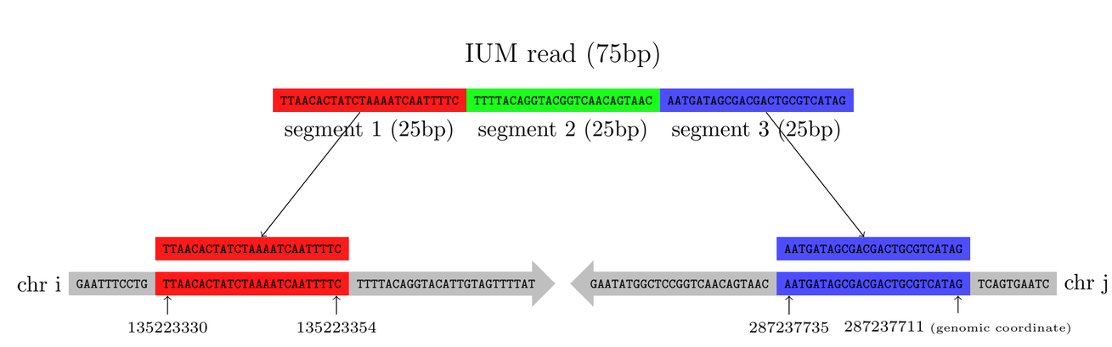
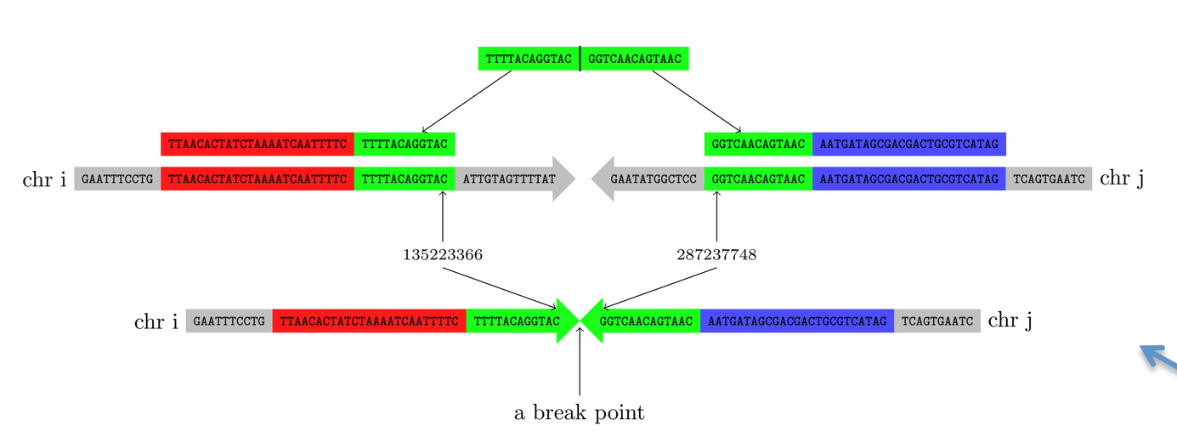
TopHat-Fusion: An algorithm for Discovery of Novel Fusion Transcripts

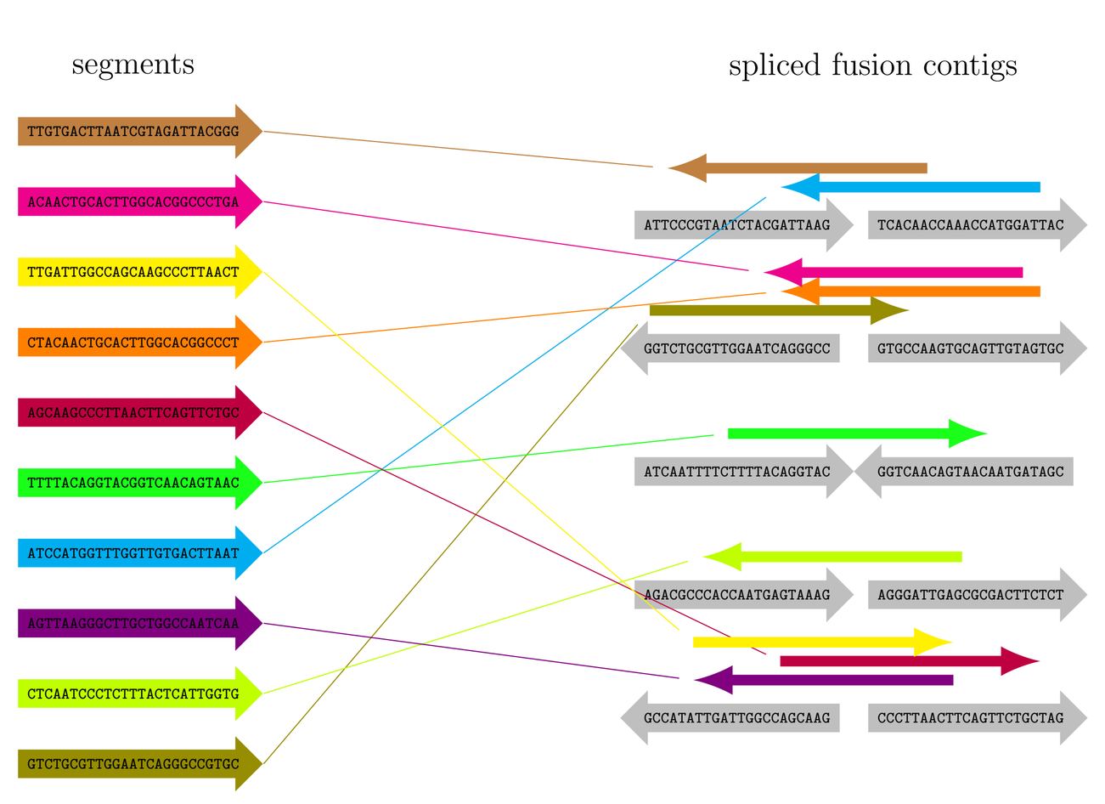
Daehwan Kim and Steven L. Salzberg
Center for Bioinformatics and Computational Biology
University of Maryland, College Park, MD 20742

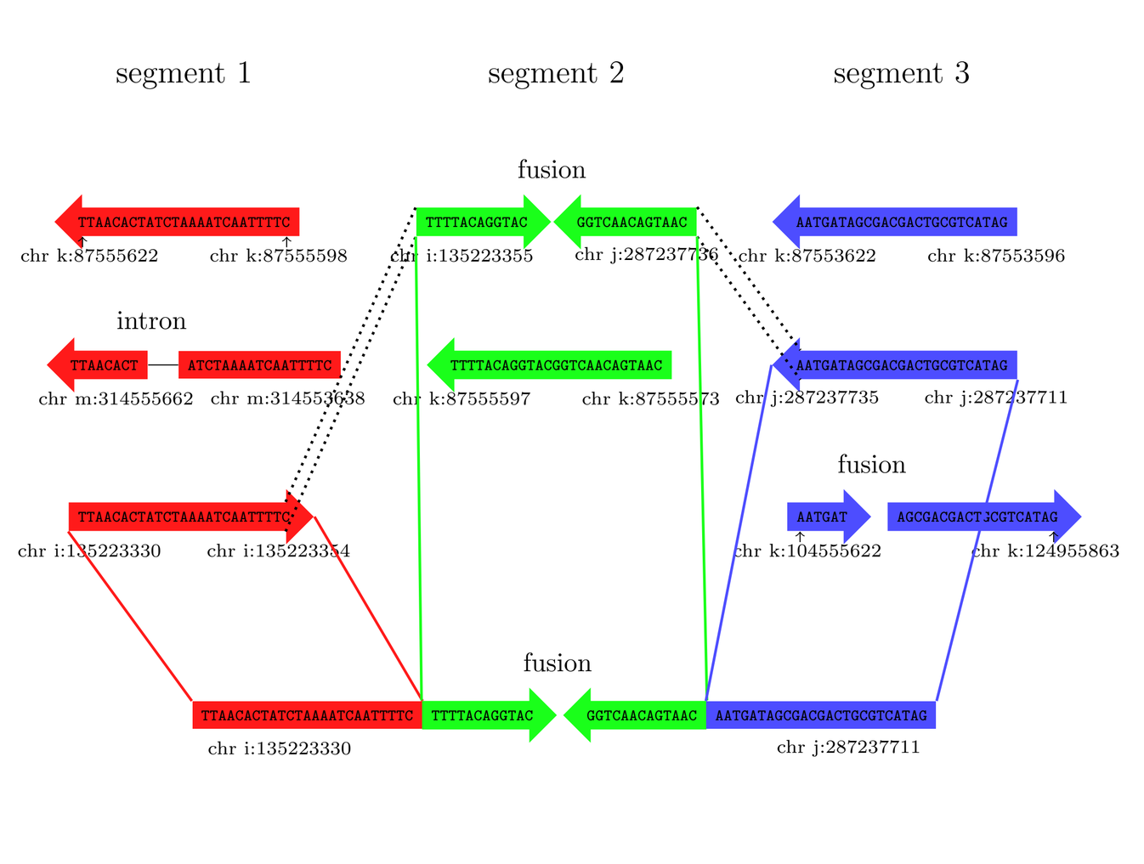
Abstract

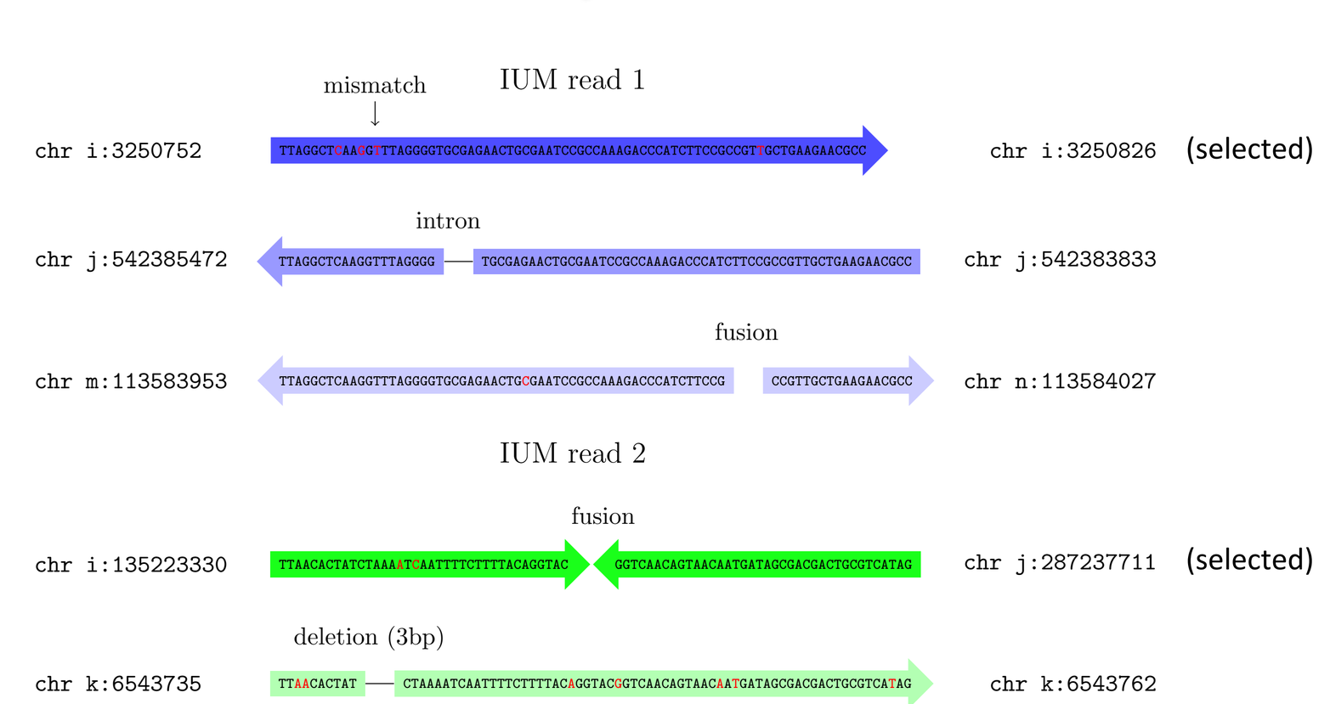
TopHat-Fusion is an enhanced version of TopHat [1] designed to handle transcripts resulting from fusion gene products. Fusion genes result from the breakage and re-joining of two different chromosomes, or from rearrangements within a chromosome. Because TopHat-Fusion does not require information about known genes, it can discover novel fusion products from known genes, unknown genes, or unannotated splice variants of known genes. Using RNA-seq data collected from breast cancer cells [2], prostate cancer cells [3], and Universal Human Reference (UHR) cells [3], we detected multiple fusion genes including novel and previously-reported ones. TopHat-Fusion is free, open-source software available from tophat.cbcb.umd.edu.









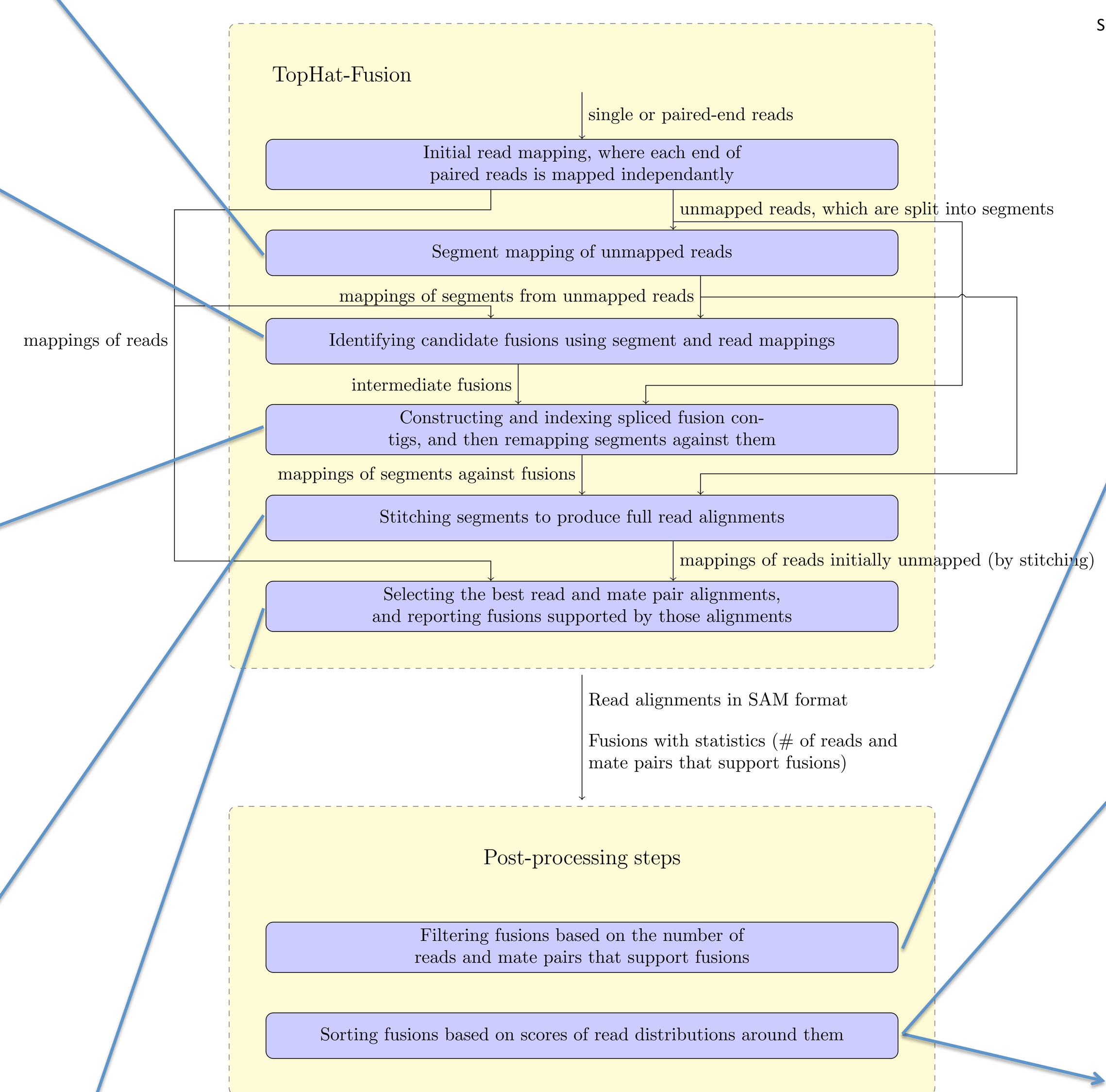


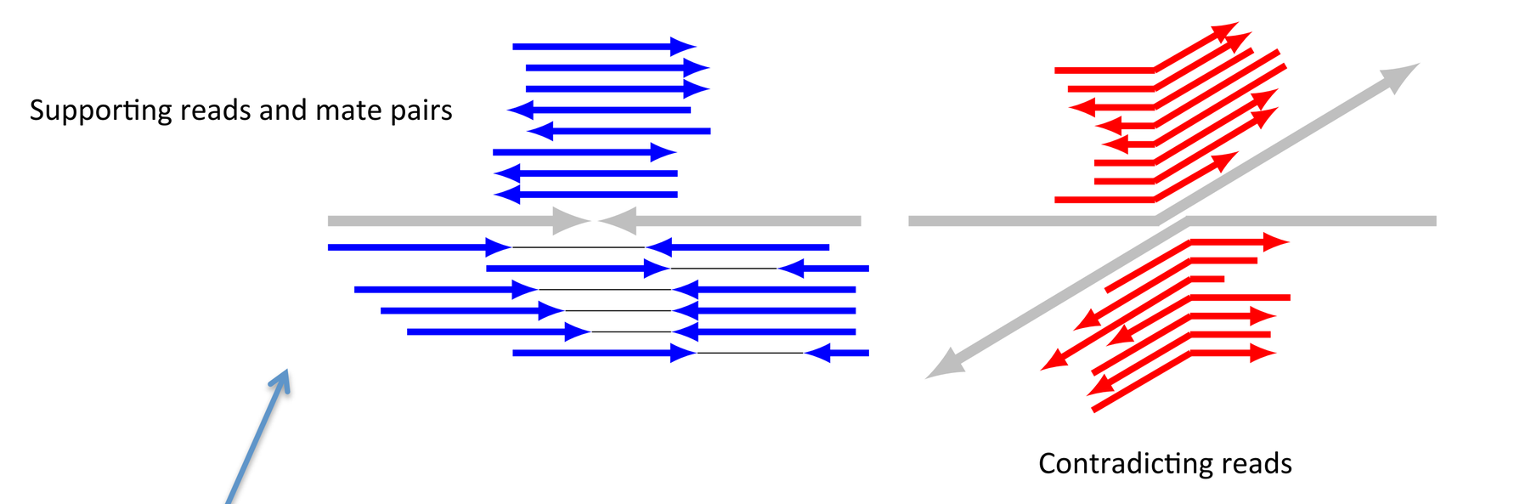
Methods

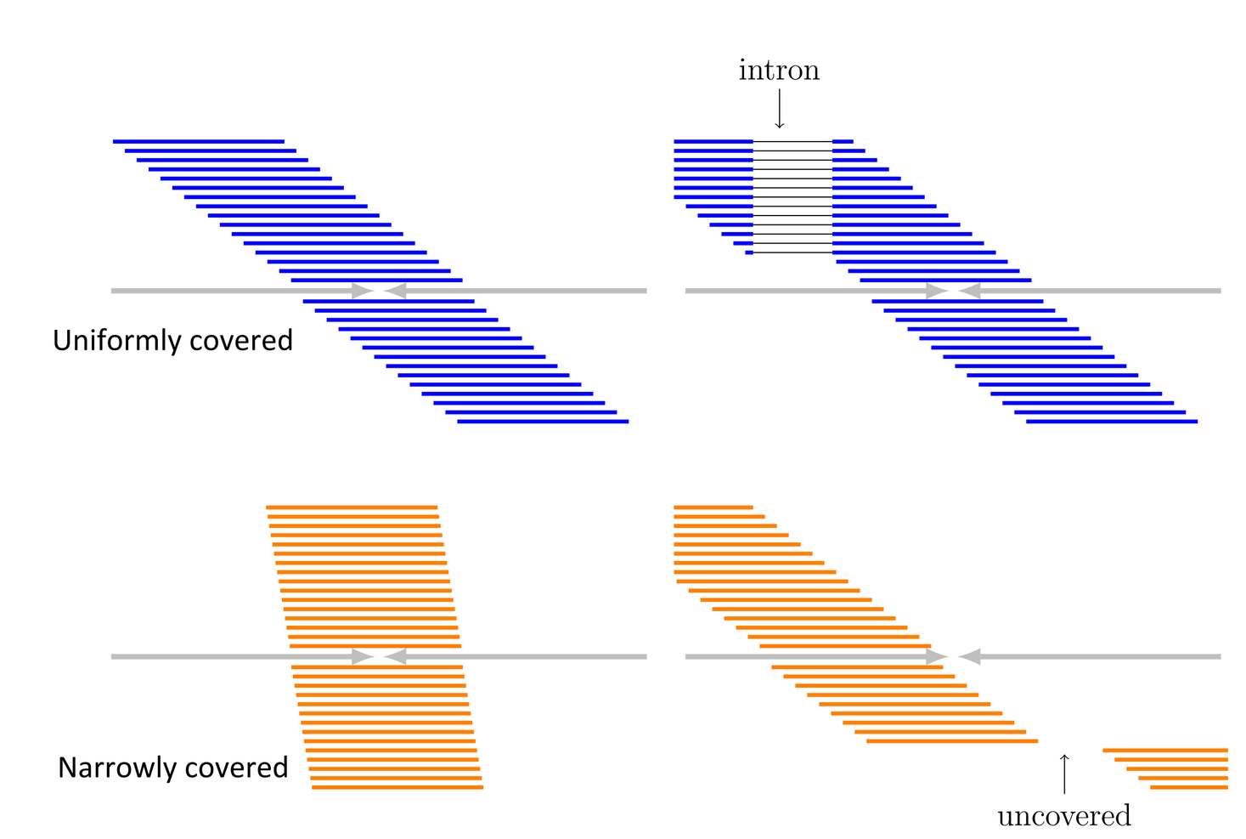
TopHat-Fusion implements several major changes to the original TopHat algorithm, all designed to enable discovery of fusion transcripts as shown in the below figure. The first step in analysis of an RNA-seq data set is to align (map) the reads to the genome. For those initially unmapped (IUM) reads, we split each read into multiple of 25-bp pieces, and then use Bowtie to map the 25-bp segments to the genome. TopHat-Fusion looks for cases where the first and last segments are mapped to either (a) two different chromosomes, in the case of inter-chromosomal fusions; or (b) two locations on the same chromosome separated by a user-defined distance. The whole read is then used to identify a fusion point by re-aligning it starting from the mapped first and last segments.

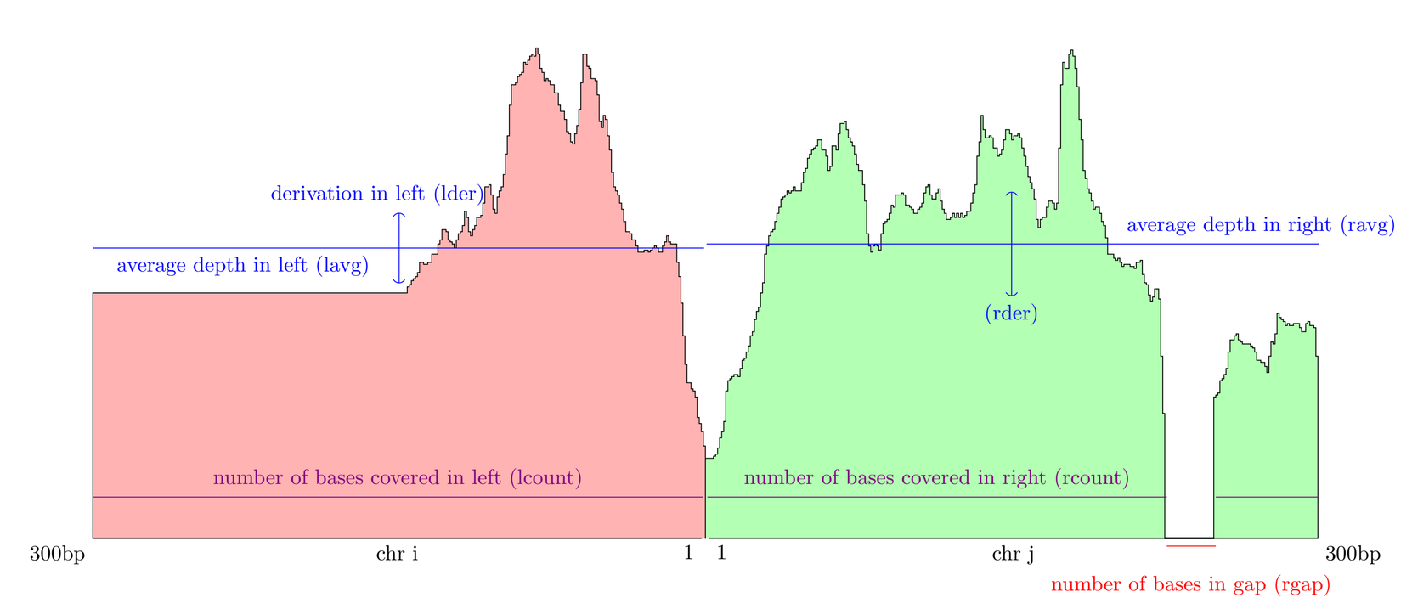
data	sample	type	fragment length	read length	number of fragments (or reads)
Edgren et al.	BT474	paired	100 & 200	50	21,423,697
	SKBR3	paired	100 & 200	50	18,140,246
	KPL4	paired	100	50	6,796,443
	MCF7	paired	100	50	8,409,785
Maher et al.	VCaP	paired	300	50	16,894,522
	UHR	paired	300	50	25,294,164
	UHR	single		100	56,129,471

After finding an initial set of fusion candidates, all the segments from IUM reads are remapped against those fusion points (as well as intron boundaries and indels), and the resulting mappings are combined together to produce full read alignments.









$$score = lcount + rcount + min(max\_avg, lavg) + min(max\_avg, ravg) - |lcount - rcount|$$
$$- min(max\_avg, |lavg - ravg|) - (lgap + rgap) - (lder + rder) * max\_avg + rate - min(1000, dist)$$

, where $max\_avg = 300$, $rate = \frac{number\ of\ supporting\ mate\ pairs}{number\ of\ contradicting\ reads + 0.5}$, $dist$ is how far the closest paired-end is from a fusion point,

$lder = \sqrt{\sum_{n=1}^{lwindow} \frac{(\frac{lavg - ldepth_n}{lavg})^2}{lwindow}}$, where $lwindow$ is the size of the left window (300bp),

and rder is similarly defined.

In case of single-end reads, $rate$ is defined as $\frac{number\ of\ supporting\ reads}{number\ of\ contradicting\ reads + 0.5}$.

Penality for expression difference $|lavg - ravg|$ may be modified or dropped as alternative splicing results in considerable difference in expression levels.

TopHat-Fusion's initial output includes many fusion candidates, most of which are induced by repeats or simply due to chance, with few supporting reads and mate pairs. We applied different filtering parameters for each data set:

(1) At least 5 fusion spanning reads or supporting mate pairs for the breast cancer cell lines (BT474, SKBR3, KPL4, MCF7).
(2) At least 10 supporting reads or mate pairs for the VCaP (prostate cancer cell) and UHR paired-end reads.
(3) At least 2 supporting reads for UHR single-end reads.

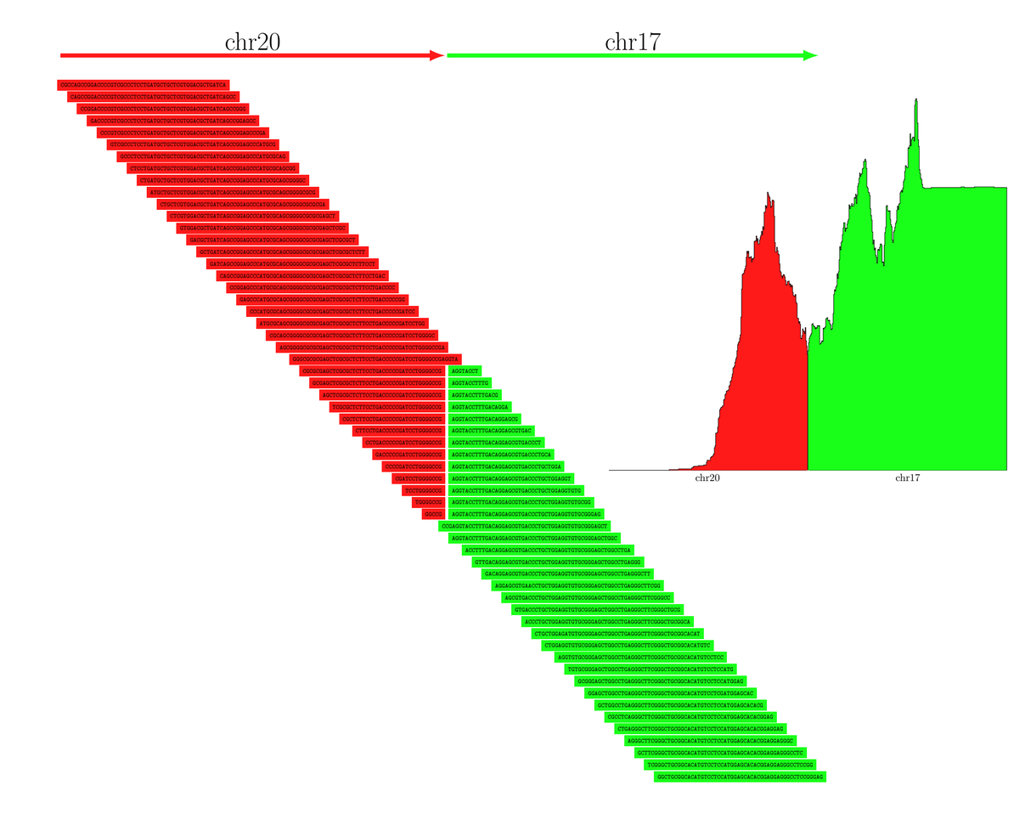
As pointed out in Edgren et al. [2], true fusion transcripts have reads mapping uniformly in a wide window across the fusion point, whereas false positive fusions are narrowly covered. Using this criterion, we examine a 600-bp window around each fusion (300-bp each side), and we reject fusion candidates for which the reads fail to cover this window without any "big" gaps. The final step is to sort candidates based on how well reads are distributed around a fusion point.

Results

Four breast cancer cell lines (BT474, SKBR3, KPL4, MCF7) from Edgren [2]

sample	5′ gene	5′ chromosome	5′ position	3′ gene	3′ chromosome	3′ position	spanning reads	supporting pairs	found*
BT474	TRPC4AP	chr20	33665850	MRPL45	chr17	36476499	2	9	
BT474	ENSG00000141232	chr17	48943418	SYNRG	chr17	35880750	26	47	
SKBR3	TATDN1	chr8	125551264	GSDMB	chr17	38066175	311	555	yes
BT474	THRA	chr17	38243102	SKAP1	chr17	46384689	28	46	
MCF7	BCAS4	chr20	49411707	BCAS3	chr17	59445685	105	284	yes
BT474	ACACA	chr17	35479452	STAC2	chr17	37374425	57	59	yes
BT474	STX16	chr20	57227142	RAE1	chr20	55929087	6	24	
BT474	MED1	chr17	37595419	ENSG00000167107	chr17	48548386	10	12	
SKBR3	ANKHD1	chr5	139825557	PCDH1	chr5	141234002	4	15	yes
BT474	ZMYND8	chr20	45852972	CEP250	chr20	34078459	10	53	yes
BT474	ENSG00000153207	chr1	247094879	NAAA	chr4	76846963	10	42	
SKBR3	SUMF1	chr3	4418012	LRRFIP2	chr3	37170638	3	12	yes
KPL4	BSG	chr19	580779	NFIX	chr19	13135832	12	27	yes
BT474	VAPB	chr20	56964574	IKZF3	chr17	37922743	4	14	yes
BT474	ENSG00000150672	chr11	85195025	HFM1	chr1	91853144	2	10	
SKBR3	CSE1L	chr20	47688988	ENSG00000236127	chr20	47956855	13	31	yes
BT474	MED13	chr17	60129899	BCAS3	chr17	59469335	3	14	
MCF7	ARFGEF2	chr20	47538545	SULF2	chr20	46365686	17	20	yes
BT474	HFM1	chr1	91853144	ENSG00000225630	chr1	565937	2	43	
KPL4	MUC20	chr3	195456612	ENSG00000236833	chr3	197391649	8	15	
MCF7	RPS6KB1	chr17	57992061	TMEM49	chr17	57917126	4	3	yes
SKBR3	WDR67	chr8	124096577	ZNF704	chr8	81733851	3	3	yes
BT474	CPNE1	chr20	34243123	PI3	chr20	43804501	2	6	yes
BT474	ENSG00000229344	chr1	568361	RYR2	chr1	237766339	1	19	
BT474	LAMP1	chr13	113951808	MCF2L	chr13	113718616	2	6	yes
MCF7	SULF2	chr20	46415146	ENSG00000171940	chr20	52210647	11	32	
BT474	WBSCR17	chr7	70958325	FBXL20	chr17	37557612	2	8	
MCF7	ENSG00000224738	chr17	57184949	TMEM49	chr17	57915653	5	6	
MCF7	ANKRD30BL	chr2	133012791	RPS23	chr5	81574161	2	6	
BT474	ENSG00000251948	chr19	24184149	SLCO5A1	chr8	70602608	2	6	
BT474	GLB1	chr3	33055545	CMTM7	chr3	32483333	2	6	yes
BT474	ZNF185	chrX	152114004	GABRA3	chrX	151468336	2	3	
KPL4	PPP1R12A	chr12	80211173	SEPT10	chr2	110343414	3	8	yes
BT474	SKA2	chr17	57232490	MYO19	chr17	34863349	5	12	yes
MCF7	LRP1B	chr2	142237963	PLXDC1	chr17	37265642	2	5	
BT474	NDUFB8	chr10	102289117	TUBD1	chr17	57962592	1	49	
BT474	ENSG00000225630	chr1	565870	NOTCH2NL	chr1	145277319	1	18	
SKBR3	CYTH1	chr17	76778283	EIF3H	chr8	117768257	18	37	yes
BT474	PSMD3	chr17	38151673	ENSG00000237973	chr1	566925	1	12	
BT474	STARD3	chr17	37793479	DOK5	chr20	53259992	2	10	yes
BT474	DIDO1	chr20	61569147	TTI1	chr20	36634798	1	10	yes
BT474	RAB22A	chr20	56886176	MYO9B	chr19	17256205	8	20	yes
KPL4	PCBD2	chr5	134259840	ENSG00000240967	chr5	99382129	1	32	
SKBR3	RARA	chr17	38465535	PKIA	chr8	79510590	1	5	yes
BT474	MED1	chr17	37607288	STXBP4	chr17	53218672	13	11	
KPL4	C1orf151	chr1	19923605	ENSG00000224237	chr3	27256479	1	5	
SKBR3	RNF6	chr13	26795971	FOXO1	chr13	41192773	2	13	
SKBR3	BAT1	chr6	31499072	ENSG00000254406	chr11	119692419	2	30	
BT474	C5orf36	chr5	93904985	PCBD2	chr5	134259811	1	19	
SKBR3	PCBD2	chr5	134263179	ANKRD30BL	chr2	133012790	1	5	
BT474	ENSG00000225630	chr1	565457	MTRNR2L8	chr11	10530147	1	35	
BT474	PCBD2	chr5	134260431	ENSG00000251948	chr19	24184146	2	6	
BT474	ANKRD30BL	chr2	133012085	ENSG00000237973	chr1	567103	2	8	
KPL4	ENSG00000225972	chr1	564639	HSP90AB1	chr6	44220780	1	7	
BT474	MTIF2	chr2	55470625	ENSG00000228826	chr1	121244943	1	11	
BT474	ENSG00000224905	chr21	15457432	PCBD2	chr5	134263223	2	7	
BT474	RPS6KB1	chr17	57970686	SNF8	chr17	47021335	48	57	yes
BT474	MTRNR2L8	chr11	10530146	PCBD2	chr5	134263156	1	6	
BT474	RPL23	chr17	37009355	ENSG00000225630	chr1	565697	3	19	
BT474	MTRNR2L2	chr5	79946288	PCBD2	chr5	134259832	1	5	
SKBR3	ENSG00000240409	chr1	569005	PCBD2	chr5	134260124	2	4	
SKBR3	PCBD2	chr5	134263289	ENSG00000239776	chr12	127650986	2	3	
BT474	ENSG00000239776	chr12	127650981	MTRNR2L2	chr5	79946277	2	3	
BT474	JAK2	chr9	5112849	TCF3	chr19	1610500	1	46	
KPL4	NOTCH1	chr9	139438475	NUP214	chr9	134062675	3	5	yes
BT474	MTRNR2L8	chr11	10530594	ENSG00000240215	chr9	33657801	4	4	
BT474	MTRNR2L8	chr11	10530179	AKAP6	chr14	32953468	1	5	
BT474	ENSG00000230916	chrX	125606246	PCBD2	chr5	134263219	1	5	
MCF7	ENSG00000226505	chr2	70329650	MRPL36	chr5	1799907	5	20	
SKBR3	CCDC85C	chr14	100002351	SETD3	chr14	99880270	5	6	yes
BT474	RPL23	chr17	37009955	ENSG00000230406	chr2	222457168	109	5	

TopHat-Fusion found 25 out of 27 fusion genes reported in the four breast cancer cells from Edgren et al [2], where they are indicated by "found*" in the left table. One of the two missing fusions, DHX35-ITCH, was included in the initial output, but it was filtered out because it was only supported by one read and one mate pair.



The above figure shows read distribution around the BCAS4-BCAS3 fusion in MCF7 cell.

VCaP data from Maher et al. [3]

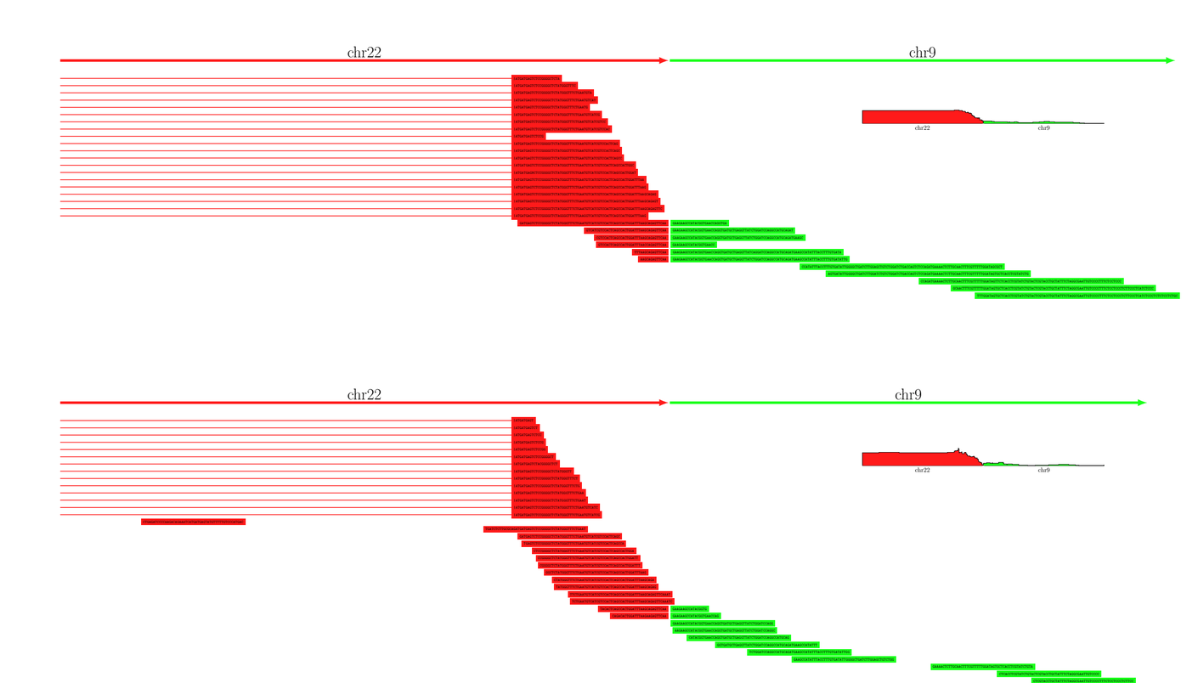
5′ gene	5′ chromosome	5′ position	3′ gene	3′ chromosome	3′ position	spanning reads	supporting pairs	found*
ZDHHC7	chr16	85023908	ABCB9	chr12	123444867	13	69	yes
TMPRSS2	chr21	42879875	ERG	chr21	39817542	7	285	yes
HJURP	chr2	234749254	EIF4E2	chr2	233421125	3	9	yes
VWA2	chr10	116008521	PRKCH	chr14	61909826	1	10	
RGS3	chr9	116299195	PRKAR1B	chr7	699055	3	11	
SPOCK1	chr5	136397966	TBC1D9B	chr5	179305324	9	31	yes
LRP4	chr11	46911864	FBXL20	chr17	37557613	5	9	
INPP4A	chr2	99193605	HJURP	chr2	234746297	6	12	yes
C16orf70	chr16	67144140	C16orf48	chr16	67700168	2	19	
NDUFV2	chr18	9102729	ENSG00000188699	chr19	53727808	1	35	
NEAT1	chr11	65190281	ENSG00000229344	chr1	568419	1	17	
ENSG00000011405	chr11	17229396	TEAD1	chr11	12883794	7	9	yes
USP10	chr16	84733713	ZDHHC7	chr16	85024243	1	22	yes
LMAN2	chr5	176778452	AP3S1	chr5	115202366	15	2	yes
WDR45L	chr17	80579516	ENSG00000224737	chr17	30439195	1	33	
RC3H2	chr9	125622198	RGS3	chr9	116299072	3	11	yes
CTNNA1	chr5	138145895	ENSG00000249026	chr5	114727795	1	12	
ENSG00000229880	chr21	46097128	IMMT	chr2	86389185	1	50	
ENSG00000214009	chrX	45918367	PCNA	chr20	5098168	1	24	

Maher et al. reported 11 fusion genes in the VCaP sample, 9 of which were identified by TopHat-Fusion. It missed a fusion between two overlapping genes ZNF577 and ZNF649 on chromosome 19. That fusion event appears to be due to read-through transcription.

UHR single-end reads vs. paired-end reads [3]

Using four previously known fusion genes GAS6-RASA3, BCR-ABL1, ARFGEF2-SULF2, and BCAS4-BCAS3, we compared the results of single and paired-end reads from the UHR RNA-seq data. The number of supporting **r**eads **p**er **m**illion reads (RPM*) indicates that single-end reads find the same fusions as paired-end reads. The two figures show read distributions around the BCR-ABL1 fusion for single (upper) and paired-end reads (lower), respectively.

type	5′ gene	5′ chromosome	5′ position	3′ gene	3′ chromosome	3′ position	spanning reads (RPM*)	supporting pairs
single	GAS6	chr13	114529968	RASA3	chr13	114751268	15 (0.267)	
paired	GAS6	chr13	114529968	RASA3	chr13	114751268	10 (0.198)	43
single	BCR	chr22	23632599	ABL1	chr9	133655755	6 (0.107)	
single	BCR	chr22	23632599	ABL1	chr9	133729450	3 (0.053)	
paired	BCR	chr22	23632599	ABL1	chr9	133655755	2 (0.040)	7
paired	BCR	chr22	23632599	ABL1	chr9	133729450	3 (0.059)	10
single	ARFGEF2	chr20	47538548	SULF2	chr20	46365683	17 (0.302)	
paired	ARFGEF2	chr20	47538545	SULF2	chr20	46365686	10 (0.198)	30
single	BCAS4	chr20	49411707	BCAS3	chr17	59445685	25 (0.445)	
paired	BCAS4	chr20	49411707	BCAS3	chr17	59445685	13 (0.257)	145



Acknowledgements

We would like to thank Ryan Kelly for his indel-finding algorithm, which he added to TopHat and which we had used as part of the fusion finding algorithm. This work is supported in part by the National Institutes of Health under grants R01-LM006845 and R01-GM083873 to SLS.

References

1. Trapnell, C., L. Pachter, and S.L. Salzberg, TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 2009. **25**(9): p. 1105-11.
2. Edgren, H., et al., Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome biology, 2011. **12**(1): p. R6.
3. Maher, C.A., et al., Chimeric transcript discovery by paired-end transcriptome sequencing. Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(30): p. 12353-8.