
100
150
200
250
300
350
400
450
500
>= n Total Reads per Peak

e


Percent Validated
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
7
8
9
10
>= n Unique Alignments per Peak

f

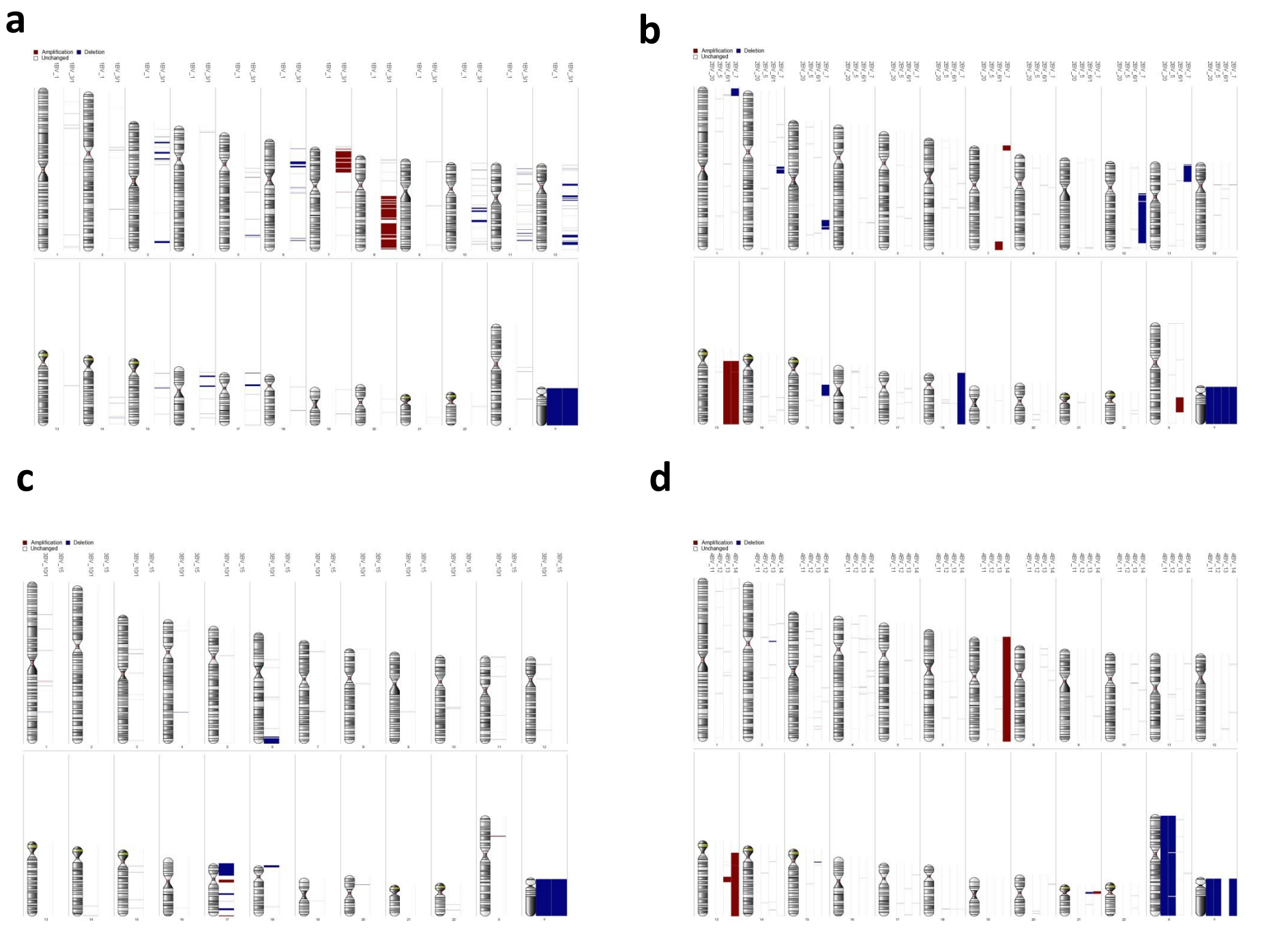

Percent Validated
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
90
110
130
150
170
190
210
230
250
270
290
>= n Bases Covered per Peak

Suppl. Fig. 1. Conditional inference trees and validation rates.

a-c: Estimate of a given insertion's likelihood to validate can be obtained by starting from the top node and making subsequent decisions based on peak characteristics. (**a**) Conditional inference trees used for colorectal cancer patients. Validation likelihoods of the four terminal nodes are 0.094; 0.5; 0.5 and 0.944 from left to right. (**b**) Conditional inference tree used for pancreatic cancer patients. Validation likelihoods of the three terminal nodes are 0.068; 0.364 and 0.909 from left to right. (**c**) Conditional inference tree used for gastric cancer patients. Validation likelihoods of the two terminal nodes are 0.11 and 0.61 from left to right. For all trees, the regression formula was: validated ~ maxcount + maxuniq + maxwidth + mapq (where validation is a categorical variable reflecting the validation result of each site). Abbreviations: maxuniq = unique alignments, maxwidth = span of alignments on reference genome, mapq = mapping quality, maxcount = total alignments.

d-f: Plots for validation rates. The validation rate is the number of sites where site-specific PCR and sequencing of PCR products was successful divided by the number of sites where a validation was attempted. Validation rate is the inverse of false positive rate (e.g. 75% validation rate = 25% false positive rate). Each plot shows the cumulative validation rate (y-axis) for peaks of greater than or equal to the range of sizes shown on the x-axis. Peak size is represented in three different ways: total number of mapped reads per peak, number of alignments that are unique (i.e. have different start positions), and number of bases covered by a peak in the reference genome (peak width). (**d**) Effect of total mapped read count on somatic insertion validation. Points on the plot represent the number of successfully validated somatic insertion candidates for peaks with n or greater total mapped reads. Colors represent tissue type: Green = Colorectal, Red = Pancreatic, Blue = Stomach. (**e**) Effect of total uniquely mapped read count on

somatic insertion validation. Points on the plot represent the number of successfully validated somatic insertion candidates for peaks with n or greater total uniquely mapped reads (i.e. each read has a different start location). Colors represent tissue type: Green = Colorectal, Red = Pancreatic, Blue = Stomach. (**f**): Effect of mapped peak width on somatic insertion validation. Points on the plot represent the number of successfully validated somatic insertion candidates for peaks covering n or greater total bases in the reference genome. Colors represent tissue type: Green = Colorectal, Red = Pancreatic, Blue = Stomach. Numbers of insertions contributing to each data point are shown in **Table S6**.


a
b
c
d
Amplification
Deletion
Unchanged

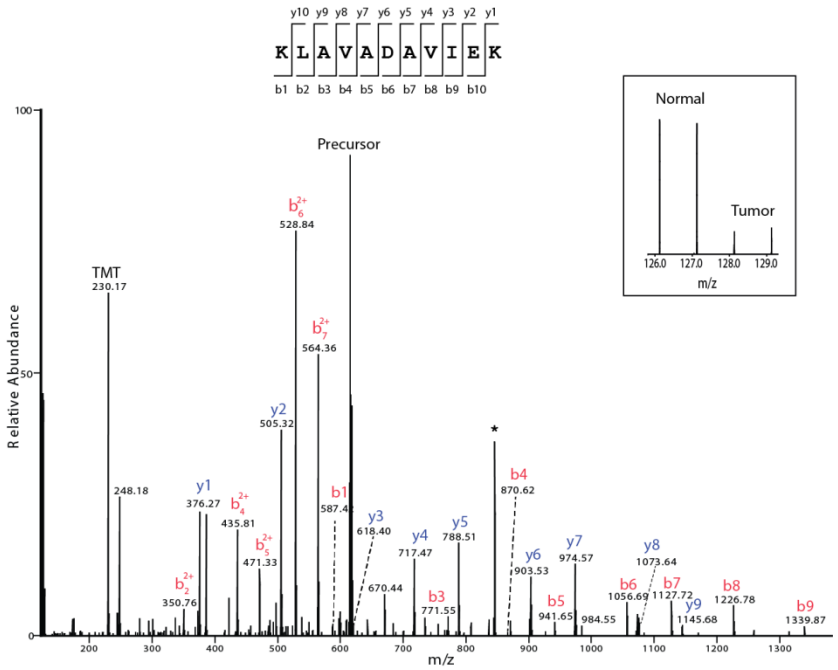
Suppl. Fig. 2. SNParray.

These sets of ideograms, **a-d**, depict the array-based copy number results of analyzed patients' DNA samples. The vertical columns to the right of each chromosome represent the CNVs of the patient's tumor sample as compared to that patient's normal sample. Full sample codes are provided in **Table 1** and **Suppl. Table 2**, sheet "j". The copy number thresholds used are <1.5 and >2.5, which are deletion and amplification, respectively, represented in blue and red, while sex-chromosome differences are analytic artifacts. (**a**): patient 1BV; (**b**): patient 2BV; (**c**): patient 3BV; (**d**): patient 4BV (note that cancer '13' is an outlier).

Of note, by our analysis of somatic L1s near or within CNVs, we did not see any overlap between these structural variations. The closest CNV was about 10 kb away. However, using SNParray, we are unable to rule out the presence of smaller CNVs that cannot be detected.

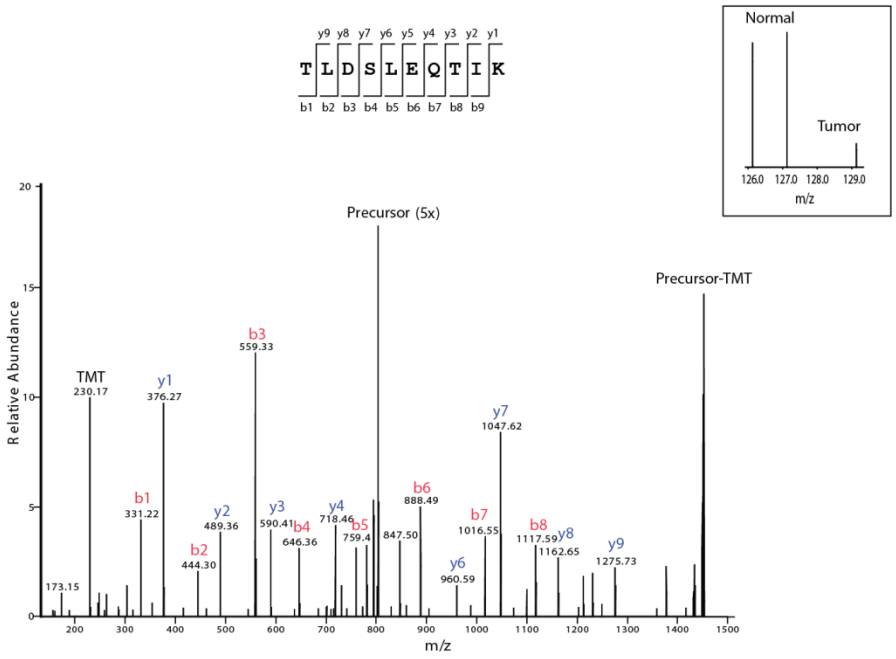
a

WARS2
Tryptophan-tRNA ligase 2, mitochondrial



b

KIAA1217
Sickle tail protein homolog



Suppl. Fig. 3. Mass spectrometry analysis.

Representative peptide-spectrum matches for peptide sequences (KLAVADAVIEK and TLDSLEQTIK) that are mapped to insertionally mutagenized genes with deregulated protein products in the polyp, namely (**a**) WARS2 (tryptophan-tRNA ligase 2, mitochondrial) and (**b**) KIAA1217 (sickle tail protein homolog) are shown, respectively. Ins. D8 in *KIAA1217* occurred in intron 1, about 150 kb away from exons 1 and 2, while ins. H12s in *WARS2* was in intron 2, about 5 kb downstream of exon 2 and 58 kb upstream of exon 3. Indexes are pictorial annotations of tandem MS fragment ions for the two peptides. Insets are zoom-in views of reporter ions derived from TMT labels reflecting the relative abundance of the peptides in normal and tumor samples, respectively. To our knowledge, this is the first mass spectrometry analysis of a colonic adenoma (a new section of sample ‘10’ in patient 3BV) and matched normal colon (a new section of normal colon, ‘15’).