

METHOD

Open Access

A simple consensus approach improves somatic mutation prediction accuracy

David L Goode^{1,5*}, Sally M Hunter^{2†}, Maria A Doyle³, Tao Ma^{3,6}, Simone M Rowley², David Choong², Georgina L Ryland^{2,4} and Ian G Campbell^{2,5,7}

Differentiating true somatic mutations from artifacts in massively parallel sequencing data is an immense challenge. To develop methods for optimal somatic mutation detection and to identify factors influencing somatic mutation prediction accuracy, we validated predictions from three somatic mutation detection algorithms, MuTect, JointSNVMix2 and SomaticSniper, by Sanger sequencing. Full consensus predictions had a validation rate of >98%, but some partial consensus predictions validated too. In cases of partial consensus, read depth and mapping quality data, along with additional prediction methods, aided in removing inaccurate predictions. Our consensus approach is fast, flexible and provides a high-confidence list of putative somatic mutations.

Background

Massively parallel sequencing (MPS) of cancer exomes is becoming a commonplace technique, and has led to the identification of genes underlying the pathogenesis of a number of cancer types [1-6]. In response to the volume of data generated by these genome-scale studies, a host of software tools has been developed to aid in distinguishing genuine somatic mutations from germline variation, alignment artifacts, and inherent MPS errors [7-11]. The rarity and diversity of somatic events that occur on a background of tumor heterogeneity, normal contamination, technical artifacts, and genomic complexity makes this task particularly challenging [1,12].

Although the methodology applied by somatic mutation algorithms varies somewhat, the aim of each program is to identify tumor-specific variants by comparing sequence data from a tumor with that generated from a normal tissue (representing the germline) from the same patient (that is, matched normal DNA). The most common application is the identification of point mutations. The germline sample is usually assumed to be free of genetic material from the tumor, although this assumption can be tested and

corrected for [12,13]. At every site where there are reads that differ from the reference genome, the probability that these reads contain legitimate genetic variants and not sequencing errors or technical artifacts is calculated. The probabilities for the tumor and germline data are compared, and a prediction about whether the site harbors a somatic mutation is made [7-11]. From this, a list of putative somatic mutations and associated confidence values is produced, which can be used in downstream analyses.

The choice of somatic mutation detection algorithm may have an important influence on the outcome of a tumor exome-sequencing study. Incorporating more information from a sequencing run (such as site-specific mapping and base qualities) improves the performance of variant detection over that of *ad hoc* metrics based on read counts alone [7,10,14,15]. Thus, it would be expected that predictions from different algorithms, which weigh different properties of the data in unique ways, may differ significantly. A conservative algorithm with high specificity may make very few incorrect predictions, but may miss many legitimate somatic mutations because of its low sensitivity. Similarly, a high-confidence set of somatic mutation predictions with a low false-positive (FP) rate is very useful in a clinical setting, but in a discovery-based research setting, it could limit the power to identify novel mutated genes and pathways [16]. This is important given the small number of recurrently mutated tumor driver genes, and the long list of infrequently mutated, yet biologically

* Correspondence: david.goode@petermac.org

†Equal contributors

¹Peter MacCallum Cancer Centre, Sarcoma Genetics and Genomics Laboratory, St. Andrew's Place, East Melbourne, Victoria, Australia

⁵Sir Peter MacCallum Department of Oncology, University of Melbourne, Parkville, Victoria, Australia

Full list of author information is available at the end of the article

important, targets identified in many cancer types to date [17,18]. Conversely, although a more inclusive algorithm may have high sensitivity, the resolution of the follow-up analyses could be diminished by the inclusion of multiple FPs, which would also strain resources dedicated to validating mutations.

In most studies, only a single algorithm is used to predict which variants are somatic, with predicted mutations validated by an orthogonal sequencing technology, such as Sanger sequencing [2,4,6]. Although a number of published papers on somatic mutation detection programs have reported robust sensitivity and low error rates, the relative advantages of selecting one program over any other are not clear, especially as few programs have been evaluated extensively for their sensitivity and specificity in an unbiased fashion in independent tumor cohorts. A naive yet logical strategy could be to employ multiple somatic mutation prediction algorithms and to select variants for follow-up based on overlapping predictions, obviating the need to make *a priori* assumptions about the relative performance of each program. Such an approach assumes that true somatic mutations will show strong evidence of their existence in multiple programs, and that errors made by individual algorithms are unlikely to intersect.

We set out to address these questions by applying three publicly available somatic single nucleotide variant (SNV) detection programs, JointSNVMix2 [7], MuTect [8], and SomaticSniper [9]. All three implement sophisticated detection algorithms, and have been used in major tumor sequencing studies [2,8,19]. We used these three programs to make somatic mutation predictions for the exomes of 27 ovarian tumors and their matched germline samples. A subset of the predictions from each program was validated by Sanger sequencing. This high-confidence set of somatic point mutations and confirmed FP predictions was then used to assess the performance of each program and the relative benefit of combining predictions from multiple programs in a consensus approach, and to compare the properties of validated and miscalled somatic mutations.

Methods

Sample cohort

Ovarian tumors were collected through two primary sources: 23 cases from the Wessex region of southern England [20], and four cases through the Australian Ovarian Cancer Study (AOCS) [21]. The cohort consisted of 3 benign, 4 borderline and 10 invasive mucinous ovarian tumors, and 10 serous borderline ovarian tumors. Representative hematoxylin and eosin (H&E)-stained sections, and the tumor pathology reports were reviewed by pathologists for all cases included in the study.

Ethics approval

The accrual of patient material used in this study was approved by the following human research ethics committees (HRECs): Southampton Hospital HREC, Peter MacCallum Cancer Centre HREC, and the HRECs of all centers participating in the AOCS. Informed consent was obtained from all patients included in the study. This project was reviewed and carried out in accordance with the ethical standards of the Peter MacCallum Cancer Centre HREC (Approval numbers 09/29 and 01/38).

DNA extraction and preparation of exome-sequencing libraries

DNA extractions were performed as described previously [4,22]. Briefly, fresh-frozen ovarian tumors were needle-microdissected to enrich for epithelial cells, to a minimum of 70% epithelial cell content. DNA was extracted using the DNeasy Blood and Tissue Kit (Qiagen Inc., Valencia, CA, USA) according to the manufacturer's recommendations. Matching peripheral blood DNA was collected and extracted from patients at the time of tumor collection, and used as the source of germline DNA.

To generate sequencing libraries, 500 ng to 1 µg of DNA were randomly sheared to approximately 200 bp using a Covaris S2 Ultrasonicator (Covaris, Woburn, MA, USA). Adapter ligated libraries were constructed and exome-enriched using either NEBNext DNA library preparation reagents (New England BioLabs, Ipswich, MA, USA) and the NimbleGen SeqCap EZ Human Exome Library (version 1; Roche Nimblegen, Madison, WI, USA) (non-indexed libraries) or TruSeq DNA Sample Preparation kits (Illumina, San Diego, CA, USA) and the NimbleGen SeqCap EZ Human Exome Library (version 2 Roche Nimblegen) (indexed libraries). Non-indexed libraries were each sequenced on one lane of an Illumina GAIIx (Illumina Inc., San Diego, CA, USA) sequencer, using 75 bp paired-end reads. Indexed libraries comprising three pooled samples were sequenced on one lane of a HiSeq 2000 sequencer (Illumina Inc.), using 100 bp paired-end reads. The mean coverage achieved ranged from 102 to 225 reads per site in the tumors and 119 to 188 reads per site in the germline.

Exome sequencing read alignment and filtering

After quality based read trimming, sequence reads were aligned to the GRCh37/hg19 human reference genome using BWA [19] and duplicate reads marked using the Picard program [23]. Aligned reads for each tumor-germline pair were combined into one alignment file in binary sequence alignment (BAM) format [24], followed by local indel realignment and base quality recalibration using the Genome Analysis Tool Kit (GATK) software [14,25]. The GATK Unified Genotyper was used to identify putative SNVs within each individual tumor and germline sample.

Identification of single nucleotide somatic mutations

MuTect (version 1.0.27783) was accessed through the MuTect website at the Broad Institute [8,26]. Because this was a development version of MuTect with limited options for setting parameters, only the default parameter set was applied, without filtering germline variants. Predictions not labeled as 'REJECT' were accepted as confident somatic mutation predictions, and considered for subsequent downstream validation and analysis steps.

JointSNVMix2 (version 0.7.5) [7] was accessed through the JointSNVMix project site [27]. Model training and somatic mutation classification were performed using the 'joint_snv_mix_two' algorithm. The default prior genotype probabilities provided in the JointSNVMix package were used for the training step. Predictions with a joint probability of having a variant in the tumor sample and no variant in the germline sample (p_{AA_AB}) of 0.9999 or greater were considered for subsequent downstream validation and analysis steps.

SomaticSniper (version 1.0.0) [9] was downloaded from the github project page for SomaticSniper [28] 0). Somatic SNVs were predicted using the joint genotyping mode (-J option) with the default prior probability of a somatic mutation (0.01). Reads with a mapping quality of 0 were filtered prior to somatic mutation identification. Predictions with a 'somatic score' of 40 or greater were considered for subsequent downstream validation and analysis steps.

For each sample, the SAMtools mpileup tool [24] was used to extract data on the mapping qualities, directionality, and depth of reads that overlapped each SNV for which at least one program had made a somatic prediction. Putative somatic SNVs meeting the following criteria were considered for validation: 1) total read depth was of 8 or greater in both the tumor and matched germline; 2) mutant allele present at a frequency of 20% or greater in the tumor and 5% or greater in the matched germline; 3) mutant allele supported by read mapping in both the forward and reverse orientations (bidirectional); and 4) variant called in only one tumor (with the exception of the *BRAF* V600 and *KRAS* G12/G13 hotspot mutations, known to be frequently mutated in these tumor types), in order to eliminate alignment and germline coverage artifacts. These criteria were chosen to restrict validation to those variants within the sensitivity of Sanger sequencing, while also minimizing the inclusion of technical artifacts. For a small proportion of variants, unidirectional variants were also considered in order to provide some assessment of the value of considering only variants with bidirectional read support.

Sanger sequencing validation

Prior to variant validation, 25 to 50 ng of DNA from each tumor sample were whole-genome amplified (WGA) using either the Repli-g Mini Kit or Repli-g Midi Kit (Qiagen). Putative somatic SNVs were validated by conventional

Sanger-based sequencing analysis of PCR products obtained from the tumor WGA DNA. Purified PCR products were directly sequenced on an ABI3130 Genetic Analyzer using BigDye Terminator (version 3.1) sequencing chemistry (both Applied Biosystems, Foster City, CA, USA). The somatic status of all validated mutations was confirmed by sequencing WGA DNA from the matching germline sample.

Comparison of true positives and false positives

The R statistical software package (version 2.15) [29] was used to perform graphical and statistical comparisons of the data for predicted somatic mutations that were validated by Sanger sequencing (true positives; TPs) and predictions that did not validate or were germline variants (FPs).

To analyze the local sequence context around TP and FP predictions, a 100 bp region flanking each site was retrieved from the human genome (GRCh37 assembly) using version 64.37 of the Ensembl database [30]. Custom Perl scripts were used to assess the GC content, and the length and abundance of homopolymers flanking each site.

For each variant where validation was attempted, we counted the numbers of reads covering the variant site that were classed as uniquely mapped, multi-mapped, or mate-rescued by the XT field in the SAM file assigned by BWA. Mismatches were counted in these reads using the NM field in the SAM file.

Results

Somatic mutation prediction

We obtained somatic mutation predictions for each of the 27 ovarian tumor samples using MuTect, JointSNVMix2, and SomaticSniper. The thresholds for defining a putative mutation as somatic we attempted to set to be as equivalent as possible between programs, and thus chose a somatic score of greater than 40 for SomaticSniper, and a joint genotype probability ($p_{AA_AB} | p_{AA_BB}$) of greater than 0.9999 for JointSNVMix2. All MuTect predictions without a 'REJECT' FILTER flag were considered in accordance with the developers' guidelines. The input for all programs was a BAM file [24] with alignment data from BWA [31] and GATK IndelRealigner [14,25].

A combined total of 9,226 somatic SNV predictions were made by the three programs, with a median of 321 predictions per sample and a range of 147 to 695. There appeared to be an association between tumor grade and the number of predicted somatic mutations (Figure 1A), with invasive mucinous tumors harboring a higher number of predicted point mutations than benign or borderline tumors (Kruskal-Wallis. $P < 0.001$). No association was found between estimates of tumor ploidy and normal contamination (as measured by ASCAT (Allele-Specific Copy number Analysis of Tumors [13])) or number of

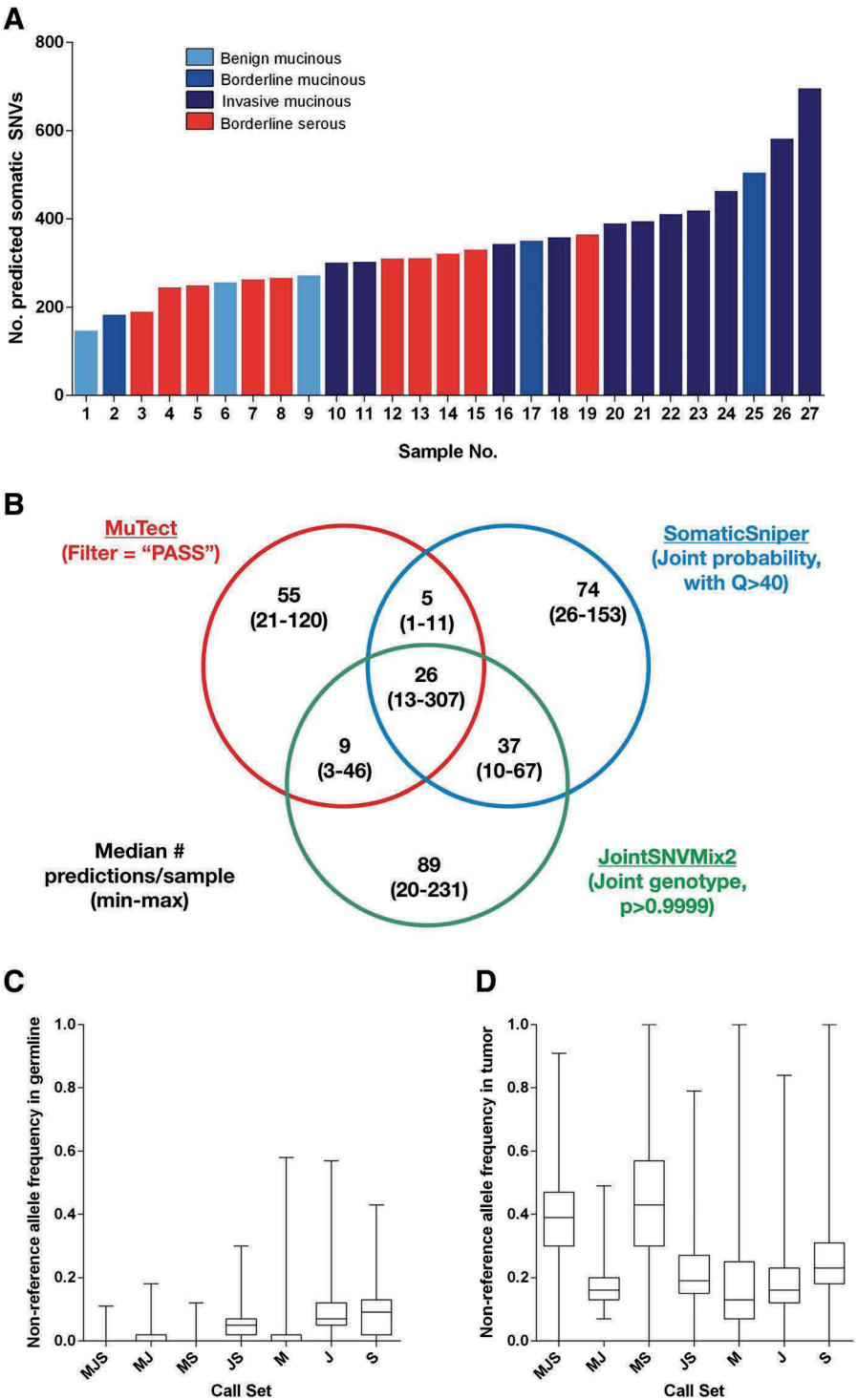


Figure 1 Frequency, concordance and read depths of somatic variant predictions in 27 ovarian tumors. **(A)** The total number of somatic SNVs predicted in each sample by the three algorithms. Bars are colored by tumor grade and histological subtype. **(B)** Concordance between somatic variant predictions from different algorithms. The minimum, maximum, and median numbers of predictions per 'call set' per sample are shown. The filtering criterion used for each program is indicated. **(C)** Non-reference allele frequency, measured as the fraction of reads in the germline samples carrying the non-reference allele, at sites predicted to be somatic variants, within each call set. Each box covers the interquartile range, with a horizontal line representing the median. Whiskers indicate the minimum and maximum values. **(D)** Fraction of reads carrying the non-reference allele in the tumor samples at sites predicted to be somatic variants, within each call set. Call sets were defined by the programs that made a prediction above threshold using the following programs: M. MuTect, J = JointSNVMix2, S. SomaticSniper.

predicted somatic SNVs (see Additional file 1: Figure S1, Table S1, and Text S1). The proportion of non-reference reads (non-reference allele frequency; NRAF), was consistently and comparably low across samples (see Additional file 1: Figure S1C), indicating that the unfiltered mutation predictions contained significant numbers of FPs, which may be due to the somatic mutation detection algorithms included in this study not accounting adequately for tumor purity and ploidy (see Additional file 1: Text S2; see Additional file 1: Figure S12). Together, these findings suggest that the differences in the amount of somatic variation between the samples analyzed here is largely driven by biological factors. Benign tumors certainly could be anticipated to harbor fewer mutations, and similar inter-tumor variability has been reported from other studies [2,5,6], consistent with the notion that tumors at a more advanced stage of malignancy have acquired more mutations.

SomaticSniper and JointSNVMix2 generally predicted the most mutations per sample, with a median of 171 and 173 per sample, respectively. MuTect was more conservative, with a median of only 115 predictions per sample. However, the majority of somatic mutations predicted by any one algorithm were not confirmed by any of the other algorithms (Figure 1B). On average, the three programs agreed only on 11.5% of the somatic predictions for a given sample, while on average more than 80% of the somatic variants predicted in a single sample were identified by only one of the three programs used (see Additional file 1: Figure S2A). This trend remained even after non-coding variants were discarded (see Additional file 1: Figure S3). It is likely that a fraction of these program-specific predictions was accounted for by FPs. However, even if only a minority of such predictions is genuinely somatic, using only one program could potentially miss many genuine mutations.

Comparison of the properties of mutation predictions from different algorithms

The poor consensus between the predictions made by each program may be explained in part by differences in their stringency settings; however, it is not clear how to directly normalize MuTect output to that of the other programs. Differences may also arise because of the distinct ways each program filters and processes sequencing reads between the tumor and germline samples, as all methods take into account base-calling and read-mapping accuracy measures when assigning probability to individual predictions.

To investigate further, we grouped the putative somatic mutations in all samples by the set of programs predicting a variant at that site (hereafter referred to as 'call sets') and compared the number of reads containing the reference and non-reference allele (that is, an SNV or an 'alternate'

allele) at each site in each sample where a prediction was made (Figures 1C,D). The property that best discriminated between prediction sets was the frequency of reads with non-reference alleles in the germline samples. Predictions that were specific to SomaticSniper or JointSNVMix2 often fell at sites where a substantial number of reads with non-reference bases had mapped (Figure 1C). These results suggest that these two programs tolerate a larger number of non-reference alleles in the germline sample, which potentially introduces a significant number of germline variants into the call set, but such variants can be eliminated by downstream filtering on frequency of reads that do not match the reference. Conversely, MuTect is much more stringent on evidence for non-reference alleles in the germline.

Additionally, predictions made by one or two programs tended to have a lower proportion of non-reference reads (NRAF) in the tumor sample than predictions made by all three programs, for which NRAF was much higher (Figure 1D), with the exception of predictions that overlapped between MuTect and SomaticSniper. The low fraction of non-reference reads may have resulted in some sites not having sufficient allelic ratios to be predicted as putative somatic mutations by one or more algorithms, but enough support to rise above the thresholds of at least one program. Many of the differences between call sets are potentially the result of such 'borderline,' low-confidence predictions. For example, the MuTect-SomaticSniper call set has a higher median NRAF than the other two-algorithm call sets because JointSNVMix2 requires higher read depths to make confident somatic mutation calls (see Additional file 1: Figure S4) and SNV calls at low-depth sites tend to have higher NRAF because a minimum number of non-reference alleles are needed to make a prediction. Total read depths in the tumor and germline samples for each call set were not consistent with differences in NRAF between groups (see Additional file 1: Figures S4A-D), suggesting that NRAF is influenced by the tendencies of each program, and is not confounded by read depth thresholds.

Validation of somatic mutation predictions

Enrichment for somatic variants

Although the variant callers used in this study were designed to identify somatic SNVs, a large number of germline variants and artifacts were anticipated to be present in each call set, based on previous analyses. We enriched for somatic SNVs using a set of minimal filters applied to all predicted SNVs. Sites covered by fewer than eight reads in the tumor and the germline samples were excluded, as were SNVs for which the non-reference allele frequency was greater than 0.05 in the germline sample or less than 0.2 in the tumor sample, and also variants that were identified in more than one sample (with the

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/367052014035006143>