Factors influencing read coverage variation in a genome assembled using short reads

Peipei Wang[1,2], Fanrui Meng[1,2], Bethany M. Moore[1,3], and Shin-Han Shiu[1,2,3,4*]

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

*Corresponding Author:
Shin-Han Shiu
Michigan State University
Plant Biology Laboratories
612 Wilson Road, Room 166
East Lansing, MI 48824-1312
517-353-7196
shius@msu.edu

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 the scripts used in this study are available on Github at:
https://github.com/peipeiwang6/Evaluating-misassembly/

Abstract

Availability of genome sequences has led to significant progress in biological sciences and beyond. With few exceptions, the great majority of existing genome assemblies are derived from short read sequencing technologies. While extremely useful, the short-read coverages across the assemblies are highly uneven, indicative of sequencing and assembly issues. To assess the underlying causes of such uneven read coverage, we used the tomato genome as an example and integrated multiple sequence features to establish machine learning models capable of predicting whether a genomic region has significantly high or low read coverage. Importantly, 0.6% (5.1Mb) and 9.7% (79.6Mb) of tomato genome assembly had significantly higher and lower coverage compared to background, respectively. By evaluating features important for the prediction, we found that GC content and high density of transposon elements are the major contributors to break points in an assembly, leading to gaps filled with Ns and the resulting low read coverages. In contrast, simple sequence repeats and tandemly duplicated genes, especially specialized metabolism genes, tend to be mis-assembled, resulting in high read coverages. We also present evidence of a misassembled regions containing tandemly duplicated specialized metabolism genes. The presence of variable coverage regions is expected to significantly impact genome-wide studies, highlighting the need to detect them in short-read based assemblies.

Keywords: genome assembly, read coverage, GC content, machine learning, specialized metabolism genes, tomato

Introduction

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 (Heather and Chain, 2016), based on which a decade was taken to deliver a draft genome of human (Consortium and International Human Genome Sequencing Consortium, 2001). The second-generation technology, i.e. “Next generation sequencing” (NGS), where thousands to millions of DNA molecules are enabled to be sequenced simultaneously dramatically shortens the time required to obtain high genome coverage (Heather and Chain, 2016). However, due to the short length of NGS reads (36bp~400bp), there are many challenges for assembling genome based on NGS reads, including the difficulty in sequencing repetitive sequences (Treangen and Salzberg, 2011), low read coverages in GC-poor or GC-rich regions (Chen et al., 2013), genome sequencing bias introduced by PCR amplification during library construction (Aird et al., 2011), and polyploidy in some species including most flowering plants (Clavijo et al., 2017). The advent of third generation sequencing, e.g., Pacific Biosciences (PacBio) single molecule real time sequencing and Oxford Nanopore sequencing, has led to another revolution in genome sequencing, where long reads up to 100 Kb can be sequenced in a single run without PCR amplification or chemical labeling of the sample. Although the much higher error rates remain an issue (Rhoads and Au, 2015; Watson and Warr, 2019), the third generation sequencing still has merit for applicants more tolerant to error rates, like structural variant calling (Sedlazeck et al., 2018) 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 (Bertioli et al., 2019; Edger et al., 2019; Zhuang et al., 2019a), 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 (Madoui et al., 2016), LTR Assembly Index (Ou et al., 2018), SQUAT (Yang et al., 2019). 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 (Kozarewa et al., 2009). Second, repetitive sequences are abundant in species with larger genomes, and have always been a major challenge for genome assemblies

(Treangen and Salzberg, 2011). 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 (Vukašinović et al., 2014).

Here, we use tomato (*Solanum lycopersicum*) as a model to assess the extent to which its assembly has variable coverage. Tomato is chosen because its assembly was built mostly with short reads. 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 (Chae et al., 2014; Moore et al., 2019) 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 (Abyzov et al., 2011; Miller et al., 2011; Li et al., 2018). Our primary goal is to use the tools developed for CNV detection for identifying genome regions with significantly high- and low- coverage of sequencing reads compared to the genome average (i.e. background), and to explore underlying factors influencing the read coverages. Most importantly, through considering a wide range of factors that can potentially influence assembly quality, machine learning models were built to predict whether a region has high, low, or background read coverage using genomic characteristics as input.

Materials and 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 (Tomato Genome Consortium, 2012). Additional SOLiD and Illumina reads were used for the base error correction. Among the scaffolds, 91 of 3223 were anchored to 12 chromosomes (Ezura et al., 2016). 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 (Tomato Genome Consortium, 2012), was sequenced with an Illumina Genome Analyzer IIx (GAIIx) sequencer, and obtained from the Solanaceae Genomics Network

(SGN, https://www.solgenomics.net) 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 genome assembly using Burrows-Wheeler Aligner (BWA-MEM) (Li and Durbin, 2010) with default parameters. Due to concern of data quality (see **Results**), dataset1 was not analyzed further.

Estimation of read depth and detection of variable coverage regions

Regions with high/low read coverage were identified using CNVnator by determining Read Depth (RD) for an optimally sized bin of the genome assembly as the number of mapped reads with ≥50% read lengths overlapping with the bin boundaries (Abyzov et al., 2011). The optimal bin size was the bin size leading to a ratio of RD average to RD standard deviation of ~4-5 as suggested (Abyzov et al., 2011). For dataset2, bin sizes from 50bp to 300bp were evaluated, and 100bp was chosen with a ratio = 5.18 (**Supplementary Table S1**). With bin size of 100bp, 20,071 and 1,385 regions were identified as low-coverage (LC) and high-coverage (HC) regions by CNVnator, respectively. The remained 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 (for details, see **Supplemental Text**) that were used to run CNVnator. For each strategy, the resultant RD for each 100bp 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). 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). In the third strategy, reads were resampled for HC, BG and LC regions based on the analysis RD values. For all three strategies, we observed very high correlation between simulated, ground truth RD and RD values resulted from simulated reads, indicating that detection of HC/LC/BG regions using CNVnator is robust.

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 ~ 64bp, 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 (Wang et al., 2018). 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 (Benson,

1999). Tandemly duplicated genes were identified using MCScanX-transposed (Wang et al., 2013), as described previously (Wang et al., 2018), 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 (Conesa and Götz, 2008), where protein sequences were searched against NCBI nr protein database using BLASTP (Camacho et al., 2009) 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 HMMER3 (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 (Benjamini and Hochberg, 1995). Log likelihood ratio was calculated as: $log_{10}\left(\frac{a/b}{c/d}\right)$, 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 (Breiman, 2001), implemented in the python package Scikit-learn (Pedregosa et al., 2011). ~7,000 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 consider each, specific SSR sequence), (7) specific SSR (2-64bp repeats, e.g. 2bp repeat: ATATATATATAT), and (8) specific *k*-mer (1-6bp, 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 (Jones E et al., 2001). Feature selection was conducted using the RandomForestClassifier function in Scikit-learn

(Pedregosa et al., 2011), 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.

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 (1,156), 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.

Results & Discussion

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

Two datasets were used to determined 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 (Tomato Genome Consortium, 2012), 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 100bp 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 (Abyzov et al., 2011), 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 ρ = 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 1,156 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. 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 mostly 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 S2**) using the Random Forest algorithm (Breiman 2001). We should emphasize that an independent, test set 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 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 contributed 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=0.46~0.58, **Fig. 2A** and **Table S2**). 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 (1-6bp, 5,460 features) may be prevalent in HC, LC, or BG regions. Because the number of these SRR 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**). 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 S3** lists the importance values of all 256 features. We found that three types of features stand out: *k*-mers (median importance

rank=57), GC contents (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 (Treangen and Salzberg, 2011). 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 1,000 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.

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 (Dohm et al., 2008; Chen et al., 2013). 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 ≥ 60%, **Table S3**). 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 S2**).

Pseudogenes (median rank=131), all protein coding genes (139), non-tandem duplicates (141), and tandem duplicates (154) also ranked in the middle (**Table S3**). Densities of these genomic features in flanking regions ranked similarly or even higher than densities in HC/LC/BG regions (**Table S3**), 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 S4**). 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 S4**). 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 S2**). 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 (**TableS2**), the comparable binary Model 35B had the same performance as Model 34B (**Fig.2A, Table S2**). However, models with only functional features had accuracy=58% and F1=0.69, which is much better than random guess (**TableS2**), 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, **TableS3,5,6,7**). 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 S2**) 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 regions 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**). 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 S2**).

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 S8**) 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 (Chae et al., 2014; Moore et al., 2019). 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 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 duplicate 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 biological process category, or TomatoCyc pathways. Given the number of domain families (**Table S9**), categories (**Table S10**) and pathways (**Table S11**) 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**).

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 (Wang et al. 2006; Zou et al. 2009) 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-Smirov 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. To further assess the relationships between SM genes and HC regions, we compared a PCR-validated region on chromosome 8 in the tomato variety M82 (Matsuba et al., 2013) with the corresponding region in the reference genome assembly from the variety Heinz. In the region from 500Kb to 590Kb, there are five tandemly duplicated genes for terpene synthases (TPS) and four cis-prenyl transferase (CPT) genes (**Fig. 5**). In the Heinz genome assembly, ψCPT8 and TPS19 were absent. In the meantime, the regions corresponding to upstream of ψCPT8 and downstream of TPS19 in Heinz were detected as HC regions, with RD=~2 (**Fig. 5**). These observations indicate that ψCPT8 and TPS19 were assembled together with CPT1 and TPS20 and then subsequently mis-annotated as spanning a ~40kb regions in the Heinz genome.

Conclusion

Although the third generation sequencing such as the PacBio (Rhoads and Au, 2015) and 10X (Mostovoy et al., 2016) are now available, the majority of existing genome

assemblies were derived from short-read based technologies. With the goal of evaluating genome assembly quality by assessing short read coverage distribution, we identified 1,156 HC and 15,034 LC regions in tomato genome assembly SL2.50 (Tomato Genome Consortium, 2012). 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 mis-assembly due to high sequence similarities. Our results highlight the extent of variable coverage in a short-read based genome assembly which contribute to underestimation of gene contents, particularly when they are flanked by TEs and sequences tandemly duplicated.

Our models performed well in predicting HC and LC regions but did so in a far from perfect fashion. Because the 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 mis-labeled and contribute to mispredictions. Furthermore, although the types and number of features we examined are extensive, we cannot rule out the possibility that additional features distinguishing HC and BG regions are missed. In this case, new approaches that do not require pre-defined features, such as deep learning, can potentially lead to better results. This is because the genome sequences can be directly used as features, such as in convolutional neural network (Zhuang et al., 2019b). In this study, we used Random Forest as the core algorithm for machine learning because it is among the most well performed algorithms. However, additional model selection by employing additional machine learning algorithms may further improve the model performance. In addition, only one region with PCR sequences is available in tomato to validate our findings, it will be intriguing to use longer reads from technologies such as PacBio to further evaluate the variable coverage regions.

Unlike methods developed for evaluating genome assembly continuity, like LTR Assembly Index (Ou et al., 2018) and MaGuS (Madoui et al., 2016), 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 (Abyzov et al., 2011; Miller et al., 2011; Li et al., 2018), we discovered potential misassembled genomic regions. Through machine learning, our study combined a large number of genomic and functional annotation features. The resulting model provides an extensive, quantitative estimate of our current state of understanding of factors contributing to variable genome coverage in assemblies based on short reads. These variable coverage regions account for ~10% of tomato genomes. Thus, their presence can impact genome-wide studies significantly and analyses involving gene families with high duplication rates. We advocate that their detection prior will further improve the findings.

Acknowledgements

We thank Christina Azodi, Siobhan Cusack, Melissa Lehti-Shiu for discussion. This work was partly supported by the National Science Foundation (IOS-1546617, DEB-1655386) and U.S. Department of Energy Great Lakes Bioenergy Research Center (BER DE-SC0018409) grant to S.-H.S.

References

Abyzov A, Urban AE, Snyder M, Gerstein M (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res **21**: 974–984

Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol **12**: R18

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Series B Stat Methodol **57**: 289–300

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res **27**: 573–580

Bertioli DJ, Jenkins J, Clevenger J, Dudchenko O, Gao D, Seijo G, Leal-Bertioli SCM, Ren L, Farmer AD, Pandey MK, et al (2019) The genome sequence of segmental allotetraploid peanut Arachis hypogaea. Nat Genet **51**: 877–884

Breiman L (2001) Machine Learning. **45**: 261–277

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics **10**: 421

Chae L, Kim T, Nilo-Poyanco R, Rhee SY (2014) Genomic signatures of specialized metabolism in plants. Science **344**: 510–513

Chen Y-C, Liu T, Yu C-H, Chiang T-Y, Hwang C-C (2013) Effects of GC bias in next-generation-sequencing data on de novo genome assembly. PLoS One **8**: e62856

Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G, Wright J, Borrill P, Kettleborough G, Heavens D, Chapman H, et al (2017) 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 **27**: 885–896

Conesa A, Götz S (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics **2008**: 619832

Consortium IHGS, International Human Genome Sequencing Consortium (2001)

Erratum: correction: Initial sequencing and analysis of the human genome. Nature **412**: 565–566

Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res **36**: e105

Edger PP, Poorten TJ, VanBuren R, Hardigan MA, Colle M, McKain MR, Smith RD, Teresi SJ, Nelson ADL, Wai CM, et al (2019) Origin and evolution of the octoploid strawberry genome. Nat Genet **51**: 541–547

Ezura H, Ariizumi T, Garcia-Mas J, Rose J (2016) Functional Genomics and Biotechnology in Solanaceae and Cucurbitaceae Crops. Springer

Heather JM, Chain B (2016) The sequence of sequencers: The history of sequencing DNA. Genomics **107**: 1–8

Jones E, Oliphant E, Peterson P (2001) SciPy: Open Source Scientific Tools for Python. http://www.scipy.org/

Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat Methods **6**: 291–295

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

Li S, Dou X, Gao R, Ge X, Qian M, Wan L (2018) A remark on copy number variation detection methods. PLoS One **13**: e0196226

Madoui M-A, Dossat C, d'Agata L, van Oeveren J, van der Vossen E, Aury J-M (2016) MaGuS: a tool for quality assessment and scaffolding of genome assemblies with Whole Genome Profiling™ Data. BMC Bioinformatics. doi: 10.1186/s12859-016-0969-x

Matsuba Y, Nguyen TTH, Wiegert K, Falara V, Gonzales-Vigil E, Leong B, Schäfer P, Kudrna D, Wing RA, Bolger AM, et al (2013) Evolution of a complex locus for terpene biosynthesis in solanum. Plant Cell **25**: 2022–2036

Miller CA, Hampton O, Coarfa C, Milosavljevic A (2011) ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. PLoS One **6**: e16327

Moore BM, Wang P, Fan P, Leong B, Schenck CA, Lloyd JP, Lehti-Shiu MD, Last RL, Pichersky E, Shiu S-H (2019) Robust predictions of specialized metabolism genes through machine learning. Proc Natl Acad Sci U S A **116**: 2344–2353

Mostovoy Y, Levy-Sakin M, Lam J, Lam ET, Hastie AR, Marks P, Lee J, Chu C, Lin C, Džakula Ž, et al (2016) A hybrid approach for de novo human genome sequence assembly and phasing. Nature Methods **13**: 587–590

Ou S, Chen J, Jiang N (2018) Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res **46**: e126

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res **12**: 2825–2830

Rhoads A, Au KF (2015) PacBio Sequencing and Its Applications. Genomics, Proteomics Bioinformatics **13**: 278–289

Sedlazeck FJ, Lee H, Darby CA, Schatz MC (2018) Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat Rev Genet **19**: 329-346

Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485**: 635–641

Treangen TJ, Salzberg SL (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet **13**: 36–46

Vukašinović N, Cvrčková F, Eliáš M, Cole R, Fowler JE, Žárský V, Synek L (2014) Dissecting a hidden gene duplication: the Arabidopsis thaliana SEC10 locus. PLoS One **9**: e94077

Wang P, Moore BM, Panchy NL, Meng F, Lehti-Shiu MD, Shiu S-H (2018) Factors Influencing Gene Family Size Variation Among Related Species in a Plant Family, Solanaceae. Genome Biol Evol **10**: 2596–2613

Wang Y, Li J, Paterson AH (2013) MCScanX-transposed: detecting transposed gene duplications based on multiple colinearity scans. Bioinformatics **29**: 1458–1460

Watson M, Warr A (2019) Errors in long-read assemblies can critically affect protein prediction. Nat Biotechnol **37**: 124–126

Yang L-A, Chang Y-J, Chen S-H, Lin C-Y, Ho J-M (2019) SQUAT: a Sequencing Quality Assessment Tool for data quality assessments of genome assemblies. BMC Genomics **19**: 238

Zhuang W, Chen H, Yang M, Wang J, Pandey MK, Zhang C, Chang W-C, Zhang L, Zhang X, Tang R, et al (2019a) The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. Nat Genet **51**: 865–876

Zhuang Z, Shen X, Pan W (2019b) A simple convolutional neural network for prediction of enhancer-promoter interactions with DNA sequence data. Bioinformatics **35**: 2899–2906

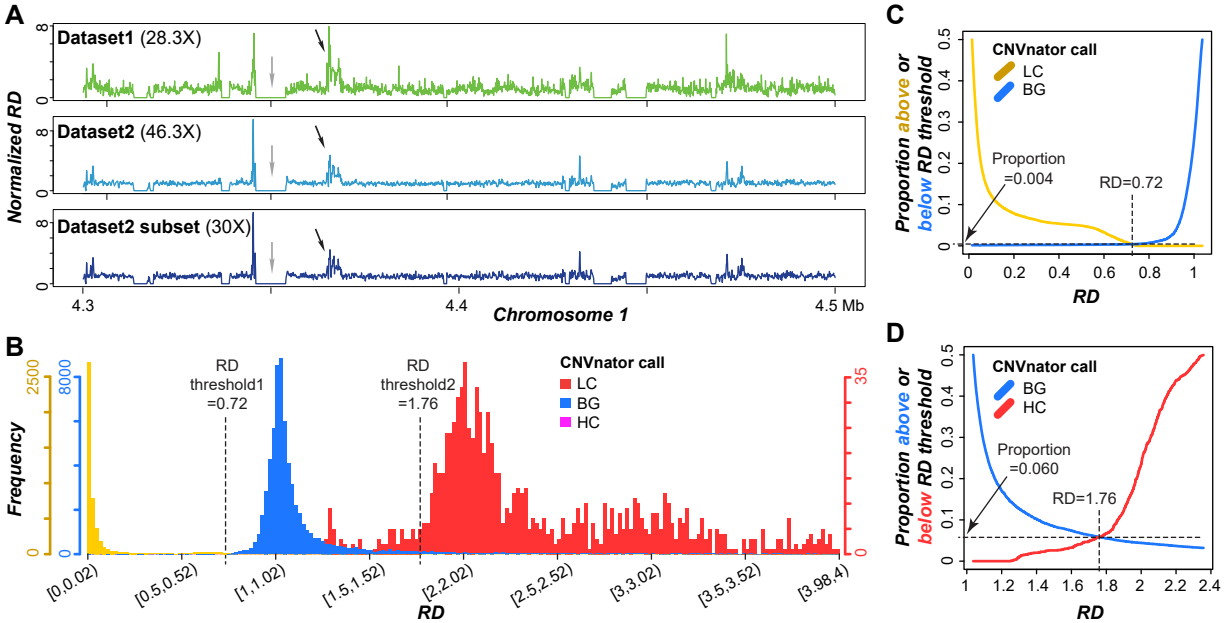

Figure 1. Identification of high, low, and background coverage regions. A, Normalized read depth (RD) of 100bp bins using reads from dataset1, dataset2, and a dataset2 subset at 30-fold coverage in regions 4.3 ~4.5Mb on Chromosome 1. Black and grey arrows show regions with high read coverage and low coverage, respectively. B, RD distributions of high coverage (HC), background (BG), and low coverage (LC) regions called by CNVnator. Dashed lines at RD=0.72 and 1.76: threshold values defining the boundaries between LC and BG regions and between BG and HC regions, respectively. C, D, Proportion of called regions above or below RD value thresholds distinguishing LC and BG regions (C) and distinguishing BG and HC regions (D). Vertical dashed line: threshold value.

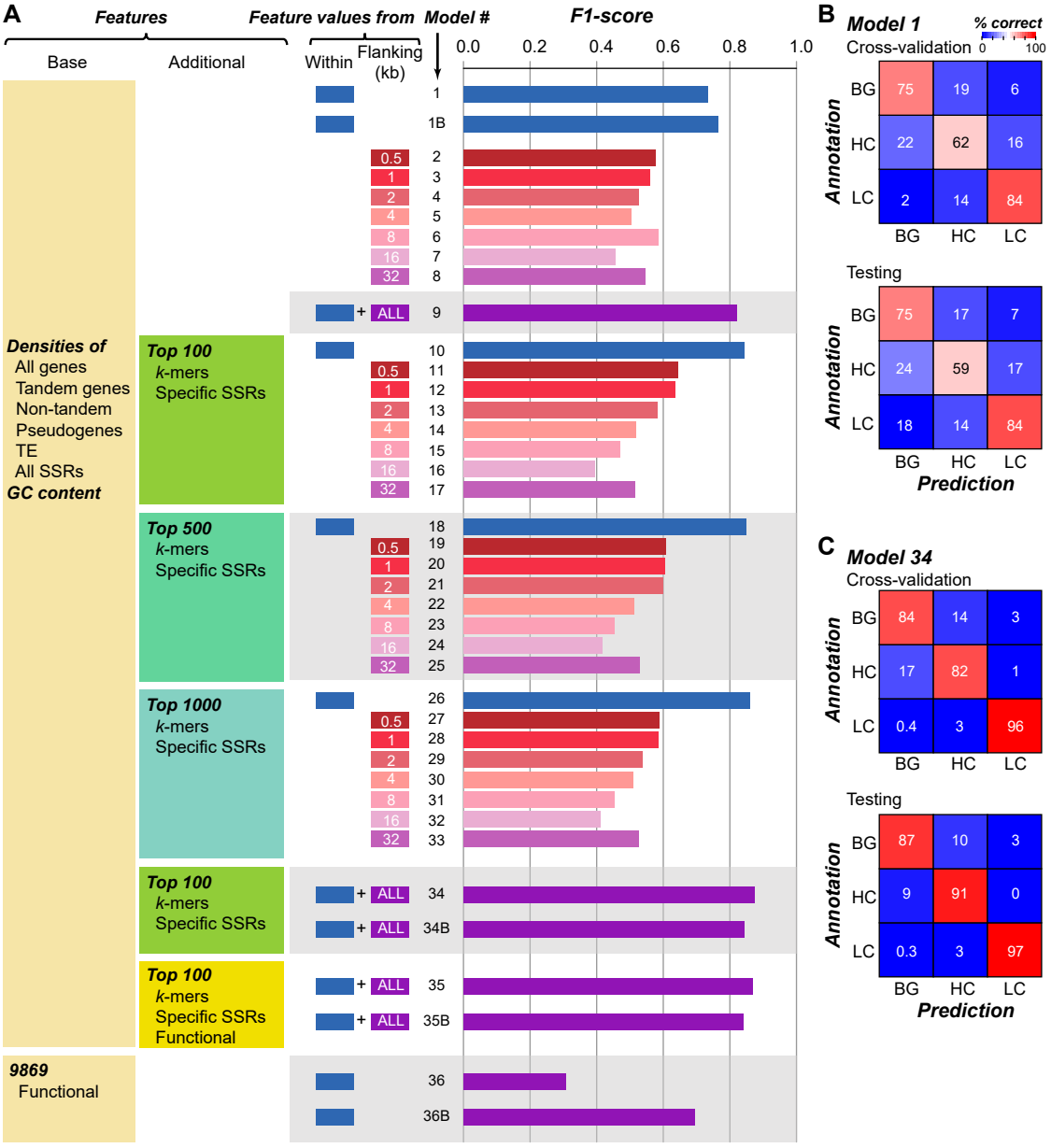

Figure 2. Prediction of HC/LC/BG regions using features of genomic sequences. A, Average F1 of prediction models using different feature sets. The first two columns indicate feature sets used including Base features used in most models and additional features for model subsets. The third and fourth columns 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, Confusing matrices of cross-validation and testing datasets based on Model 1. Color scale: % correct prediction. C, Confusing matrices of cross-validation and testing datasets based on Model 34.

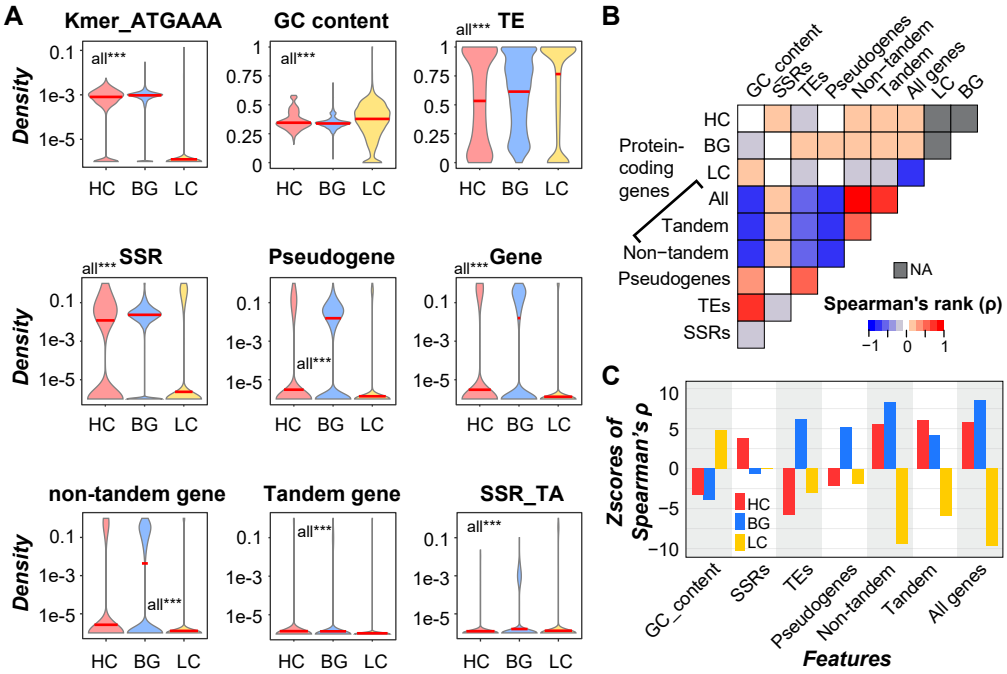

Figure 3. Relationships between genomic features and HC, LC and BG regions. A, Differences in density distributions of important features between HC/LC/BG regions. ***: p<1e-3. B, Correlation between densities of HC/LC/BG regions and other genome features across the genome in bins of 500Kb. Color scale: Spearman's rank correlation coefficient (ρ). C, Z-scores of observed ρ calculated using the distribution of 1000 random Spearman's ρ 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.

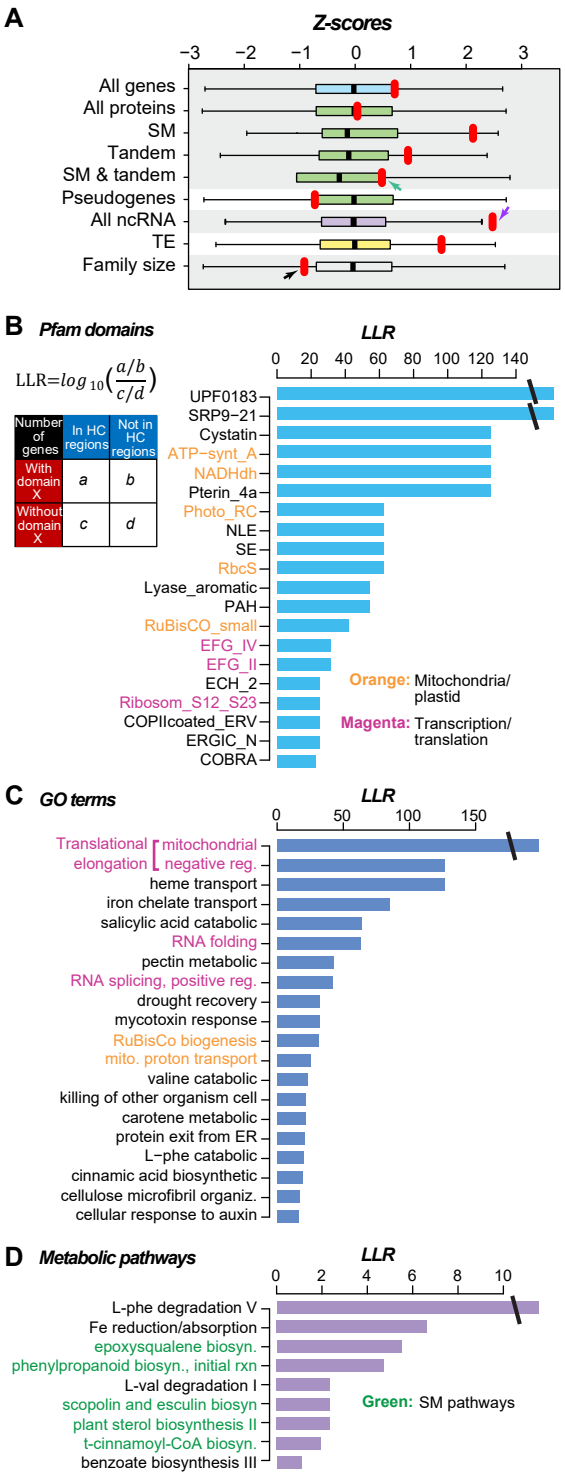

Figure 4. Sequences found in HC regions and their functions. A, Z-score calculated by comparing observed total number of features in true HC regions (red dots) to normalized distributions (boxplots) of the numbers of different genomic features overlapped with 10,000 randomly sampled BG regions (based 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.

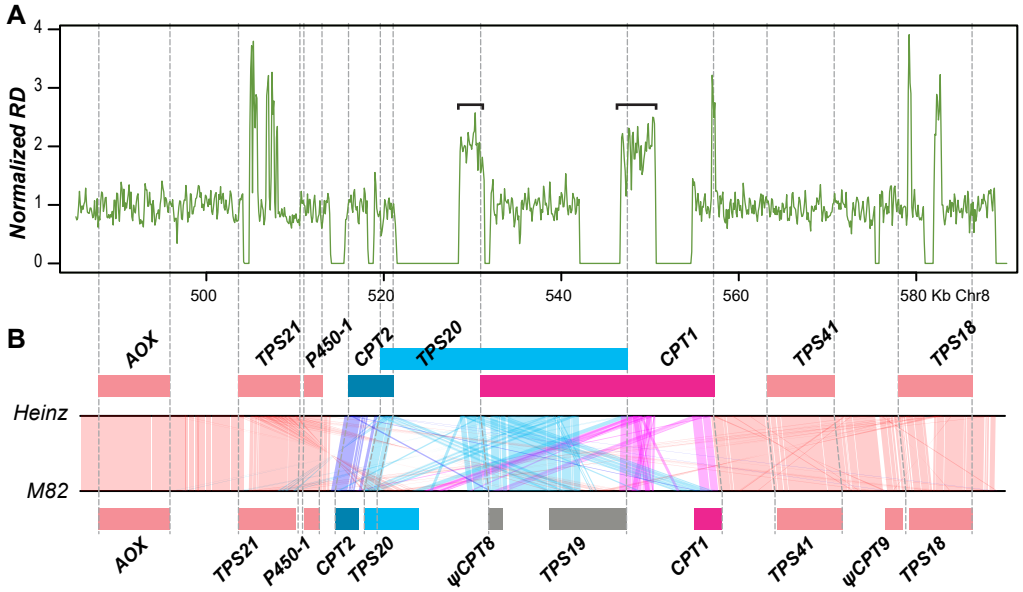

Figure 5. Example mis-assembled HC regions. A, Normalized RD in a region containing an SM gene cluster on Chromosome 8. Normalized RD was calculated for each 100bp bin. Two black brackets indicate identified HC regions, with RD around 2. Valley regions with RD=0 indicate gaps in genome assembly filled up with Ns. B, Syntenic alignments (colored regions between black lines) and gene annotation (colored boxes) of the same region in (A) between Heinz (reference genome) assembled based on shotgun sequencing and M82 assembled using Sanger sequencing reads, respectively. 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 > 50bp. AOX: alcohol oxidase; TPS: terpene synthases; P450: cytochrome P450; CPT: cis-prenyl transferase; ψ: pseudogene.