Published online 16 September 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 3 **1**
https://doi.org/10.1093/nargab/lqab082

Allele-specific assembly of a eukaryotic genome corrects apparent frameshifts and reveals a lack of nonsense-mediated mRNA decay

Raúl O. Cosentino [1,2,*], Benedikt G. Brink [1,2] and T. Nicolai Siegel [1,2,*]

[1]Division of Experimental Parasitology, Faculty of Veterinary Medicine, Ludwig-Maximilians-Universität in Munich, Lena-Christ-Str. 48, Planegg-Martinsried 82152, Germany and [2]Biomedical Center, Division of Physiological Chemistry, Faculty of Medicine, Ludwig-Maximilians-Universität in Munich, Großhaderner Str. 9, Planegg-Martinsried 82152, Germany

Received June 18, 2021; Revised August 25, 2021; Editorial Decision August 31, 2021; Accepted September 06, 2021

ABSTRACT

To date, most reference genomes represent a mosaic consensus sequence in which the homologous chromosomes are collapsed into one sequence. This approach produces sequence artefacts and impedes analyses of allele-specific mechanisms. Here, we report an allele-specific genome assembly of the diploid parasite *Trypanosoma brucei* and reveal allelic variants affecting gene expression. Using long-read sequencing and chromosome conformation capture data, we could assign 99.5% of all heterozygote variants to a specific homologous chromosome and build a 66 Mb long allele-specific genome assembly. The phasing of haplotypes allowed us to resolve hundreds of artefacts present in the previous mosaic consensus assembly. In addition, it revealed allelic recombination events, visible as regions of low allelic heterozygosity, enabling the lineage tracing of *T. brucei* isolates. Interestingly, analyses of transcriptome and translatome data of genes with allele-specific premature termination codons point to the absence of a nonsense-mediated decay mechanism in trypanosomes. Taken together, this study delivers a reference quality allele-specific genome assembly of *T. brucei* and demonstrates the importance of such assemblies for the study of gene expression control. We expect the new genome assembly will increase the awareness of allele-specific phenomena and provide a platform to investigate them.

INTRODUCTION

Genomes contain all genetic material of an organism, thus representing the ultimate template of heredity. In addition, genomes lay the foundation for most molecular research. Therefore, incomplete or erroneous genome assemblies have a broad impact, constraining the spectrum of possible analyses, masking findings and delaying research.

The advent of high-throughput short-read sequencing technologies 15 years ago has resulted in an enormous increase of draft genomes, a 1500% increase in the period between 2005 and 2012 (1). Yet, most of those assemblies were left incomplete and highly fragmented, due to the limitation of short-read sequences, failing to resolve most repetitive regions. This scenario started to change with the more recent deployment of long-read sequencing technologies such as PacBio and Nanopore, which have dramatically improved the contiguity of genome assemblies (2,3). Furthermore, complementation of long-read sequencing with methods that capture chromosome conformation, such as Hi-C, enables accurate scaffolding, often reaching chromosome-scale assemblies (4). Hi-C is a high-throughput assay that measures the physical contact frequency between pairs of DNA regions on a genome-wide scale by paired-end sequencing. Even though this approach was developed to study the three-dimensional organization of genomes, it was quickly observed that the frequency of interaction between any given pair of sequences was highly dependent on the distance between them in the linear chromosome. Thus, this information could be used to order chromosome pieces in genome assembly projects. Since then, the usage of Hi-C data has shown to be one of the most powerful scaffolding approaches (5,6).

Ploidy refers to the number of complete chromosome sets in a cell. Many eukaryotic organisms have two or more sets of chromosomes. These chromosome sets are often not exactly identical to each other and gene alleles can be differently regulated in the context of the alternative homologous chromosomes. In fact, an increasing body of evidence shows that allelic biased gene expression is an important factor in normal cell development and in the emergence of dis-

*To whom correspondence should be addressed. Tel: +49 89 2180 77098; Email: n.siegel@lmu.de
Correspondence may also be addressed to Raúl O. Cosentino. Tel: +49 89 2180 71156; Email: raul.cosentino@lmu.de

ease. For example, a recent study showed that monoallelic gene expression is prevalent in human brain cells, possibly playing a role in the complexity and diversity of brain cell functions, which may also explain the variable penetrance of neuronal disorder-related mutations (7).

Yet, to date, most reference genome assemblies of diploid (and polyploid) organisms represent a mosaic consensus sequence for which the different alleles have been collapsed into one sequence per chromosome, not corresponding to any of the true chromosomal alleles. This procedure introduces sequence artefacts in the assembly, leading to annotation and analysis errors (8,9). Furthermore, it disregards the information of true allelic variants and the association between them, essential for the analysis of allele-specific mechanisms (10,11), the identification of recombination events (12) or evolutionary studies (13).

To recover the haplotypes from a collapsed assembly, typically a two-step procedure is employed (14). First, the positions and the possible alleles for each of the heterozygous sites are determined. Second, the co-occurrence of alleles in neighbouring sites along the chromosomes is defined. The first step can be achieved with regular short-read sequencing data and established variant callers, such as FreeBayes (15) or GATK (16). However, tools to address the second step, the linkage of neighbouring variant sites, have only emerged recently and include (i) long-read sequencing technologies such as PacBio or Nanopore, (ii) short-read sequencing-based assays capable of connecting distal regions from the same chromosomal allele, such as Hi-C, and (iii) computational tools able to integrate information from different sequencing technologies to link variants, such as HapCUT2 (17). Together, these technological advances are starting to make the generation of haplotype-specific genome assemblies possible.

In this paper, we describe the generation of a fully phased genome assembly of *Trypanosoma brucei*. *T. brucei* is a diploid eukaryotic parasite, the causal agent of sleeping sickness in humans and nagana in cattle, and a model organism for the study of antigenic variation. To generate the original genome assembly from the *T. brucei* TREU 927 isolate (18), both haplotypes were collapsed into one mosaic sequence. Recently, combining PacBio sequencing and Hi-C data, we assembled the genome of the *T. brucei* Lister 427 isolate (Tb427v9) (19), the isolate most widely used in laboratory settings. While the Tb427v9 assembly contained phased subtelomeric regions, the central part of the chromosomes encoding most of the transcriptome remained collapsed, disregarding the information of allelic variants and the association between them. To address these limitations, we aimed to reveal the allelic variation in the *T. brucei* Lister 427 isolate genome and to reconstruct the haplotypes, generating an allele-specific genome assembly. The newly assembled phased genome will open opportunities to the trypanosome community to explore a new level of gene regulation, namely the allele-specific expression.

MATERIALS AND METHODS

Error correction of PacBio reads

Error-corrected PacBio reads were generated with Proovread (20) using as input PacBio raw reads and a set of gDNA-seq short reads from *T. brucei* Lister 427 (used datasets are available in Supplementary Table S1).

Variant detection

Error-corrected PacBio reads were mapped to the genome assemblies with Minimap (v2.10) (21), while short reads were mapped with bwa-mem (v0.7.16) (22). Supplementary, secondary and low-quality alignments (mapq < 10) were filtered with samtools (v1.8) (23). Variant detection was performed with FreeBayes (v1.2.0) (15) and the resulting variant files (VCFs) were filtered by QUAL > 20.

Assembly error correction

Variants suggesting errors (i.e. homozygous for an alternative allele, or heterozygous for two alternative alleles) in the *T. brucei* Lister 427 v9 genome assembly, identified from mapping error-corrected PacBio reads, were corrected using a Perl (v5.26.2) script parsing the VCF file. This error-correction step led to the generation of the *T. brucei* Lister 427 genome assembly v10 (Tb427v10).

Annotation

Genome assembly versions were annotated primarily with Companion v1.0.1 (24). Additionally, variant surface glycoprotein (VSG) genes were annotated based on BLAST results to the manually curated VSGnome (25), as done previously (19). We reannotated genes indicated by Companion to have introns as pseudogenes because (i) with the exception of two genes, *T. brucei* is thought to lack introns (26), and (ii) after manually inspecting several genes annotated by Companion as containing introns, we observed no evidence of introns but of frameshifts.

Comparison of annotated proteome

Coding sequences from the annotated protein-coding genes from the chromosomal 'cores' for each genome assembly version were extracted with BEDTools (v2.26.0) (27) 'getfasta' function and translated with fastaq (v3.17.0) 'translate' function. Each proteome multifasta was reciprocally aligned to the proteome from *T. brucei* TREU 927 [version 42, extracted from TriTrypDB (28)] using BLAST (v2.7.1+) (29) with settings to generate a tabular extended output '-outfmt "7 std qlen slen"'. For each query protein, the best hit in the subject proteome was extracted based on the maximum bit score. The analysis of query over subject protein size ratio was done using Python (v3.6), with the module pandas (v1.03), and the respective plots were made using matplotlib (v3.2.1) (30).

Density distribution analysis

The Tb427v10 genome assembly was divided into 10 kb overlapping bins, with 1 kb sliding window with BEDTools (v2.26.0) 'makewindows' function. For each DNA-seq dataset, the number of heterozygote variants within each bin was counted with BEDTools 'coverage' function, using as input the variant files (VCFs) previously generated with FreeBayes (see the 'Variant detection' section

in the 'Materials and Methods' section). These VCF files contained the coordinates of the heterozygote variants for each DNA-seq dataset. For the same bins, the number of mapped reads for the different sequencing datasets analysed (for the list of datasets used, see Supplementary Table S1) was counted. GC content was calculated with BEDTools 'nuc' function. 'Mappability' was defined as the percentage of mapping reads obtained with reads filtered by mapping quality (mapq > 10) from the mapping reads obtained without filtering. Mappability serves as an indicator on how repetitive (low mappability) or unique (high mappability) regions are. Pearson correlation between the different datasets was calculated with the module pandas from Python and the heatmap was generated with the module seaborn (v0.10.0).

Ploidy analysis

To estimate the ploidy for each chromosome in each gDNA-seq datasest, the median coverage of the chromosome was divided by the median coverage for all the chromosomes. For the specific analysis of candidate chromosomes with trisomy, their median coverage was divided by the median coverage of the rest of the chromosomes.

Analysis of the effect of ploidy on transcript and translation levels

RNA-seq and ribosome profiling data were mapped to the Tb427v10 genome assembly with bwa-mem (v0.7.16) and converted to bam files with samtools (v1.8). Gene counts were obtained with subread (v1.6.2) (31) function 'featureCounts' and gene expression fold change between aneuploid and diploid cell lines was calculated in R (v3.5.2) with DESeq2 (v1.22.2) (32) using default parameters. Box plots of $\log_2$ fold change distributions considering all genes within each chromosome were generated using ggplot2 (v3.3.0).

Allele phasing

The heterozygote variants identified in the diploid 'core' of the Tb427v10 genome assembly were phased with HapCUT2 (17) using raw PacBio reads and Hi-C data as input with settings to phase not only single-nucleotide variants, but also insertions or deletions (INDELs; '–indels 1'). The phasing output of HapCUT2 was used to generate the allele 'A' and the allele 'B' fasta genome file using a custom-made Perl script. HapCUT2 generates phased 'blocks' for each chromosome. In case of several blocks, only the block with the highest number of variants phased was considered.

Minichromosomal VSG identification

To identify putative minichromosomal contigs, we first searched for contigs containing a curated set of minichromosomal VSGs (25). At the same time, we searched for contigs containing a minichromosomal signature, a 177 bp repeat (33). We selected two regions of the 177 bp repeat that were highly conserved (34), performed BLAST searches and extracted the contigs that had at least one hit. The search with both regions returned us the same set of contigs. Additionally, we searched for contigs containing telomeric repeats (we used five times 'TTAGGG' as query in the BLAST search) and an annotated VSG gene. Finally, we combined the set of putative minichromosomal contigs identified with the different approaches.

To assess the completeness of VSGs, we first generated a protein multifasta of all annotated VSGs combining BEDTools 'getfasta' function and fastaq 'translate' function. We predicted the presence of peptide signals using SignalP (v5.0) (35) and GPI anchor using netgpi (v1.1) (36), and defined loose VSG protein size limits (between 400 and 600 aa) based on size ranges previously observed (25). Given that we noticed that start and stop codons were frequently not correctly annotated for the VSGs (e.g. CDS annotation was starting after the 'ATG' and finishing before the stop codon), we decided to make a more sensitive selection of CDSs. We extracted from each annotated VSG its coordinates extending 300 bp on each side. Then, we determined the longest ORF and performed the same predictions as before. A VSG was determined as functionally complete if it had (either in the standard or in the sensitive approach) a predicted peptide signal, a predicted GPI anchor and a size between 400 and 600 aa.

Hi-C interactions to the fully phased genome

Hi-C paired-end reads were truncated with HiCUP (v0.8.0) (37) and each read of the pairs was mapped independently with bwa-mem to the *T. brucei* Lister 427 allele-specific genome assembly. Normalized interaction frequency matrices were generated using HiC-Pro (v2.11.4) (38) and heatmaps were obtained with HICsuntdracones (v0.2.0, doi: 10.5281/zenodo.3570497), following the same pipeline we have used previously (19).

Allele-specific quantification

Allele-specific read counts for RNA-seq and ribosome profiling data from *T. brucei* Lister 427 (39) were obtained using GATK (v4.1.8.1) (16) 'ASEReadCounter' function with '–min-mapping-quality 10', and the variant file previously generated for the error-corrected PacBio reads on Tb427v10 (see the 'Variant detection' section in the 'Materials and Methods' section), filtered to contain only SNPs. Then, the counts for each allele were assigned to haplotypes based on the HapCUT2 phasing output, using a Python script. Bar plots for the allelic counts in selected genes with premature termination codons (PTCs) were generated using matplotlib.

RESULTS

Outline of the genome phasing strategy

In order to generate an allele-specific genome assembly of the *T. brucei* Lister 427 isolate, we established the following strategy: (i) error-corrected PacBio reads were mapped to our previously published, collapsed genome assembly (Tb427v9) (19), (ii) variant positions were identified and those variants suggesting errors in the assembly were corrected, (iii) the distribution of variants across the chromosomes was analysed, as well as the ploidy based on different

gDNA-seq datasets from Lister 427 clones, and (iv) using Hi-C data and raw PacBio reads, the variants were linked, to generate an allele-specific assembly (outlined in Figure 1).

Error-correction step improves accuracy of protein-coding gene annotation

Genome assemblies based on long-read technologies are usually more contiguous than those based on short reads, but at the nucleotide level they are less accurate (40). Moreover, the predominant errors that these technologies introduce are insertions or deletions of bases (INDELs) (41), which, if located within an open reading frame (ORF), often lead to artificial truncation and the incorrect annotation of protein-coding genes as pseudogenes (42).

Thus, in order to increase the accuracy of the Lister 427 genome assembly, we performed an error-correction step before generating an allele-specific genome assembly. To this end, we first produced highly accurate long reads by correcting 5.36 Gb of raw PacBio reads with 31.83 Gb of Illumina short reads (20), producing 4.75 Gb of error-corrected PacBio reads (72× genome coverage). The recommended coverage for Illumina reads to produce highly accurate PacBio reads (∼99.98% accuracy) is >50× (20,43). Here, we used ∼478× coverage of Illumina reads, clearly exceeding the recommended minimum value. Then, we mapped the error-corrected PacBio reads to the collapsed Lister 427 genome assembly v9 (Tb427v9), identified and classified variant positions. The genome sequence was corrected in the positions where the mapped reads suggested errors and the impact of the correction in terms of protein-coding gene annotation was assessed (Figure 2A). Using the error-corrected PacBio reads, we identified >98 000 variant positions in the diploid 'core' of the Tb427v9 genome assembly. Even though most of these variants were of the 'expected type', i.e. heterozygotes between the allele present in the assembly and an alternative allele (variants nomenclated as 'Ref/Alt1'), around 7000 variants suggested errors in the genome assembly (Figure 2B). Those 'error-suggesting variants' were either homozygous for an alternative allele (nomenclated as 'Alt1/Alt1') or heterozygous between two alternative alleles (nomenclated as 'Alt1/Alt2'). As expected by the error bias of long-read-based assemblies, most (98%) of the error-suggesting variants were INDELs, or complex [i.e. a combination between single-nucleotide polymorphisms (SNPs) and INDELs] (Figure 2B). In contrast, the Ref/Alt1 variants were mostly SNPs (Supplementary Figure S1A). Therefore, we proceeded to correct all error-suggesting variants, generating genome assembly Tb427v10. As expected, variant detection on the error-corrected Tb427v10 genome assembly showed a strong (98%) decrease in error-suggesting variants and a concomitant increase in Ref/Alt1 variants by ∼5000, proportional to the number of Alt1/Alt2 variants corrected in Tb427v9. In total, we detected 96 060 heterozygote variants (Figure 2B), indicating an average of 4.21 heterozygote variants per kb in the core genome of *T. brucei* Lister 427.

Next, to assess whether the introduced sequence corrections were actually improving the genome assembly accuracy, we compared the annotation of protein-coding genes between the assembly versions. Given that most of the errors that we had corrected were INDELs or complex errors, we rationalized that if the error-suggesting variants were real errors (i.e. not technical artefacts), correcting them would 'repair' broken ORFs, increasing the number of annotated protein-coding genes and decreasing the number of truncated protein-coding genes. The opposite effect would be expected should the corrections have introduced errors.

Thus, we annotated the Tb427v10 assembly with the same pipeline used to annotate the Tb427v9 assembly. Comparing the two annotations, we observed an increase of 4% (305 genes) in the total number of annotated protein-coding genes in Tb427v10 (Figure 2C), indicating that the errors were real and that the correction step increased the accuracy of the genome assembly sequence. To further evaluate the quality of the genome and its annotation, we implemented a previously described strategy of ORF size comparison (44). The strategy involves alignment of all annotated protein sequences from a query genome to a curated database of proteins followed by calculation of the size ratio between each query protein and its best hit in the database. If the query protein is of the expected length, the size ratio to its best hit in the database will be ∼1 (∼0 in $\log_2$ scale); if the protein in the assembly is truncated, the size ratio to its best hit will be <1 (<0 in $\log_2$ scale). We compared the annotated proteome of the Lister 427 genome assembly versions, before (Tb427v9) and after (Tb427v10) the correction step, to the proteome of the TREU 927 isolate of *T. brucei*, whose genome annotation has been curated extensively. Our comparison revealed a strong decrease in the number of genes with size ratios below 1 (from 387 to 146 genes for size ratio <0.9) and an increase in the number of genes with size ratio ∼1 (from 7255 to 7761 genes for size ratio between 0.95 and 1.05) after the correction of the genome (Figure 2D), again supporting the importance of the error-correction approach.

Besides serving as a genome improvement assessment, the size ratio analysis provided us with additional information. We identified genes where the size in our Lister 427 isolate assembly was larger than its best hit in the TREU 927 isolate assembly (size ratio >1). It is possible that these genes are larger in the Lister 427 isolate. However, a more plausible explanation seems to be that, given that in our assembly we used a long-read-based approach, we could resolve the extension of repeat regions that were collapsed in the Sanger sequencing-based TREU 927 isolate assembly. Indeed, we analysed four of the cases with 'size ratio >2', and found that three of them were clear examples of extreme repeat collapsing (Figure 2E). The remaining one was an incident where our assembly suggests that two genes are in fact one (Figure 2F). The later finding is supported by microscopy data from the TrypTag project, an initiative to tag and visualize all *T. brucei* proteins (45,46). N-terminal tagging of the first fraction of the gene (Tb927.2.5860) and C-terminal tagging of the second fraction of the gene (Tb927.2.5870) show a very specific localization in the flagellar tip of the parasite, while tags that break the actual gene (C-terminal tagging and N-terminal tagging of Tb927.2.5860 and Tb927.2.5870, respectively) show cytoplasmic (mis)localization (Figure 2F, lower panel). The complete list of size ratio between protein-

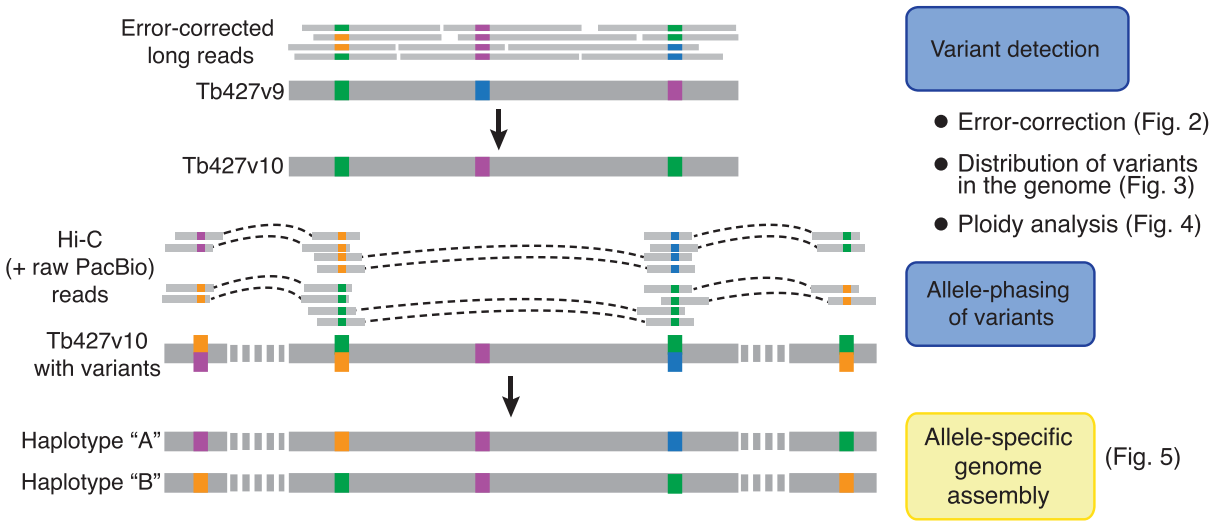

Figure 1. Haplotype phasing procedure. Short-read error-corrected PacBio reads were used to identify variants and correct errors in the *T. brucei* Lister 427 Tb427v9 genome assembly. The green, orange, violet and blue blocks on the read and assembly representations indicate the presence of different sequences on variant loci. Then, the distribution of heterozygous variants along the genome and the ploidy of *T. brucei* Lister 427 clones were analysed. Finally, Hi-C data and raw PacBio reads were used to link variants and reconstruct the haplotypes of the homologous chromosomes.

coding genes from Tb427v10 and the TREU 927 isolate assembly is available in Supplementary Table S2.

In summary, the usage of error-corrected PacBio reads led to an increase of 7% of protein-coding genes with the expected length, suggesting a significant improvement of overall genome sequence and enabled the identification of ∼96 000 heterozygote variants.

Uneven distribution of variants across the genome

Following the improvement of the assembly accuracy, we proceeded to analyse the genome-wide distribution of the heterozygote variants identified. On average, we observed 4.21 variants per kb, yet variant distribution was very uneven, ranging from 0 to >150 variants per 10 kb window (Figure 3A–F and Supplementary Figure S4). As expected, we obtained a very similar genomic distribution of variants from other sequencing datasets of different Lister 427-derived clones. Interestingly, we also observed a similar variant distribution in other *T. brucei* strains. Thus, we speculated the defined genomic distribution of variants to be caused by differences in mutation permissiveness, with non-coding regions, such as transcription start and transcription termination regions, possibly harbouring a higher variant density. However, we observed no clear correlation with any genomic feature analysed, such as H2A.Z and H3.V distributions, enriched at transcription start and termination sites, respectively (47) (Supplementary Figure S2).

Noticeably, several chromosomes contained long regions where heterozygosity levels drop to near zero, representing loss-of-heterozygosity (LOH) regions. These regions were mostly located towards chromosome ends. For example, on chromosomes 2 and 3 LOH regions extended for >300 and >600 kb inwards from the 3' telomeres (Supplementary Figure S4 and Figure 3B, respectively). On other chromosomes, LOH regions were also present internally (Figure 3C). To be able to exclude the possibility that LOH regions are the result of local haploidy, we analysed available DNA-seq data. No drop in coverage was observed for these regions that would have pointed to the partial loss of one of the two homologous chromosomes (Supplementary Figure S3). Thus, the observed LOH regions are most likely the result of recent recombination events between the two homologous alleles, which have reset heterozygosity.

To investigate how recently these recombination events may have occurred, we searched for LOH regions in different lines of the Lister 427 isolate. The *T. brucei* Lister 427 was isolated >50 years ago (48) and clones adapted to culture in mammal-stage condition (also referred to as bloodstream form clones, 'BSF') or insect-stage condition (also referred to as procyclic form clones, 'PCF') have been selected early on (49). As expected for cells that expand clonally in culture, we observed a very similar heterozygosity pattern across the chromosomes (Figure 3 and Supplementary Figure S4). Most of the LOH regions were common to all clones tested. However, some were specific to insect-stage-adapted clones (Figure 3D and E) and one LOH region was only present in a specific insect-stage-adapted clone (Figure 3F). These observations suggest that most of the recombination events leading to LOH regions occurred before the cell culture adaption. In addition, our data suggest that the insect-stage parasites are either more prone to recombination events than mammalian-stage parasites or that the specific LOH events increased their fitness. Work in yeast has shown that LOH can increase the fitness of an organism by removing one or several suboptimal gene alleles (50).

Trisomy leads to proportional increase in transcript levels

While analysing DNA-seq data to exclude haploidy in the LOH regions, we observed for a 'wild-type' Lister 427 mammal-stage-adapted clone from our laboratory ['Tb427 BSF WT (Siegel 2014)'], originally acquired from the Cross

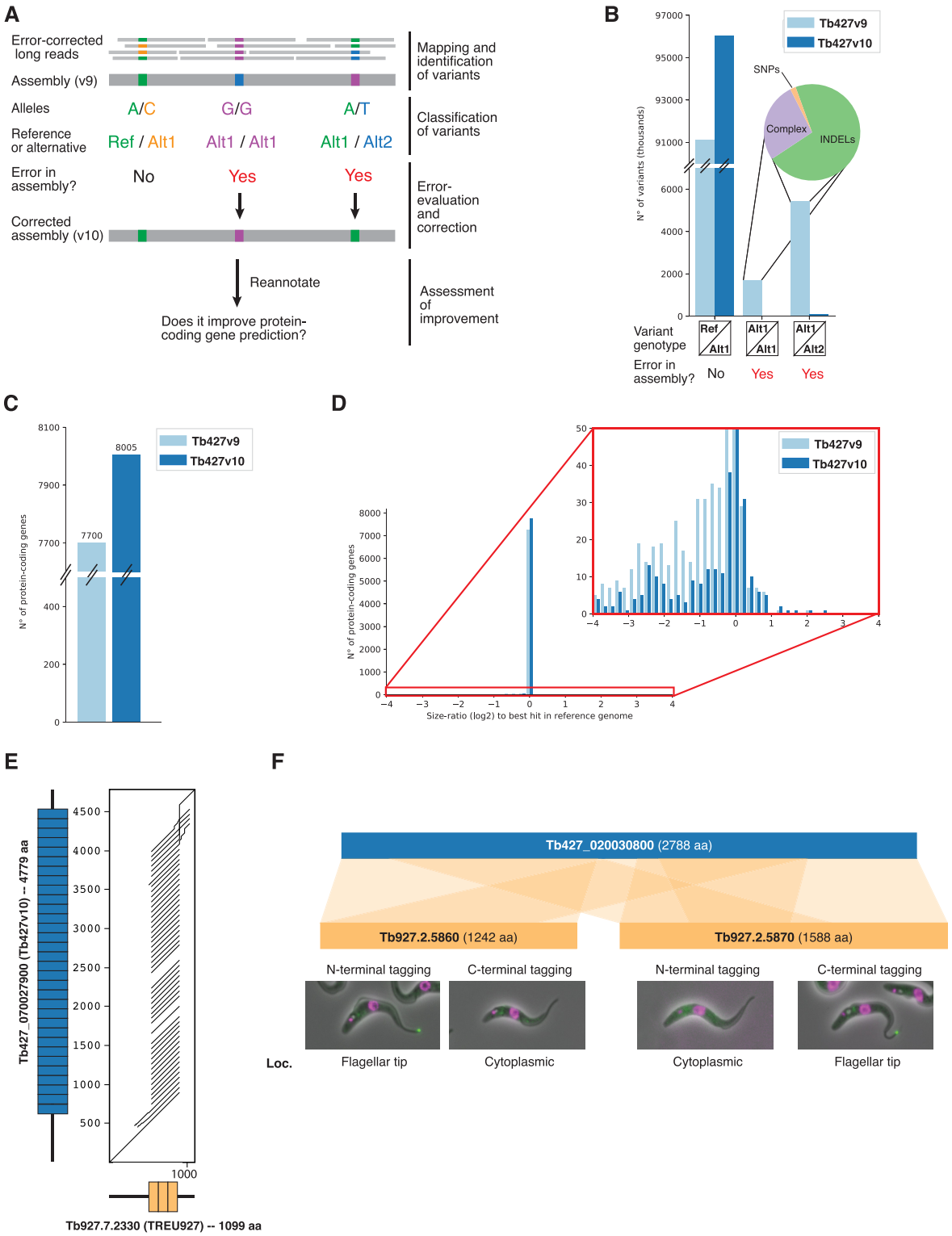

Figure 2. Genome error correction. (**A**) Error-correction approach. (**B**) Number of variants identified before (Tb427v9) and after error correction (Tb427v10) of the genome, grouped by variant genotype. The pie chart shows the proportion of SNPs, complex and INDELs among the error-suggesting variants (the sum of 'Alt1/Alt1' + 'Alt1/Alt2' variants) in Tb427v9. (**C**) Number of protein-coding genes annotated. (**D**) Size ratio distribution ($\log_2$ scale) of protein-coding genes compared to the *T. brucei* TREU927 genome assembly. The right panel is a zoom-in to the region selected in red. (**E**) Alignment dot plot between syntenic orthologs in Tb427v10 (Tb427_070027900) and *T. brucei* TREU927 (Tb927.7.2330), showing a large repeat-array length difference, illustrated by the number of blue and orange boxes in the gene representations on the axes. (**F**) Alignment of Tb427_020030800 from Tb427v10 to two contiguous protein-coding genes annotated in the *T. brucei* TREU927 genome assembly (Tb927.2.5860 and Tb927.2.5870), suggesting the gene was wrongly split in the latter assembly. The lower panel shows the subcellular localization (in green) of the N- and C-terminal tagged versions of Tb927.2.5860 and Tb927.2.5870 (data provided by TrypTag.org).

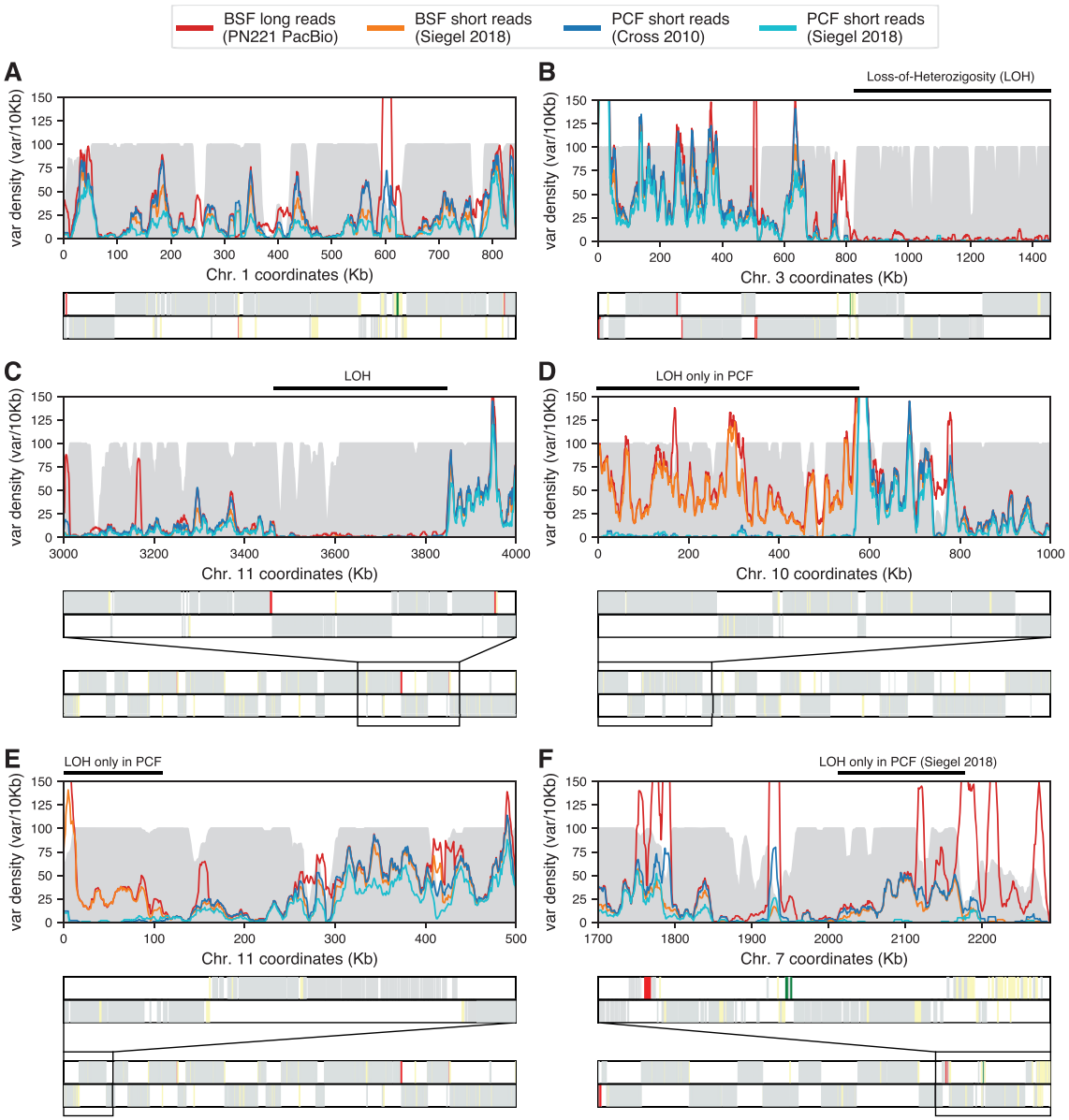

Figure 3. Variant density and LOH regions in the *T. brucei* Lister 427 genome. (**A–E**) Heterozygosity from different *T. brucei* Lister 427 sequencing datasets on selected chromosomal regions of the Tb427v10 genome assembly. In the background, a filled histogram (grey) represents 'mappability' (in range 0–100) for short-read sequencing data (see the 'Materials and Methods' section). LOH regions are indicated by a black line above the plot. The lower panel shows the location and coding strand of protein-coding genes (grey), pseudogenes (yellow), VSG genes (red) and rRNA genes (green) as vertical lines, and zoom-out representation of the complete chromosome when a smaller region was selected.

laboratory (The Rockefeller University, New York), a ∼50% increase in DNA-seq coverage across chromosome 5 compared to the remaining chromosomes (Figure 4A and B), pointing to a third copy of that chromosome. Other datasets derived from the same Lister 427 clone also showed increased coverage for chromosome 5, albeit some to a lower level, possibly the result of a mixed population. No increase in chromosome 5 coverage was observed for the cell line used to assemble the Tb427v10 genome ('Tb427 BSF PN221'), which also derives from the Lister 427 isolate (Figure 4A), suggesting that the observed aneuploidy was incorporated during the *in vitro* culture after the generation of the Tb427 BSF PN221 cell line. To test whether aneuploidy was common among *T. brucei* cell lines, we analysed available gDNA-seq datasets from other Lister 427 clones, from other *T. brucei brucei* isolates and from the related subspecies *T. brucei gambiense* and *T. brucei rhodesiense*. We did not observe any substantial changes in coverage between chromosomes for the datasets belonging to *T. brucei* TREU 927, *T. brucei gambiense* and *T. brucei rhodesiense* clones, suggesting normal ploidy in all the chromosomes (Figure 4A). Interestingly, we found two additional datasets from the *T. brucei* Lister 427 isolate suggesting aneuploidies, both from insect-stage-adapted clones. One showed increased coverage across chromosomes 2 and 6, while the other showed increased coverage across

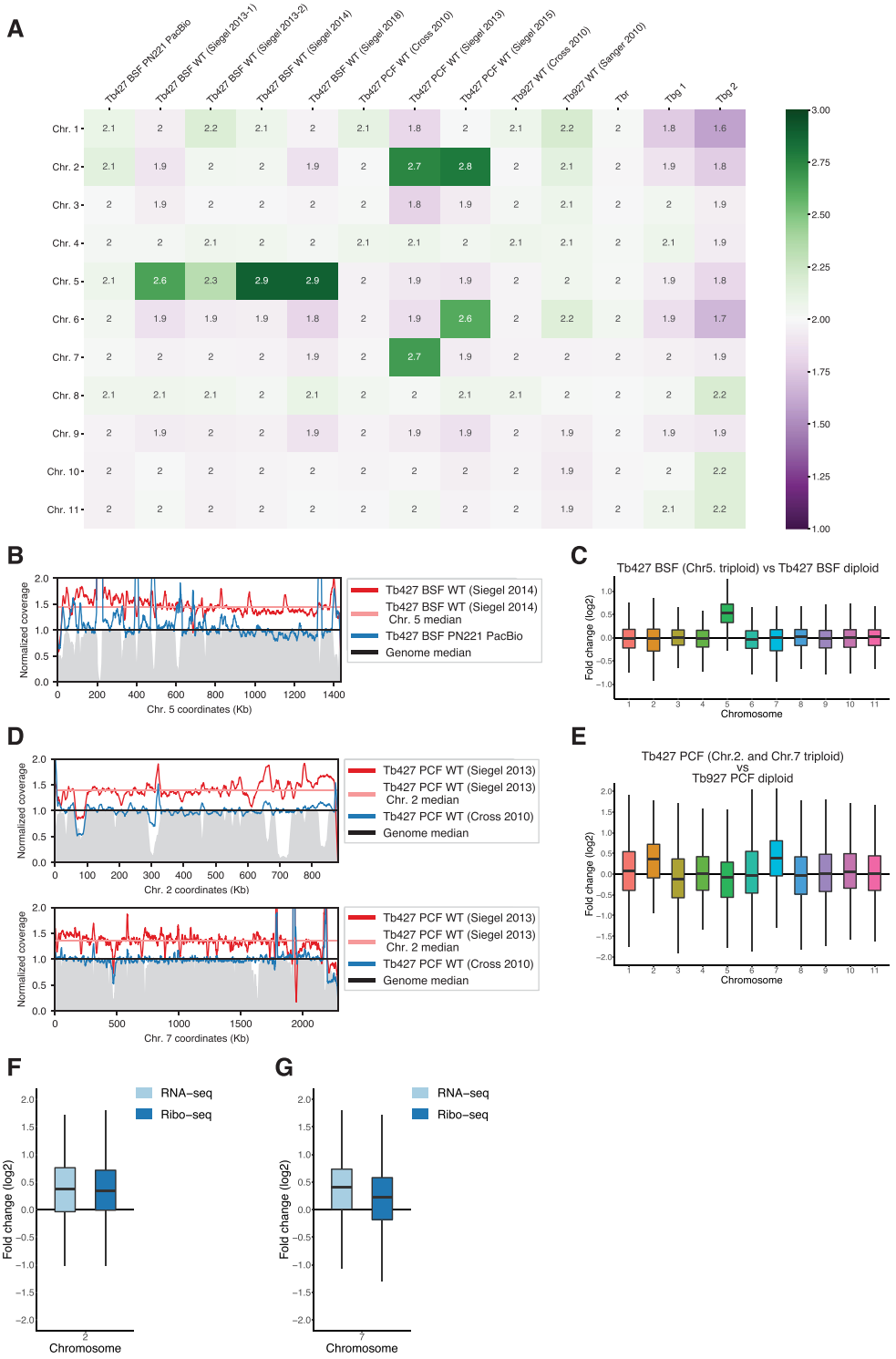

Figure 4. The effect of aneuploidy on gene expression in *T. brucei*. (**A**) Median coverage per chromosome for different *T. brucei* DNA-seq datasets, normalized to genome median and centred at two (to illustrate expected diploidy). In the labels, 'Tb427' indicates a Lister 427-derived clone, 'Tb927' a TREU927-derived clone, 'Tbr' *T. brucei rhodesiense* and 'Tbg' *T. brucei gambiense*. (**B**) Normalized coverage density in chromosome 5 for Tb427 BSF WT (Siegel 2014) clone (dark red line) and its median (straight light red line) compared to Tb427 BSF PN221 PacBio clone (dark blue line). The genome median is set to 1 (straight black line). Mappability is shown as a grey filled line (in range 0–1). (**C**) RNA-seq fold change ($\log_2$) pooled by chromosome, from a *T. brucei* Lister 427 clone triploid for chromosome 5 over a diploid *T. brucei* Lister 427 clone. (**D**) Normalized coverage density in chromosome 2 (upper panel) and chromosome 7 (lower panel) for Tb427 PCF WT (Siegel 2013) clone (dark red line) and its median (straight light red line) compared to Tb427 PCF WT (Cross 2010) clone (dark blue line). The genome median is set to 1 (straight black line). Mappability (in range 0–1) is indicated in grey. (**E**) RNA-seq fold change ($\log_2$) pooled by chromosome, from a *T. brucei* Lister 427 PCF triploid for chromosomes 2 and 7 over a diploid *T. brucei* TREU927 clone. (**F**, **G**) RNA-seq and Ribo-seq fold change ($\log_2$) between the same clones as (e), for chromosomes 2 and 7, respectively.

chromosomes 2 and 7 (Figure 4A), suggesting trisomy on these chromosomes. In all observed cases, the increase in coverage occurred across the entire chromosome (Figure 4B–D and Supplementary Figure S5), eliminating the possibility of partial chromosome amplifications. To our knowledge, aneuploidy has not been previously reported in *T. brucei*.

The organization of genes in polycistronic transcription units (18) and the lack of canonical RNA pol II promoters (51) have led to the assumption that RNA pol II transcription is not regulated in *T. brucei*. If true, the increase in chromosome copy number should thus result in a proportional increase in transcript levels. To test this assumption, we compared available RNA-seq data from a Lister 427 clone with trisomy 5 to a diploid clone carrying two copies of chromosome 5. Our analysis showed a ∼50% increase in the transcript level for the genes in chromosome 5 in the cell line with trisomy 5 (Figure 4C), indicating a direct correlation between gene copy number and transcript levels. These observations support the assumption that *T. brucei* lacks transcriptional regulation, but also suggest the absence of a general dosage compensation mechanism that senses and regulates transcript levels post-transcriptionally.

Similar results were obtained when comparing RNA-seq data from the insect-stage-adapted Lister 427 clone (52) containing trisomies 2 and 7 to RNA-seq data from a regular diploid *T. brucei* TREU 927 isolate insect-stage-adapted clone (39) (Figure 4E). Interestingly, comparing ribosome profiling data from both clones, we observed a negligible reversion of expression levels (Figure 4F and G and Supplementary Figure S6), suggesting minimal feedback regulation at the translational level.

Haplotype phasing of the Lister 427 genome assembly

With the knowledge of the uneven genome-wide variant distribution and having confirmed diploidy for all chromosomes in the clone used for the genome assembly, we went ahead to generate an allele-specific genome assembly of *T. brucei* Lister 427. To this end, we combined raw PacBio reads and Hi-C data, together with HapCUT2 (17), a tool capable of integrating information from diverse sequencing technologies to link variants into haplotypes. Following this approach, we could assign 99.55% of the total heterozygote variants (95 628 / 96 060 variants) to a specific chromosomal haplotype, enabling the almost complete reconstruction of the two alleles for each chromosome 'core' (Figur 5).

To evaluate the accuracy of the haplotyping and to reconstruct the full chromosome alleles, we quantified the Hi-C interactions between the chromosome 'core' alleles and the haploid-like subtelomeres. We observed that each alternative subtelomere interacted strongly with only one core allele, in a distance-dependent fashion (Supplementary Figure S7). This allowed us to scaffold the full chromosome alleles (Figure 5) and also served as validation of the phasing approach. An inaccurate variant phasing of the chromosomal cores would have led to indistinguishable or patched interactions between the haploid-like subtelomeres and both core alleles, which was not observed.

Allele-specific premature termination codons (PTCs) affect translation but not transcript levels

To investigate the biological significance of haplotype variations, we annotated each of the haplotypes independently and characterized a selection of allelic differences. In total, we detected 8030 and 8007 genes for haplotypes A and B, respectively. Both haplotypes contained more annotated genes than the collapsed Tb427v10 genome assembly (Supplementary Figure S1B), suggesting that haplotype phasing corrected errors present in the collapsed assembly. In order to identify specific differences in the annotated proteome between haplotypes, and in comparison to the collapsed Tb427v10 assembly, we performed reciprocal all versus all BLAST searches between the three assemblies and analysed the size ratio to best hits, similar to our analysis following the initial error correction.

First, to understand why some gene ORFs were of the expected length only after haplotyping, we investigated several exemplary genes. We found that one reason for ORF truncations in the collapsed assembly was the presence of internal repeats of varying length between the alleles, leading to frameshifts during collapsing. Another apparent reason for ORF truncation in some genes was the presence of nearby allelic variants, which generated a shift in sequence homology between the alleles, leading to wrong read alignments and the introduction of INDEL errors in the collapsed assembly (e.g. see Supplementary Figure S8). These findings highlight the importance of allele phasing for the accurate identification of protein-coding genes in heterozygous organisms.

We next investigated whether allelic variants were introducing PTCs in protein-coding genes, leading to different CDS length between the alleles. We analysed the reciprocal BLAST results searching for genes for which the CDS in one of the alleles was >10% smaller or was directly absent. Based on this analysis, we identified 123 putative genes with different CDS lengths between the alleles (Supplementary Table S3). Finally, to investigate the implications of PTCs on gene expression, we analysed allele-specific transcript and translation levels for genes with an allele-specific PTC. We used RNA-seq and ribosome profiling data from a Lister 427 mammal-stage-adapted clone (39) and quantified the number of reads mapping to each allele for each variant position in the genes. Two examples are illustrated in Figure 6. Gene Tb427_090051600 (v10), annotated as hypothetical protein, encodes a protein of 379 aa in haplotype 'B', identical to its orthologs in other *T. brucei* strains, while haplotype 'A' contains a 1 bp insertion shortly after the start of the CDS, which changes the reading frame and leads to a short ORF (71 aa) followed by a second ORF (300 aa) just downstream. The second ORF contains a complete functional domain, based on InterPro (53), potentially coding a 'P-loop containing nucleoside triphosphate hydrolase'. Our allele-specific expression analysis indicates that both alleles are fully transcribed, but that for haplotype 'A' only transcripts from the first ORF are being translated, leading to a short, probably non-functional, polypeptide (Figure 6A). Another example is the gene Tb427_110082100. It encodes a putative electron-transfer flavoprotein alpha of 324 aa in haplotype 'A', identical to its orthologs in other *T. brucei*

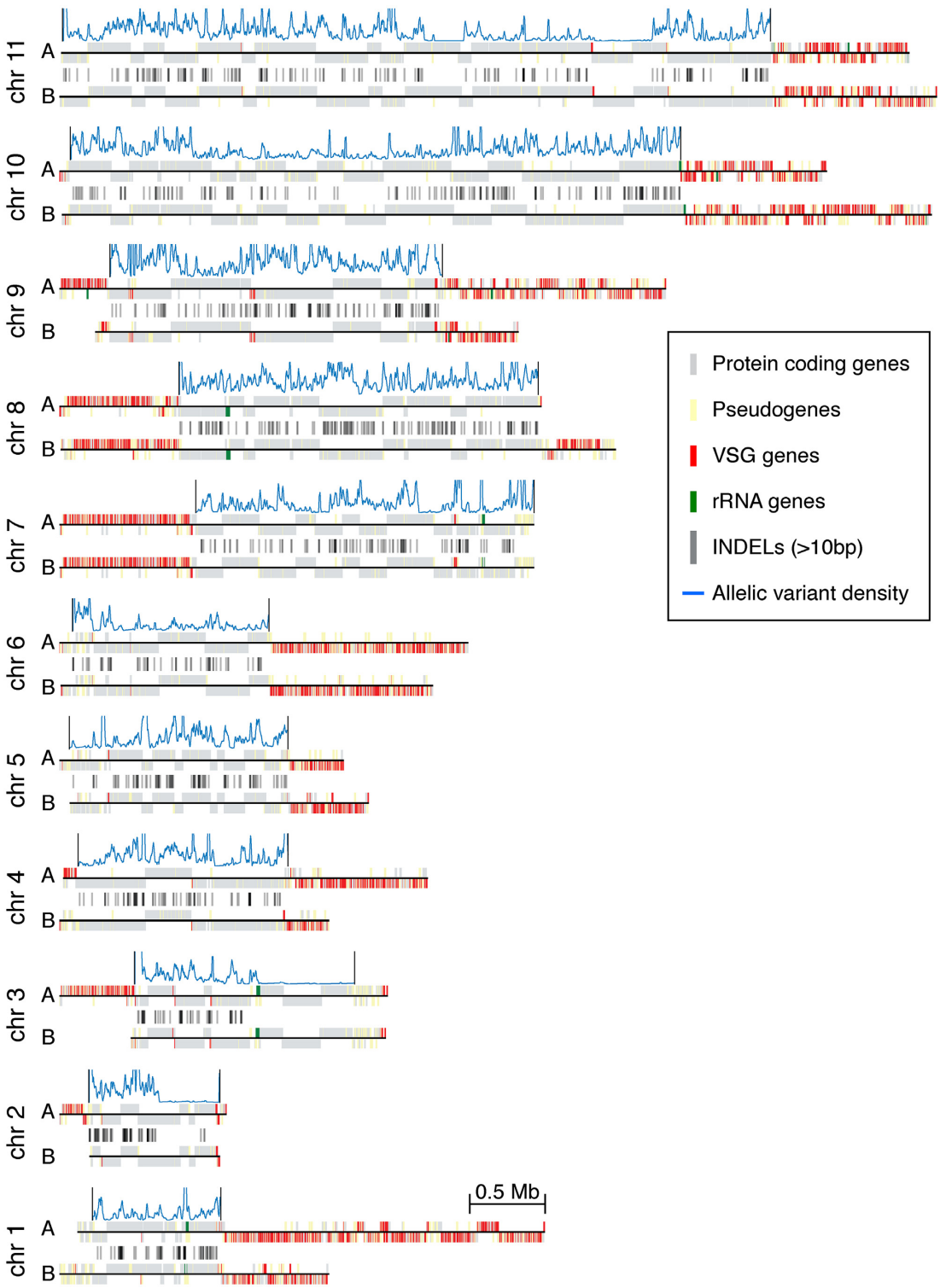

Figure 5. Fully phased *T. brucei* Lister 427 genome. Both alleles are plotted for each chromosome. Protein-coding genes (grey), pseudogenes (yellow), VSG genes (red) and rRNA genes (green) are indicated by vertical lines on top or bottom (depending on the coding strand) of the black lines representing the chromosomal sequence. Variant density is indicated with a blue histogram on top of each chromosomal 'core' region. INDELs (>10 bp) are indicated by dark grey lines between the 'cores'.

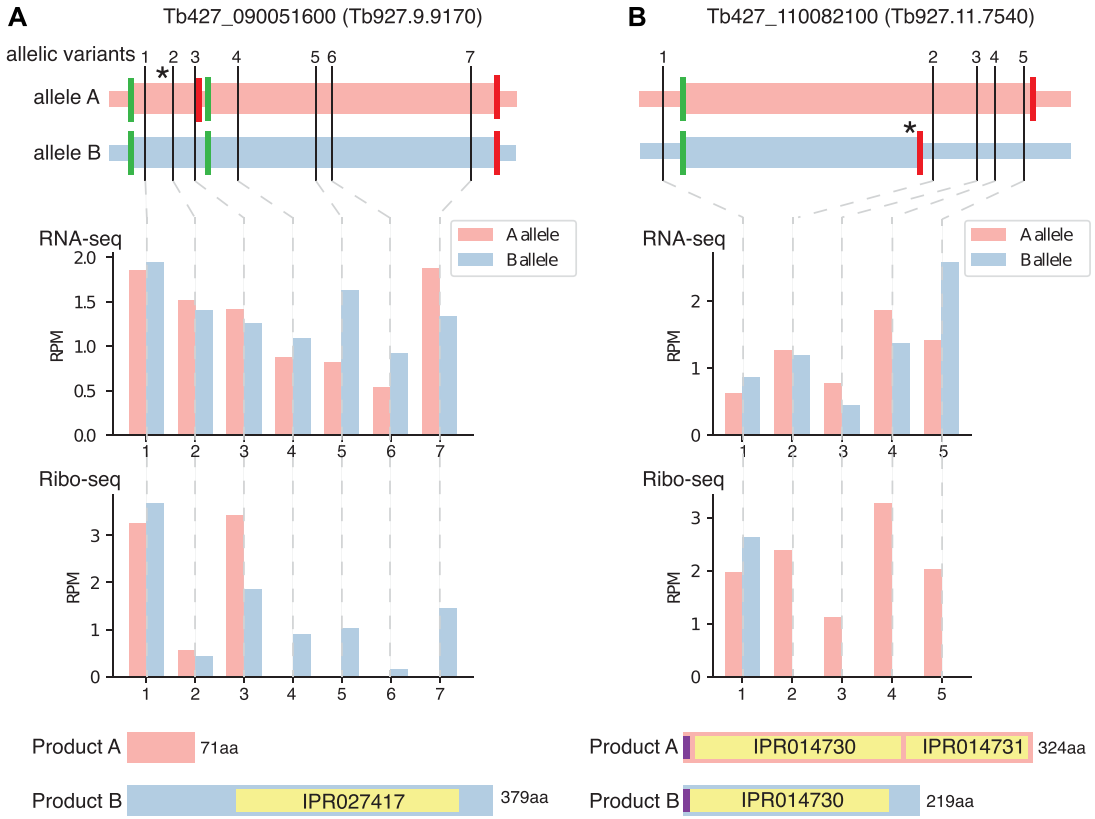

Figure 6. Allele-specific transcript and translation levels in genes with allele-specific PTCs. (**A**, **B**) Examples of genes with allele-specific variants leading to a PTC. In the upper panels, the thin pink and blue horizontal lines represent the sequence of the two alleles, while the thicker line on top indicates the ORFs. Variant positions are indicated by black vertical lines and numbered. An asterisk indicates the position of a frameshifting INDEL variant. In green and red vertical lines, start and stop codons are indicated, respectively. The middle panels show RNA-seq and Ribo-seq reads per million counts for both alleles in each of the variant positions. The lower panels show the expected protein size for each allele. Yellow boxes indicate InterPro domains and violet boxes signal peptides.

strains. On haplotype 'B', the gene has a 1 bp insertion towards the middle of the actual ORF, leading to a shorter ORF of 219 aa, without any other possible start codon downstream. The quantification of the reads on the variant positions indicated that both alleles are being transcribed at similar levels. However, as expected, translation of the 3' region only occurs for transcripts of haplotype 'A' (Figure 6B). It seems likely that haplotype 'B' is producing a truncated protein containing the signal peptide and the first flavoprotein domain (IPR014730). Interestingly, electron-transfer flavoproteins are heterodimers composed of an alpha and a beta subunit (54). The alpha subunit contains the domains IPR014730 and IPR014731, while the beta subunit contains only the IPR014730 domain. Thus, our observations raise the possibility that the truncated protein encoded in haplotype 'B' mimics the beta subunit of the electron-transfer flavoprotein.

In summary, for these two genes, as well as for four additional genes with PTCs (Supplementary Figure S9), we observed that transcript levels are unaffected by PTCs, while ribosome occupancy drops directly downstream of the PTCs with no indication of translation being reinitiated. We obtained identical results with independent transcriptome and ribosome profiling data generated from the same strain (Supplementary Figure S10) by another group (55).

Location and completeness of the VSG gene repertoire

T. brucei harbours a vast repertoire of genes coding for VSGs, the major surface antigen (56). While most of the VSG gene repertoire is located in subtelomeric arrays of the 11 megabase chromosomes (19), a part of it is located in so-called minichromosomes, 30–150 kb long chromosomes (57). The relevance of having some VSG genes located on minichromosomes is unclear. To help elucidate the significance of minichromosomal VSG genes, we decided to identify minichromosomes among the unassembled contigs. We searched for specific minichromosomal signatures (33) and for minichromosomal VSG genes previously identified in this strain (25), leading to the identification of 60 minichromosomal contigs. On them we found a total of 51 VSG genes, all of them located in different contigs, except for one case where we found 2 VSG genes on the same contig. The number of VSG genes in the different genomic locations of Tb427v10 genome assembly is shown in Table 1. Previous studies based on different genome assemblies indicate that a very low portion of the VSG gene repertoire is fully functional, i.e. most of them are pseudogenes or gene fragments (18,25,58). We analysed the 2872 VSG genes annotated, searching for the presence of required features for a VSG to be functional (e.g. peptide signal, GPI

Table 1. VSG location and completeness: number of VSGs in the different genomic locations in the *T. brucei* Lister 427 genome assembly and the number of VSGs predicted to be fully functional

Location	No. of VSGs	Fully functional
Subtelomeric	2634	346
Chromosomal core	27	3
Minichromosomal	51	43
Bloodstream form expression site (telomeric)	13	13
Bloodstream form expression site (non-telomeric)	4	0
Metacyclic form expression site	8	7
Unassigned contig	135	21
Total	2872	433

anchor signal) and found that only 433 (15%) can be considered as fully functional, in line with previous estimations for *T. brucei* Lister 427 (25). Interestingly, as shown before, we found that minichromosomes are enriched in fully functional VSGs (43 out of 51, 84%). The complete list of VSG genes, their location in the genome and which ones are predicted to be fully functional is available in Supplementary Table S4.

DISCUSSION

The most frequent errors in long-read sequencing technologies are INDELs (41). This has been shown to particularly affect the accuracy of protein-coding gene annotation, by introducing frameshifts and premature stop codons (42). Here, we implemented an error-correction step to the *T. brucei* Lister 427 genome assembly using short-read error-corrected PacBio reads. To evaluate the relevance of the correction process, we compared the protein-coding gene annotation before and after correction. We observed a considerable improvement, correcting >500 protein-coding genes, representing 7% of the whole proteome. Our results suggest that the application of error-correction steps is crucial to achieve high-quality genome assemblies and that, as has been shown before for other organisms (59), measuring protein annotation accuracy is a very informative way to assess it. We believe that a process similar to the one we performed should be applied to correct available non-hybrid genome assemblies from long-read technologies. In the meantime, caution should be taken when analysing protein-coding genes in long-read-based assemblies, especially when observing apparently absent genes, pseudogenes or truncated genes compared to closely related organisms. For future genome assemblies, this problem could be considered from the beginning by either implementing a hybrid strategy in the assembly (e.g. generating short-read error-corrected long reads and using them as input in the assembly) or using the latest HiFi reads from PacBio, which have been shown to reduce INDEL errors (60).

The average heterozygosity of the *T. brucei* Lister 427 genome was 4.2 variants per kb. This is almost twice the number predicted for the *T. brucei* TREU 927 genome (2.2 variants per kb) (61). In comparison, heterozygosity in *Leishmania* is in general very low (62), ranging from almost complete homozygosity to 1.6 variants per kb in an extremely heterozygous outlier from *Leishmania donovani* (63). In *Trypanosoma cruzi*, heterozygosity was shown to be around 2 variants per kb (64), with the exception of strains belonging to hybrid lineages, where heterozygosity can reach levels of 20 variants per kb (65). Importantly, the average heterozygote variant density within CDSs in the *T. brucei* Lister 427 genome was 2.5 variants per kb and almost 50% of the CDSs had two or more phased variants (>25% had four or more), making our phased genome a very interesting template to study genome-wide allele expression bias. A phased genome with similar variant density (66) was successfully used to study allele-specific expression control in the fungal pathogen *Candida albicans* (67).

Long regions with LOH were identified in different chromosomes of the Lister 427 genome. Most of them were conserved across clones laboratory-adapted to mammal-stage and insect-stage conditions, suggesting that the recombination events leading to them occur prior to clonal adaptation. LOH regions are thought to arise after 'gene conversion' compensatory mechanisms for counteracting deleterious mutations in asexual species (68). However, all subspecies of *T. brucei*, with the exception of the human infective *T. brucei gambiense* Group 1, have the ability to undergo sexual reproduction (69). Thus, the high number of LOH regions identified in the Lister 427 genome is unexpected. A plausible explanation is that the recombination events leading to the LOH regions occurred in the early phases of cell culture, possibly playing a role in the adaptation to laboratory conditions. An ancestral line of present Lister 427 clones was shown to be fly transmissible and able to undergo meiosis and genetic exchange (48,69), while all the Lister 427-derived clones we analysed are monomorphic, i.e. lost the ability to differentiate into other forms. Sequence comparison among this ancestral line and the current ones might shed light on the origin of the LOH regions and allow to pinpoint the genetic changes responsible for losing the ability to complete the natural development cycle.

We identified *T. brucei* Lister 427 clones with trisomy in 1 or 2 of the 11 chromosomes. This is, to our knowledge, the first time that aneuploid *T. brucei* clones have been described. Chromosomal aneuploidy was shown to be extensive in other kinetoplastids, such as *Leishmania* (63,70,71) or *T. cruzi* (72), but there was still no evidence that this could occur in Kinetoplastida from the Salivarian evolutionary branch. However, given that we only analysed clones adapted to laboratory conditions, we cannot rule out that the aneuploidies observed are a laboratory artefact, not occurring in the natural life cycle of *T. brucei*. Indeed, recent efforts analysing gDNA-seq data from multiple field isolates of *T. brucei* could not detect any aneuploidy (73). Despite the fact that the observed aneuploidies could be only an adaptation to laboratory conditions, it is interesting that different chromosomal trisomies were selected in clones adapted to different life cycle stages. Stage-specific aneuploidies were previously observed in *L. donovani* (74). This suggests that either genes located in distinct chromosomes promote growth in the alternative life cycle stages or the tolerance for overexpression of genes in distinct chromosomes is different, or, probably, a combination of both. Data from other laboratory-adapted clones and field isolates may shed light on this matter. In fact, while this manuscript was under revision, an article reporting the sequencing of two

T. brucei field isolates from cows was made available on bioRxiv (75). The authors found that one of the strains was triploid for chromosome 5, the same aneuploidy we observed for Lister 427-derived clones adapted to mammal-stage conditions. These findings suggest that this particular aneuploidy is positively selected for growth in this condition and is occurring not only in cell culture adapted cells but also in the field.

The availability of transcriptomic data from aneuploid clones allowed us to assess the impact of ploidy on transcript levels. We observed a proportional increase of mRNA transcript levels throughout the trisomic chromosomes, supporting the absence of transcriptional regulation, a general dosage compensation mechanism and global allele-specific expression in *T. brucei*. It would be interesting to analyse quantitative proteomic data in the aneuploid clones to see whether there is a compensatory mechanism at this stage or whether the parasite can deal with a chromosome-wide 50% increase in protein level. In *Leishmania* parasites, several studies have also shown an overall correlation of ploidy and transcript levels (74,76,77), albeit with gene dosage-independent fluctuations for some chromosomes and indications of regulatory mechanisms compensating at the protein expression level (77). On a practical matter, the observation that there is a direct correlation of mRNA transcript level and ploidy indicates that comparative RNA-seq data could be used to directly determine deviations from euploidy in *T. brucei* field isolates.

There was an increase in protein-coding genes annotated for each haplotype after phasing in comparison to the collapsed assembly (Tb427v10), suggesting that this procedure is also important for accurate identification of protein-coding genes. Indeed, we identified several single copy genes that were only corrected after phasing. Furthermore, we identified gene alleles encoding proteins of different sizes and using the haplotype information in combination with RNA-seq and ribosome profiling data we could validate the predicted size differences at the translational level, illustrating the usefulness of generating haplotype-specific assemblies to study allele-specific behaviours.

In many eukaryotes, mRNAs containing PTCs are rapidly degraded by a process named 'nonsense-mediated decay' (NMD), which requires the activity of the RNA helicase Upf1. A previous study analysing this process in *T. brucei* observed a reduction of mRNA levels after introducing PTCs in an endogenous gene and a reporter gene (78). They also found that the level of reduction was higher when the PTC occurred early in the ORF, but found that this reduction was not dependent on Upf1. These findings suggested that if there was a NMD mechanism triggered by PTCs, it was a non-canonical one. Here, in all six genes analysed with one allele containing a PTC, we see no evidence of NMD, i.e. transcript levels remain even for both alleles, even for genes where the PTC occurs very early on. Thus, our findings argue against the presence of a fully active NMD mechanism in *T. brucei*. The different results between the studies regarding the influence of PTC on mRNA levels might be related to gene-specific sequences left exposed when not translated, which could allow binding of proteins inducing degradation in some cases but not in others.

We believe that the availability of an accurate allele-specific genome assembly of *T. brucei* will serve as a better platform for the design of constructs in gene-specific studies but will also allow to address multiple questions at the genome-wide level. Especially in combination with single-cell RNA-seq approaches, it will be possible to address questions about transcriptional kinetics or allele-specific expression. Furthermore, the first allele-specific genome assembly of *T. brucei* will serve as a benchmark in the development of much needed fully automatic pipelines for the genome assembly of field isolates with extreme levels of heterozygosity in order to fully understand the diversity and evolution of this parasite. The development of such pipelines will be useful not only for the assembly of trypanosome genomes, but in general for animal or plant wild isolates with high levels of heterozygosity.

CONCLUSION

Here, we report the generation of the first full allele-specific genome assembly of a protozoan parasite and find that thorough error correction and variant phasing are key to accurate gene annotation in heterozygote organisms. Comparative expression analysis of genes with allele-specific PTCs shows that translation but not the overall transcript level is being affected, suggesting the absence of NMD in *T. brucei* and possibly in other kinetoplastids.

We expect that the full gene allelic resolution provided here will empower not only the analysis of allele-specific expression in this important pathogen, but also the genome-wide identification of sequence determinants of transcript and translation levels. This genome assembly with a striking uneven heterozygosity will serve as an ideal (small yet complex) diploid genome to weigh automatic genome assembly pipelines directed towards wild eukaryotic organisms.

DATA AVAILABILITY

Published datasets used in this study are described in Supplementary Table S1. All the new sequencing data are available in the European Nucleotide Archive under the primary accession number PRJEB43606. The different genome assembly versions (Tb427v10, Tb427v10_phaseA, Tb427v10_phaseB and Tb427v10_phased_diploid) and their respective annotation files can be found in Zenodo (https://doi.org/10.5281/zenodo.4674607). The Tb427v10 genome assembly is already integrated into TriTrypDB (28); the new datasets will be made available soon. All the supplementary tables, as well as the workflows and custom-made Unix Shell, Perl, Python and R scripts, have been deposited in the Zenodo repository made for this manuscript (https://doi.org/10.5281/zenodo.4674607).

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

Localization data for Tb927.2.5860 and Tb927.2.5870 were provided by the TrypTag project (TrypTag.org). We would

like to thank Frank Förster (University of Würzburg, Germany) for generating the error-corrected PacBio reads. We also would like to thank the Bioinformatics Core Facility from the Biomedical Center at Ludwig-Maximilians-Universität Munich for access to the computing cluster, Tobias Straub and all members of the Siegel laboratory for valuable discussions, and Fernán Agüero, Vanessa Luzak and Kirsty McWilliam for critical reading of the manuscript.

FUNDING

European Research Council [3D_Tryps 715466 to T.N.S.]; Humboldt Foundation [Georg Forster Fellowship to R.O.C.].

Conflict of interest statement. None declared.

REFERENCES

1. Bertelli,C. and Greub,G. (2013) Rapid bacterial genome sequencing: methods and applications in clinical microbiology. *Clin. Microbiol. Infect.*, **19**, 803–813.
2. Gordon,D., Huddleston,J., Chaisson,M.J.P., Hill,C.M., Kronenberg,Z.N., Munson,K.M., Malig,M., Raja,A., Fiddes,I., Hillier,L.W. *et al.* (2016) Long-read sequence assembly of the gorilla genome. *Science*, **352**, 6281.
3. Jain,M., Fiddes,I.T., Miga,K.H., Olsen,H.E., Paten,B. and Akeson,M. (2015) Improved data analysis for the MinION nanopore sequencer. *Nat. Methods*, **12**, 351–356.
4. Kaplan,N. and Dekker,J. (2013) High-throughput genome scaffolding from *in vivo* DNA interaction frequency. *Nat. Biotechnol.*, **31**, 1143–1147.
5. Dudchenko,O., Batra,S.S., Omer,A.D., Nyquist,S.K., Hoeger,M., Durand,N.C., Shamim,M.S., Machol,I., Lander,E.S., Aiden,A.P. and Aiden,E.L. (2017) *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, **356**, 92–95.
6. Bickhart,D.M., Rosen,B.D., Koren,S., Sayre,B.L., Hastie,A.R., Chan,S., Lee,J., Lam,E.T., Liachko,I., Sullivan,S.T. *et al.* (2017) Single-molecule sequencing and chromatin conformation capture enable *de novo* reference assembly of the domestic goat genome. *Nat. Genet.*, **49**, 643–650.
7. Zhao,D., Lin,M., Pedrosa,E., Lachman,H.M. and Zheng,D. (2017) Characteristics of allelic gene expression in human brain cells from single-cell RNA-seq data analysis. *BMC Genomics*, **18**, 860.
8. Korlach,J., Gedman,G., Kingan,S.B., Chin,C.-S., Howard,J.T., Audet,J.-N., Cantin,L. and Jarvis,E.D. (2017) *De novo* PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *GigaScience*, **6**, 1–16.
9. Koren,S., Rhie,A., Walenz,B.P., Dilthey,A.T., Bickhart,D.M., Kingan,S.B., Hiendleder,S., Williams,J.L., Smith,T.P.L. and Phillippy,A.M. (2018) *De novo* assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.*, **36**, 1174–1182.
10. Wang,B., Tseng,E., Baybayan,P., Eng,K., Regulski,M., Jiao,Y., Wang,L., Olson,A., Chougule,K., Buren,P.V. *et al.* (2020) Variant phasing and haplotypic expression from long-read sequencing in maize. *Commun. Biol.*, **3**, 78.
11. Fan,J., Hu,J., Xue,C., Zhang,H., Susztak,K., Reilly,M.P., Xiao,R. and Li,M. (2020) ASEP: gene-based detection of allele-specific expression across individuals in a population by RNA sequencing. *PLoS Genet.*, **16**, e1008786.
12. Dréau,A., Venu,V., Avdievich,E., Gaspar,L. and Jones,F.C. (2019) Genome-wide recombination map construction from single individuals using linked-read sequencing. *Nat. Commun.*, **10**, 4309.
13. Leitwein,M., Duranton,M., Rougemont,Q., Gagnaire,P.-A. and Bernatchez,L. (2020) Using haplotype information for conservation genomics. *Trends Ecol. Evol.*, **35**, 245–258.
14. Zhang,X., Wu,R., Wang,Y., Yu,J. and Tang,H. (2019) Unzipping haplotypes in diploid and polyploid genomes. *Comput. Struct. Biotechnol. J.*, **18**, 66–72.
15. Garrison,E. and Marth,G. (2012) Haplotype-based variant detection from short-read sequencing. arXiv doi: https://arxiv.org/abs/1207.3907v2, 20 July 2012, preprint: not peer reviewed.
16. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M. *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
17. Edge,P., Bafna,V. and Bansal,V. (2017) HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.*, **27**, 801–812.
18. Berriman,M., Ghedin,E., Hertz-Fowler,C., Blandin,G., Renauld,H., Bartholomeu,D.C., Lennard,N.J., Caler,E., Hamlin,N.E., Haas,B. *et al.* (2005) The genome of the African trypanosome *Trypanosoma brucei*. *Science*, **309**, 416–422.
19. Müller,L.S.M., Cosentino,R.O., Förstner,K.U., Guizetti,J., Wedel,C., Kaplan,N., Janzen,C.J., Arampatzi,P., Vogel,J., Steinbiss,S. *et al.* (2018) Genome organization and DNA accessibility control antigenic variation in trypanosomes. *Nature*, **563**, 121.
20. Hackl,T., Hedrich,R., Schultz,J. and Förster,F. (2014) proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics*, **30**, 3004–3011.
21. Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.
22. Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv doi: https://arxiv.org/abs/1303.3997, 26 May 2013, preprint: not peer reviewed.
23. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000,Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
24. Steinbiss,S., Silva-Franco,F., Brunk,B., Foth,B., Hertz-Fowler,C., Berriman,M. and Otto,T.D. (2016) Companion: a web server for annotation and analysis of parasite genomes. *Nucleic Acids Res.*, **44**, W29–W34.
25. Cross,G.A.M., Kim,H.-S. and Wickstead,B. (2014) Capturing the variant surface glycoprotein repertoire (the VSGnome) of *Trypanosoma brucei* Lister 427. *Mol. Biochem. Parasitol.*, **195**, 59–73.
26. Siegel,T.N., Hekstra,D.R., Wang,X., Dewell,S. and Cross,G.A.M. (2010) Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites. *Nucleic Acids Res.*, **38**, 4946–4957.
27. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
28. Aslett,M., Aurrecoechea,C., Berriman,M., Brestelli,J., Brunk,B.P., Carrington,M., Depledge,D.P., Fischer,S., Gajria,B., Gao,X. *et al.* (2010) TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res.*, **38**, D457–D462.
29. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
30. Hunter,J.D. (2007) Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95.
31. Liao,Y., Smyth,G.K. and Shi,W. (2019) The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.*, **47**, e47.
32. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
33. Wickstead,B., Ersfeld,K. and Gull,K. (2004) The small chromosomes of *Trypanosoma brucei* involved in antigenic variation are constructed around repetitive palindromes. *Genome Res.*, **14**, 1014–1024.
34. Sloof,P., Bos,J.L., Konings,A.F.J.M., Menke,H.H., Borst,P., Gutteridge,W.E. and Leon,W. (1983) Characterization of satellite DNA in *Trypanosoma brucei* and *Trypanosoma cruzi*. *J. Mol. Biol.*, **167**, 1–21.
35. Armenteros,J.J.A., Tsirigos,K.D., Sønderby,C.K., Petersen,T.N., Winther,O., Brunak,S., von Heijne,G. and Nielsen,H. (2019) SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.*, **37**, 420–423.
36. Gíslason,M.H., Nielsen,H., Almagro Armenteros,J.J. and Johansen,A.R. (2021) Prediction of GPI-anchored proteins with pointer neural networks. *Curr. Res. Biotechnol.*, **3**, 6–13.

37. Wingett,S., Ewels,P., Furlan-Magaril,M., Nagano,T., Schoenfelder,S., Fraser,P. and Andrews,S. (2015) HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research*, **4**, 1310.
38. Servant,N., Varoquaux,N., Lajoie,B.R., Viara,E., Chen,C.-J., Vert,J.-P., Heard,E., Dekker,J. and Barillot,E. (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.*, **16**, 259.
39. Jensen,B.C., Ramasamy,G., Vasconcelos,E.J.R., Ingolia,N.T., Myler,P.J. and Parsons,M. (2014) Extensive stage-regulation of translation revealed by ribosome profiling of *Trypanosoma brucei*. *BMC Genomics*, **15**, 911.
40. Quail,M.A., Smith,M., Coupland,P., Otto,T.D., Harris,S.R., Connor,T.R., Bertoni,A., Swerdlow,H.P. and Gu,Y. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**, 341.
41. Weirather,J.L., de Cesare,M., Wang,Y., Piazza,P., Sebastiano,V., Wang,X.-J., Buck,D. and Au,K.F. (2017) Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*, **6**, 100.
42. Watson,M. and Warr,A. (2019) Errors in long-read assemblies can critically affect protein prediction. *Nat. Biotechnol.*, **37**, 124.
43. Koren,S., Schatz,M.C., Walenz,B.P., Martin,J., Howard,J.T., Ganapathy,G., Wang,Z., Rasko,D.A., McCombie,W.R., Jarvis,E.D. *et al.* (2012) Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nat. Biotechnol.*, **30**, 693–700.
44. Stewart,R.D., Auffret,M.D., Warr,A., Walker,A.W., Roehe,R. and Watson,M. (2019) Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat. Biotechnol.*, **37**, 953–961.
45. Dean,S., Sunter,J.D. and Wheeler,R.J. (2017) TrypTag.org: a trypanosome genome-wide protein localisation resource. *Trends Parasitol.*, **33**, 80–82.
46. Halliday,C., Billington,K., Wang,Z., Madden,R., Dean,S., Sunter,J.D. and Wheeler,R.J. (2019) Cellular landmarks of *Trypanosoma brucei* and *Leishmania mexicana*. *Mol. Biochem. Parasitol.*, **230**, 24–36.
47. Siegel,T.N., Hekstra,D.R., Kemp,L.E., Figueiredo,L.M., Lowell,J.E., Fenyo,D., Wang,X., Dewell,S. and Cross,G.A. (2009) Four histone variants mark the boundaries of polycistronic transcription units in *Trypanosoma brucei*. *Genes Dev.*, **23**, 1063–1076.
48. Peacock,L., Ferris,V., Bailey,M. and Gibson,W. (2008) Fly transmission and mating of *Trypanosoma brucei brucei* strain 427. *Mol. Biochem. Parasitol.*, **160**, 100–106.
49. Cross,G.A.M. and Manning,J.C. (1973) Cultivation of *Trypanosoma brucei* sspp. in semi-defined and defined media. *Parasitology*, **67**, 315–331.
50. Smukowski Heil,C.S., DeSevo,C.G., Pai,D.A., Tucker,C.M., Hoang,M.L. and Dunham,M.J. (2017) Loss of heterozygosity drives adaptation in hybrid yeast. *Mol. Biol. Evol.*, **34**, 1596–1612.
51. Wedel,C., Förstner,K.U., Derr,R. and Siegel,T.N. (2017) GT-rich promoters can drive RNA pol II transcription and deposition of H2A.Z in African trypanosomes. *EMBO J.*, **36**, 2581–2594.
52. Vasquez,J.-J., Hon,C.-C., Vanselow,J.T., Schlosser,A. and Siegel,T.N. (2014) Comparative ribosome profiling reveals extensive translational complexity in different *Trypanosoma brucei* life cycle stages. *Nucleic Acids Res.*, **42**, 3623–3637.
53. Blum,M., Chang,H.-Y., Chuguransky,S., Grego,T., Kandasaamy,S., Mitchell,A., Nuka,G., Paysan-Lafosse,T., Qureshi,M., Raj,S. *et al.* (2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.*, **49**, D344–D354.
54. Roberts,D.L., Frerman,F.E. and Kim,J.-J.P. (1996) Three-dimensional structure of human electron transfer flavoprotein to 2.1-Å resolution. *Proc. Natl Acad. Sci. U.S.A.*, **93**, 14355–14360.
55. Antwi,E.B., Haanstra,J.R., Ramasamy,G., Jensen,B., Droll,D., Rojas,F., Minia,I., Terrao,M., Mercé,C., Matthews,K. *et al.* (2016) Integrative analysis of the *Trypanosoma brucei* gene expression cascade predicts differential regulation of mRNA processing and unusual control of ribosomal protein expression. *BMC Genomics*, **17**, 306.
56. Cross,G.A. (1975) Identification, purification and properties of clone-specific glycoprotein antigens constituting the surface coat of *Trypanosoma brucei*. *Parasitology*, **71**, 393–417.
57. Van der Ploeg,L.H.T., Schwartz,D.C., Cantor,C.R. and Borst,P. (1984) Antigenic variation in *Trypanosoma brucei* analyzed by electrophoretic separation of chromosome-sized DNA molecules. *Cell*, **37**, 77–84.
58. Marcello,L. and Barry,J.D. (2007) Analysis of the VSG gene silent archive in *Trypanosoma brucei* reveals that mosaic gene expression is prominent in antigenic variation and is favored by archive substructure. *Genome Res.*, **17**, 1344–1352.
59. Florea,L., Souvorov,A., Kalbfleisch,T.S. and Salzberg,S.L. (2011) Genome assembly has a major impact on gene content: a comparison of annotation in two *Bos taurus* assemblies. *PLoS One*, **6**, e21400.
60. Wenger,A.M., Peluso,P., Rowell,W.J., Chang,P.-C., Hall,R.J., Concepcion,G.T., Ebler,J., Fungtammasan,A., Kolesnikov,A., Olson,N.D. *et al.* (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.*, **37**, 1155–1162.
61. Jackson,A.P., Sanders,M., Berry,A., McQuillan,J., Aslett,M.A., Quail,M.A., Chukualim,B., Capewell,P., MacLeod,A., Melville,S.E. *et al.* (2010) The genome sequence of *Trypanosoma brucei gambiense*, causative agent of chronic human African trypanosomiasis. *PLoS Negl. Trop. Dis.*, **4**, e658.
62. Rogers,M.B., Downing,T., Smith,B.A., Imamura,H., Sanders,M., Svobodova,M., Volf,P., Berriman,M., Cotton,J.A. and Smith,D.F. (2014) Genomic confirmation of hybridisation and recent inbreeding in a vector-isolated *Leishmania* population. *PLoS Genet.*, **10**, e1004092.
63. Franssen,S.U., Durrant,C., Stark,O., Moser,B., Downing,T., Imamura,H., Dujardin,J.-C., Sanders,M.J., Mauricio,I., Miles,M.A. *et al.* (2020) Global genome diversity of the *Leishmania donovani* complex. *eLife*, **9**, e51243.
64. Franzén,O., Talavera-López,C., Ochaya,S., Butler,C.E., Messenger,L.A., Lewis,M.D., Llewellyn,M.S., Marinkelle,C.J., Tyler,K.M., Miles,M.A. *et al.* (2012) Comparative genomic analysis of human infective *Trypanosoma cruzi* lineages with the bat-restricted subspecies *T. cruzi marinkellei*. *BMC Genomics*, **13**, 531.
65. Ackermann,A.A., Panunzi,L.G., Cosentino,R.O., Sánchez,D.O. and Agüero,F. (2012) A genomic scale map of genetic diversity in *Trypanosoma cruzi*. *BMC Genomics*, **13**, 736.
66. Muzzey,D., Schwartz,K., Weissman,J.S. and Sherlock,G. (2013) Assembly of a phased diploid *Candida albicans* genome facilitates allele-specific measurements and provides a simple model for repeat and indel structure. *Genome Biol.*, **14**, R97.
67. Muzzey,D., Sherlock,G. and Weissman,J.S. (2014) Extensive and coordinated control of allele-specific expression by both transcription and translation in *Candida albicans*. *Genome Res.*, **24**, 963–973.
68. Weir,W., Capewell,P., Foth,B., Clucas,C., Pountain,A., Steketee,P., Veitch,N., Koffi,M., De Meeûs,T., Kaboré,J. *et al.* (2016) Population genomics reveals the origin and asexual evolution of human infective trypanosomes. *eLife*, **5**, e11473.
69. Peacock,L., Bailey,M., Carrington,M. and Gibson,W. (2014) Meiosis and haploid gametes in the pathogen *Trypanosoma brucei*. *Curr. Biol.*, **24**, 181–186.
70. Mannaert,A., Downing,T., Imamura,H. and Dujardin,J.-C. (2012) Adaptive mechanisms in pathogens: universal aneuploidy in *Leishmania*. *Trends Parasitol.*, **28**, 370–376.
71. Negreira,G.H., Monsieurs,P., Imamura,H., Maes,I., Kuk,N., Yagoubat,A., Broeck,F.V.d., Sterkers,Y., Dujardin,J.-C. and Domagalska,M.A. (2020) Exploring the evolution and adaptive role of mosaic aneuploidy in a clonal *Leishmania donovani* population using high throughput single cell genome sequencing. bioRxiv doi: https://doi.org/10.1101/2020.03.05.976233, 06 March 2020, preprint: not peer reviewed.
72. Reis-Cunha,J.L., Baptista,R.P., Rodrigues-Luiz,G.F., Coqueiro-Dos-Santos,A., Valdivia,H.O., de Almeida,L.V., Cardoso,M.S., D'Ávila,D.A., Dias,F.H.C., Fujiwara,R.T. *et al.* (2018) Whole genome sequencing of *Trypanosoma cruzi* field isolates reveals extensive genomic variability and complex aneuploidy patterns within TcII DTU. *BMC Genomics*, **19**, 816.
73. Almeida,L.V., Coqueiro-dos Santos,A., Rodriguez-Luiz,G.F., McCulloch,R., Bartholomeu,D.C. and Reis-Cunha,J.L. (2018) Chromosomal copy number variation analysis by next generation sequencing confirms ploidy stability in *Trypanosoma brucei* subspecies. *Microb. Genom.*, **4**, e000223.

74. Dumetz,F., Imamura,H., Sanders,M., Seblova,V., Myskova,J., Pescher,P., Vanaerschot,M., Meehan,C.J., Cuypers,B., Muylder,G.D. *et al.* (2017) Modulation of aneuploidy in *Leishmania donovani* during adaptation to different *in vitro* and *in vivo* environments and its impact on gene expression. *mBio*, **8**. e00599-17.
75. Mulindwa,J., Ssentamu,G., Matovu,E., Marucha,K.K., Aresta-Branco,F., Helbig,C. and Clayton,C. (2021) The effect of *in vitro* culture on unicellular eukaryotes: adaptation of *Trypanosoma brucei brucei* bloodstream forms results in gene copy-number changes. bioRxiv doi: https://doi.org/10.1101/2021.06.10.446608, 10 June 2021, preprint: not peer reviewed.
76. Barja,P.P., Pescher,P., Bussotti,G., Dumetz,F., Imamura,H., Kedra,D., Domagalska,M., Chaumeau,V., Himmelbauer,H., Pages,M. *et al.* (2017) Haplotype selection as an adaptive mechanism in the protozoan pathogen *Leishmania donovani*. *Nat. Ecol. Evol.*, **1**, 1961–1969.
77. Piel,L., Rajan,K.S., Bussotti,G., Varet,H., Legendre,R., Proux,C., Douche,T., Gianetto,Q.G., Chaze,T., Vojtkova,B. *et al.* (2021) Post-transcriptional regulation of *Leishmania* fitness gain. bioRxiv doi: https://doi.org/10.1101/2021.03.22.436378, 22 March 2021, preprint: not peer reviewed.
78. Delhi,P., Queiroz,R., Inchaustegui,D., Carrington,M. and Clayton,C. (2011) Is there a classical nonsense-mediated decay pathway in trypanosomes? *PLoS One*, **6**, e25112.