Bioinformatics, 33(22), 2017, 3511–3517
doi: 10.1093/bioinformatics/btx482
Advance Access Publication Date: 29 July 2017
Original Paper

OXFORD

Genome analysis

Reference genome assessment from a population scale perspective: an accurate profile of variability and noise

José Carbonell-Caballero[1,*], Alicia Amadoz[1], Roberto Alonso[1], Marta R. Hidalgo[1], Cankut Çubuk[1], David Conesa[2], Antonio López-Quílez[2] and Joaquín Dopazo[1,3,4,5,*]

[1]Computational Genomics, Principe Felipe Research Centre, Valencia 46012, [2]Estadística e investigación Operativa, Universitat de València, Burjassot 46100, [3]Clinical Bioinformatics Area, Fundación Progreso y Salud, Hospital Virgen del Rocio, Sevilla 46100, [4]Functional Genomics Node (INB), Fundación Progreso y Salud, Hospital Virgen del Rocio, Sevilla 46100 and [5]Bioinformatics in Rare Diseases (BiER), Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Fundación Progreso y Salud, Hospital Virgen del Rocio, Sevilla 46100, Spain

*To whom correspondence should be addressed.
Associate Editor: Bonnie Berger

Received on March 3, 2017; revised on July 10, 2017; editorial decision on July 25, 2017; accepted on July 28, 2017

Abstract

Motivation: Current plant and animal genomic studies are often based on newly assembled genomes that have not been properly consolidated. In this scenario, misassembled regions can easily lead to false-positive findings. Despite quality control scores are included within genotyping protocols, they are usually employed to evaluate individual sample quality rather than reference sequence reliability. We propose a statistical model that combines quality control scores across samples in order to detect incongruent patterns at every genomic region. Our model is inherently robust since common artifact signals are expected to be shared between independent samples over misassembled regions of the genome.

Results: The reliability of our protocol has been extensively tested through different experiments and organisms with accurate results, improving state-of-the-art methods. Our analysis demonstrates synergistic relations between quality control scores and allelic variability estimators, that improve the detection of misassembled regions, and is able to find strong artifact signals even within the human reference assembly. Furthermore, we demonstrated how our model can be trained to properly rank the confidence of a set of candidate variants obtained from new independent samples.

Availability and implementation: This tool is freely available at http://gitlab.com/carbonell/ces.

Contact: jcarbonell.cipf@gmail.com or joaquin.dopazo@juntadeandalucia.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Next-generation sequencing (NGS) studies have experienced a considerable decrease in cost/throughput relation, which has certainly changed the design and scope of current plant and animal genomic studies. The number of available samples has been proportionally increased, opening the door to deal with large population scale studies, beyond the remarkable international consortiums such as the 1000 genomes project (Abecasis *et al.*, 2012), EXAC (Lek *et al.*, 2016) or The Cancer Genome Atlas (Weinstein *et al.*, 2013).

During the last decade, population scale studies have provided successful results about underlying variability of hundreds of species, even about specific subgroups of individuals like human subpopulations (Moorjani *et al.*, 2013; Xing *et al.*, 2010) or particular geographic regions under study (Abecasis *et al.*, 2012; Boomsma *et al.*, 2014; Dopazo *et al.*, 2016; Gudbjartsson *et al.*, 2015; Nagasaki *et al.*, 2015; Tishkoff *et al.*, 2009). However, few of them have taken advantage of the numerous quality related scores computed during variant analysis to evaluate the reliability of the reference genome sequence in itself. In some cases, assembly errors are responsible for a great number of unexpected results, especially for non-model organisms, where reference assembly has not been properly consolidated. Evolutionary studies also constitute a sensible context, since sample reads are often mapped against a reference genome that belongs to a related but distinct species.

There are few available standard protocols to evaluate the confidence of a given reference genome assembly (RGA). Commonly, a set of simple descriptive measurements is used to evaluate the fragmentation degree and the percentage of genome recovered by a given assembly, where misassembled regions cannot be easily identified. Good examples of this philosophy can be seen at the Assemblathon contest (Bradnam *et al.*, 2013) or the GAGE (Salzberg *et al.*, 2012) initiative, where state-of-the-art scores were used to compare the reliability of different assemblers over a set of real and simulated datasets.

Some recent tools (Clark *et al.*, 2013; Hunt *et al.*, 2013; Rahman and Pachter, 2013; Vezzi *et al.*, 2012) extended the classic approach to a more detailed region-based evaluation. In particular REAPR (Hunt *et al.*, 2013) uses a pair-end mapped sample of similar characteristics to the evaluated reference genome in order to detect incongruous genomic patterns that are directly related to assembly artifacts. More recently, misFinder (Zhu *et al.*, 2015) combined a similar approach with the help of a near species reference genome, also provided by QUAST (Gurevich *et al.*, 2013). Additionally, some other tools have appeared to cover specific contexts like bacterial (Walker *et al.*, 2014) or metagenomic (Mikheenko *et al.*, 2016) assembly evaluation.

Existing tools do not provide assembly evaluation metrics when a population (or a group of samples) of interest is sequenced for genotyping purposes. This scenario matches with a great percentage of current genomic studies where a variant-discovery oriented protocol is implemented to detect those genomic features potentially related to phenotypic traits of interest. When a non-model organism is studied, the absence of a valid reference genome is replaced by a *de novo* assembled sequence often limited in quality, whose misassembled regions inevitably lead to false-positive associations. To restrain this bias, a set of quality control metrics is usually obtained to evaluate the confidence of every predicted sample variant. However, these metrics are never used to evaluate the reference genome. In this case, if we summarize and project the quality control scores against the reference sequence we can construct a statistical model that characterizes in detail every region or nucleotide of the genome. We propose this kind of model, which is naturally able to capture unstable regions since similar quality patterns are expected to be found across different samples. That provides statistically robust evidence supported by several independent observations.

In this work, a novel RGA evaluation protocol is presented. Our methodology is based on an empirical model constructed from a set of selected quality control measurements obtained after mapping the reads of a population of interest, allowing local evaluation of the RGA without needing the support of a near species reference genome. Finally, the quality control scores are extended with a set of population genetics metrics to evaluate the reference genome in terms of allelic variability, providing a valuable portrait about the underlying evolutionary processes that the studied samples could have recently experimented as a species or clade.

2 Methods

The evaluation of a RGA is performed through the construction of a local genomic profile (LGP). The LGP is based on a sliding-window protocol that dissects the RGA into windows of a specific size. Inside each window, allelic variability and noise susceptibility are measured and summarized by using different statistical scores. The LGP is composed of a set of the empirical distributions (one per score) obtained by combining all computed window values along the genome.

Region-based characterization

At every single window W, a set of quality control scores (Supplementary Table S1) are computed by using the sample reads that specifically cover the window location. Then, the obtained scores are summarized to provide a representative value per window and score (Fig. 1).

The window value is computed depending on the score nature. In particular, we define two types of score: (i) those naturally defined at every genomic location (like base or mapping quality) and (ii) those exclusive of some type of locus (like variant-derived quality control scores). In the first case, the window value (x) is computed as:

$$x = \sum_{i \in W} r_i / l, \tag{1}$$

where

$$r_i = \sum_{j \in S} y_{ij} / s, \tag{2}$$

and l corresponds to the window length, and r_i to the summary computed at the relative window position i, being y_{ij} the score value

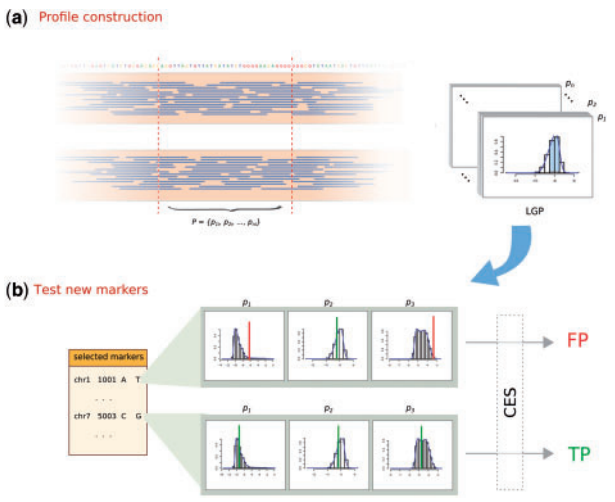

Fig. 1. General scheme of the methodology. (**a**) The LGP is constructed from sample reads that cover regions across the genome. (**b**) Then, specific markers of interest can be evaluated by contrasting their corresponding window value against the stored empirical distributions. Finally, the CES is computed to obtain the definitive diagnosis

obtained for the sample j at this position, S the total set of samples and s the total number of samples of S.

When the score is only computable at certain nucleotides or samples (like variant-derived scores) the summary is constrained to those specific evaluated elements:

$$x = \sum_{i \in V} r_i / v, \tag{3}$$

where

$$r_i = \sum_{j \in S_v} y_{ij} / s_v, \tag{4}$$

and V corresponds to the set of evaluated positions and S_v to the set of evaluated samples, being v and s_v their respective sizes.

In the general case, all base pairs within a window contribute equally to the summarized value. However, if we are interested in evaluating a set of specific genomic locations (like mutated loci), the window positions can be centered around the corresponding coordinates of interest, assigning more relevance to their nearer nucleotides. In particular, the summary is computed as a weighted sum where each r_i contributes to the window value depending on the distance to its center (where the base of interest is located).

$$x = \sum_{i \in W} r_i w_i, \tag{5}$$

being

$$\sum_{i \in W} w_i = 1, \tag{6}$$

where w_i corresponds to the weight assigned to the relative position i. In particular, w_i is computed as:

$$w_i = \frac{W_c - |i - W_c|}{\sum_{k \in W} W_c - |k - W_c|}, \tag{7}$$

where W_c corresponds to the window center. Also in this case, the weighted scores can be adapted to variant-derived scores substituting W for V, that is:

$$x = \sum_{i \in V} r_i w_i, \tag{8}$$

where

$$w_i = \frac{W_c - |i - W_c|}{\sum_{k \in V} W_c - |k - W_c|}. \tag{9}$$

Combined error score

Quality control scores describe noise artifacts from different points of view. If base quality or strand bias can predict false positives caused by errors during sequencing, mapping quality or indels frequency can detect abnormal mappings when two different homologous regions are inconsistently merged. Although poorly assembled regions often provide simultaneously extreme values for several noise estimators, it is not strictly necessary to find multiple artifact signals when a specific kind of noise is present at a given region. Due to this, we combine the set of empirical P-values obtained from all scores into a single and more accurate artifact estimator, that we call combined error score (CES).

In order to define a robust estimator, the CES is computed by using the Fisher's method for combining P-values (Fisher, 1925), where a combined score

$$x = -2 \sum_{q=1}^{m} \log (p_q), \tag{10}$$

is assumed to be distributed according to a χ^2 distribution with $2m$ degrees of freedom, being m the number of quality control scores and p_q the corresponding P-value for a quality control score q.

In this way, the CES is computed as

$$\mathrm{CES} = f(x), \tag{11}$$

where f corresponds to the χ^2 cumulative density function.

General overview (guide to users)

The input of our protocol is mainly composed of two pieces: (i) the reference genome that we want to evaluate and (ii) the set of samples used to perform the evaluation. Also, the user must define the prefered window length (l). The window length represents an heuristic parameter that must be coherently defined. Despite its heuristic character, it has natural limits: too small window values will not take advantage on neighborhood bases while large window values will dilute too much the error estimation. Without needing optimization, a good approach can be to define l to a value close to used read length in sample sequencing, or otherwise, pair-end size (see Supplementary Fig. S3), since they define the core of the sample profiling.

After tool execution, the CES is obtained, providing a quantitative estimator that reflects the reliability of every region of the evaluated genome. This value can be used under different strategies. In the general case, the CES can be applied to directly filter those regions with statistically significant values, where the presence of assembly artifacts are robustly proved. But also, it can be used as a ranking criteria to establish which obtained polymorphisms or genomic features should be firstly validated or selected for subsequent analysis. These two strategies can be also combined, reducing hence the expected number of false-positive findings, and reinforcing the final study conclusions about samples of interest.

Validation and use cases

The proposed methodology can be applied to a broad range of cases. To illustrate this, several experiments have been designed. In particular, a set of selected organisms, representing different useful scenarios, were chosen to perform the evaluation. In all cases, selected organisms have an available stable reference sequence, that is used to detect the location of misassembled regions within the corresponding assembly under evaluation. The comparison between the stable reference sequence and the assembled genome is based on a BLAST protocol that estimates the degree of similarity between the sequence of a specific window and its corresponding region into the original (reference) genome. We refer to this metric as similarity score.

Similarity score computing

The reliability of an evaluated *de novo* assembly is assessed by comparing its sequence against the corresponding reference genome, which is considered the ground truth. The evaluation is performed in a region-based manner. Concretely, the *de novo* assembly is divided in regions of a specific size and its sequences mapped against the reference sequences by using a BLAST protocol. The quality of the hit (BLAST *bit score*) obtained by a given region is used to define

its integrity. Due to a given region sequence can hit multiple times the reference sequence (as repetitive elements), the similarity score is computed by comparing the two best hits as:

$$s = \sqrt{b_1^2 - b_2^2}, \quad (12)$$

where s corresponds to the similarity score, and b_1 and b_2 the BLAST *bit scores* corresponding to the first and the second best hit, respectively. This approach allows us to estimate whether a region of interest should be unequivocally assembled or not, reflecting in that case the repetitive nature of this loci.

Arabidopsis thaliana

Arabidopsis thaliana (*Ath*) represents one of the most widely studied organisms in plant biology. The latest version of its RGA (TAIR10., Berardini *et al.*, 2015) contains almost 136 Mb and can be considered a quite stable assembled genome. In general, plant genomics is an interesting case of use of our methodology since recently assembled genomes are extensively used to detect which polymorphisms are behind desirable phenotypic traits in crops. In this experiment, two consecutive assembled references of *Ath* were downloaded (ftp://ftp.arabidopsis.org/home/tair/Sequences/whole_chromosomes/) and compared. In particular, TAIR8 release (http://www.ncbi.nlm.nih.gov/assembly/GCF_000001735.1/) was analysed by our protocol, where potentially misassembled regions were obtained by comparing its sequence against the newer version of the genome TAIR9/10 (http://www.ncbi.nlm.nih.gov/assembly/GCF_000001735.3/). This comparison was done through the similarity score obtained from the BLAST-based protocol. In order to test our methodology, a set of 16 *Ath* NGS samples were downloaded from the European Nucleotide Archive (ENA) (Leinonen *et al.*, 2011a) (study accession PRJEB2457) and mapped with BWA software (Li and Durbin, 2010) (in *mem* mode) against the TAIR8 sequence. Then, a LGP was constructed ($l = 100$ bp) and the obtained quality control scores from each window were compared against the similarity score obtained to the same regions. Also, the coordinates of regions that were updated in TAIR8 assembly were evaluated against a set of randomly selected regions in order the estimate the sensitivity of our methodology to detect proved assembly artifacts.

Saccharomyces cerevisiae

Saccharomyces cerevisiae (*Sce*), a species of yeast, is an eukaryotic model organism widely used in molecular biology. Its genome contains approximately 12 Mb and it has been extensively tested in order to discover putative protein–protein interactions, single gene knock-down effects or synthetic lethality gene combinations, among others.

For this experiment, a set of 79 yeast samples were downloaded from the Sequence Read Archive (SRA) (Leinonen *et al.*, 2011 b) (study accession SRP091984) and subsequently mapped by using BWA software (Li and Durbin, 2010) (in *mem* mode) into a *de novo* assembly obtained *ad hoc* by using the reads of a selected individual (SRR4446970), representing the case where assembly is addressed only by using NGS reads. The assembly was performed by using Spades (Bankevich *et al.*, 2012) tool (a *kmer* size of 33). Finally, the LGP was carried out ($l = 100$ bp), and the obtained scores were also compared against the corresponding similarity score obtained by comparing the *de novo* assembly against the known reference genome (GCF_000146045.2 NCBI accesion, a good description can be found at Saccharomyces Genome Database at http://www.yeastgenome.org/cgi-bin/chromosomeHistory.pl).

Aeromonas hydrophilia

Aeromonas hydrophilia (*Ahy*) is a heterotrophic bacteria present in many human related environments, including sources of fresh water. It is resistant to most common antibiotics and causes several human diseases (like gastroenteritis), also, is considered one of the most virulent fish pathogens. Its genome contains approximately 5 Mb, and was included within the GAGE-B initiative (Magoc *et al.*, 2013) where several bacterial organisms where assembled by different available tools under study. For this experiment, we downloaded from the GAGE-B repository (https://ccb.jhu.edu/gage_b/genomeAssemblies/index.html) the *Ahy* assembly made by Abyss (Simpson *et al.*, 2009) tool. In this case, the LGP (with $l = 100$ bp) was constructed by using a set of NGS samples simulated (see Supplementary Materials) from the official *Ahy* reference genome (NC_008570 accession at NCBI, O'Leary *et al.*, 2016 repository). As previously, the quality control scores from each window were compared against the corresponding similarity scores obtained between Abyss assembly and the official *Ahy* reference genome.

Homo sapiens

Homo sapiens (*Hsa*) genomics is one of the most important fields in molecular biology research. Since the first draft (Lander et al., 2001), to its first stable assembled sequence in 2003 (International Human Genome Sequencing Consortium, 2004), it has been updated dozens of times (https://genome.ucsc.edu/FAQ/FAQreleases.html). At the moment, the human RGA is considered a high quality assembled sequence, with very few updates at every new release. Over this conservative scenario, two different experiments were designed to evaluate the accuracy of our methodology to detect putative misassembled regions in human genome. To do this, human reference genome version 37 (GRCh37, GCA_000001405.1) was downloaded from the Genome Reference Consortium official repository (https://www.ncbi.nlm.nih.gov/grc/human). In the first experiment, a set of well-known inconsistent loci in human RGA (fixed patched regions at GRCh37.p13 genome release, GCA_000001405.14) was compared against a set of randomly selected positions representing the background state of human genome in terms of error probability. To do this, 50 whole genome sequenced samples were downloaded from 1000 genomes project (Abecasis *et al.*, 2012) repository and used to construct a LGP ($l = 200$ bp, selected due to mean exon size). Then, the obtained quality control scores were compared between the two types of regions (patched and random). A second experiment was designed to evaluate the accuracy of our methodology under a genotyping context. Concretely, 30 selected exome samples (Supplementary Table S2) were downloaded from 1000 genomes project (Abecasis *et al.*, 2012) repository and used to construct a LGP ($l = 100$ bp). In this case, a second group of independent samples were also downloaded and genotyped by using GATK (McKenna *et al.*, 2010), a widely used variant calling NGS predictor (http://www.broadinstitute.org/gatk/guide/best-practices). NGS derived genotypes were compared against those predicted by a SNP validation array included within the 1000 genomes project official repository. The number of mismatches between the two standard protocols were used as a measure of noise degree and compared against the quality control scores initially obtained from the first group of samples. In this case, the two groups of samples allow us to evaluate whether a LGP constructed from a set of reference samples can be used later to evaluate the error probability of a new set.

3 Results

The assembly evaluation of *Ath* showed a high degree of similarity between the older version (TAIR8) and the current reference sequence (TAIR9/10) (Supplementary Fig. S1a), which properly agrees with the few number of updates accumulated between the two sequences during the last years. Likewise, the obtained *de novo* assembly for *Sce* described similar results. With 2337 scaffolds and a size of 11 669 271 bp (95.9% of the original genome, N50 = 61 488 bp) it showed a distribution of similarity scores mainly centered at higher values (Supplementary Fig. S1b), which suggests that NGS reads provided an assembly of reasonable quality. Contrarily, the evaluation of the downloaded *Ahy de novo* assembly (Supplementary Fig. S1c) showed a great density of similarity scores spread over lower values, which suggests that a significant portion of the assembly contains chimeric pieces of the original genome.

The profiled individual quality control scores showed strong differences between highly similar and poorly assembled regions in *Ath*, *Sce* and *Ahy* (Supplementary Fig. S2). Notably, mapping error probability (MEP) and Mann–Whitney derived scores (ME-MWZ, CE-MWZ) exhibited a clear descending trend when the similarity score showed an increment. Also, allelic variability scores (AF, ND, H and PI) showed a similar trend in all cases, especially strong in *Ahy* genome, demonstrating a robust relation between local quality and density of non-reference alleles. Interestingly, we can appreciate in some cases a different trend below a similarity score threshold (≈120), suggesting that this kind of estimators are especially useful when assembly artifacts are present at a subtle scale. On the other hand, CF, PP, MUF and IF (indicators of strong assembly errors) are especially sensitive in *Ahy* and *Sce*, but with less power in *Ath*, reflecting the differences between a high quality assembled genome and the tested *de novo* assemblies.

The computed CES exhibited a good concordance with similarity scores (Fig. 2) for all organisms. Particularly, we can observe how the CES keeps a value close to 1 when the similarity scores falls below a specific threshold (≈120), indicating an unequivocal presence of assembly artifacts. Then, it progressively decreases as similarity score reaches higher values. Summarizing the CES, we found that statistically significant windows in *Ath* described 2 187 900 bp (≈1.8% of genome) with strong signals of artifact presence (adjusted CES < 0.01). Also, 321 800 bp ≈ (2.8% genome) in *Sce* and 434 900 bp (≈8.7% genome) in *Ahy* were marked for posterior revision. Computed REAPR error scores also showed a good coincidence, specially for unequivocally altered regions. However, it showed a sensitivity loss in those regions where artifacts are partially present (high degree of similarity). These results are reflected in the lower statistical correlations (Table 1) obtained when compared with our methodology. Also, evaluation of patched regions throwed lower differences between the different loci type compared with CES.

On the other hand, *Ath* patched regions depicted a different pattern of CES to randomly selected regions (Fig. 2d), more separable than REAPR score. This pattern was reproduced for the three types of patches (insertions, deletions and modifications), also confirmed for almost all LGP scores. The equivalent analysis shows similar results in *Hsa*, where the difference between patched and random regions was also evident (Fig. 3a), including allele variability related profiles that showed a clear excess of variants at inconsistent regions.

At the second experiment in human RGA evaluation, the CES showed a considerable growing trend when the number of misgenotyped individuals also increased (Fig. 3b). This evidence was also supported by the majority of profiled scores, including those related to allelic variability. Furthermore, we evaluated the ability of our methodology to properly rank the likelihood of the individual variants of a set of independent samples. Concretely, the variants obtained from each sample were ranked according to the CES obtained from the LGP. Then, we checked the distribution of false-positive variants along the defined rank. The Figure 3c shows how the majority of false-positive variants are restricted to the end of the rank, demonstrating hence the suitability of the computed rank to separate true positive from false-positive findings. Figure 3d also describes this global effect where REAPR showed higher spreading of false-positive variants along the rank.

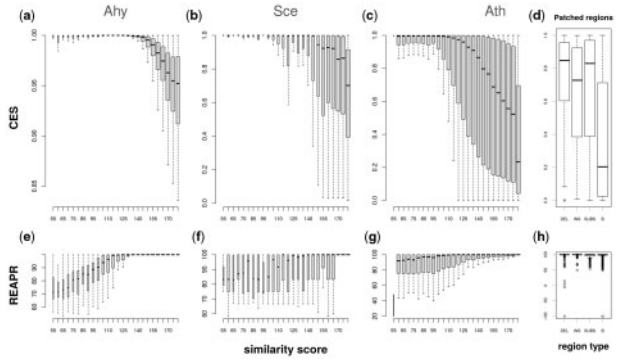

Fig. 2. Distribution of CES values depending on similarity score for *Ahy* (**a**), *Sce* (**b**) and *Ath* (**c**). CES was also plotted for *Ath* patched regions (**d**) and splitted in deletions (DEL), insertions (INS), substitutions (SUBS) and the set of randomly selected loci (B) that represents the background variability state of the genome. Distribution of REAPR values are also represented for the same categories: *Ahy* (**e**), *Sce* (**f**), *Ath* (**g**) and *Ath* patches (**h**)

Table 1. Correlations between BLAST-based similarity score and REAPR/log(CES) for *Ath*, *Sce* and *Ahy*

	REAPR	CES
Ath	0.30	0.48
Ahy	0.55	0.62
Sce	0.37	0.41

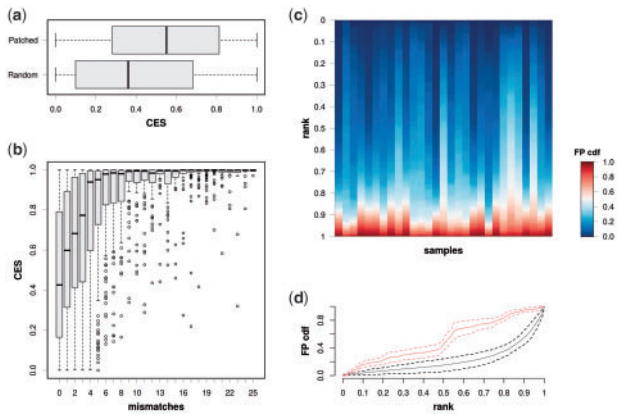

Fig. 3. CES distribution values for *Hsa* analysis. Clear differences are shown between patched and random regions of the genome (**a**). Also, CES showed a clear correlation with the number of mismatches between the NGS protocol and the validation SNP array (**b**). Interestingly, the false-positive variants of an independent set of samples fall at the end of the rank (**c**). The mean cumulative density function (cdf) of false positives is depicted (**d**) with clear differences between REAPR (light red curve) and our methodology (black curve)

4 Discussion

In this work, an effective methodology to characterize a RGA from a population perspective has been presented and evaluated. Across different experiments we have demonstrated how our protocol robustly detects both highly variable regions and noisy pieces of tested genomes. It is important to note that this protocol can be easily integrated in a real study since the statistical inference is constructed from the variability and quality-related scores usually obtained during a conventional NGS genotyping pipeline.

In general, the evaluation of a RGA is usually not undertaken due to the lack of proper bioinformatic tools, being the assembly necessarily considered correct as a whole. Our approach is especially useful when a non-model organism is under study, since the sequence is not usually well consolidated. In this case, misassembled regions will lead to false-positive differences when comparing species, varieties or groups of interest. Nevertheless, our results conclude that reproducible patterns of noise can be found even within a high-quality assembly such as human reference genome, suggesting that genome evaluation should be ordinarily applied in a broad range of studies.

Despite the huge effort made by bioinformaticians in the last decade to evaluate in detail the plethora of incoming genome assemblies (Bradnam *et al.*, 2013; Salzberg *et al.*, 2012), there still persists an important lack of standard methodologies to provide region wise measurements of a given RGA, the primary framework for any kind of downstream sequencing analysis. Although some methods provide local error assessment (Clark *et al.*, 2013; Rahman and Pachter, 2013; Vezzi *et al.*, 2012), new insights are needed to obtain a more robust noise susceptibility evaluation in newly assembled genomes when a specific region of interest is selected. The results presented here are inherently robust since significant quality or allele variability patterns are well supported by a set of independent observations provided by the population in itself. This results in a relevant improvement compared with currently available tools, without needing a close species reference genome to support the inference. This point can be easily proved through the sensitivity differences obtained with REAPR in those regions that partially contain assembly artifacts (high similarity scores in Fig. 2). Also patched regions analysis in *Ath* showed clear improvement of our methodology compared with REAPR.

Our approach is based on the empirical analysis of a set of selected noise estimators that allows a coarse-to-fine evaluation. While some estimators (like pair-end integrity descriptors) are able to describe large assembly inconsistences, some others (like variant-based comparators) are able to capture subtle differences, such as base changes in patched regions of *Ath* genome, between evaluated assemblies and reference genomes. It is important to note that our methodology could be easily extended in the future by including new noise estimators with the ability to add or improve any noise source detection. Also, inherent heuristic parameters of the method such as sample size of window length effect have been properly evaluated in order to provide more descriptive using guide to those users interested in evaluate their reference genomes.

In this study, we have presented an important case of use of our methodology, that is the preventive evaluation of a set of selected markers obtained from a population of individuals (selected 1000 genomes samples). Here, our computed score (CES) allows to properly rank the obtained candidate variants, separating true positive from false-positive markers, which would drastically optimize the true positive Sanger validation rate, and therefore, the consumed resources. Interestingly, the rank is effective even when the model has been constructed by using an independent population of samples, which demonstrates the robustness of our proposal. False-positive finding has been also exemplify through the comparison of patched against random regions (both in *Hsa* and *Ath* genomes) and the correlation between CES and similarity scores, showing in all cases a good degree of concordance, improving REAPR results.

As we have demonstrated, allele variability and noise susceptibility scores can be synergistically combined in order to improve the detection of inconsistent regions of the genome that can be proposed to be avoided at any further analysis. Furthermore, region-based allelic variability measurements could be in the future easily used to evaluate the variability patterns of different genomic substructures such as coding or intronic regions, intra or intergenic loci, or allelic variability patterns of different protein families evolved under different conditions.

Finally, both the source code and the tool description are available at the official code repository http://gitlab.com/carbonell/ces where the user can easily understand the details of our protocol.

Funding

This work was supported by grants [BIO2014-57291-R] from the Spanish Ministry of Economy and Competitiveness and 'Plataforma de Recursos Biomoleculares y Bioinform$_4$ticos' [PT13/0001/0007] from ISCIII, both co-funded with European Regional Development Funds (ERDF); and [EU H2020-INFRADEV-1-2015-1 ELIXIR-EXCELERATE (ref. 676559)]. David Conesa and Antonio López-Quílez would like to thank the Spanish Ministry of Economy and Competitiveness via research grant MTM2016-77501-P (jointly financed with the European Regional Development Fund).

Conflict of Interest: none declared.

References

Abecasis, G.R. *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Bankevich, A. *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19**, 455–477.

Berardini, T.Z. *et al.* (2015) The arabidopsis information resource: making and mining the 'gold standard' annotated reference plant genome. *Genesis*, **53**, 474–485.

Boomsma, D.I. *et al.* (2014) The genome of the Netherlands: design, and project goals. *Eur. J. Hum. Genet.*, **22**, 221–227.

Bradnam, K.R. *et al.* (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience*, **2**, 10.

Clark, S.C. *et al.* (2013) ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics*, **29**, 435–443.

Dopazo, J. *et al.* (2016) 267 Spanish exomes reveal population-specific differences in disease-related genetic variation. *Mol. Biol. Evol.*, **33**, 1205–1218.

Fisher, R. (1925) *Statistical Methods for Research Workers*. Oliver and Boyd Ltd., Edinburgh.

Gudbjartsson, D.F. *et al.* (2015) Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.*, **47**, 435–444.

Gurevich, A. *et al.* (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.

Hunt, M. *et al.* (2013) REAPR: a universal tool for genome assembly evaluation. *Genome Biol.*, **14**, R47.

International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Leinonen, R. *et al.* (2011a) The European nucleotide archive. *Nucleic Acids Res.*, **39** (Suppl. 1), 44–47.

Leinonen, R. *et al*. (2011b) The sequence read archive. *Nucleic Acids Res*., **39** (Suppl. 1), 2010–2012.

Lek, M. *et al*. (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.

Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.

Magoc, T. *et al*. (2013) GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics*, **29**, 1718–1725.

McKenna, A. *et al*. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*., **20**, 1297–1303.

Mikheenko, A. *et al*. (2016) MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics*, **32**, 1088–1090.

Moorjani, P. *et al*. (2013) Genetic evidence for recent population mixture in India. *Am. J. Hum. Genet*., **93**, 422–438.

Nagasaki, M. *et al*. (2015) Rare variant discovery by deep whole-genome sequencing of 1, 070 Japanese individuals. *Nat. Commun*., **6**, 8018.

O'Leary, N.A. *et al*. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*., **44**, D733–D745.

Rahman, A. and Pachter, L. (2013) CGAL: computing genome assembly likelihoods. *Genome Biol*., **14**, R8.

Salzberg, S.L. *et al*. (2012) GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Res*., **22**, 557–567.

Simpson, J.T. *et al*. (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res*., **19**, 1117–1123.

Tishkoff, S.A. *et al*. (2009) The genetic structure and history of Africans and African Americans. *Science*, **324**, 1035–1044.

Vezzi, F. *et al*. (2012) Reevaluating assembly evaluations with feature response curves: GAGE and assemblathons. *PLoS One*, **7**, e52210.

Walker, B.J. *et al*. (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, **9**, e112963.

Weinstein, J.N. *et al*. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet*., **45**, 1113–1120.

Xing, J. *et al*. (2010) Genetic diversity in India and the inference of Eurasian population expansion. *Genome Biol*., **11**, R113.

Zhu, X. *et al*. (2015) misFinder: identify mis-assemblies in an unbiased manner using reference and paired-end reads. *BMC Bioinformatics*, **16**, 386.