Integrating WES and RNA-Seq Data For Short Variant Discovery

Rosa Barcelona-Cabeza
Sequentia Biotech SL
Walter Sanseverino
Sequentia Biotech SL
Riccardo Aiese Cigliano (raiesecigliano@sequentiabiotech.com)
Sequentia Biotech SL

Research Article

Keywords: WES, RNA-seq data, discovery, ASE

Posted Date: September 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-847277/v1

Abstract

The integration of omics has enormous potential that can be exploited for variant discovery. Several algorithms have been developed to detect somatic variants in an integrated fashion but, to our knowledge, there is still no strategy for germline variant calling. On this basis, we have developed a strategy to identify germline variants by integrating both WES and RNA-seq data. This integrated strategy identifies short variants (SNPs and indels) from raw sequence data, which are classified into six groups to improve variant interpretation: strong-evidence, DNA-only, RNA-only, allele-specific expression (ASE), RNA-editing and RNA-rescue variants. Four samples were analyzed and we show an increase in the number of identified variants without a great effect on performance compared to the exclusive use of WES data, allowing the validation of the variants identified by both types of data (strong-evidence variants), and the identification of RNA-editing and ASE. This integrated strategy provides a method to identify germline SNPs and indels from WES and RNA-seq data taking full advantage of both omics to broaden the range of identified variants and perform variant validation.

Introduction

Next-generation sequencing has greatly facilitated the discovery of genetic variants by enabling interrogation of the human genome with high throughput at relatively low cost. Whole genome sequencing (WGS) allows to identify any type of genetic variation in the entire genome, while whole exome sequencing (WES) covers only the exonic regions of the genome. However, WES is the preferred method for both research and clinical use due to its lower price and faster data analysis[1–3].

RNA-sequencing (RNA-seq) can also be applied to discover variants within the expressed regions of the genome[4,5]. RNA-seq provides some advantages over WES as it allows the identification of new variants in highly expressed genes or outside the target regions of the WES analysis. Nevertheless, RNA-seq also has drawbacks as it is not capable of detecting variants in non-transcribed or low-expressed genes and is more prone to false positive calls due to errors during RNA to cDNA conversion, mapping mismatches, alternative splicing or gene fusion[4,6].

Discovery of genomic variants using both RNA-seq and WES data can increase the target regions where variants are called and provides an orthogonal method to validate variations by complementing WES analysis with RNA-sEq. The incorporation of RNA-seq data also enables the detection of the preferential expression of a parental allele in heterozygous variants, also called allele-specific expression (ASE), and of RNA editing events which have been implicated in several disorders including cancer[7–9] and neurodegenerative diseases[10]. Moreover, the availability of RNA-seq data allows for additional analyses such as measuring transcript expression levels or detecting novel fusion genes. All of the above can make using RNA-seq and WES more cost effective than just using WGS data.

Currently, there are tools to identify short somatic variants using both WES and RNA-seq data such as RADIA[11] or VaDiR[12] but, to our knowledge, not for the detection of short germline variants. For this

reason, we have developed a strategy to identify short germline variation integrating both WES and RNA-seq data. This integrated strategy has been applied for the discovery of germline variants in four samples, showing an increase in the number of identified variants over using only WES data with virtually no performance impact.

Results

Short variant discovery from only WES calling

In the short variant discovery using only WES data, a total of 620,959 variants were identified from four samples, including 543,510 SNPs and 77,449 indels. The genotype match and the allele match of the identified variants were evaluated and a higher performance was found for SNPs than for indels in both matches (Figs. 1 and 2). The genotyping of SNPs reached a recall, precision and F-score greater than 99%, 87% and 92% respectively in all samples, while the genotyping of indels reached a recall, precision and F-score greater than 88%, 41% and 58% respectively (Fig. 1). The identification of alleles in SNPs achieved a recall greater than 97%, a precision greater than 88% and an F-score greater than 92% in all samples (Fig. 2). In the identification of indel alleles, a precision, recall and F-score greater than 73%, 43% and 60%, respectively, were found in all samples (Fig. 2).

An integrated approach for short variant discovery

The integrated approach described here provides a method for the discovery of short germline variants from the integration of WES and RNA-seq data. It is mainly based on the individual calling of the WES and RNA-seq data in GVCF mode and their subsequent joint genotyping. The comparison of WES calls and RNA-seq calls allows the validation of variants, enables the identification of RNA-editing and ASE events, and allows the classification of variants into six groups based on their quality, genotype or expression in the different types of source data: strong-evidence, DNA-only, RNA-only, ASE, RNA-editing and RNA-rescue variants.

A total of 718,239 variants were identified using the integrated approach from four samples, of which 622,229 were SNPs and 96,010 were indels (Table 1). These variants were classified into six groups consisting of 44,264 strong-evidence, 578,359 DNA-only, 68,193 RNA-only, 589 ASE, 24,082 RNA-editing and 2,752 RNA-rescue variants (Table 1). To evaluate the identified variants, we considered two types of matches, genotype match (GT) and allele match (AL), for each type of source data. Therefore, we evaluated the allele and genotype information obtained from WES data, the allele and genotype information from the RNA-seq data, and the allele information obtained by jointly considering both WES and RNA-seq alleles.

Table 1
Overview of total variants identified in all samples by variant type using the integrated approach.

Variant type	Total	SNPs	Indels
ALL	718,239	622,229	96,010
Strong-evidence	44,264	41,363	2,901
DNA-only	578,359	501,496	76,863
RNA-only	68,193	57,666	10,527
RNA-rescue	589	318	271
RNA-editing	24,082	19,041	5,041
ASE	2,752	2,345	407

For the evaluation of genotyping from WES data, RNA-only and RNA-rescue variants were excluded from evaluation because they had a low DNA genotype quality, as well as RNA-editing and ASE variants because they had a different allele expression. For the assessment of genotype from RNA-seq data, DNA-only variants were excluded as they had a low RNA genotype quality, as well as RNA-editing and ASE variants because they had a different allele expression. For the evaluation of the allele information, RNA-editing variants were excluded from the analysis in all cases (DNA, RNA and both).

In general, genotyping from the integrated approach achieved similar recall, precision and F-score regardless of source data (DNA or RNA) (Fig. 1). Regarding allele identification metrics, recall, precision and F-score were also similar when WES data, RNA-seq data, or both were used (Fig. 2). Furthermore, we observed a higher performance of the integrated approach in the identification of SNPs than of indels (Figs. 1 and 2).

Regarding the variants obtained from WES data, the identification of alleles and the genotyping in SNPs achieved a recall greater than 97%, a precision greater than 87% and an F-score greater than 92% in all samples (Figs. 1 and 2). Concerning the indels, their genotypes were identified with a recall greater than 88%, a precision greater than 39% and an F-score greater than 56% (Fig. 1), while the allele identification achieved a recall greater than 72%, a precision greater than 40% and an F-score greater than 57% in all samples (Fig. 2).

With respect to the short variants identified from the RNA-seq data, the genotyping of SNPs reached a recall greater than 99%, a precision greater than 93% and an F-score greater than 96% in the four samples, while indel genotyping achieved recall, precision, and F-score greater than 90%, 40% and 56%, respectively, in all samples (Fig. 1). The identification of the alleles of SNPs achieved a recall greater than 97%, a precision greater than 90% and an F-score greater than 93% (Fig. 2). The identification of the

alleles of indels achieved a recall greater than 78%, a precision greater than 37% and an F-score greater than 54% in all samples (Fig. 2).

The alleles of the SNPs identified jointly considering both the WES alleles and the RNA-seq alleles achieved a recall greater than 97%, a precision greater than 88%, and an F-score greater than 92% in all samples (Fig. 2). For indels, there was a recall greater than 73%, a precision greater than 40%, and an F-score greater than 57% in all samples (Fig. 2).

Strong-evidence variants

Strong-evidence variants are variants with the same genotype in the DNA and RNA-seq data, as well as a reliable quality in both. Thus, the same alleles and genotypes are obtained regardless of the type of data and therefore the same metrics.

There was a high recall, precision and F-score in genotyping and identification of alleles for strong-evidence SNPs and indels (Supplementary Fig. 1 online). Strong-evidence SNPs were identified with a recall, precision and F-score greater than 98% in all samples (Supplementary Fig. 1 online). Strong-evidence indels did not achieve as high metrics as SNPs but they did achieve the highest metrics of all types of variants (Supplementary File S1).

DNA-only variants

DNA-only variants have an RNA coverage lower than or equal to 6 and/or a strand bias in RNA greater than or equal to 2 and/or an RNA GQ lower than or equal to 20. Because of this, the alleles and genotype from the RNA data cannot be trusted, and therefore RNA-editing and ASE cannot be detected.

For these variants we observed low recall, precision and F-score in the genotyping from RNA data (Supplementary Fig. 2 online). From the DNA data, there was a reliable but lower recall, precision, and F-score than observed in the strong-evidence variants (Supplementary File S1). DNA-only SNPs were genotyped with a F-score greater than 92% in all samples (Supplementary Fig. 2 online). Furthermore, there was a decrease in recall, precision and F-score in genotyping of DNA-only variants compared to the final GT metrics from DNA data of the integrated approach, which includes strong-evidence and DNA-only variants (Supplementary File S1).

Regarding the allele identification of SNPs and indels from DNA data, RNA data and jointly considering both (DNA + RNA), a reliable recall, precision and F-score was observed (Supplementary Fig. 2 online). Alleles of DNA-only SNPs were identified with a F-score greater than 91% in all the cases (Supplementary Fig. 2 online). There was a decrease in recall, precision and F-score in identifying DNA-only SNPs alleles relative to total SNPs alleles, including all variant types except RNA-editing variants, when using DNA data, RNA data and both (Supplementary File S1). However, in general, there was a higher precision and F-score in the allele identification of DNA-only indels than of total indels in DNA, RNA and with the joint use of both (Supplementary File S1).

RNA-only variants

RNA-only variants are variants with a DNA coverage lower than or equal to 6, therefore the alleles and genotype from DNA data cannot be trusted and ASE and RNA-editing cannot be identified. The presence of ASE variants or RNA editing variants can negatively affect the metrics when evaluating genotypes in RNA data since the truth dataset was derived from DNA data.

We observed an expected low recall, precision and F-score in genotyping from DNA data (Supplementary Fig. 3 online) for RNA-only variants. Concerning SNPs and indels genotyping from RNA data, we observed a good recall, precision and F-score, but lower than those observed in strong-evidence variants (Supplementary File S1). Genotyping of RNA-only SNPs from RNA data had a F-score greater than 94% in all samples (Supplementary Fig. 3 online). Moreover, we observed lower recall, precision and F-score in genotyping of RNA-only variants relative to total variants in RNA data, including Strong-evidence, RNA-only and RNA-rescue variants (Supplementary File S1).

The recall, precision and F-score of the allele identification of RNA-only SNPs were greater than 96.8%, 90.6%, and 93.7%, respectively, in all samples when considering the alleles from RNA data and considering both DNA and RNA alleles (Supplementary Fig. 3 online). The alleles of RNA-only indels were identified with a recall greater than 79.5%, a precision greater than 39.8% and an F-score greater than 56.8% in all samples when considering the alleles from RNA data and considering both DNA and RNA alleles (Supplementary Fig. 3 online). Regarding the identification of alleles from DNA data, metrics were lower with an F-score greater than 89.3% for RNA-only SNPs and greater than 49.8% for RNA-only indels in all samples (Supplementary Fig. 3 online). There was an increase in precision and F-score in detecting alleles for RNA-only SNPs than for total SNPs, which includes all variant types except RNA-editing variants, when using RNA data and RNA with DNA data, but a decrease when using DNA data (Supplementary File S1). However, there was a decrease in precision and F-score for allele identification in RNA-only in relation to the total indels in all cases (DNA, RNA or DNA + RNA) (Supplementary File S1).

ASE variants

ASE variants are heterozygous for DNA but homozygous for RNA because only one of the two alleles has been expressed. The recall, precision and F-score in the AL match were the same for the DNA and RNA data (Supplementary Fig. 4 online). In the sample NA12878, allele identification reached an F-score of 87.4% for ASE SNPs and 21.4% for ASE indels when considering DNA alleles, RNA alleles, or jointly considering both (Supplementary Fig. 4 online). For sample HG00171, we observed an F-score of 44.9% for the allele identification of ASE SNPs and of 52.5% for the allele identification of ASE indels in all cases (DNA, RNA and DNA + RNA) (Supplementary Fig. 4 online). In HG00378, we observed an F-score of 25.7% for the alleles of the ASE SNPs and an F-score of 33.8% for the alleles of the ASE indels, while in NA20509, the F-scores were 57% and 49.4% for the alleles of ASE SNPs and ASE indels, respectively, when considering DNA alleles, RNA alleles, or jointly considering both (Supplementary Fig. 4 online).

Concerning the genotyping of ASE variants, there was an improvement in the metrics when using DNA data instead of RNA data (Supplementary Fig. 4 online), this is because the truth dataset was derived from DNA data so the expression of the variants was not considered. In NA12878, we observed an F-score of 87.4% in the genotyping of ASE SNPs from DNA data and an F-score of 0% in the genotyping of ASE SNPs from RNA data, while the genotyping of ASE indels only achieved an F-score of 16.6% when using DNA data and an F-score of 13.3% when using RNA data (Supplementary Fig. 4 online). In the sample HG00171, genotyping from DNA data reached an F-score of 40.2% for ASE SNPs and an F-score of 46.5 for ASE indels, whereas genotyping from RNA data achieved an F-score of 1.8% and 11.7% for ASE SNPs and ASE indels, respectively (Supplementary Fig. 4 online). For samples HG00378 and NA20509, we observed an F-score of 21.6% and 48.3%, respectively, in the genotyping of ASE SNPs from DNA and an F-score of 26.9% and 32.2%, respectively, in the genotyping of ASE indels from DNA. In the genotyping of samples HG00378 and NA20509 from RNA data, we observed an F-score of 2.4% and 7.3%, respectively, for the ASE SNPs and an F-score of 13.3% and 17.3%, respectively, for the ASE indels (Supplementary Fig. 4 online).

RNA-editing variants

RNA-editing variants are the variants present in RNA but absent in DNA. They cannot be evaluated as our truth dataset was derived from DNA data where RNA-editing variants were not present. For this reason, we observed very low recall, precision and F-score in genotype and allele identification from RNA data (Supplementary Fig. 5 online).

RNA-rescue variants

RNA-rescue variants meet the minimum thresholds of coverage and strand bias in DNA and RNA, have an RNA GQ greater than 50 but a DNA GQ lower than or equal to 20. Therefore, the genotype information from RNA data can be trusted but not the genotype information from DNA data. In addition, RNA-editing and ASE variants cannot be recognized and will be present within the RNA-rescue variants, which may affect their evaluation.

We observed an F-score greater than 70% in all samples in the genotype and allele identification of RNA-rescue SNPs in RNA data (Supplementary Fig. 6 online). However, the identification of indels in RNA data had very low metrics, we observed an F-score in genotyping lower than 22% and an F-score in allele identification lower than 40% in all the samples (Supplementary Fig. 6 online).

Comparison of variant discovery with the integrated approach versus WES calling

The total number of identified variants increased when WES and RNA-seq data were integrated (WES + RNA calling) instead of using only WES data (WES calling) (Table 2). Specifically, a total of 718,239 variants were identified in all samples with the integrated approach (Table 1), while 620,959 variants were identified in all samples with only WES calling (Table 2). Two types of variant matches were considered for evaluation (GT match and AL match) and each type of match includes different variants. Because of

this, the number of variants obtained from WES + RNA calling was evaluated depending on the type of match and compared with the total variants obtained from WES calling.

Table 2
Overview of the total variants identified by sample using the integrated approach and using only WES data.

Sample	Analysis	Variants		
		Total	**SNPs**	**Indels**
NA12878	WES + RNA calling	300,597	259,503	41,094
	WES calling	259,692	224,840	34,852
HG00171	WES + RNA calling	142,760	122,761	19,999
	WES calling	130,068	113,298	16,770
HG00378	WES + RNA calling	150,370	129,610	20,760
	WES calling	133,523	116,342	17,181
NA20509	WES + RNA calling	124,512	110,355	14,157
	WES calling	97676	89030	8646

The evaluation of the genotype information extracted from DNA data included two types of variants: strong-evidence and DNA-only variants. These two types of variants were enough to exceed the total number of variants obtained with WES calling in all samples except NA20509 (Table 3). The evaluation of genotype information from RNA data included strong-evidence, RNA-only and RNA-rescue variants. The number of variants detected for genotype matching from RNA data was much lower than those obtained from DNA data when using both WES and RNA-seq data (Table 3).

Table 3
Overview of the number of variants identified by sample using the integrated approach and using only WES data.

Sample	Analysis	Data type	Match	Variants		
				Total	**SNPs**	**Indels**
NA12878	WES + RNA calling	DNA	GT	260535	224987	35548
			AL	266317	229854	36463
		RNA	GT	47288	42359	4929
			AL	68881	60859	8022
		DNA + RNA	AL	287333	248646	38687
	WES calling	DNA	-	259692	224840	34852
HG00171	WES + RNA calling	DNA	GT	131453	113844	17609
			AL	134158	116030	18128
		RNA	GT	13295	11425	1870
			AL	19922	17403	2519
		DNA + RNA	AL	140257	120943	19314
	WES calling	DNA	-	130068	113298	16770
HG00378	WES + RNA calling	DNA	GT	134219	116216	18003
			AL	138657	120023	18634
		RNA	GT	19761	17454	2307
			AL	27187	24154	3033
		DNA + RNA	AL	147339	127262	20077
	WES calling	DNA	-	133523	116342	17181
NA20509	WES + RNA calling	DNA	GT	96406	87806	8600
			AL	101784	92342	9442
		RNA	GT	32702	28109	4593
			AL	42550	37076	5474
		DNA + RNA	AL	119228	106337	12891
	WES calling	DNA	-	97676	89030	8646

Regarding the assessment of allele matches, all variant types except RNA-editing variants were considered. Here, the number of variants detected with our integrated strategy (WES + RNA calling) from DNA data was higher than with WES calling and even higher when the allelic information of both DNA and RNA was considered (Table 3). Specifically, a total of 640,916 variants were identified in all samples with the integrated strategy for the allele match from DNA data (266,317 in NA12878, 134,158 in HG00171, 138,657 in HG00378 and 101,784 in NA20509) and a total of 694,157 variants when considering both DNA and RNA data (287,333 in NA12878, 140,257 in HG00171, 147,339 in HG00378 and 119,228 in NA20509), while a total of 259,692 variants were identified in the allele match with only WES calling (259,692 in NA12878, 130,068 in HG00171, 133,523 in HG00378 and 97,676 in NA20509) (Table 3).

Evaluation of recall, precision and F-score in the identification of variants was performed with respect to the genotype information (GT match) and the allele information (AL match). There was an improvement in SNP genotyping from DNA and RNA data in most samples (HG00171, HG00378 and NA20509), but a decrease in the precision and F-score when indels genotyping from DNA (NA12878, HG00171 and HG00378) and RNA (NA12878, HG00171, HG00378 and NA20509) (Fig. 1). Concerning the SNPs genotyping from DNA data, the recall was the same with the integrated approach or with WES calling for all samples: NA12878 (99.6%), HG00171 (99.3%), HG00378 (99.3% and) and NA20509 (99.1%) (Fig. 1). Precision was the same in NA12878 (96.5%) (Fig. 1 – A) but higher in WES + RNA calling than in WES calling for HG00171 ($Precision_{WES+RNA-calling}$ = 88.5% and $Precision_{WES-calling}$ = 88.4%), HG00378 ($Precision_{WES+RNA-calling}$ = 87.9% and $Precision_{WES-calling}$ = 87.7%) and NA20509 ($Precision_{WES+RNA-calling}$ = 87.6% and $Precision_{WES-calling}$ = 87.2%) (Fig. 1 – B, C, D). F-score was higher in WES + RNA calling than in WES calling for HG00171 ($F\text{-}score_{WES+RNA-calling}$ = 93.6% and $F\text{-}score_{WES-calling}$ = 93.5%), HG00378 ($F\text{-}score_{WES+RNA-calling}$ = 93.3% and $F\text{-}score_{WES-calling}$ = 93.1%) and HG00378 ($F\text{-}score_{WES+RNA-calling}$ = 93% and $F\text{-}score_{WES-calling}$ = 92.7%) (Fig. 1 – B, C, D), but lower for NA12878 ($F\text{-}score_{WES+RNA-calling}$ = 98.0% and $F\text{-}score_{WES-calling}$ = 98.1%) (Fig. 1 – A). Regarding the genotyping of indels from DNA data, recall, precision and F-score were higher with WES + RNA calling than with WES calling for NA20509 ($Recall_{WES+RNA-calling}$ = 92%, $Recall_{WES-calling}$ = 91.9%, $Precision_{WES+RNA-calling}$ = 66.4%, $Precision_{WES-calling}$ = 66.1%, $F\text{-}score_{WES+RNA-calling}$ = 77.1% and $F\text{-}score_{WES-calling}$ = 76.9%) (Fig. 1 – D), while for the rest of samples, they were lower (with the exception of the recall of NA12878 which is the same) (Fig. 1). Concerning the SNPs genotyping from RNA data, there was an improvement in recall, precision and F-score for HG00171 ($F\text{-}score_{WES+RNA-calling}$ = 96.3% and $F\text{-}score_{WES-calling}$ = 93.5%), HG00378 ($F\text{-}score_{WES+RNA-calling}$ = 96.8% and $F\text{-}score_{WES-calling}$ = 93.1%) and NA20509 ($F\text{-}score_{WES+RNA-calling}$ = 97% and $F\text{-}score_{WES-calling}$ = 92.7%) with the exception of the recall for NA20509 which was the same (99.1%) (Fig. 1 – B, C, D). However, all metrics for NA12878 were lower with WES + RNA calling than with WES calling ($Recall_{WES+RNA-calling}$ = 99.1%, $Recall_{WES-calling}$ = 99.6%, $Precision_{WES+RNA-calling}$ = 96.1%, $Precision_{WES-calling}$ = 96.5%, $F\text{-}score_{WES+RNA-calling}$ = 97.1% and $F\text{-}score_{WES-calling}$ = 98.1%) (Fig. 1 – A). Regarding the genotype of indels from RNA data, there was a decrease in precision and F-score for all samples (Fig. 1) and an increase in recall for NA12878

($Recall_{WES+RNA-calling}$ = 97.4% and $Recall_{WES-calling}$ = 95.6%), HG00171 ($Recall_{WES+RNA-calling}$ = 92.8% and $Recall_{WES-calling}$ = 89.6%) and HG00378 ($Recall_{WES+RNA-calling}$ = 92.7% and $Recall_{WES-calling}$ = 88.9%) (Fig. 1 – A, B, C), but there was also a decrease in recall for NA20509 ($Recall_{WES+RNA-calling}$ = 90.9% and $Recall_{WES-calling}$ = 91.9%)(Fig. 1 – D).

Overall, the precision and F-score in the allele identification of SNPs and indels decreased when the integrated approach was applied instead of performing WES calling (Fig. 2). There was a decrease in the precision and F-score of allele identification of SNPs and indels from DNA data in all samples: NA12878 ($F\text{-}score_{WES+RNA-calling}$ = 97.6% and $F\text{-}score_{WES-calling}$ = 98.4%), HG00171 ($F\text{-}score_{WES+RNA-calling}$ = 92.6% and $F\text{-}score_{WES-calling}$ = 93.2%), HG00378 ($F\text{-}score_{WES+RNA-calling}$ = 92.3% and $F\text{-}score_{WES-calling}$ = 92.8%) and NA20509 ($F\text{-}score_{WES+RNA-calling}$ = 92.3% and $F\text{-}score_{WES-calling}$ = 92.8%) (Fig. 2). Concerning the alleles obtained from RNA data, there was a decrease in the precision and F-score to identify SNPs and indels in all samples, but they increased for the identification of SNPs in HG00378 ($F\text{-}score_{WES+RNA-calling}$ = 93.8% and $F\text{-}score_{WES-calling}$ = 92.8%) and NA20509 ($F\text{-}score_{WES+RNA-calling}$ = 95.8% and $F\text{-}score_{WES-calling}$ = 92.8%) (Fig. 2). Regarding the alleles obtained jointly considering DNA and RNA data, the recall, precision and F-score in the identification of SNPs in NA20509 increased ($F\text{-}score_{WES+RNA-calling}$ = 92.9% and $F\text{-}score_{WES-calling}$ = 92.8%) (Fig. 2 – D), however, the precision and F-score decreased in the detection of SNPs in the rest of samples: NA12878 ($F\text{-}score_{WES+RNA-calling}$ = 97.6% and $F\text{-}score_{WES-calling}$ = 98.4%), HG00171 ($F\text{-}score_{WES+RNA-calling}$ = 92.8% and $F\text{-}score_{WES-calling}$ = 93.2%) and HG00378 ($F\text{-}score_{WES+RNA-calling}$ = 92.5% and $F\text{-}score_{WES-calling}$ = 92.8%) (Fig. 2 – A, B, C). In the identification of alleles in indels there was a decrease in the precision and F-score in all samples when considering jointly DNA and RNA data (Fig. 2).

Discussion

The integration of genomic data with transcriptomic data offers new opportunities for variant discovery. These new opportunities range from improving the accuracy of variant identification in highly expressed genes to finding new variants only detectable from RNA-seq data, such as RNA-editing variants or ASE variants, as well as offering an additional method of variant validation. There are already tools that integrate these two types of data, such as RADIA[11] or VaDiR[12]. However, these tools are based on the detection of somatic short variants and do not detect germline short variants. The main objective of RADIA is to use the RNA-seq data to validate the somatic variants found in DNA[11]. All somatic variants called by RADIA are supported by DNA, and RNA-only variants are not called. In addition, it does not include the detection of RNA-editing variants or ASE variants, so part of the benefits of using RNA-seq are not being exploited. VaDiR calls somatic variants only using RNA-seq data, DNA data is only used to filter germline variants[12]. Because of this, VaDiR misses mainly low-frequency RNA-seq variants and some of the potential of using both types of data is not exploited.

Taking into account all the above, we have developed a strategy to identify germline short variants by integrating the variants identified by both DNA and RNA-seq, thus making the most of these two omics. To our knowledge, there is no other published strategy to date that allows the identification and classification of germline variants by integrating both DNA and RNA-seq data. Specifically, we have focused on integrating RNA-seq with WES data to offer a cost-efficient alternative to using WGS data, as WES data is currently preferred in both clinical and research use[1–3] due to its lower price. By integrating RNA-seq data into WES data, there was an improvement in genotyping, but a decrease in allele match metrics. In any case, these metric changes were minor, so it can be concluded that the integrated approach increased the number of detected variants without a great effect on performance. The integrated approach also allows the validation of variants identified by both types of data and improves prioritization of variants by classifying them into six groups: strong-evidence, DNA-only, RNA-only, ASE, RNA-editing and RNA-rescue variants. However, it is important to note that the RNA-editing variants and ASE variants identified here could not be validated due to the lack of validated variants from RNA-seq data. The presence of ASE variants can negatively affect the metrics when evaluating genotypes in RNA data since the truth dataset was derived from DNA data where the expression is unknown. The presence of RNA editing variants can also negatively affect the metrics when evaluating both genotype and alleles in RNA data since RNA editing occurs after DNA transcription and synthesis. Finally, an additional benefit of this integrated approach is that RNA-seq data can be used for further analyses on gene expression levels or gene fusions.

Methods

Datasets

Four samples were selected to evaluate the performance of this strategy: NA12878 [or SAME123392], HG00171 [or SAME124961], HG00378 [or SAME124745] and NA20509 [or SAME124354]. All of them have publicly available WES data, RNA-seq data and high-quality variant information.

NA12878 sample is derived from the GM12878 cell line from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research. WES and RNA-seq data of NA12878 were obtained from the Sequence Read Archive (SRA)[13] under accession numbers SRR2106342 and SRX082565, respectively. Truth data of NA12878 was generated by the Genome in a Bottle (GiaB) consortium, led by the National Institute of Standards (NIST)[14].

HG00171, HG00378 and NA20509 samples are available in the International Genome Sample Resource (IGSR)[15]. The three of them are part of the 1000 Genomes Project[16] where their WES, RNA-seq and truth data were obtained. The RNA-seq data was generated by the Geuvadis consortium[17] and the truth data correspond to the integrated phase III variants of the 1000 Genomes Project.

Short variant discovery from only WES data

Reads were aligned to the hg19 human reference genome using the BWA-MEM algorithm[18] from the Sentieon utilities v202010.02[19]. After sorting the alignment BAM files, duplicate reads were removed and BQSR was performed using the Sentieon utilities v202010.02[19]. Finally, variant calling was performed using Sentieon Haplotyper v202010.02[19] in VCF mode.

The identified variants were compared to the truth variant information using the hap.py framework[20] with the vcfeval comparison tool[21].

An integrated approach for short variant discovery

The strategy for variant discovery from WES and RNA-seq data consist of three main steps: WES calling, RNA-seq calling and joint variant calling. Joint variant calling includes genotyping, filtering and classification of short germline variations (Fig. 3).

WES calling

WES reads were aligned to the hg19 human reference genome using the BWA-MEM algorithm[18] implemented in the Sentieon utilities v202010.02[19]. The resulting alignment BAM files were then sorted and indexed using the Sentieon sort utility v202010.02[19]. Duplicate reads were removed and base quality score recalibration (BQSR) was performed using the Sentieon utilities v202010.02[19]. Finally, variant calling was performed using Sentieon Haplotyper v202010.02[19] in GVCF mode.

RNA-seq calling

RNA-seq reads were trimmed using BBDuk v35.85[22] setting a minimum length of 35 bp and minimum base quality score of 25. Then, the trimmed reads were aligned to the hg19 human reference genome using STAR v2.7.3a[23] in two-pass mode to improve alignments around novel splice junctions[24]. Read groups were added to the sorted alignment BAM file using Picard's AddOrReplaceReadGroups[25]. Next, duplicate reads were removed, reads were split at splicing junctions into exon segments and base quality score recalibration (BQSR) was performed using the Sentieon utilities v202010.02[19]. Finally, variant calling was performed using Sentieon Haplotyper v202010.02[19] in GVCF mode and with the “trim_soft_clip” option activated to exclude the soft clipped bases from the analysis.

Genotyping

Both WES and RNA-seq GVCF were collected and passed together to the joint genotyping tool, GVCFtyper from Sentieon v202010.02[19] where the minimum phred-scaled confidence thresholds for calling and for emitting variants were adjusted to 20.

Filtering of genomic variants

Only variants with a depth of coverage (DP) greater than 6 reads and a strand bias (SB) lower than 2 in both WES and RNA-seq data were retained. The SB filter was only applied if the read counts for both the major allele and the minor allele were greater than or equal to 10. DP information was automatically calculated by Sentieon during variant calling whereas SB was calculated following the formula described previously in a mitochondrial heteroplasmy study[26]:

$$\left| \frac{b}{a+b} - \frac{d}{c+d} \right| / \left(\frac{b+d}{a+b+c+d} \right)$$

where a, c represent the forward and reverse reads counts of the major allele, and b, d represent the forward and reverse reads counts for the minor allele.

RNA-based variants require more filters than WES-based variants due to its greater susceptibility to false positives. For this reason, RNA-based variants were removed if they were found in homologous regions, interspersed repeats or low-complexity sequences, in the inaccessible genome, or within 5bp upstream of an exon start or downstream of an exon end site. To perform such RNA-based filtering, genomic variations must be annotated first. This annotation of variants was performed using ANNOVAR software[27] and different databases: (i) RepeatMasker track from the UCSC Genome Browser[28] to identify repetitive regions, (ii) the genomicsuperDups database[29] to obtain homologous regions and (iii) the ENCODE Blacklist[30] to annotate variants that are not in the accessible genome. Finally, the distance of each variant to its closest exon boundary was calculated using BEDTools utilities v2.29.2[31], pybedtools Python library v0.8.1[32] and the Reference sequence (RefSeq) Gene database[33].

Classification of genomic variants

Variants were classified in six groups to facilitate the prioritization of variants: Strong-evidence, DNA-only, RNA-only, ASE, RNA-editing and RNA-rescue variants (Fig. 4). RNA-only variants do not meet the minimal threshold for DNA but meet the RNA thresholds (DP, SB and RNA-based filters) and have an RNA genotype quality (GQ) greater than 20. DNA-only variants meet the DNA thresholds (DP and SB) and have a DNA GQ greater than 20 but do not meet the RNA thresholds (DP and SB) and/or have an RNA GQ lower than or equal to 20. Strong-evidence variants meet the DNA and RNA thresholds, have a DNA GQ and an RNA GQ greater than 20 and the same genotype in DNA and RNA data. ASE variants meet the DNA and RNA thresholds, have a DNA GQ and an RNA GQ greater than 20 and are heterozygous for DNA but homozygous for RNA. RNA-editing variants meet the DNA and RNA thresholds, have a DNA GQ and an RNA GQ greater than 20 and are present in RNA but absent in DNA. RNA-rescue variants meet the DNA and RNA thresholds, have an RNA GQ greater than 50 but a DNA GQ lower than or equal to 20.

Benchmark

The evaluation of performance was conducted according to the Global Alliance for Genomics and Health (GA4GH) best practices[34] using hap.py framework[20] with the vcfeval comparison tool[21] and the truth

variant information publicly available for each sample.

Two types of variant matches were considered for evaluation: (i) genotype (GT) match when the unphased genotype and alleles of a variant match in the truth and query set, and (ii) allele (AL) match when the truth and query set contain the same allele regardless of genotype. This evaluation was applied to the allele and genotype information obtained from WES data, to the allele and genotype information from the RNA-seq data, and to the allele information obtained by jointly considering both WES and RNA-seq alleles.

It should be taken into account that not all variants were considered for all evaluations. The variants to consider changed depending on the type of match (AL or GT match) and the source of information (DNA, RNA, or both). To evaluate the genotype information obtained from WES data, the RNA-only, RNA-rescue, RNA-editing and ASE variants were removed as they did not meet the minimum DNA GQ criteria (RNA-only and RNA-rescue) or they had a different allele expression (RNA-editing and ASE). To assess the genotype information obtained from RNA-seq data, the DNA-only variants were removed as they had an RNA GQ lower than or equal to 20 and the RNA-editing and ASE variants were also removed as they had a different allele expression. To evaluate the allele information, RNA-editing variants were removed from the analysis in all cases (DNA, RNA and both).

Additionally, an evaluation was carried out by type of variant according to our new classification system: strong-evidence, DNA-only, RNA-only, ASE, RNA-editing and RNA-rescue variants.

Data availability

NA12878 sample is available from the Sequence Read Archive (SRA) [https://www.ncbi.nlm.nih.gov/sra] under accession numbers SRR2106342 and SRX082565 for WES and RNA-seq data, respectively. All raw sequence data for HG00171, HG00378 and NA20509 samples are available from the International Genome Sample Resource (IGSR) [https://www.internationalgenome.org/data-portal/sample].

Declarations

Author contributions

R.B.-C. designed and developed the approach, conducted the analyzes, drafted the manuscript, and interpreted the data. R.A.C. conceived the work, and interpreted the data. W.S. contributed to the conception of the work. All authors reviewed the manuscript.

Funding

This work was supported by the Spanish Government with an industrial doctorate fellowship from MINECO (DI-17-09652) awarded to R.B.-C.

Competing interests

Authors are part of the Sequentia Biotech company. R.A.C. and W.S. are the co-founders of the company.

Additional information

Correspondence and requests for materials should be addressed to R.A.C.

References

1. Dewey, F. E. *et al.* Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study.Science (80-.).354,(2016).
2. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease.Nature526,82–89(2015).
3. Yang, Y. *et al.* Molecular findings among patients referred for clinical whole-exome sequencing.JAMA - J. Am. Med. Assoc.312,1870–1879(2014).
4. Piskol, R., Ramaswami, G. & Li, J. B. Reliable identification of genomic variants from RNA-seq data.Am. J. Hum. Genet.93,641–651(2013).
5. Lam, S. *et al.* Development and comparison of RNA-sequencing pipelines for more accurate SNP identification: Practical example of functional SNP detection associated with feed efficiency in Nellore beef cattle.BMC Genomics21,1–17(2020).
6. Lee, J. H., Ang, J. K. & Xiao, X. AnalysisanddesignofRNAsequencingexperimentsforidentifyingRNAeditingandothersingle-nucleotidevariants.*RNA*vol.19725–732(2013).
7. Paz-Yaacov, N. *et al.* Elevated RNA Editing Activity Is a Major Contributor to Transcriptomic Diversity in Tumors.Cell Rep.13,267–276(2015).
8. Han, L. *et al.* The Genomic Landscape and Clinical Relevance of A-to-I RNA Editing in Human Cancers.Cancer Cell28,515–528(2015).
9. Kung, C. P., Maggi, L. B. & Weber, J. D. The Role of RNA Editing in Cancer Development and Metabolic Disorders.Frontiers in Endocrinologyvol.9762(2018).
10. Krestel, H. & Meier, J. C. RNA editing and retrotransposons in neurology.Frontiers in Molecular Neurosciencevol.11163(2018).
11. Radenbaugh, A. J. *et al.* RADIA:RNA and DNA integrated analysis for somatic mutation detection.PLoS One9,(2014).
12. Neums, L. *et al.* VaDiR: An integrated approach to Variant Detection in RNA.Gigascience7,1–13(2018).
13. Leinonen, R., Sugawara, H. & Shumway, M. The sequence read archive.Nucleic Acids Res.39,D19(2011).
14. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls.Nat. Biotechnol.32,246–251(2014).

15. L, C. *et al.* The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data.Nucleic Acids Res.45,D854–D859(2017).
16. Auton, A. *et al.* A global reference for human genetic variation.Nature526,68–74(2015).
17. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans.Nature501,506–511(2013).
18. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform.Bioinformatics25,1754–1760(2009).
19. Freed, D., Aldana, R., Weber, J. A. & Edwards, J. S. The Sentieon Genomics Tools - A fast and accurate solution to variant calling from next-generation sequence data.bioRxiv115717(2017)doi:10.1101/115717.
20. Krusche, P. Haplotypecomparisontools/hap.py.
21. Cleary, J. *et al.* ComparingVariantCallFilesforPerformanceBenchmarkingofNext-GenerationSequencingVariantCallingPipelines.*bioRxiv*023754(2015)doi:10.1101/023754.
22. Bushnell, B. (2014).
23. Dobin, A. *et al.* STAR:Ultrafast universal RNA-seq aligner.Bioinformatics29,15–21(2013).
24. Veeneman, B. A., Shukla, S., Dhanasekaran, S. M., Chinnaiyan, A. M. & Nesvizhskii, A. I. Two-pass alignment improves novel splice junction quantification.Bioinformatics32,43–49(2016).
25. Picard toolkit.(2019).
26. Guo, Y. *et al.* The use of next generation sequencing technology to study the effect of radiation therapy on mitochondrial DNA mutation.Mutat. Res. - Genet. Toxicol. Environ. Mutagen.744,154–160(2012).
27. Wang, K., Li, M. & Hakonarson, H. A. N. N. O. V. A. R. Functional annotation of genetic variants from high-throughput sequencing data.Nucleic Acids Res.38,(2010).
28. Kent, W. J. *et al.* The Human Genome Browser at UCSC.Genome Res.12,996–1006(2002).
29. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: Organization and impact within the current human genome project assembly.Genome Res.11,1005–1017(2001).
30. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome.Sci. Rep.9,1–5(2019).
31. Quinlan, A. R. & Hall, I. M. A flexible suite of utilities for comparing genomic features.Bioinformatics26,841–842(2010).
32. Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: A flexible Python library for manipulating genomic datasets and annotations.Bioinformatics27,3423–3424(2011).
33. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation.Nucleic Acids Res.44,D733–D745(2016).
34. Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human genomes.Nat. Biotechnol.37,555–560(2019).

Figures

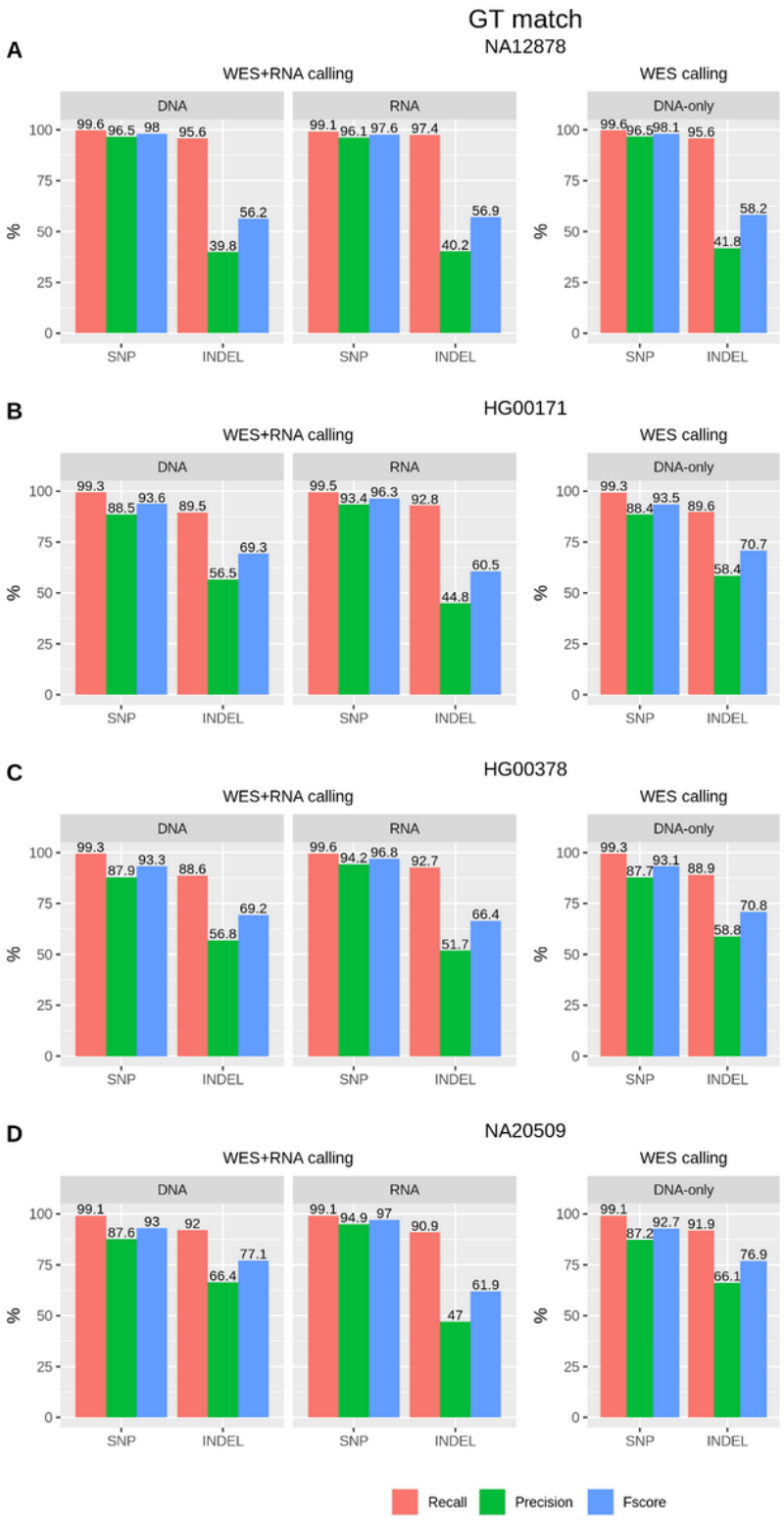

Figure 1

Benchmark results of the genotype match with the integrated approach and with WES calling for different samples. Recall, precision and F-score in the identification of the genotype of SNPs and indels with the

integrated approach and with WES calling for NA12878 (A), HG00171 (B), HG00378 (C) and NA20509 (D).

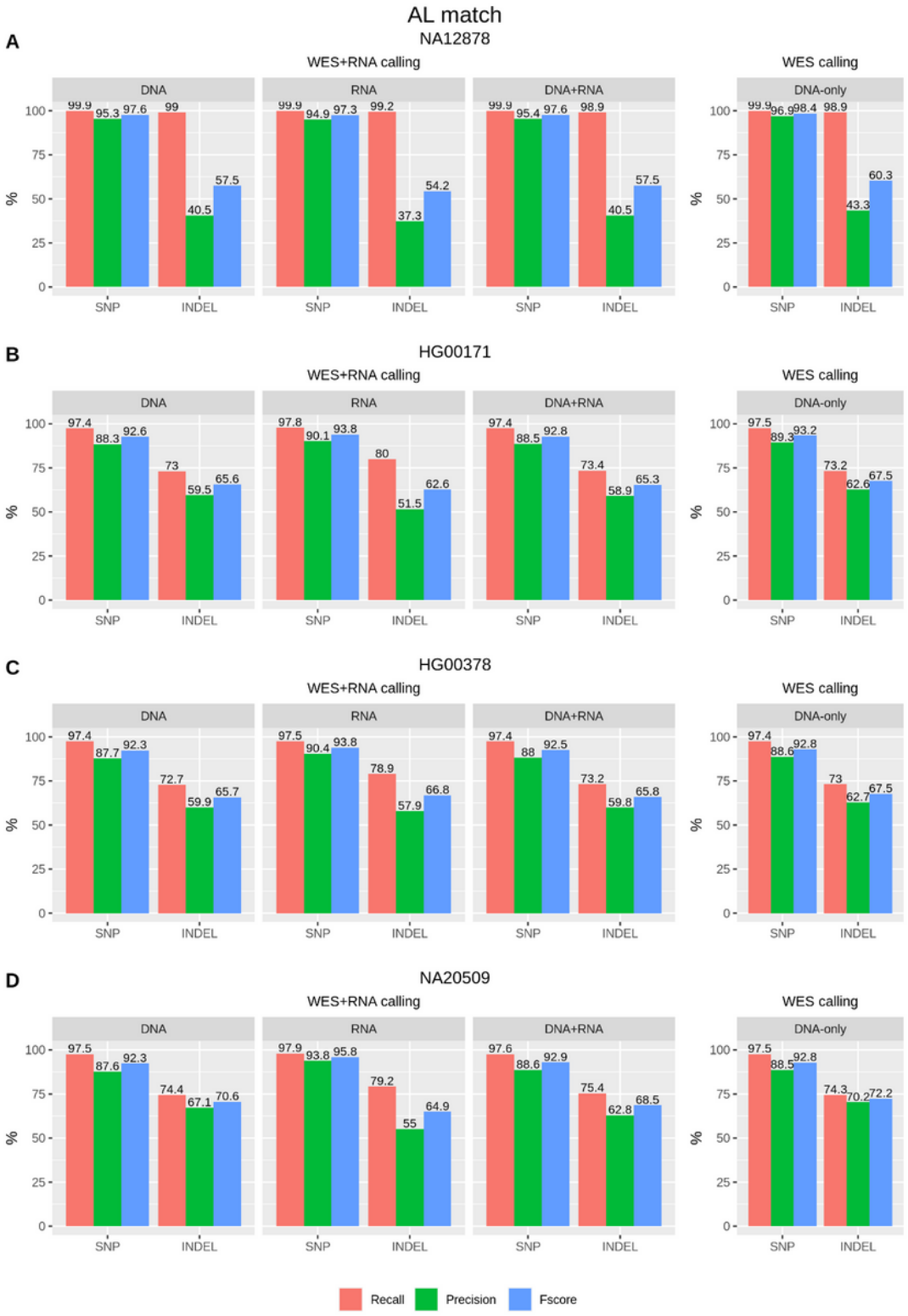

Figure 2

Benchmark results of the allele match with the integrated approach and with WES calling for different samples. Recall, precision and F-score in the identification of the alleles of SNPs and indels with the

integrated approach and with WES calling for NA12878 (A), HG00171 (B), HG00378 (C) and NA20509 (D).

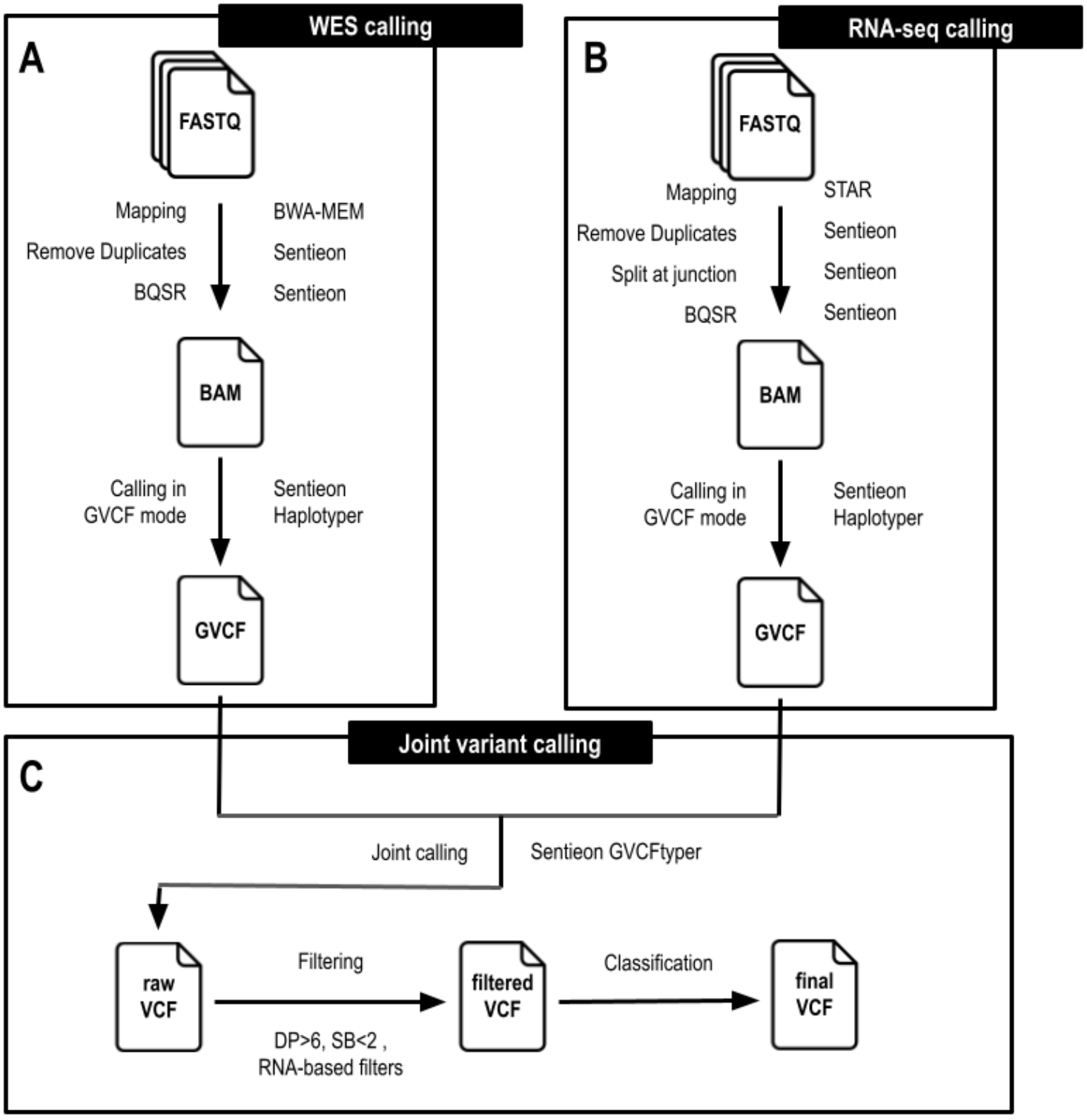

Figure 3

Workflow for variant identification using both WES and RNA-seq data. The workflow mainly consists of three steps: A) WES mapping and calling in GVCF mode. B) RNA-seq mapping and calling in GVCF mode. C) Joint calling genotyping of WES and RNA-seq data, filtering and classification of variants.

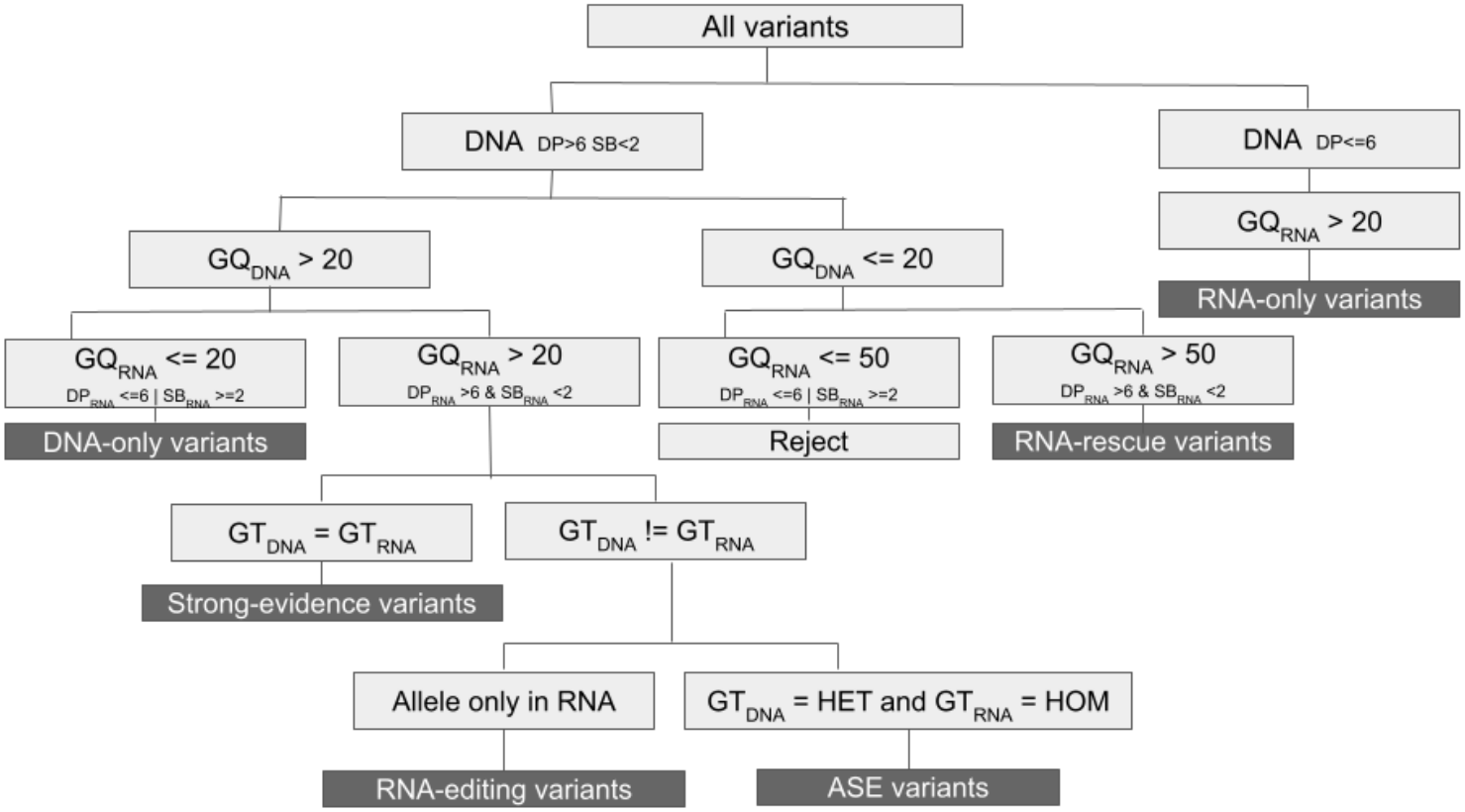

Figure 4

Workflow for variant classification.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFiguresandTables.pdf