Targeted Assembly of Short Sequence Reads

René L. Warren [1,*], Robert A. Holt[1,2]

[1]British Columbia Cancer Agency, Genome Sciences Centre, 675 W. 10th Avenue, Vancouver, BC V5Z 1L3 Canada

[2]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, Canada

As next-generation sequence (NGS) production continues to increase, analysis is becoming a significant bottleneck. However, in situations where information is required only for specific sequence variants, it is not necessary to assemble or align whole genome data sets in their entirety. Rather, NGS data sets can be mined for the presence of sequence variants of interest by localized assembly, which is a faster, easier, and more accurate approach. We present TASR, a streamlined assembler that interrogates very large NGS data sets for the presence of specific variants, by only considering reads within the sequence space of input target sequences provided by the user. The NGS data set is searched for reads with an exact match to all possible short words within the target sequence, and these reads are then assembled stringently to generate a consensus of the target and flanking sequence. Typically, variants of a particular locus are provided as different target sequences, and the presence of the variant in the data set being interrogated is revealed by a successful assembly outcome. However, TASR can also be used to find unknown sequences that flank a given target. We demonstrate that TASR has utility in finding or confirming genomic mutations, polymorphism, fusion and integration events. Targeted assembly is a powerful method for interrogating large data sets for the presence of sequence variants of interest. TASR is a fast, flexible and easy to use tool for targeted assembly.

INTRODUCTION

The revolution in DNA sequencing technologies has enabled faster and cheaper data generation, to the point where data collection is becoming a less concerning bottleneck than data storage and analysis [1]. This is especially true for laboratories with limited informatics resources. TASR (Targeted Assembly of Sequence Reads) builds upon our previous SSAKE assembler [2] but unlike its predecessor, it only considers reads for assembly that have a perfect 15nt word match to input target sequences. Thus, TASR has particular utility for finding or confirming, through local assembly, the presence of specific sequences or sequence variants of interest. To our knowledge, de Bruijn graph assemblers published to date do not have this functionality. Here we demonstrate its utility for discriminating real from artifactual variant calls in tumour genomes, which may facilitate large-scale validation efforts. Further, we show that by targeted assembly it is possible to identify tumour-associated fusion transcripts and, finally, we demonstrate the utility of targeted assembly for identifying polymorphisms of interest in ancient human DNA.

RESULTS

Algorithm

DNA sequence reads in a fastq or fasta format are fed into the algorithm via a file of filenames, using the –f option. DNA sequence targets, used to interrogate all raw reads in a sequence data set, are supplied as a multi fasta file using the –s option. Sequence targets are read first. From each target, every possible 15-character word from the plus and minus strand are extracted and stored in a hash table. Next, reads from the NGS data set are interrogated as described in [2], except that rather than using a greedy algorithm, any read with an exact match of its first 15 bases to any of the 15-mer words from the target sequence, is retained. These reads are collected in an array for subsequent assembly, thereby limiting the sequence space of the assembly to that of the target region. Note that low-complexity and large DNA sequence targets will draw in more reads, which will impact the performance of TASR. The identity and coverage of every base, within and beyond the user-provided target sequence, is stored in a hash table c. The sequence within the bounds of the user-supplied target sequence will exactly match the target itself, but recruited sequence reads will typically extend beyond the boundaries of the target sequence, and this flanking sequence may also be included in the assembly. In some instances the identity of the sequence that flanks the target may be unknown and may in fact be of greatest interest to the user. A consensus sequence is derived, taking exactly matching bases at each position within the target region, and extended outward, bi-directionally, to include the most represented base at positions outside the target sequence. In this regard, TASR is unchanged from the most recent version of SSAKE (v3.7) where consensus bases, situated outside the target region are derived using a majority-rule approach analogous to that of VCAKE [3]. For extension, each base has to be covered by user-defined -o (set to 2 by default) and its abundance relative to the next most called base equal or above the user-defined ratio –r (0.7). Extension is terminated when a position is encountered that does not meet these user-specified criteria. This process is repeated for each target sequence supplied in the –s file. TASR outputs target-derived contigs in fasta format, read positions and base coverage in text files and per-position information in the pileup format [4].

Implementation

TASR is implemented in PERL and will run on any platform where PERL is installed. It is available (under GPL licence) from: http://www.bcgsc.ca/bioinfo/software/tasr

* To whom correspondence should be addressed. e-mail: rwarren@bcgsc.ca

and supplemental data files from: ftp://ftp.bcgsc.ca/supplementary/TASR

Testing

TASR runs were executed on a shared computer running CentOS 5.5, with 2 Intel Xeon X5680 CPUs at 3.33Ghz (12 core, 24 threads) and 48GB of RAM. At most, using our larger data set (3.5B Saqqaq Paleo-Eskimo NGS reads), TASR required <10M RAM and ran for 12 hours.

Verifying candidate SNVs in a lobular breast cancer genome

Shah and co-workers [5] reported 32 confirmed somatic coding single nucleotide variants (SNVs) in a metastatic lobular breast cancer specimen. These confirmed variants resulted from testing, by capillary re-sequencing, of a larger number of putative variants that were originally suggested by whole tumour genome shotgun sequencing (WGSS) using the Illumina platform and running Maq [6] and SNVmix [7]. We interrogated 31 of these verified SNVs using 51nt sequence targets containing either the mutant or the HG18 reference base. We also selected, at random, 31 SNVs that had been tested by capillary re-sequencing, but not verified (Shah et al., unpublished data). The sequence data, providing up to 36-fold coverage of the human genome, was analyzed incrementally (Figure 1) using TASR (default options). Maximum sensitivity was reached at moderate coverage (ca. 36-fold) where 29/31 (93.5%) of previously verified SNV were positive for the variant in question. Interestingly 30/31 (96.8%) also showed the reference base, reflecting the cellular heterogeneity of the tumour. At this same coverage only 9/31 of variants that failed previous verification by capillary sequencing showed the SNV, and of these, 7 showed the reference base. This offers an improved level of discrimination for putative SNVs and may streamline ongoing efforts to verify putative tumour mutations.

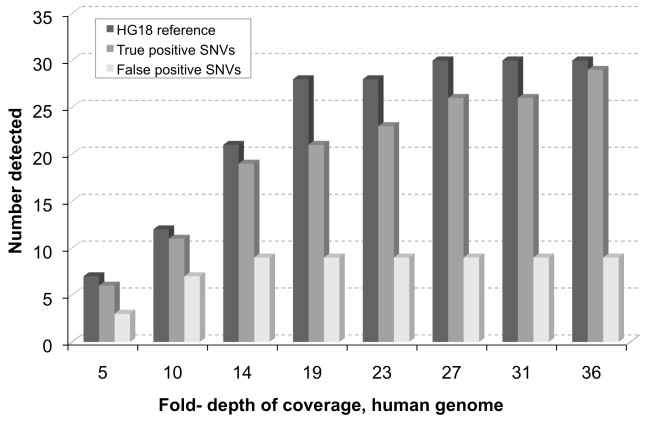

Figure 1. Detection of true positive versus false positive SNVs in lobular breast cancer (LBC). TASR was run incrementally on up to 2 billion, 51 and 76nt lobular breast cancer NGS whole-genome shotgun reads, providing 5 to 36-fold coverage of the 3Gbp human genome. We used as targets 51nt sequences containing one of 31 SNVs detected by NGS read alignment and confirmed by Sanger sequencing (true positive), 31 matching sequences containing the reference base instead (reference) and 31 detected by NGS read alignment but not confirmed by Sanger sequencing (false positive). Although close to twice as much WGSS data had been generated from the LBC sample, we see that a fraction of that (~19-fold) is sufficient for confirming most (68%) true positive SNVs.

Detecting an HPV fusion in HeLa cells

As part of the original development of NGS RNA-seq methodology, Morin and colleagues [8] evaluated HPV18 transcription in HeLa cell lines, which are known to carry this viral integration. We set out to determine the human-HPV fusion site by targeted assembly of these RNA-seq reads. Using four 38nt sequence targets each comprising the same 37 HPV-specific bases preceding one of four possible DNA base as the 38th base, respectively, we used TASR (default options) to interrogate 37.4M RNA-seq reads. A single sequence target was extended into a contig that also comprised human cDNA sequences. Overall, 288 NGS reads co-assembled and 51 chimaeric reads, each having 1 or more HPV18- and human-only base(s) covered the fusion site.

Detecting fusion transcripts in prostate adenocarcinoma

For each of the RNA-seq data sets corresponding to three adenocarcinoma and 3 non-tumour adjacent prostate tissue samples, we looked for the presence of a fusion gene, TMPRSS2:ERG, which is known to be common in prostate adenocarcinoma, and is a strong prognostic indicator [9]. Using two 50nt target sequences, one containing the last 36bp of exon 1 and the first 14bp of exon 2 and the other the last 36bp of TMPRSS2 exon1 and first 14bp of ERG exon 4, we ran TASR on each set (-m 15 –c 1 other options defaulted). In another experiment, we used two 38nt sequence targets that differed only by their last 3' base, simulating a scenario where very little information is available about a given event (Figure 2). We designed both experiments with the aim of detecting portion of the TMPRSS2 and TMPRSS2:ERG transcripts and ran TASR under the same conditions. In both experiments, we found the fusion in 2/3 adenocarcinoma samples and 1/3 adjacent normal samples with reads spanning the fusion coordinate and containing both unique ERG and TMPRSS2 bases. The TMPRSS2-only target also yields a contig for a non-fusion transcript. The number of fusion reads is generally lower than that of the TMPRSS2 transcript (adenocarcinoma NGS data SRX027125, Figure 2) and may be an indicator of lower expression of the fusion, and/or cellular mosaicism. Although we used the entire available SRA data for each corresponding sample, we noticed that a single sequence run (e.g. SRR066437 ~4.7M spots) was sufficient for detection of the fusion in positive samples.

Detecting SNPs in ancient human DNA

Rasmussen and colleagues [10] used a 4,000 year old sample of perma-frost-preserved hair to obtain the genome sequence of an early Paleo-Eskimo settler of Greenland, and reported single nucleotide polymorphisms (SNPs) including those known to confer phenotypes such as black hair color, dry ear wax, higher % fat mass, cold adaptation, not European light skin, and thick hair/shovel shaped upper front teeth. We interrogated the sequence data (3.5 billion 70nt

reads obtained from the SRA) to determine whether targeted assembly could recover these specific, but no other, SNPs. We ran TASR (default options) using sequence targets that were each 70nt in length and contained the variant or reference alleles. By providing a comprehensive target sequence file that accounts for all known polymorphisms within the alleles tested, we hypothesized that only reads having the legitimate base will be recruited and co-assemble with the appropriate target sequence. We found read support for all six variants and no evidence of other alleles at these positions. Read coverage over each SNP is variable with the cold adaptation, high % fat mass, not European light skin, black hair, dry earwax and thick hair-associated polymorphisms covered by 3, 7, 7, 17, 22, 34 reads, respectively. None of the negative controls (sequence targets comprised of any possible alternate alleles from dbSNP) co-assembled sequence reads over the base under scrutiny.

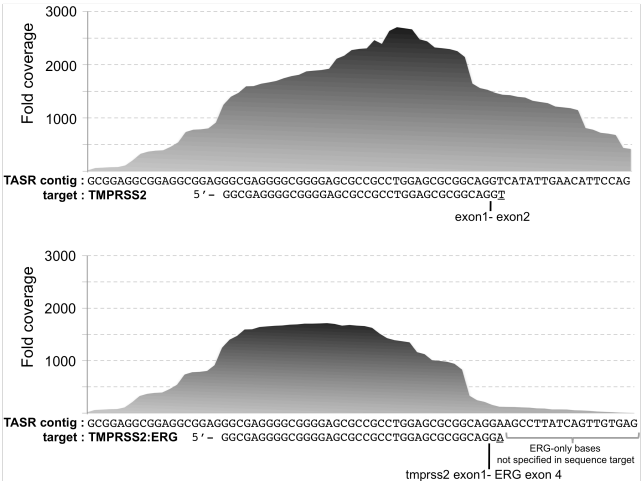

Figure 2. De novo assembly of prostate carcinoma RNA-seq data. Using a TMPRSS2:ERG target sequence that differs from a TMPRSS2 target by a single base (underlined), TASR generated a contig, which captures 18 ERG-specific bases fused to exon 1 of TMPRSS2 in a prostate adenocarcinoma sample (SRA accession SRX027125). These bases were not specified in the target sequence and thus, unknown from the original hypothesis. A total of 121 reads span the TMPRSS2:ERG fusion coordinate (underlined base). Higher base coverage is expected in the middle of the contig where 15-mer read recruitment reaches a maximum for both strand and is unaffected by the limiting effects of the minimum overlap (-m) option on the edge of the sequence target. This highlights the importance of using a sequence target that is sufficiently long and at least the same length as the input reads. From this result, it is very likely that the prostate adenocarcinoma sample contains an admixture of TMPRSS2 transcripts, including the TMPRSS2{NM_005656.2}:r.1_71_ERG{NM_004449.3}:r.226_3097 fusion and that those have varied abundance, as reflected by high depth of coverage.

DISCUSSION

TASR is a targeted *de novo* assembler that uses supplied sequences to target initial read recruitment and assembly. The targets can be any sequence, actual read, existing reference or synthetic sequence. For example, when testing for fusion transcripts, or any other unknown sequence flanking a target, NGS data sets could be interrogated using four distinct target sequences, each with one of the four possible nucleotides as their last 3' base. This is a key advantage of targeted assembly, since alignment of NGS sequence to a complete reference would not be expected to return sequence reads that contained a significant number of non-reference bases. Likewise, reads representing a rare fusion or insertion event may be excluded from whole genome *de novo* assemblies if they have low representation in the NGS raw data. Further, most large-scale *de novo* assemblies are precluded by the shear data size, such as those processed in our study (e.g. 3.5B Saqqaq Paleo-Eskimo NGS reads from 238 Illumina sequence lanes). Although the development of *de novo* assemblers such as ABySS [11] and SOAPdenovo [12], now makes whole-genome and whole-transcriptome human NGS read assembly a reality, researchers would still have to sift through thousands, if not millions, of contigs for sequences of interest. We provide a solution that allows relatively quick (~3.5B reads from 238 sequencing lanes providing 81X average coverage of a 3GB human genome, processed in 12 hrs) and flexible hypothesis testing that is targeted to a genomic region of interest, whether it comprises a fusion, translocation, SNV or SNP.

In TASR, the assembly results are influenced strongly by the design of the target sequences used as input. Targets that are as long as the shortest read being considered and shorter than two read sizes (e.g. a 70bp target could be extended by 2 * (70 – m option) will produce the best results. This length will ensure that all overlapping 15-mer from a given target recruit the maximum number of candidate reads for assembly. Also, the use of longer NGS reads (>70) will increase the chance of generating longer contigs with more novel or previously unknown bases, which is instrumental in the characterization of previously unknown events, such as the detection of a viral integration site. For example, using 33nt NGS reads and setting the minimum overlap to 15 could extend the target on each side by at most 18bp (33-15) whereas the use of 150nt reads with 150nt targets could yield 420bp contigs having up to 270 previously uncharacterized bases.

Frequently, it may be of interest to mine a NGS data set for the presence of one or more single nucleotide variants of interest; for example, variants that have known associations to specific traits or genetic disorders. For SNV or SNP detection, it is prudent to have the base under scrutiny in the middle of the target, to increase the chances of recruiting candidate reads that have the particular base at any possible position. It may seem peculiar to use a *de novo* assembler to help validate single-base changes in genomes, especially now that fast and large-scale read alignment methods, such as bwa [13], exist. TASR has the advantage of 1) conducting targeted assembly to a specific region and thus, alleviate the need of sifting through large alignment or assembly files. Also, 2) it is very stringent in that it will only recruit and co-assemble NGS reads whose bases overlapping the target sequence are in perfect agreement. This has the advantage of rapidly testing a simple hypothesis such as whether a locus has the reference base, a variant or both, by looking at a read pileup over the base under scrutiny. Lastly, 3) it performs *de novo* assemblies, such that overlapping bases that fall outside the target region have the potential to characterize a novel sequence.

CONCLUSIONS

The utility of TASR is to interrogate specific target sequences by local assembly. The targets can be any sequence, an actual read, a reference or synthetic sequence. Here we demonstrate, using TASR, the mining of NGS data sets for fusion transcripts and two types of single nucleotide variants (SNV), somatic mutations in tumour genomes and polymorphisms in ancient DNA. TASR uses a stringent target-targeted assembly scheme, where the more complex and unique a target sequence is, the less likely non-specific reads are to co-assemble, facilitating variant detection. TASR may be used in a manner that is complementary to NGS read alignment tools, in order to help confirm false-positive events that may result from NGS sequencing or mapping errors. As it performs a *de novo* assembly of reads outside the target region, it may be used for the targeted assembly of chimaeric reads whose bases may help characterize novel fusion, translocation or integration events. Since TASR does not require indexed databases or multi-staged runs it is easy to use. It runs on commodity hardware with a low computer resource footprint.

METHODS

Lobular breast cancer (LBC) whole-genome sequence data were previously described [5]. We processed ~2B paired-end reads (76 and 51bp) which provided ~36-fold coverage of the human genome. The HeLa RNA-seq data (37.4M single-end 31bp reads) was obtained from Morin and colleagues [8]. RNA-seq data from 3 human prostate adenocarcinoma and 3 matched adjacent normal samples was obtained from the SRA (SRP003611). An average of 32M reads (33bp) was downloaded for each of 5 samples, and from the 6th sample, SRX027125, we obtained 65.5M reads. All of the 3.5B whole-genome shotgun sequences (70bp reads in 238 fastq files totalling 878 Gbytes) from the extinct Saqqaq Paleo-Eskimo [10] was obtained from the SRA (SRP001453), for the purpose of targeted assembly of SNPs (dbSNP:rs5746059, rs17822931, rs16891982, rs1426654, rs3827760, rs1042522).

ACKNOWLEDGEMENTS

We thank Marco Marra and Samuel Aparicio from the BC Cancer Agency for sharing the HeLa and LBC NGS data, respectively.

REFERENCES

1. Stein LD: The case for cloud computing in genome informatics. Genome Biol. 2010, 11:207
2. Warren RL, Sutton GG, Jones SJ, Holt RA: Assembling millions of short DNA sequences using SSAKE. Bioinformatics 2007, 23:500-501
3. Jeck WR, Reinhardt JA, Baltrus DA, Hickenbotham MT, Magrini V, Mardis ER, Dangl JL, Jones CD: Extending assembly of short DNA sequences to handle error. Bioinformatics 2007, 23:2942-2944
4. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup: The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009, 25: 2078-2079
5. Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, Burleigh A, Delaney A, Gelmon K, Guliany R, Senz J, Steidl C, Holt RA, Jones S, Sun M, Leung G, Moore R, Severson T, Taylor GA, Teschendorff AE, Tse K, Turashvili G, Varhol R, Warren RL, Watson P, Zhao Y, Caldas C, Huntsman D, Hirst M, Marra MA, Aparicio S: Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature 2009, 461:809-813
6. Li H, Ruan J, Durbin R: Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008, 18:1851–1858
7. Goya R, Sun MG, Morin RD, Leung G, Ha G, Wiegand KC, Senz J, Crisan A, Marra MA, Hirst M, Huntsman D, Murphy KP, Aparicio S, Shah SP: SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics 2010, 26:730-736
8. Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, McDonald H, Varhol R, Jones S, Marra M: Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques 2008, 45:81-94
9. Nam RK, Sugar L, Yang W, Srivastava S, Klotz LH, Yang LY, Stanimirovic A, Encioiu E, Neill M, Loblaw DA, Trachtenberg J, Narod SA, Seth A: Expression of the TMPRSS2:ERG fusion gene predicts cancer recurrence after surgery for localised prostate cancer. Br. J. Cancer 2007, 97:1690–1695
10. Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, Gupta R, Bertalan M, Nielsen K, Gilbert MT, Wang Y, Raghavan M, Campos PF, Kamp HM, Wilson AS, Gledhill A, Tridico S, Bunce M, Lorenzen ED, Binladen J, Guo X, Zhao J, Zhang X, Zhang H, Li Z, Chen M, Orlando L, Kristiansen K, Bak M, Tommerup N, Bendixen C, Pierre TL, Grønnow B, Meldgaard M, Andreasen C, Fedorova SA, Osipova LP, Higham TF, Ramsey CB, Hansen TV, Nielsen FC, Crawford MH, Brunak S, Sicheritz-Pontén T, Villems R, Nielsen R, Krogh A, Wang J, Willerslev E: Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature 2010, 463:757-762
11. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: ABySS: a parallel assembler for short read sequence data. Genome Res. 2009, 19:1117–1123
12. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J: De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010, 20:265-272
13. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009, 25:1754-1760