

RESEARCH ARTICLE

Open Access



Impact of short-read sequencing on the misassembly of a plant genome

Peipei Wang^{1,2}, Fanrui Meng^{1,2}, Bethany M. Moore^{1,3} and Shin-Han Shiu^{1,2,3,4*} 

Abstract

Background: Availability of plant genome sequences has led to significant advances. However, with few exceptions, the great majority of existing genome assemblies are derived from short read sequencing technologies with highly uneven read coverages indicative of sequencing and assembly issues that could significantly impact any downstream analysis of plant genomes. In tomato for example, 0.6% (5.1 Mb) and 9.7% (79.6 Mb) of short-read based assembly had significantly higher and lower coverage compared to background, respectively.

Results: To understand what the causes may be for such uneven coverage, we first established machine learning models capable of predicting genomic regions with variable coverages and found that high coverage regions tend to have higher simple sequence repeat and tandem gene densities compared to background regions. To determine if the high coverage regions were misassembled, we examined a recently available tomato long-read based assembly and found that 27.8% (1.41 Mb) of high coverage regions were potentially misassembled of duplicate sequences, compared to 1.4% in background regions. In addition, using a predictive model that can distinguish correctly and incorrectly assembled high coverage regions, we found that misassembled, high coverage regions tend to be flanked by simple sequence repeats, pseudogenes, and transposon elements.

Conclusions: Our study provides insights on the causes of variable coverage regions and a quantitative assessment of factors contributing to plant genome misassembly when using short reads and the generality of these causes and factors should be tested further in other species.

Keywords: Genome misassembly, Read coverage, Machine learning, *Solanum lycopersicum*

Background

The number of whole genome sequences has increased dramatically in the last decades due to the development of new generations of sequencing technologies and reduced cost. The “first” generation was Sanger sequencing technology [1], based on which a decade was taken to deliver a draft genome of human [2]. The second-generation technology—i.e., “Next generation sequencing” where thousands to millions of DNA molecules are enabled

to be sequenced simultaneously dramatically shortens the time required to obtain high genome coverage [1]. However, due to the short length of these reads (36 bp ~ 400 bp), there are many challenges for assembling genome based on short reads, including the difficulty in sequencing repetitive sequences [3], low read coverages in GC-poor or GC-rich regions [4], genome sequencing bias introduced by PCR amplification during library construction [5], and polyploidy in some species including most flowering plants [6]. The advent of third generation sequencing, e.g., Pacific Biosciences (PacBio) single molecule real time sequencing [7] and Oxford Nanopore sequencing [8], has led to another revolution in genome sequencing, where long reads up to 100 Kb can be sequenced in a single

* Correspondence: shius@msu.edu

¹Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA

²DOE Great Lake Bioenergy Research Center, Michigan State University, East Lansing, MI 48824, USA

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



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

run without PCR amplification or chemical labeling of the sample. Although the much higher error rates remain an issue [9], the third generation sequencing still has merit for applicants more tolerant to error rates, like structural variant calling [10] and, combined with short-read sequencing, is overtaking projects that focus on short reads only.

Although the number of genome sequences which take the advantages of both second and third generation sequencing are increasing [11–14], the majority of genome assemblies available from the National Center of Biotechnology Information (NCBI) were generated predominantly using short reads from the second-generation technology. Before the third-generation sequencing is more widely applied to improve these genome assemblies, it remains important to assess the quality of existing short-read based assemblies. Several methods have been developed for this purpose, like Scaffold N50, MaGuS [15], LTR Assembly Index [16], SQUAT [17]. In addition to these methods, another strategy is to assess how well an assembly is covered by the reads used for building the assembly.

For an ideal genome assembly, the sequencing reads would be uniformly distributed across the genome. However, in the real world, when sequencing reads are mapped back to the genome, the read coverage varies across genome due to multiple reasons. First, regions with extremely high or low GC content may not be sequenced equally compared with other GC-balanced regions, leading to low or even no coverage of reads [18]. Second, repetitive sequences are abundant in species with larger genomes, and have always been a major challenge for genome assemblies [3]. Repeats longer than read length would lead to gaps in the genome assembly due to uncertainty in assembly of these regions. This would break down the genome into pieces, leading to the loss of linkage information among genetic markers. Third, repeats may also be led to misassembly where two unlinked regions were joined together and resulted in higher than usual read coverages. In the case of repetitive sequences containing genes, such as tandemly duplicated genes and retrogenes, such misassembly would reduce the gene copy number estimation. These missing genes not only make it challenging to account for all the genes in a genome but also create problems for functional genomic studies by impacting gene expression level estimates or loss-of-function studies. For example, the annotated SEC10a gene from *Arabidopsis thaliana* and its recent tandem duplicate copy SEC10b, were assembled together and annotated as a single gene, which explains why homozygous T-DNA insertion mutant of either copy has no phenotypic change compared to wild-type [19].

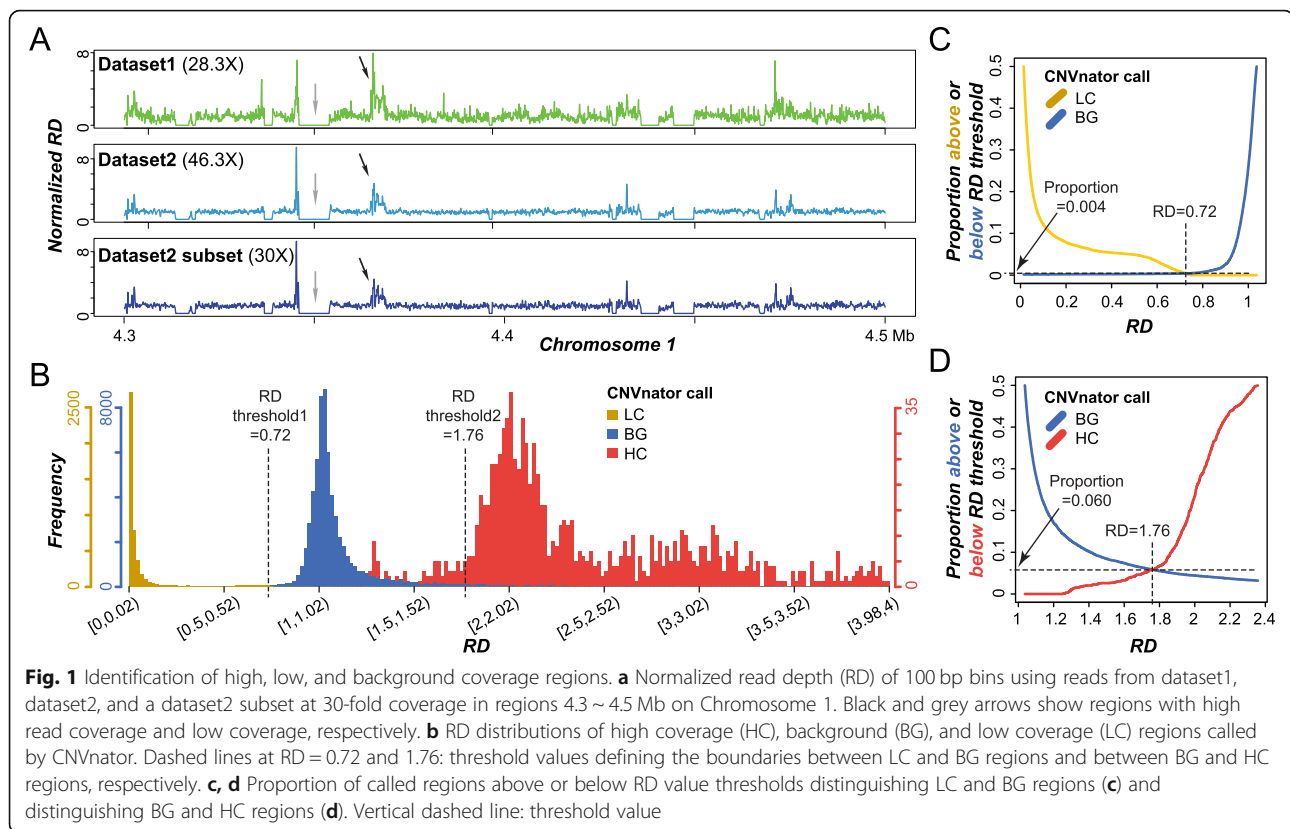
Here, we use tomato (*Solanum lycopersicum*) as a model to assess the extent to which its assembly has

variable coverage and the extent to which the misassembly is associated with variable read coverage. Tomato is chosen because assemblies built with short reads, as well as PacBio long reads are both available. In addition, it is an important crop and a major model species for studying specialized metabolism, particularly considering specialized metabolism genes tend to be duplicated tandemly [20, 21] and may tend to be misassembled. The idea of searching for regions with significantly high or low read coverages has been extensively applied in estimating copy number variation (CNV) among species or populations [22–24]. Here, we aimed to investigate whether the read coverage can be used in detecting misassembled regions. Using CNV detection tools, we identified genome regions with significantly high- and low-coverage of sequencing reads compared to the genome average (i.e., background). Based on genomic sequence information on regions with variable coverage, our primary goals were to explore underlying factors influencing the read coverages through machine learning approaches. Most importantly, assembly quality was assessed through comparison between short and long-read assemblies. Finally, factors (i.e., sequence features) informative to predict misassembled regions were also investigated.

Results and discussion

Abundance of tomato genomic regions with higher and lower than average coverage

Two datasets were used to determine how well different genomic regions were covered and to define variable coverage regions: genomic regions with significantly higher coverage (HC) or lower coverage (LC) than average (referred to as background, BG). The first, dataset1, was generated with Illumina Genome Analyzer IIX (GAIIx) sequencer with 90-bp paired-end and 54-bp mate pair reads (~28x coverage) used in the original genome assembly [25], and the second, dataset2, was generated with Illumina HiSeq 2000 sequencer with 101-bp paired-end reads (~46-fold coverage). To assess the qualities of these two datasets and to see if both datasets should be analyzed, the tomato genome assembly was split into 100 bp bins and, for each bin, the read depth (RD) was determined using either dataset1 or 2 (see [Methods](#)). After correcting the RD values for GC content bias [22], the median RDs for dataset1 and dataset2 are 1.04 and 1.05, respectively (an example region on chromosome 1, Fig. 1a). The RD values from these two datasets are significantly correlated (Spearman's $\rho = 0.40$, $p < 2.2e-16$), and consistently revealed bins with substantial deviations from the median value in both directions indicating the presence of HC and LC regions (e.g., grey and black arrows in Fig. 1a). However, RDs of dataset1 was significantly more variable across genome



(variance = 0.25 using 0 ~ 99 percentile values) compared to those of dataset2 (variance = 0.15, F-test, $p < 2.2e-16$, Fig. 1a). This is not simply due to the higher genome coverage of dataset2 (~ 46.3x) compared to dataset1 (~ 28.6x), because a subset of randomly sampled reads from dataset2 to ~30x genome coverage has much more similar RD estimates as dataset2 (Spearman's $\rho = 0.87$, $p < 2.2e-16$, Fig. 1a).

The comparably higher RD variance in dataset1 may be due to lower sequencing quality and shorter read lengths that contribute to erroneous read mapping and may lead to overestimates of HC and LC regions. Thus, only results based on dataset2 were discussed further. Based on the RD values, HC, LC, and BG regions were identified with CNVnator (see [Methods](#)). However, we found that RD distributions of HC, LC and BG regions called by CNVnator were overlapping (Fig. 1b). Given our goal is to identify HC and LC regions with high confidence, two threshold RD values were chosen: 0.72 and 1.76 that minimize the overlap between LC and BG regions (Fig. 1c), and that between BG and HC regions (Fig. 1d), respectively. Thus, HC regions were defined as regions with RD > 1.76, and LC regions were those with RD < 0.72. BG regions had RD values between 0.72 and 1.76. This resulted in 1156 HC, 19,451 BG and 15,034 LC regions. The HC and LC regions account for 0.6% (5.1 Mb) and 9.7% (79.6 Mb) of the genome, respectively.

The regions that are not classified as HC, LC, or BG due to this dual thresholding scheme are referred to as “other” regions. As expected, 95.5% of LC regions contain gaps filled with Ns, whereas only 18.2% of HC and 42% of BG contain Ns (Fig. S1a). In addition, the median lengths for LC, BG, and HC regions are 1.70, 16.30, and 2.80Kb, respectively (Fig. S1b).

Prediction of HC, LC, and BG regions with a multi-class model using seven genomic features

After the HC/LC/BG regions were defined, we established machine learning model predicting which regions would be HC, LC, or BG regions. By predicting HC/LC/BG regions, we would have a better understanding of what the contributing genomic characteristics were, especially for HC regions where misassembly most likely have occurred. Starting out, we used seven features (referred to as base features, yellow box, Fig. 2a), including GC content, density values: all genes, tandem genes, non-tandem genes, pseudogenes, transposable elements (TEs), and simple sequence repeats (SSRs) of genomic regions to build a 3-class (HC, LC, or BG) model (referred to as Model 1, Fig. 2a, Table S1) using the Random Forest algorithm [26]. We should emphasize that an independent, test set (10%) of HC, LC, and BG regions were set aside that was not used for model building. Thus, the test set was ideal for validating our

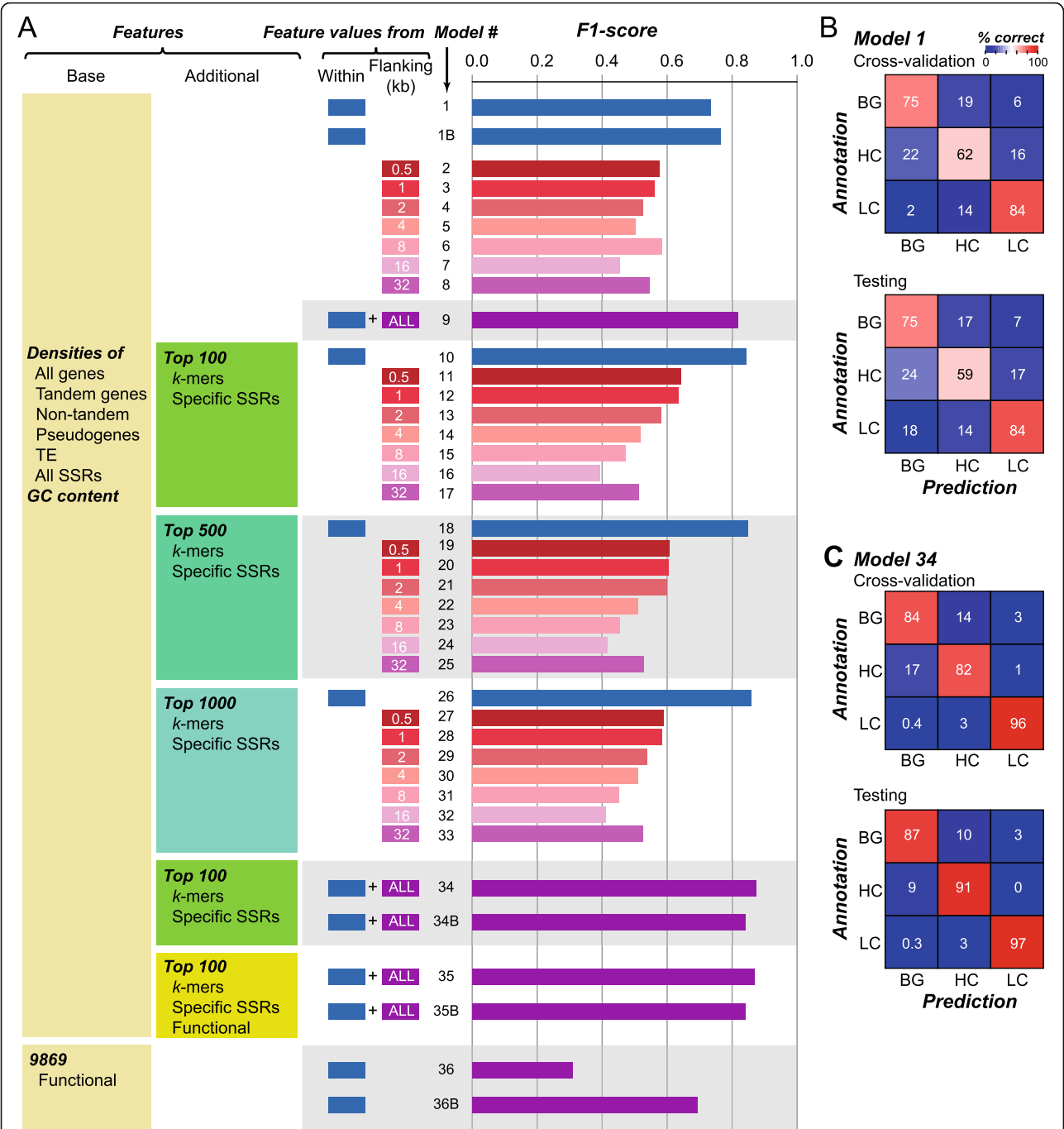


Fig. 2 Prediction of HC/LC/BG regions using features of genomic sequences. **a** Average F1 of prediction models using different feature sets. The first two columns indicate feature sets used including Base features used in most models and additional features for model subsets. The third and fourth columns indicate whether feature values were derived from sequences within HC/LC/BG regions and/or flanking regions (0.5 ~ 32Kb), respectively. In the flanking region column, “all” refers to a combined set of values from all seven flanking regions. The fifth column contains model number (#), where B indicates binary model. **b** Confusion matrices of cross-validation and testing datasets based on Model 1. Color scale: % correct prediction. **c** Confusion matrices of cross-validation and testing datasets based on Model 34

models. Using Model 1, 61.9, 83.9, and 74.8% of HC, LC, and BG regions, respectively, were correctly predicted (Fig. 2b). Here the percentage true cases predicted correctly is defined as the recall value. Importantly, the testing set not used for model training were predicted with a similar recall (Fig. 2b). To jointly consider both recall and precision (% predictions that are correct), we determined the F1-score that is the harmonic mean of

precision and recall. In our machine learning pipeline, we started with equal numbers of training and testing HC, LC, and BG regions (33% each). Thus, random guess would lead to an accuracy of 33% and $F1 = 0.33$. On the other hand, a perfect model would have an accuracy and an $F1$ of 1. Model 1's $F1 = 0.73$ (Fig. 2a), while it was much better than random guess, the $F1$ was far from perfect.

Defining HC, LC, and BG regions with additional features

To improve upon Model 1, we included additional features from two sources. The first was the same seven base features but with values from flanking regions. The rationale was that the regions right next to HC, LC, and BG regions may have similar properties which can contribute to a better model. To assess this, we first build prediction models using only sequences flanking HC, LC, and BG regions by 0.5, 1, 2, 4, 8, 16, and 32kbs to build seven models (Model 2–8) and found that the performance of these models was worse than that of Model 1 (accuracy = 46 ~ 58%, $F1 = 0.46 \sim 0.58$, Fig. 2a and Table S1). In addition, as the sizes of the flanking regions increased, the prediction performance decreased (Fig. 2a) This is likely because flanking regions can be of different types, i.e., a region flanking an HC region may be LC and/or BG regions. However, this is not because these regions are not important. When the features used for building Model 1 were combined with those for Model 2–8, the resulting model (Model 9) had a substantially improved $F1 = 0.82$ (Fig. 2a) compared to Model 1 ($F1 = 0.74$). This finding suggests that sequences flanking the HC/LC/BG regions, by themselves insufficient, have information that are useful for the prediction task.

In addition to flanking region, we focused on dissecting if HC, LC, and BG regions have different sequence composition—instead of compositions of much longer sequences (genes and transposons) or single nucleotides (GC content), we investigated whether specific SSRs (repeats with 2–64 bp units, 156,444 features) and/or k -mer (1–6 bp, 5460 features) may be prevalent in HC, LC, or BG regions. Because the number of these SSR and k -mer features was large, we first identified a subset of SSR and k -mer features with p -value < 0.05 (Kruskal-Wallis H test) among HC/LC/BG regions. Top 100 SSR and k -mers were further selected with a feature selection algorithm (see Methods). Feature selection is the practice of selecting the most informative features for model building. Because of the number of features (over 160,000) is much higher than the number of instances used for model training, we run the risk of grossly overfitting the model. Thus feature selection was applied, in addition to the practice of cross-validation and setting aside a testing set, to alleviate the issue of overfitting. By

incorporating these top SSR and k -mer features with seven base features to predict HC/LC/BG regions (Model 10), the performance of Model 10 ($F1 = 0.84$, Fig. 2a) was even better than the Model 9 ($F1 = 0.82$) that did not consider SSRs and k -mers but flanking regions. Thus, there exist substantial differences in the short sequence compositions among HC/LC/BG regions. We also included the top 500 and the top 1000 k -mers/SSRs to create Model 18 and Model 26 that improved performance further with $F1 = 0.85$ and 0.86 , respectively (Fig. 2a). Although additional k -mers/SSRs may further improve predictions, they likely have diminishing contribution judging from the small $F1$ differences between Model 10, Model 18 and Model 26 (blue bars, Fig. 2a). Next, we combined the features used in Model 9 and those from used in Model 10 to establish an all-inclusive Model 34 with 256 features that had $F1 = 0.87$ (Fig. 2a). Importantly, $> 84\%$ BG, $> 82\%$ HC and $> 96\%$ LC regions are correctly predicted in both training and test (not used in model training) datasets (Fig. 2c).

Features important for the prediction of HC/LC/BG regions

Model 34 has the highest $F1 = 0.87$ using 256 features (56 base features from 8 regions, 100 top k -mers, and 100 top SSRs, Fig. 2a). We next evaluated which features were among the most informative in distinguishing HC/LC/BG regions (highest feature importance values, see Methods). Table S2 lists the importance values of all 256 features. We found that three types of features stand out: k -mers (median importance rank = 57), GC content (median rank = 66), and density of TEs (median rank = 70). TEs have long been implicated in their contributions to misassembly due to their lengths and high degree of similarities [3]. Interestingly, the HC regions tend to have a significantly lower TE density compared to BG regions (Fig. 3a), likely reflecting genomic regions differing in recent transposition events. In contrast, LC regions have the highest TE density, although distribution of LC regions across the genome is not correlated with TE distribution (Spearman's $\rho = 0.09$, Fig. 3b). Furthermore, when we randomly reshuffled HC/LC/BG regions designations 1000 times and determined the correlation distribution between prevalence of TEs and random genomic regions, the observed correlation value of LC regions was significantly lower than that of the random expectation (z -score = -3.0 , Fig. 3c). One potential reason is that the assembler may be confused by the repetitive nature of TE and short length of sequencing reads to assemble sequences correctly, resulting in gaps filled with Ns in LC regions (Fig. S1a) with TE sequences at the breakpoints in one or both ends, which in turn led to higher TE density in LC regions.

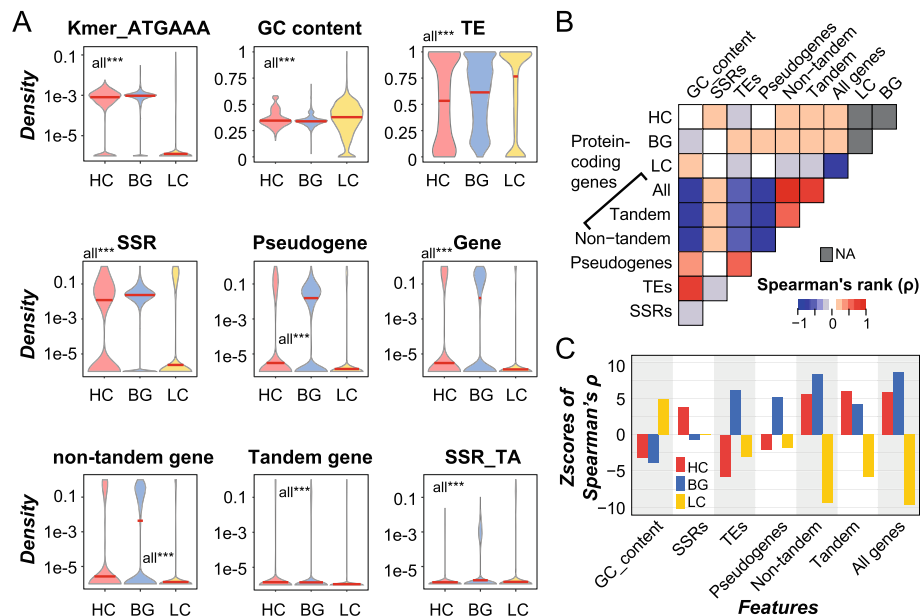


Fig. 3 Relationships between genomic features and HC, LC and BG regions. **a** Differences in density distributions of important features between HC/LC/BG regions. ***: $p < 1e-3$. **b** Correlation between densities of HC/LC/BG regions and other genome features across the genome in bins of 500Kb. Color scale: Spearman's rank correlation coefficient (ρ). **c** Z-scores of observed p calculated using the distribution of 1000 random Spearman's p as the null distribution. First, regions were selected so they were the same number and length as true HC/LC regions (not overlapping with each other). The remaining genomic regions not occupied by the selected random regions were taken as randomly expected BG regions. Then correlation between densities of randomly selected regions and a genomic feature was calculated to establish the null distribution

As for GC content, it is well documented that, specifically for short read sequencing with the Illumina platform, GC-rich and GC-poor regions tend not to be sequenced and thus contribute to regions with low coverage or breakpoints in assemblies [4, 27]. Consistent with this, LC regions have significantly higher GC content compared to HC and BG regions (Fig. 3a). We also found that a major reason that the top 100 k -mers were important was because of their high AT content (88% with AT content ≥ 80 and 100% with AT content $\geq 60\%$, Table S2). In addition, the top two SSRs with highest importance in the prediction were “AT” (ranked 152) and its reverse complement “TA” (ranked 153). Contrary to densities of individual SSRs which generally had very low ranks (i.e., less important), densities of all SSRs ranked in the middle (108), indicating that it is more informative in distinguishing HC/LC/BG regions to consider SSRs as a whole. Consistent with this, when only the top 128 features were used, where the density of all SSRs was considered but no individual SSR feature, the model's performance didn't decline (Model 34_4 in Table S1).

Pseudogenes (median rank = 131), all protein coding genes (139), non-tandem duplicates (141), and tandem duplicates (154) also ranked in the middle (Table S2). Densities of these genomic features

in flanking regions ranked similarly or even higher than densities in HC/LC/BG regions (Table S2), suggesting the differences in genomic environment around HC/LC/BG regions. By determining the correlations between the prevalence of HC/LC/BG regions and the prevalence of genomic features in corresponding regions, we found that genomic regions with high densities of not only HC, but also BG regions tend to have higher gene density, regardless if tandem and non-tandem genes were separated or not (all $\rho > 0.12$, p -values $< 2.0e-6$, Fig. 3b and Table S3). Because LC regions tend to contain Ns, regions with higher LC density are expected to have lower gene density (all $\rho < -0.17$, p -values $< 2.2e-11$, Fig. 3b and Table S3). Given that density of protein coding genes is informative in distinguishing among HC/LC/BG regions, we next asked whether the types of genes, in terms of functional aspects (e.g., Pfam domains, biological processes and metabolic pathways) also impact RD. To test this, top 100 functional characteristics (domains, biological processes and pathways) of genes within HC/LC/BG regions (9869 features in total) were also combined with Model 34 to build Model 35 (Fig. 2a, Table S1). However, there was no apparent improvement compared to model 34 (model 35's overall accuracy = 87%, F1 = 0.87).

Features important in binary classification model distinguishing HC and BG regions

Although the importance analysis allows us to pinpoint the features crucial for Model 34's performance, Model 34 is a 3-class model and thus it is not straightforward to tell if a feature is important because it allows us to distinguish HC from BG and LC regions or other scenarios. Because we are mostly interested in assessing why HC regions exist, we next establish a binary classification model to distinguish HC from background regions. Using the same features as in Model 1, we established a new Model 1B (B = binary) with HC and BG regions as classes and found that it has an accuracy = 84% and F1 = 0.76 (Fig. 2a). As expected, Model 34B that used the same feature set as Model 34 for binary classification had even better accuracy = 92% and F1 = 0.84 (Fig. 2a). Like Model 35 which didn't lead to improved classification among HC/LC/BG regions by including functional features (Table S1), the comparable binary Model 35B had the same performance as Model 34B (Fig. 2a, Table S1). However, models with only functional features had accuracy = 58% and F1 = 0.69, which is much better than random guess (Table S1), suggesting that functional features are still informative in distinguishing HC from BG regions.

As expected, the important features for binary classification of HC and BG regions differ from those for the 3-class model. For example, densities of *k*-mer in Model 34B (median rank = 73) and in Model 35B (median rank = 72) were no longer the most important feature categories as in Model 34 (median rank = 57) and 35 (median rank = 55, Table S2,4,5,6). In contrast, GC content, density of TE and SSRs had the highest median ranks (12, 13, and 57 in Model 35B, respectively). For HC regions, one hypothesis for their presence is due to the presence of multiple copies of highly similar sequences arranged in tandem that are misassembled. If this is true, one would expect that SSRs and tandem genes would tend to be co-localized with HC regions compared to BG regions. Consistent with the above hypothesis, although the density of SSRs in HC regions was slightly lower than BG regions (Fig. 3a), it was significantly higher than randomly expected (z-score = 3.81, Fig. 3c). In contrast, density of SSRs in BG regions was slightly lower than random expectation (z-score = -0.69 Fig. 3c). In addition, the density of SSRs across genome is positively correlated with HC, not BG, regions (Fig. 3b), and the flanking regions of HC also have higher density of SSRs than those of BG regions (Fig. S2). These results suggest the potential contribution of SSRs to misassembly in HC regions, which resulted in underestimation of SSRs density in HC regions. The situation is similar for tandem genes, although it is not as important as SSRs (median rank = 155). The observed correlation

value (ρ) for HC regions was significantly higher than random expectation (z-score = 6.0) compared to that for BG regions (z-score = 4.1, Fig. 3c). Note that, although both have positive z-scores due to consideration of LC regions also, the higher z-score for HC regions indicates that tandem gene density is more prevalent in HC than in BG regions. Conversely, compared to BG regions, HC regions tend to have fewer non-tandem genes (Fig. 3c). Thus, the presence of tandem genes also contributes to misassembly.

Properties of genes located in HC regions

In earlier section, functional characteristics (domains, biological processes and pathways) of genes within HC/LC/BG regions were also combined with the seven base features to build Model 35 and 35B (Table S1) that resulted in no apparent improvement compared to model 34 and 34B that did not incorporate functional characteristics (Fig. 2a). This may be because properties contributing to the enriched presence of genes with certain functional characteristics were already considered, it is also possible that, due to the large number of features considered and the fact that functional characteristics tend to be lower ranked, the contribution of functional characteristics was not apparent in Model 35 and 35B because other features dominated. To assess the extent to which functional characteristics could be used to predict whether a genomic region would be BG, HC, or LC, we established three-class models using only functional features and found that it had accuracy = 41% and F1 = 0.31, very close to random guess, no matter how many features were selected (Model 36, Fig. 2a, Table S1). However, binary model for classifying HC and BG regions using only functional features had accuracy = 57.9% and F1 = 0.69, indicating that they were informative (Model 36B, Fig. 2a, Table S1).

To assess what types of genes tend to be located in BG and HC regions, we first determined if the numbers of different types of genes (Table S7) were over or under-represented in HC compared to BG regions. By generating 10,000 datasets with randomized HC locations, we established the randomly expected numbers of different gene types and the resulting null distributions were used to assess the statistical significance of observed numbers of different gene types (Fig. 4a). In this analysis, two types of genes stand out, specialized metabolism (SM) protein coding genes and RNA genes. SM genes has a z-score = 2.1, indicating that SM genes tend to be found in HC regions and thus misassembled. This is consistent with the findings that SM genes tend to belong to large gene families, located in tandem clusters, and be recently duplicated [20, 21]. However, genes in larger families are not necessarily in HC regions (black arrow, Fig. 4a) and number of SM genes that are tandemly duplicated is not

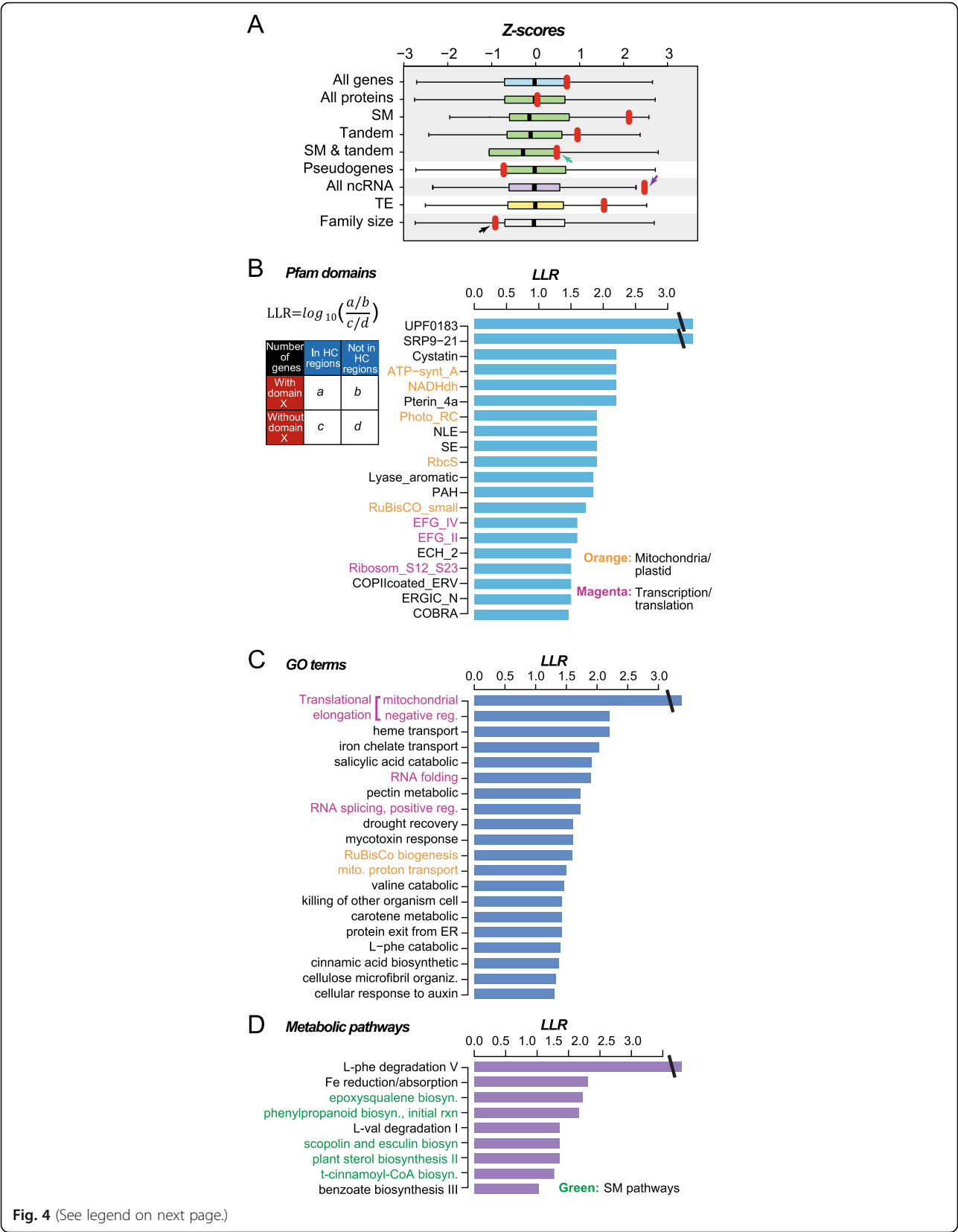


Fig. 4 (See legend on next page.)

(See figure on previous page.)

Fig. 4 Sequences found in HC regions and their functions. **a** Z-score calculated by comparing observed total number of features in true HC regions (red dots) to normalized distributions (boxplots) of the numbers of different genomic features overlapped with 10,000 randomly sampled BG regions (based on the number and length distribution of HC regions). **b** Top 20 Pfam domains with highest Log Likelihood Ratio (LLR, see upper-left insert and Methods), indicating enrichment within HC regions. **c** Top 20 GO terms with highest LLR. **d** 9 metabolic pathways with LLR > 1. Orange and magenta fonts indicate mitochondria/plastid and transcription/translation related processes, respectively, and green font shows specialized metabolism pathway

significantly higher than random expectations (green arrow, Fig. 4a). Thus, it is likely that the over-representation of SM genes in HC regions is due to their higher duplication rate, but not always through tandem duplication, resulting in closely related copies that were misassembled. It also can be because tandem duplicated SM genes were misassembled together, which makes the number of tandem duplicated SM genes underestimated. In addition to SM genes, surprisingly, non-coding RNAs (ncRNAs) tend to be enriched within HC regions (z-score = 2.5, purple arrow, Fig. 4a). We speculate that tomato ncRNA regions may have a higher-than-average rate of recent duplications, which would indicate there are more ncRNA regions than annotated and ncRNA expression levels may be overestimated because multiple ncRNA regions are assembled together.

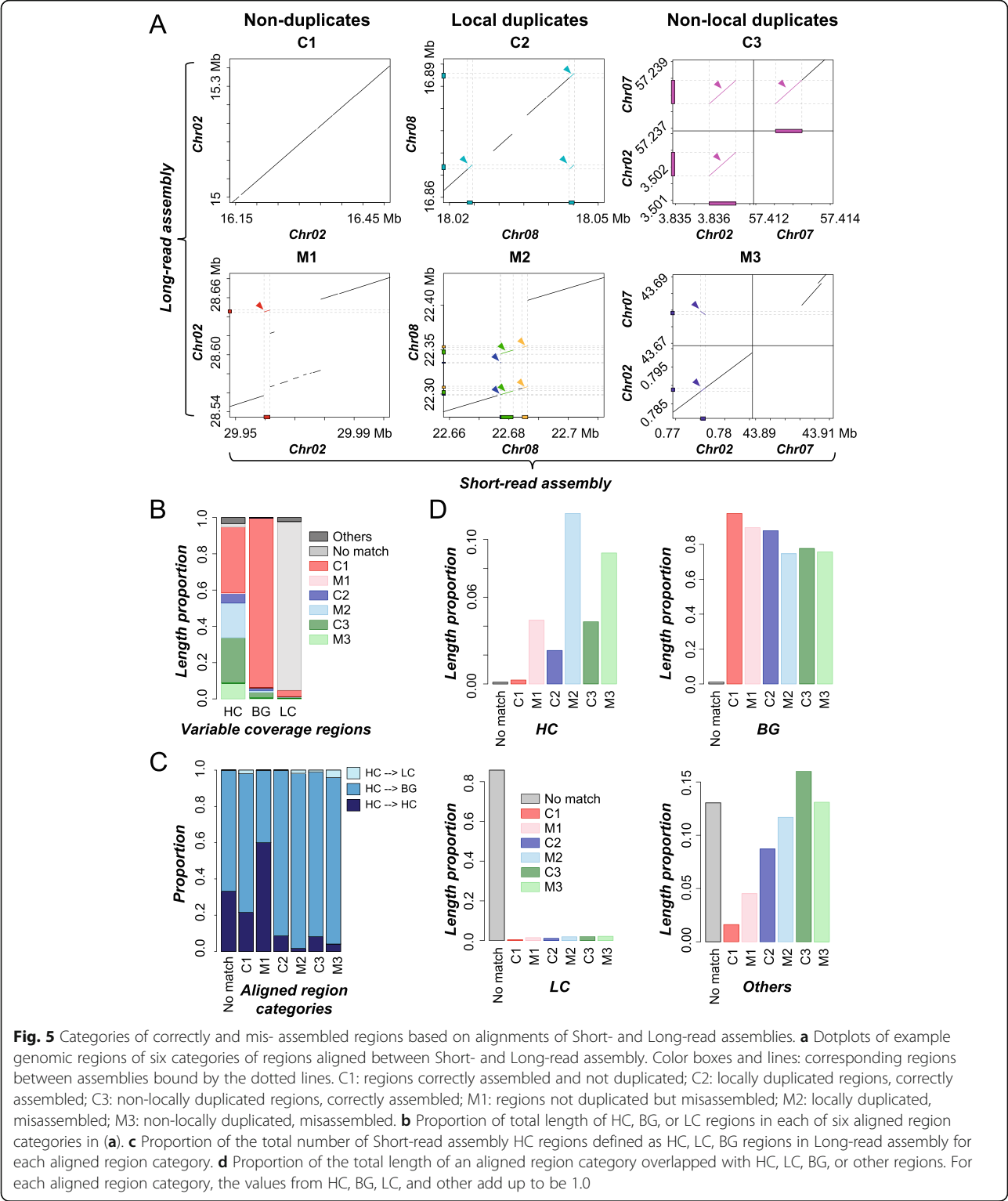
Our finding that SM genes tend to be over-represented in HC regions suggests that genes with other functions may have similar behaviors. To address this, we asked if there was enrichment of any Pfam domain family, Gene Ontology (GO) biological process category, or TomatoCyc pathways. Given the number of domain families (Table S8), categories (Table S9) and pathways (Table S10) were large, multiple testing correction was applied and resulted in only one statistically significantly enriched entry (salicylic acid catabolic process). To assess if there are general patterns we may have missed due to the stringency of the multiple testing corrections, we examined the Log Likelihood Ratio (LLRs, see Methods) between the numbers of genes with or without a protein domain X and the numbers of genes within or out of HC regions (inserted table, Fig. 4b). Similarly, we examined the LLRs for biological processes (Fig. 4c) and pathways (Fig. 4d). Here the HC regions were compared with the whole genome. We have also conducted the same analysis but between HC and BG regions that produced similar results (Table S11–13).

There are three general patterns that emerge. The first is the prevalence of nuclear encoded proteins responsible for mitochondrial and plastid functions among the Pfam domains and the GO categories with the highest LLRs—including ATP-synt_A: ATP synthase A chain, NADHdh: NADH dehydrogenase, Photo_RC: photosynthesis reaction center, and RbcS and RuBisCO_small:

Ribulose-1,5-bisphosphate carboxylase small subunit (Fig. 4b), as well as mitochondrial proton transport and RuBisCo biogenesis (Fig. 4c). The second general pattern is the occurrence of domain/process related to transcription and translations—including various translational elongation factor G (EFG) domains, translational elongation-related functions, ribosomal proteins, RNA splicing (Fig. 4b,c). One outstanding property of genes that fit these two general patterns is their extremely high level of expression. Such high level of expression is known to lead to the generation of retrogenes and retro-pseudogenes with highly similar sequences that littered around various parts of the genomes [28, 29] thus higher coverage within genomes. Consistent with this hypothesis, the average number of introns in genes found in HC regions was significantly lower than that in BG regions (2.13 vs. 4.38 in average, Kolmogorov-Smirnov test, $p = 4.3e-09$). The third general pattern is revealed from the few metabolic pathways with LLR value > 1 where five out of nine were SM pathways (Fig. 4d), as expected.

Evaluation of HC region misassembly by comparing short-read and long-read assemblies

To assess the extent to which HC regions tend to be misassembled, Short-read assembly (query) was aligned to Long-read assembly (subject) using MUMmer [30] (see Methods), and the aligned regions were shown in Table S14–16. Aligned regions were classified into six categories (see Methods, Fig. 5a): 1) non-duplicated, Correctly assembled (C1, 681.0 Mb); 2) non-duplicated, misassembled (M1, 0.6 Mb); 3) locally duplicated (i.e., on the same chromosome of the Long-read assembly), correctly assembled (C2, 11.5 Mb); 4) locally duplicated, misassembled (M2, 8.2 Mb); 5) non-locally (on different chromosome) duplicated, correctly assembled (C3, 29.5 Mb); and 6) non-locally duplicated, misassembled (M3, 4.9 Mb). We found that 86.3 Mb of Short-read assembly regions, mostly consisted of Ns, that could not be aligned to the Long-read assembly. Since LC regions tend to consist of Ns, it is not surprising that 93% of the total length of LC regions had no match in Long-read assembly (Fig. 5b). Thus, LC regions were not further examined. Of the 5.1 Mb HC regions, 0.03 Mb (0.53%), 0.97 Mb (19.1%) and 0.44 Mb (8.7%) were in misassembled



categories M1, M2 and M3, respectively (Fig. 5b). Compared to HC regions, the proportion of misassembled regions in BG were significantly lower (0.08, 0.9 and 0.5% for M1, M2, and M3, respectively; Fisher's Exact tests, all $p < 2.8\text{e-}09$). Among the three categories of misassembled regions, only 1.9% HC regions (M1) did not have duplicated subject regions in the Long-read assembly. The majority of misassembly (67.4 and 30.8%) was due to

duplications (M2 and M3, respectively), suggesting that HC regions were much more likely to be misassembled due to duplications, especially when duplications occurred on the same chromosome.

Thus far, HC regions were defined by mapping short reads to the Short-read assembly. To evaluate whether these Short-read assembly-based HC regions were still classified as HC in the Long-read assembly, the short reads were also mapped to the Long-read assembly to determine variable coverage regions. This resulted in 297 HC, 4479 BG and 1971 LC regions based on the Long-read assembly. Importantly, among 1156 Short-read assembly-based HC regions, once we map the reads to the Long-read assembly, only 88 (7.6%) overlapped with the Long-read assembly-based HC regions. In addition, among misassembled HC regions (coverage defined using the Short-read assembly), 96.5 and 91.8% of M2 and M3 were identified as BG based on the Long-read assembly (Fig. 5c, Table S17). These findings further suggest that higher than usual read coverage is a good indicator of misassembly.

HC regions tend to be misassembled compared to BG or LC regions (Fig. 5b). If we broke down the six aligned region categories (Fig. 5a), it was clear that HC regions have higher proportion of M2 (11.8%) and M3 (9.1%) compared to other categories (0.3–4.4%, Fig. 5d). Nonetheless, 74.6% of M2 and 75.7% of M3 were identified in BG regions (Fig. 5d). One potential reason is that some true HC regions were identified as BG in our analysis. If that was the case, we would expect misassembled BG regions (which presumably were HC) would have significantly higher read coverage compared to correctly assembled BG regions. Consistent with this expectation, the median RDs of 100 bp BG region bins that were misassembled (1.13 and 1.28 for M2 and M3, respectively) were higher than the median RDs of correctly assembled BG bins (1.03 and 1.06 for C2 and C3, respectively; Wilcoxon signed-rank tests, both $p < 2.2e-16$, Fig. S3). In addition to read coverage differences, we found that misassembled BG regions tend to be much shorter (median lengths = 698 bp for M2/M3 combined) than misassembled HC regions (2328 bp; Fig. S1c; Wilcoxon signed-rank test, $p < 2.2e-16$). This is likely because CNVnator [22] merges adjacent bins based on read depth similarity and in doing so, shorter regions with variable coverage may not be identified. In any case, the read coverage difference is small. Thus, if we relaxed the HC detection threshold, it would significantly increase the false HC calls by calling true BG regions as HC.

An example HC containing region, which is from 500Kb to 590Kb on chromosome 8, had further supported our assumption above (Fig. S4). In this region, there are five tandemly duplicated genes for *terpene synthases* (TPS) and four *cis-prenyl transferase* (CPT)

genes. By comparing the genomic sequences of the short-read and long-read assemblies, and the Polymerase Chain Reaction (PCR)-validated sequence in the tomato variety M82 [31] (Fig. S4b), the two HC regions (Fig. S4a) were identified as to be mis-assembled in the short-read assembly (case 3 and 4 in Fig. S4c). In addition, two BG regions (case 1 and 5, Fig. S4c), where the read coverages were ~ 2 times higher than the background, were also identified as misassembly, supporting the idea that the number of HC regions were underestimated using the current criteria. These results also suggest that for a HC region, if the long-read assembly is not available, one can validate the sequence in the region using the genomic PCR approach, as stated in Matsuba et al., (2013) [31].

Genome features distinguishing correctly and mistakenly assembled regions

To understand why some HC regions were not identified as misassembled, using the same genome features as in Model 35 for classifying HC, BG, and LC regions, a binary classification model (Model 37) was built to distinguish HC regions consisted of mainly M2/M3 (> 50%, referred to as the HC_M2/M3 class) or mainly C2/C3 (> 50%, referred to as the HC_C2/C3 class). Model 37 resulted in an F1 = 0.79 (balanced positive and negative classes, thus the background F1 was 0.5), indicating that mis- and correctly assembled HC regions were significantly distinct from each other in certain genome features. Among the top 20 most important features (Fig. 6a and Table S18, detailed distribution of feature values was shown in Fig. S5), interestingly the most informative ones were those of regions flanking the misassembled regions. The flanking sequences of HC_M2/M3 regions tend to have higher densities of SSR, pseudogenes, TE, and non-tandem genes. At first it was surprising that it was the features in the flanking regions that were informative. In hindsight, if a region was misassembled, the distinguishing signature would likely be buried with it. From the flanking regions, one can better defined whether the sequence in between is problematic, i.e., in our case, misassembled. Within HC_M2/M3 regions, there tend to be higher densities of four types of *k*-mers including TATTTC, TGTAAG, ATACTT, and GATTTT. However, it is not clear why these *k*-mers are informative.

In the above modeling exercise, we were able to distinguish HC regions that were likely misassembled from those that were not. Recall that not only HC regions contain misassembled sequences, in fact BG regions have higher proportion of M2/M3 regions (Fig. 5d). To further dissect their differences and to understand why some misassembled regions were not detected as HC, another model (Model 38) was built to distinguish HC_

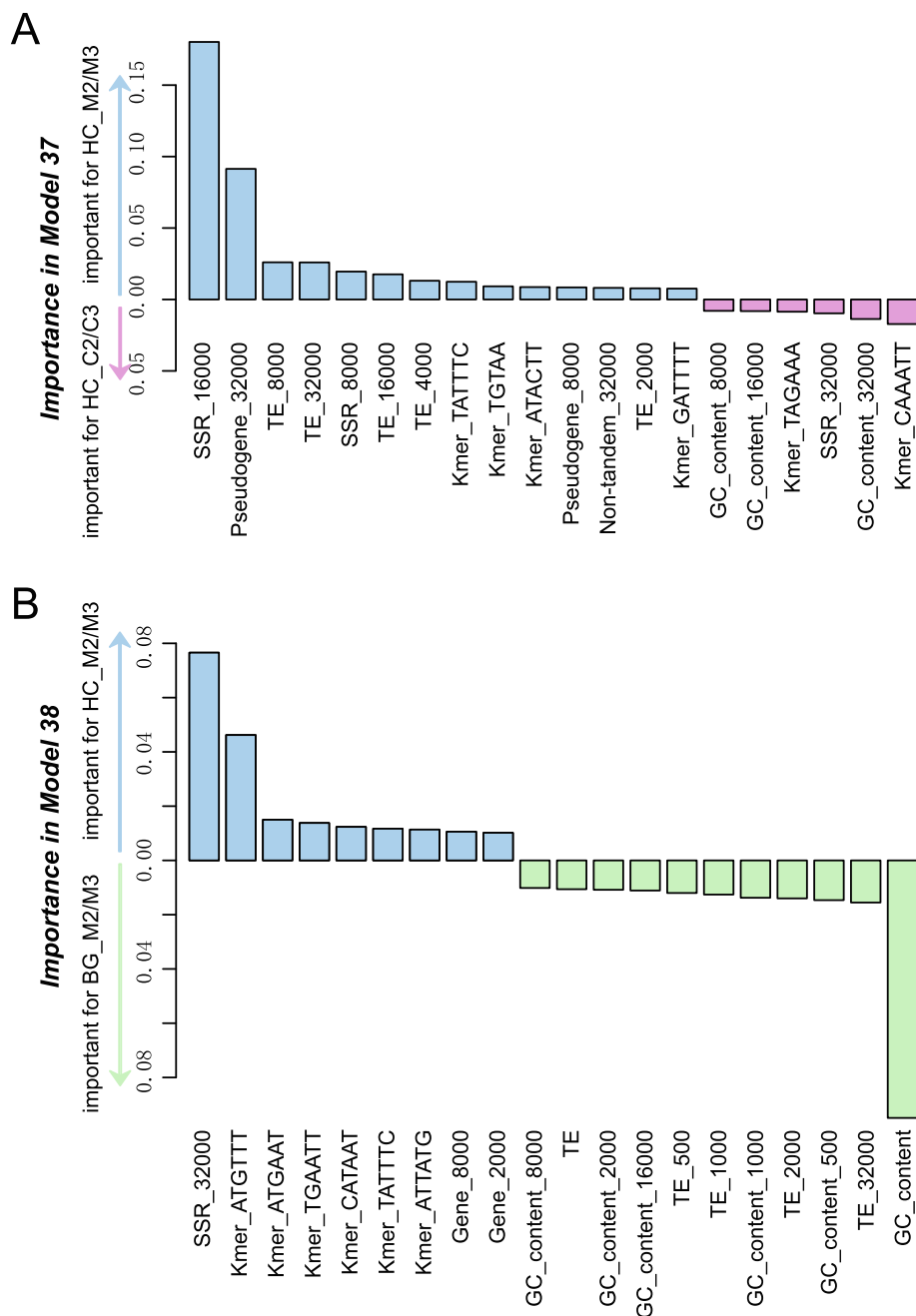


Fig. 6 Important features in Model 37 and Model 38. **a** Model 37 is for classifying HC_M2/M3 and HC_C2/C3 to assess the features important for predicting misassembled HC regions. **b** Model 38 is for classifying HC_M2/M3 and BG_M2/M3 to assess if misassembled HC and BG regions have distinguishable features. Bar height: importance value in the Random Forest model. Blue, red, green: median values of features are higher in HC_M2/M3, HC_C2/C3, and BG_M2/M3, respectively. The distributions of feature values were shown in Fig. S5 and Fig. S6 for Model 37 and Model 38, respectively

M2/M3 and BG_M2/M3 (> 50% of a HC or BG region overlapped with M2/M3), using same features as in Model 35. The resulting F1 was 0.77. Among the top 20 most important features (Table S19), BG_M2/M3 regions tend to have higher GC content (41.3%) and TE density (0.92) compared to HC_M2/M3 regions (GC = 33.6%, TE

density = 0.61, Fig. 6b, feature value distributions shown in Fig. S6). This trend is also true when comparing BG_M2/M3 and HC_M2/M3 flanking regions (Fig. 6b, Fig. S6). Interestingly, the comparatively lower GC content in HC_M2/M3 regions (33.6%) is more similar to the 36.6% overall GC content in the tomato genome. In addition, the GC

content in genic region is at 42.4%, suggesting BG_M2/M3 and HC_M2/M3 may be located in relatively gene-rich and poor regions, respectively. Contrary to this expectation, however, HC_M2/M3 regions tend to have significantly higher gene density (average = 0.16, rank = 142) compared to BG_M2/M3 regions (average = 0.04, Wilcoxon signed-rank test, $p = 1.3e-15$). This is also true when comparing flanking regions (Fig. 6b and Fig. S6). With regard to TE, we have already shown that HC regions tend to have a significantly lower TE density compared to BG regions regardless whether they are misassembled or not (Fig. 3a). Taken together, these predictive models perform well for distinguishing mis- from correctly assembled HC regions and for predicting whether a misassembled region lies in BG or HC regions. Using model interpretation strategies, we are able to identify salient genome features underlying the models' ability to make good quality predictions.

Conclusions

Although the third-generation sequencing such as the PacBio [7] and 10X [8, 32] are now available, the majority of existing genome assemblies are currently derived from short-read based technologies. With the goal of evaluating genome assembly quality by assessing short read coverage distribution, we identified 1156 HC and 15,034 LC regions in tomato genome assembly SL2.50 [25]. These variable coverage regions collectively accounted for ~10% of the genome assembly, indicating the severity of the issue. By applying machine learning methods, we found that HC and LC regions can be predicted with high accuracy. High GC content and TE density are the major factors contributing to the low read coverage or the break point of assembly, while SSRs and tandem duplicates, especially specialized metabolism genes, tend to be in HC regions, potentially leading to misassembly due to high sequence similarities. By comparing Short- and Long-read assemblies, 27.8% of HC regions were potentially misassembled due to duplications. In addition, 91.8% of misassembled HC regions no longer defined as high coverage when we mapped the short reads to the Long-read assembly. Our results highlight the extent to which variable coverage in a Short-read assembly contribute to misassembly, particularly when they are flanked by TEs and tandemly duplicated sequences.

Misassembled regions that are duplicated were detected in both HC and BG regions. It is straightforward to appreciate why misassembled HC region strongly correlated with duplication in the Long-read assembly—higher read coverage is a strong indication that more than one genome region are likely assembled together. However, it is not as obvious why BG regions would be duplicated. There are four explanations. First, HC

regions could be underestimated in our approach. Misassembled BG regions tend to have slightly higher read coverages compared to correctly assembled ones. Second, related to the first explanation, after the partitioning of the genome to HC/LC/BG regions, read depth varies continuously across the genome, and there are no sharp boundaries between HC and BG regions (as opposed to between LC and BG), we established a threshold to define HC and BG regions. As a result, regions with coverage near the defined threshold may be mislabeled. Third, we defined a genomic region into six categories based on whether it is misassembled or not, duplicated or not, and, if duplicated, locally or not. This analysis is based on anchored matches of the Short and Long-read assemblies and thus alignment methods and their parameter choice is expected to impact our findings. Finally, the current tomato Long-read assembly still has scaffolds that cannot be mapped to any chromosomes, which may also contribute to an underestimate of misassembled regions using read coverage. Although there remain areas for further improvement, our results highlight the utility of detecting HC regions in short-read based assemblies for identifying potential misassembled regions. Although not all HC regions have evidence of misassembly based on the Long-read assembly, we showed that, with the machine learning model, misassembled HC regions can be readily distinguished from those that are correctly assembled.

Unlike methods developed for evaluating genome assembly continuity, like LTR Assembly Index [16] and MaGuS [15], here we focused on identifying misassembled regions based on variation in read coverage across the genome, and uncovering the underlying contributors in genome sequences using machine learning. Using tools for identifying regions with significantly high or low read coverages in estimating CNVs among individuals [22] and comparison to Long-read assembly, we discovered potential misassembled genomic regions. Even though the repeats and tandem duplicated genes were known to contribute to genome misassembly with short reads, our study extensively explored the contribution of a large number of genomic and functional annotation features through machine learning. The resulting model provides a comprehensive, quantitative estimate of our current state of understanding of factors contributing to variable genome coverage and assembly issues in short read assemblies. These variable coverage regions account for ~10% of tomato genomes. In addition, HC regions tend to be misassembled. Our approach can be used to assess the extent to which a region of a short-read based plant genome assembly may be misassembled based on read coverage. Considering that the presence of misassembled regions can impact genome-wide studies significantly, their detection prior to genome-wide analysis

should be conducted to reduce the impact of misassembly. Furthermore, the goal of our study is a cursory survey of the potential issues using tomato as an example, it would be meaningful to conduct such analysis thoroughly among other organisms in the future to understand whether our findings are tomato-specific or more general.

Methods

Genome assembly and sequencing reads

The genome sequence assembly SL2.50 of tomato cultivar ‘Heinz 1706’ was downloaded from NCBI (<https://www.ncbi.nlm.nih.gov/>). The genome was assembled mainly using 454 reads, Sanger sequencing reads of two sets of Bacterial Artificial Chromosome (BAC) clone pools, and BAC and Fosmid clone end sequences [25], and was referred to as the Short-read assembly. Additional SOLiD and Illumina reads were used for the base error correction. Among the scaffolds, 91 of 3223 were anchored to 12 chromosomes [33]. To evaluate the extent of misassembly, tomato assembly SL4.0, which was assembled using 80X PacBio sequences (referred to as the Long-read assembly), was also downloaded from Solanaceae Genomics Network (SGN, <https://www.solgenomics.net>).

There are two batches of Illumina genomic sequencing reads available in tomato. The first batch of reads (referred to as dataset1), used for base error correction in genome assembly with a ~28-fold coverage of the genome [25], was sequenced with an Illumina Genome Analyzer IIx (GAIIx) sequencer, and obtained from the SGN in the form of BAM file (version SL2.9). The other read batch (dataset2) was sequenced with an Illumina HiSeq 2000 sequencer, with a ~46-fold coverage of the genome (SRP010718). Reads of these two datasets were remapped to the Short-read assembly and Long-read assembly using Burrows-Wheeler Aligner (BWA-MEM) [34] with default parameters. BWA-MEM was selected for read mapping because it is one of the most accurate and time-efficient tools [35]. To eliminate the impact of bias in PCR amplification on read coverage calling, duplicate reads (identical reads with same mapped location) were marked and removed using Picard (<http://broadinstitute.github.io/picard>). Due to concern of data quality (see **Results**), dataset1 was not analyzed further. Note that mapping and assembly tools both impact the quality of assemblies. The reason we did not explore the impact of these tools is because our goal is to assess variable coverage in assemblies that already exist and used by the communities.

Estimation of read depth and detection of variable coverage regions

Regions with high/low read coverage were identified using CNVnator [22] by determining Read Depth (RD)

for an optimally sized bin of the genome assembly as the number of mapped reads with $\geq 50\%$ of read lengths overlapping with the bin boundaries [22]. The optimal bin size was the bin size leading to a ratio of RD average to RD standard deviation of ~4–5 as suggested [22]. For dataset2, bin sizes from 50 bp to 300 bp were evaluated, and 100 bp was chosen with a ratio = 5.18 (Table S20). With bin size of 100 bp, 20,071 and 1385 regions were identified in Short-read assembly as low-coverage (LC) and high-coverage (HC) regions by CNVnator, respectively. The remaining 20,743 regions were treated as background (BG). The RD values from this CNVnator run for further analysis is referred to as “analysis RD” values.

To assess the sensitivity and accuracy of HC/LC region detection, reads were resampled with three strategies to generate simulated RD values that were used to run CNVnator. For each strategy, the resultant RD for each 100 bp bin was compared to the simulated RD values (ground truth). In the first strategy, reads were resampled for HC, BG and LC regions, based on idealized RD values (input RDs for HC, BG and LC regions were assigned as 2, 1 and 0, respectively, no decimal point values) (Fig. S7a,b). In the second strategy, reads were resampled for HC, BG and LC regions based on the rounded analysis RD values (e.g., for regions with analysis RD values of 0.88 and 2.31, 1x and 2x reads were resampled for these two regions, respectively) (Fig. S7c). In the third strategy, reads were resampled for HC, BG and LC regions based on the analysis RD values (Fig. S7d). For all three strategies, we observed very high correlation between simulated, ground truth RD and RD values resulted from simulated reads ($PCC \geq 0.97$), indicating that detection of HC/LC/BG regions using CNVnator is robust.

Impact of genome coverages on RD values

As shown in Fig. 1a, subset of dataset2 at ~30-fold coverage had very similar RD distribution as dataset2. To assess the extent to which read coverage impacts the detection of HC/LC/BG regions, we randomly resampled reads at 5-fold, 10-fold, 20-fold coverages from dataset2. The resultant RDs of all subsets of dataset2 (5 ~ 30-fold coverage) were compared to analysis RDs using all reads in dataset2 (~46-fold coverage). The correlation decreased as the read coverage decreased (PCC from 0.99 to 0.89; Fig. S7e-h). Genome sequencing reads at ≥ 20 -fold coverage may provide very similar information for HC/LC/BG region detection ($PCC = 0.98$, $\rho \geq 0.92$), while reads at 10- or 5-fold coverage likely have some error rates in HC/LC/BG region detection ($PCC \leq 0.95$, $\rho \leq 0.86$). These results also suggest that the RD variation in detected HC/LC/BG regions reflect potential assembly issues or sequencing bias, instead of being noise

introduced by random sampling of reads. To test this, a fake dataset, where the reads were randomly sampled from tomato genome to ~46-fold coverage, was used, and as expected, there is no LC or HC region detected.

In addition, to address the impact of artifacts produced by read mapping using BWA-MEM on RD values, the original reads (~46-fold) were re-mapped to the Short-read assembly and the resultant RD values were almost the same as original values (both PCC and $\rho = 1.0$; Fig. S7i), suggesting that the impact, if any, is negligible.

Choice of q-value threshold in HC/LC/BG region detection

To get HC/LC/BG regions with high confidence, regions with q_0 (proportion of reads with multiple matches across genome in a region) > 0.5 [22] were filtered out because they likely represented repetitive sequences. F measure (F1) values were used to measure the consistency between true HC/LC/BG region designations and new HC/LC/BG regions determined using resampled reads. F1 were calculated as: $F\text{-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$, where $\text{precision} = \frac{TP}{TP+FP}$; $\text{recall} = \frac{TP}{TP+FN}$; TP = true positive, FP = false positive, FN = false negative (Fig. S8). *P*-values of identified HC and LC regions were adjusted to account for multiple testing [36]. To choose an adjusted *p*-value (*q*-value) to maximize F1 scores, HC and LC regions identified using reads resampled by three strategies and reads at 30-fold coverage as Fig. S7 were compared to HC and LC regions detected using dataset2. F1 score varies when different *q*-values were used as thresholds to call HC/LC/BG regions (Fig. S8). In Fig. S8a-c and e-g, F1 fitted curve arrives a platform after *q*-value > 0.06 , whereas in Fig. S8d and h, the break point of *q*-value is 0.08. Therefore, only regions with *q*-value < 0.08 were retained, resulting in 1227 HC and 15,095 LC regions.

Genome and functional annotations, definitions of genome features, and gene set enrichment analysis

Tomato gene annotation version SL2.50 was downloaded from NCBI (<https://www.ncbi.nlm.nih.gov/>). Aside from gene annotations, we defined or obtained additional genome features including pseudogenes, transposable elements (TE), simple sequence repeat (SSR; stretch of DNA, 2~64 bp, repeated > 1 time and the repetitions are immediately adjacent to each other), and tandemly duplicated genes. Pseudogenes were defined as genomic regions with significant similarity to protein-coding genes had premature stops/frameshifts and/or were truncated as described in [37]. Transposable element (TE) annotation was based on SGN ITAG2.4 release. SSRs were detected using Tandem Repeats Finder with recommended parameters with Match = 2, Mismatch = 7, Delta = 7, PM = 80, PI = 10. Minscore = 50, Maxperiod = 500 [38]. Tandemly duplicated genes were

identified using MCSanX-transposed [39], as described previously [37], where paralogs are directly adjacent to each other, or separated by ≤ 10 nonhomologous genes.

Three types of functional annotation data were used including Gene Ontology (GO) terms, Metabolic pathway annotation, and Pfam domain annotation. GO terms were inferred using blast2go [40], where protein sequences were searched against NCBI nr protein database using BLASTP [41] with an E-value cut-off of $1e-5$. Tomato metabolic pathway annotation V3.0 was downloaded from Plant Metabolic Network (<https://pmn.plantcyc.org/>). Genes in specialized metabolic pathways were annotated as specialized metabolism (SM) genes. Pfam domains in tomato annotated protein sequences were identified by searching against Pfam Hidden Markov Models (<https://pfam.xfam.org>, v.29.0) using HMME R3 (<http://hmmer.org>) with the trusted cutoff.

Gene set enrichment analysis was performed using Fisher's exact test and Likelihood Ratio test, *p*-values were adjusted for multiple testing [36]. Log likelihood ratio was calculated as: $\log_{10}(\frac{a/b}{c/d})$, where *a*, *b*, *c* and *d* are the numbers in a 2×2 contingency table as in Fisher's exact test. For example, for testing whether genes with the GO term *G* tend to be in HC region: *a* - the number of genes with GO term *G* in HC regions, *b* - the number of genes that don't have GO term *G* but are in HC regions, *c* - the number of genes with GO term *G* but not in HC regions, and *d* - the number of genes that don't have GO term *G* and are not in HC regions.

Multi-class machine learning models for predicting whether a genomic region has high, low, or background coverage

To classify a genomic region into one of the following three classes: HC, LC, and BG, three-class models were established using Random Forest [26], implemented in the python package Scikit-learn [42]. ~7000 properties of genomic sequences within HC, LC and BG regions, and in their corresponding flanking regions (in bins of 0.5, 1, 2, 4, 8, 16, and 32 Kb, including both upstream and downstream) were used as features for building machine learning models. There were three types of features. The first was GC content. The second includes densities of: (1) all genes, (2) tandemly duplicated genes, (3) non-tandem genes, (4) pseudogenes, (5) transposable elements, (6) all SSRs (without considering each, specific SSR sequence), (7) specific SSR (2-64 bp repeats, e.g., 2 bp repeat: ATATATATATAT), and (8) specific *k*-mer (1-6 bp, e.g., 5-mer, GGCGG). Density was calculated as the proportion of each HC/LC/BG region occupied by the feature in question. The last type was presence/absence of genes with a particular functional annotation in a given region. These functional annotations

include: GO term, Pfam domain, and metabolic pathway annotations.

In addition to presence/absence, numbers of annotated entries were also used as features to build models, which didn't differ significantly in performance from models built with presence/absence features and were not discussed further. Kruskal-Wallis H test was done to determine if there are statistically significant differences of each density among HC/LC/BG regions using SciPy [43]. Feature selection was conducted using the RandomForestClassifier function in Scikit-learn [42], and potentially informative features were selected based on their importance determined by the entropy criterion which measures the quality of tree split according to the information gain when each feature was used. For each class, 10% of the regions were held out from the model training/validation process to serve as independent test data. The rest 90% were used as training/validation data for model training.

To avoid the potential data leakage in the model training (i.e., accidental sharing of information between the data for training the model and the data for evaluating/testing the model), we assessed whether the distances between regions of test and training sets would impact the model performance by creating two training/test sets. The first is by randomly selecting 10% of regions from the whole set, thus some regions in test set can be close to regions in training set. The second is by randomly selecting two genomic spans from each of the 12 chromosomes. We then selected test HC/LC/BG regions from within the genomic spans (24 total). In the meantime, the training HC/LC/BG regions were selected outside of the 24 spans. Thus, regions in test set were close to each other, but distant from regions in the training set. In the second data set, the distances of HC regions between training and test set had a median of 1081.2 kb (interquartile range of 73.2 ~ 2686.8 kb), LC regions with a median of 1176.9 kb (interquartile range of 528.1 ~ 2080.4 kb), and BG regions with a median of 1239.6 kb (interquartile range of 486.0 ~ 2083.1 kb). For Model 34 trained and evaluated with the first dataset, the model performance was $F1_{CV} = 0.87$, $F1_{test} = 0.79$. In the model trained and evaluated with the second dataset that maximized distance between training and test sets, the performance was nearly identical ($F1_{CV} = 0.87$, $F1_{test} = 0.80$). Thus, it is unlikely that test data is contaminated by adjacent training sequences. Therefore, only models based on the first dataset was reported throughout.

For model training, we used equal numbers of instances from each class (HC, LC, or BG) to create balanced datasets that facilitate interpretation of model performance. Because HC regions were in the minority (1156), LC and BG regions were randomly sampled till

they were the same numbers as HC regions. In total, 100 random balanced datasets were generated. Using each of the first 10 balanced dataset, the grid search approach as implemented in the GridSearchCV function in Scikit-Learn was used to determine the best combination of parameters. In this approach, each balanced dataset was split into training (90%) and validation (10%) subsets following the 10-fold cross-validation scheme. The best set of parameters for Random Forest (*max_depth*, *max_feature*, and *n_estimators*) were identified according to the mean of F1-macro (average F1 score for three classes) across the 10 GridSearchCV runs. Then each of the 100 balanced datasets was used to establish a three class (HC/LC/BG) prediction model using the best parameter set with 10-fold cross validation to assess the robustness of classification results. The average true positive rate and the average F1 score across 100 runs were used to evaluate model performance.

Identification of potentially misassembled regions

The Short-read assembly (query) was aligned to the Long-read assembly (subject) with the NUCmer algorithms using default settings (`--mumreference --delta --breaklen 200 --mincluster 65 --diagfactor 0.12 --max-gap 90 --minmatch 20`), as implemented in MUMmer 4.0.0beta2 [30]. Chromosome coordinates and sequence similarity of aligned regions were produced by mummer-plot utility. Only aligned regions with identities $\geq 95\%$ were used for the downstream analysis which may lead to false negatives.

Before identifying if a Short-read assembly region is misassembled, we first ask, in the MUMmer alignment, if a query region of the Short-read assembly was duplicated in the Long-read assembly or not. A query region was classified as having duplicated subjects if it had ≥ 2 aligned subject regions in the Long-read assembly, regardless of these regions are on the same chromosome or not. Otherwise, it is regarded as non-duplicated. A query region without duplicated subject was defined as misassembled if the subject region it aligned to was in a different location on the same or on different chromosome compared to the location of query region. A query region with duplicated subject regions was defined as misassembled if it was: 1) ≥ 100 bp, and 2) the lengths of overlaps between duplicated subject regions $< 50\%$ of the query, and 3) the subject regions aligned to only one Short-read assembly region. We considered misassembled regions that had one copy in Short-read assembly while ≥ 2 in Long-read assembly. Thus, misassembled regions with ≥ 2 copies in Short-read assembly and even more copies in Long-read assembly were not analyzed because they were minor cases and the challenges in defining additional unifying categories for these cases.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-021-07397-5>.

Additional file 1: Figure S1. Properties of HC/LC/BG regions with high confidence. **a** Proportion of Ns in HC/LC/BG regions. **b** Length distribution of HC/LC/BG regions. **c** Length distribution of overlapped regions between HC/LC/BG and M2/M3 regions.

Additional file 2: Figure S2. Genomic feature distributions in flanking regions of HC/LC/BG regions of different length (0.5 ~ 32Kb). Violin plots showing distributions of GC content, and densities of genes, tandemly duplicated genes, pseudogenes, transposable element and SSRs in HC, BG, and LC regions. Red line indicates median value.

Additional file 3: Figure S3. RD distribution of 100 bp BG bins overlapped with aligned region categories. For each plot, left Y-axis: number of BG bins in correctly assembled category (C1, C2 or C3); right Y-axis: number of BG bins in mis-assembled category (M1, M2 or M3). The categories are defined in Fig. 5a.

Additional file 4: Figure S4. Example mis-assembled HC regions. **a** Normalized RD in a region containing an SM gene cluster on Chromosome 8. Normalized RD was calculated for each 100 bp bin. Two black brackets indicate identified HC regions, with RD around 2. Valley regions with RD = 0 indicate gaps in genome assembly filled up with Ns. **b** Syntenic alignments (colored regions between black lines) and gene annotation (colored boxes) of the same region in **(a)** between short-read and long-read assemblies of Heinz and the PCR-validated sequence in M82. Corresponding positions in **(a)** and **(b)** were delineated with grey dashed lines. Colored regions: BLASTn matches between Heinz and M82, with E-value < 1e-20 and alignment length > 50 bp. AOX: alcohol oxidase; TPS: terpene synthases; P450: cytochrome P450; CPT: cis-prenyl transferase; ψ : pseudogene. **c** Dotplot of the region in **(a)** between Short- and Long-read assemblies. Five cases were indicated using arrow heads, and all these five short regions were identified to be in BG regions in long-read based assembly ($\rightarrow$ BG).

Additional file 5: Figure S5. Important features in model distinguishing HC_M2/M3 and HC_C2/C3. Violin plots show distributions of each features or GC contents within regions or in flanking regions. Lines within violin plots indicate median values.

Additional file 6: Figure S6. Distributions of important features in model distinguishing HC_M2/M3 and BG_M2/M3. Violin plots show distributions of each features or GC contents within regions or in flanking regions. Lines within violin plots indicate median values.

Additional file 7: Figure S7. Sensitivity and accuracy of CNVnator in RD value calculation, and impact of genome coverages on RD values. **a** RD distribution of new CNVnator runs by mapping re-sampled reads from the tomato genome based on simulated input RD, where the only possible RD values were 0 (LC), 1 (BG), or 2 (HC). **b-d** Correlation between known, simulated input RDs and new RD values from new CNVnator run using the resampled reads. In **(b)**, the simulated RD values were generated as in **(a)**. In **(c)**, the analysis RD values (those generated with CNVnator by mapping dataset2 reads on to the tomato genome, see [Methods](#)) were first discretized (rounded) to their closest integers, then the rounded RD values were used for resampling reads for determining new RD values. In **(d)**, the analysis RD values were directly used for resampling reads for determining new RD values. **e-i** Correlation between RD values using all reads (46X coverage) and RD values using subsets of reads at variable coverage: **(e)** 30X, **(f)** 20X, **(g)** 10X, **(h)** 5X, **(i)** 46X.

Additional file 8: Figure S8. F1 of HC/LC/BG region calling. **a-d** CNVnator runs were conducted using resampled reads based on different starting RD values as in Fig. S7 using read dataset 2. As in Fig. S7, the simulated RD values include: **(a)** possible RD values of only 0 (LC), 1 (BG), or 2 (HC); **(b)** analysis RD values discretized/rounded to integers; **(c)** analysis RD without discretization; **(d)** analysis RD without discretization but down-sampled to 30X (note the y-axis range is much smaller compared to **(a-c)**). Each dot indicates an F1 value (y-axis) at a given q-value threshold (x-axis), where F1 was calculated using: (1) numbers of nucleotides in overlapping regions between the HC/LC region designations based on the analysis RD and new HC/LC regions determined using

resampled reads (True Positive), (2) numbers of nucleotides in true HC/LC regions but determined as BG regions in new run (False Negative), and (3) numbers of nucleotides in true BG regions but determined as HC/LC regions in new run (False Positive, see [Methods](#)). Orange line: LOESS fitted curve. **e-h** Same as **(a-d)** except that the F1 was determined based on numbers of regions as opposed to numbers of nucleotides.

Additional file 9: Table S1. Performance of 3-class or binary prediction models. **Table S2.** Importance values and Kruskal test of features in Model 34. **Table S3.** Correlation among densities of HC/LC/BG regions and genomic features. **Table S4.** Importance values of features in Model 35. **Table S5.** Importance values of features in Model 34B. **Table S6.** Importance values of features in Model 35B. **Table S7.** Number of annotated sequences in HC regions and 10,000 randomly sampled BG regions (based on the number and length distribution of HC regions). **Table S8.** Gene set enrichment analysis for Pfam domains for HC regions vs. whole genome. **Table S9.** Gene set enrichment analysis for GO terms for HC regions vs. whole genome. **Table S10.** Gene set enrichment analysis for metabolic pathways for HC regions vs. whole genome. **Table S11.** Gene set enrichment analysis for Pfam domains for BG regions vs. whole genome. **Table S12.** Gene set enrichment analysis for GO terms for HC regions vs. BG regions. **Table S13.** Gene set enrichment analysis for metabolic pathways for HC regions vs. BG regions. **Table S14.** Aligned regions between Short-read assembly and Long-read assembly. **Table S15.** Categories of regions in Short-read assembly. **Table S16.** Corresponding variable coverage regions between two assemblies. **Table S17.** HC regions in Short-read assembly and corresponding variable coverage regions in Long-read assembly. **Table S18.** Importance values of features in Model 37. **Table S19.** Importance values of features in Model 38. **Table S20.** Choice of bin size.

Abbreviations

PacBio: Pacific Biosciences; NCBI: National Center of Biotechnology Information; CNV: Copy number variation; HC: Regions with significantly higher read coverage; LC: Regions with significantly lower read coverage; BG: Regions with genome-wide average read coverage; RD: Read depth; TE: Transposable elements; SSRs: Simple sequence repeats; SM: Specialized metabolism; ncRNAs: Non-coding RNAs; LLRs: Log Likelihood Ratio; GO: Gene Ontology; C1: Regions correctly assembled; M1: Non-duplicated, misassembled regions; C2: Locally duplicated and correctly assembled regions; M2: Locally duplicated but misassembled regions; C3: Non-locally duplicated, correctly assembled regions; M3: Non-locally duplicated but misassembled regions

Acknowledgements

We thank Melissa Lehti-Shiu, Christina Azodi and Siobhan Cusack for helpful discussions.

Authors' contributions

P.W. and S.-H.S. conceived the original research plans; F.M. and P.W. conducted the bulk of the computational analysis; B.M. assisted with the machine learning models; P.W. and S.-H.S. wrote the manuscript with contributions of all the authors; S.-H.S. agrees to serve as the author responsible for contact and ensures communication. All authors read and approved the final manuscript.

Funding

This work was partly supported by the National Science Foundation [IOS-1546617, DEB-1655386 to S.-H.S.]; and U.S. Department of Energy Great Lakes Bioenergy Research Center [BER DE-SC0018409 to S.-H.S.]. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and its additional files. All the scripts used in this study are available on Github at: <https://github.com/ShiuLab/Evaluating-misassemblies>

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA. ²DOE Great Lake Bioenergy Research Center, Michigan State University, East Lansing, MI 48824, USA. ³The Ecology, Evolution, and Behavioral Biology Program, Michigan State University, East Lansing, MI 48824, USA. ⁴Department of Computational Mathematics, Science, and Engineering, Michigan State University, East Lansing, MI 48824, USA.

Received: 27 May 2020 Accepted: 19 January 2021

Published online: 02 February 2021

References

- Heather JM, Chain B. The sequence of sequencers: the history of sequencing DNA. *Genomics*. 2016;107(1):1–8.
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860–921.
- Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*. 2011;13(1):36–46.
- Chen YC, Liu TL, Yu CH, Chiang TY, Hwang CC. Effects of GC bias in next-generation sequencing data on De novo genome assembly. *PLoS One*. 2013;8(4):e62856.
- Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol*. 2011;12(2):R18.
- Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G, Wright J, et al. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Res*. 2017;27(5):885–96.
- Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. *Science*. 2009;323(5910):133–8.
- Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol*. 2016;17(1):239.
- Watson M, Warr A. Errors in long-read assemblies can critically affect protein prediction. *Nat Biotechnol*. 2019;37:124–6.
- Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet*. 2018;19(6):329–46.
- Bertioli DJ, Jenkins J, Clevenger J, Dudchenko O, Gao DY, Seijo G, et al. The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. *Nat Genet*. 2019;51(5):877–84.
- Edger PP, Poorten TJ, VanBuren R, Hardigan MA, Colle M, McKain MR, et al. Origin and evolution of the octoploid strawberry genome. *Nat Genet*. 2019;51(4):765.
- Zhang L, Chen F, Zhang X, Li Z, Zhao Y, Lohaus R, et al. The water lily genome and the early evolution of flowering plants. *Nature*. 2020;577(7788):79–84.
- Zhuang WJ, Chen H, Yang M, Wang JP, Pandey MK, Zhang C, et al. The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nat Genet*. 2019;51(5):865–76.
- Madoui MA, Dossat C, d'Agata L, van Oeveren J, van der Vossen E, Aury JM. MaGuS: a tool for quality assessment and scaffolding of genome assemblies with whole genome profiling (TM) data. *BMC Bioinformatics*. 2016;17:115.
- Ou SJ, Chen JF, Jiang N. Assessing genome assembly quality using the LTR assembly index (LAI). *Nucleic Acids Res*. 2018;46(21):e126.
- Yang LA, Chang YJ, Chen SH, Lin CY, Ho JM. SQUAT: a sequencing quality assessment tool for data quality assessments of genome assemblies. *BMC Genomics*. 2019;19(Suppl 9):238.
- Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods*. 2009;6(4):291–5.
- Vukasinovic N, Cvrckova F, Elias M, Cole R, Fowler JE, Zarsky V, et al. Dissecting a hidden gene duplication: the *Arabidopsis thaliana* SEC10 locus. *PLoS One*. 2014;9(4):e94077.
- Chae L, Kim T, Nilo-Poyanco R, Rhee SY. Genomic signatures of specialized metabolism in plants. *Science*. 2014;344(6183):510–3.
- Moore BM, Wang PP, Fan PX, Leong B, Schenck CA, Lloyd JP, et al. Robust predictions of specialized metabolism genes through machine learning. *Proc Natl Acad Sci U S A*. 2019;116(6):2344–53.
- Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*. 2011;21(6):974–84.
- Miller CA, Hampton O, Coarfa C, Milosavljevic A. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One*. 2011;6(1):e16327.
- Li S, Dou X, Gao R, Ge X, Qian M, Wan L. A remark on copy number variation detection methods. *PLoS One*. 2018;13(4):e0196226.
- Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*. 2012;485(7400):635–41.
- Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
- Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*. 2008;36(16):e105.
- Wang W, Zheng HK, Fan CZ, Li J, Shi JJ, Cai ZQ, Zhang GJ, Liu DY, Zhang JG, Vang S et al. High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell*. 2006;18:1791–802.
- Zou C, Lehti-Shiu MD, Thibaud-Nissen F, Prakash T, Buell CR, Shiu SH. Evolutionary and expression signatures of pseudogenes in *Arabidopsis* and rice. *Plant Physiology*. 2009;151:3–15.
- Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. *Genome Biol*. 2004;5(2):R12.
- Matsuba Y, Nguyen TTH, Wiegert K, Falara V, Gonzales-Vigil E, Leong B, et al. Evolution of a complex locus for terpene biosynthesis in *Solanum*. *Plant Cell*. 2013;25(6):2022–36.
- Leggett RM, Clark MD. A world of opportunities with nanopore sequencing. *J Exp Bot*. 2017;68(20):5419–29.
- Ezura H, Arizumi T, Garcia-Mas J, Rose J. Functional genomics and biotechnology. In: *Solanaceae and cucurbitaceae crops*. Vol. 70. Berlin, Heidelberg: Springer; 2016.
- Li H, Durbin R. Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics*. 2010;26(5):589–95.
- Otto C, Stadler PF, Hoffmann S. Lacking alignments? The next-generation sequencing mapper segemehl revisited. *Bioinformatics*. 2014;30(13):1837–43.
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B*. 1995;57:289–300.
- Wang PP, Moore BM, Panchy NL, Meng FR, Lehti-Shiu MD, Shiu SH. Factors influencing gene family size variation among related species in a plant family, Solanaceae. *Genome Biol Evol*. 2018;10(10):2596–613.
- Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 1999;27(2):573–80.
- Wang Y, Li J, Paterson AH. MScanX-transposed: detecting transposed gene duplications based on multiple colinearity scans. *Bioinformatics*. 2013;29(11):1458–60.
- Conesa A, Götz S. Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int J Plant Genom*. 2008;2008:619832.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825–30.
- Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. Author Correction: SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*. 2020;17(3):352.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.