Repetitive Elements in the genome of *Nomia melanderi*

Methods

We detected and quantified repetitive elements in the genome of *Nomia melanderi* (Nme) *de novo* from short reads, and annotated the genome sequence assemblies (contigs and scaffolds >= 500 bp; Nme: n=7918, L=301.741 Mbp). The short read based approach used a random subsample of 5 million short reads from Nme to analyze with DnaPipeTE 1.1 (Goubert et al. 2015) and following Stolle et al. (2018) by depleting reads from mitochondrial sequences to avoid biasing the detection of highly repetitive sequences. For mitochondrial sequence depletion, we first aligned reads to the respective genome sequence assembly (bwa-mem, Li 2013) and identified scaffolds and contigs as potential mitochondrial by assessing read depth (bedtools, Quinlan & Hall 2010, cutoff: 500x coverage), assuming the number of sequenced mitochondrial copies is much higher than for the nuclear genome. The identified high coverage contigs and scaffolds were further analyzed for sequence similarity (blastn v2.2.28+) to the mitogenome of the closest available bee species, the halictid *Halictus rubicundus* (KT164656.1). Contigs and scaffolds identified as putatively mitochondrial (Nme: scaffold235256, scaffold241193, scaffold252191, scaffold252994, scaffold257806) were used to deplete the short reads from mitochrondrial sequences by alignment (bwa-mem) to these scaffolds, and extraction of unmapped reads (bedtools). The extracted forward reads were used for repeat analysis in DnaPipeTE (5 iteration per run) with different numbers of reads further subsampled to represent a genome sequence assembly length coverage of 0.20x to 0.40x (steps of 0.05x). This series of repeat content estimates allowed to overcome the technical limitation of assemblying repetitive elements at too low coverage and thus determine the input data amount at which the total genomic repeat content remained stable (without further increase). The final set of repetitive elements for Nme (from 0.30x) was then used to annotate the genome sequence assembly of Nme (RepeatMasker v4.0.7, 10% sequence divergence cut-off, Smit AFA, Hubley R, Green P: RepeatMasker Open-4.0 http://www.repeatmasker.org).

Results

Nme

In a genome-assembly independent approach using short reads and DnaPipeTE, we assembled 54236 repetive elements (total length 19.61 Mb) representing the repetitive content of 37.5% of the Nme genome (Suppl. File “S3.Nme.Repeats.tar.gz”: Nme.TE.fa, Nme.TE.gff) (Fig. Repeats1). While annotated transposable elements (7866 repeats were annotated, total length 3.28 Mb) from all major groups are present among the repetitive elements, elements of unknown type are the three by far most abundant repetitive sequences, representing about a third of all repeats (no similarities to known repetitive elements, conserved domains or sequences in NCBI’s non-redundant nt database). We detected all major groups of repeats (Fig. Repeats2). Of annotated transposable elements, LINE retrotransposons were the most abundant, followed by LTR retrotransposons and small amounts of DNA or other transposons. Of LTRs, Gypsy were most frequent, followed by Copia and BelPao as well as very few ERV/Retrovirus, Penelope or DIRS annotations. LINE were mostly represented by I and Jockey elements, additionally by few L1, R2 and RTE. DNA elements were mostly represented by Tc1-Mariner, PiggyBac, hAT and Kolobok families, additionally by few Sola, P, PIF-Harbinger, Merlin, CACTA, Mutator, Transib and Chapaev. Some annotations suggest the presence of Crypton, Helitron and Maverick elements as well as 5S/tRNA SINE (Suppl. File “S3.Nme.Repeats.tar.gz”: Nme.basepairs.by.type.txt, Nme.TE.groups.counts.txt, Nme.TE.elements.counts.txt, Nme.TE.RM.annotation.report.txt).

A majority of the detected retroelements show little sequence divergence, indicating recent activity, particularly Gypsy (LTR), Copia (LTR), I (LINE) and R2 (LINE) (Fig. Repeats3).

Annotation of the genome sequence assembly yielded 25.93 Mbp of masked sequences (8.59% at 10% sequence divergence), less than the repetitive fraction of >37% inferred by DnaPipeTE (even at a 20% sequence divergence threshold, only 43.36 Mbp (14.37%) were masked), suggesting that a

substantial fraction of the repetitive part of the genome is not part of the genome sequence assembly. This is likely due to the technical limitations to assembly repetitive elements from short reads.

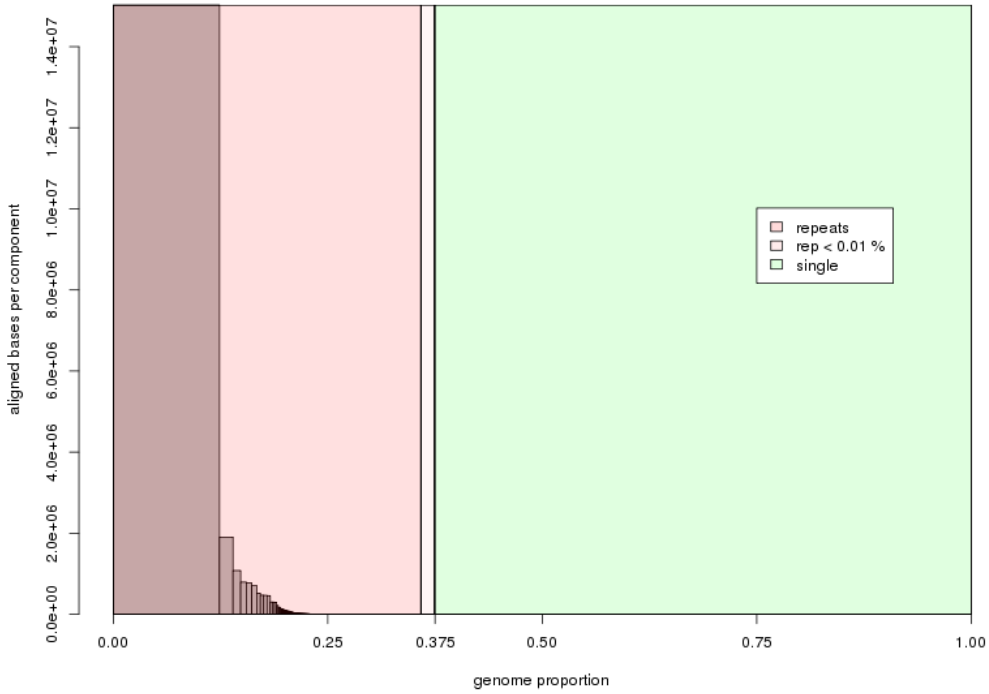

Fig. Repeats1. Total repeat content in the Nme genome, based on short reads representing 0.30x reads coverage. Each bar represents a single repetitive element, with its width showing the total genomic fraction and its height showing the number of bases aligned to this element.

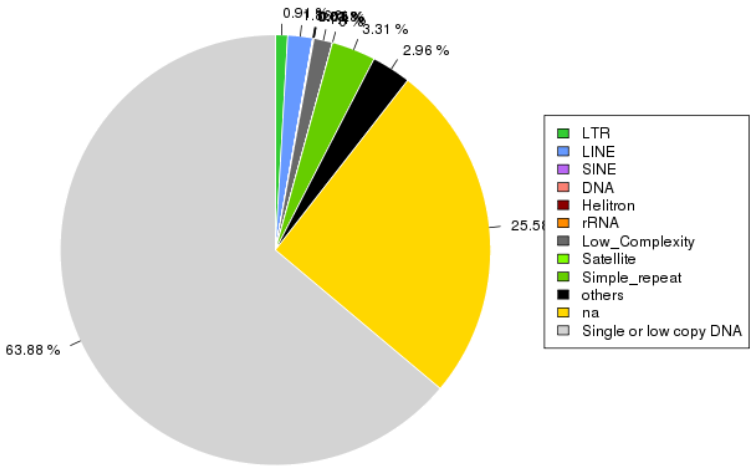

Fig. Repeats2. Types of repetitive elements found in the Nme genome and their proportions.

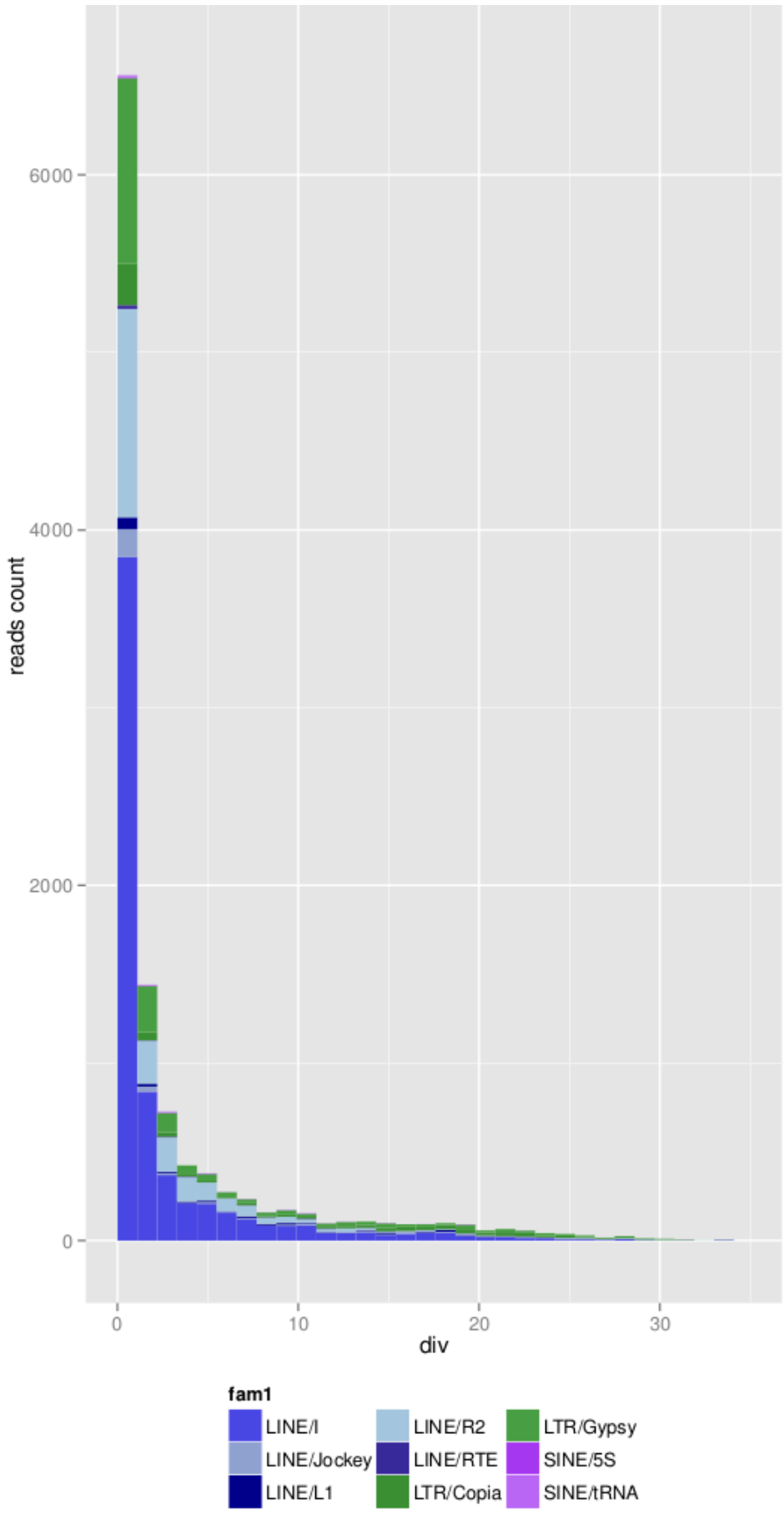

Fig. Repeats3. Counts of reads from Nme transposable families by sequence divergence (div).

References

Flutre T, Duprat E, Feuillet C, Quesneville H: Considering transposable element diversification in de novo annotation approaches. PloS one 2011, 6:e16526.

Goubert C et al. 2015 De novo assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti). Genome Biol Evol. 2015 7(4):1192-205. doi: 10.1093/gbe/evv050.

Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN]

Quinlan AR, Hall IM (2010): BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15; 26(6): 841–842.

Stolle E, Pracana P, Howard P, Paris CI, Brown SJ, Castillo-Carrillo CA, Rossiter SJ, Wurm Y. 2018. Degenerative expansion of a young supergene, bioRxiv. doi: https://doi.org/10.1101/326645

Suppl. File "S3.Nme.Repeats.tar.gz" containing files:

S3.Nme.Repeats.report.pdf

Nme.TE.fa	Nme repetitive elements (fasta sequences)
Nme.TE.gff	Nme sequence assembly annotation (repetitive elements, gff)
Nme.basepairs.by.type.txt	Nme basepairs per repeat type
Nme.TE.groups.counts.txt	Nme counts per TE group
Nme.TE.elements.counts.txt	Nme counts per TE elements
Nme.TE.RM.annotation.report.txt	Nme repeats annotated with RepeatMasker