Supplemental Methods

Genome sequencing

Paired end (PE) and mate pair (MP) libraries were prepared using Illumina library preparation protocols. Whole genome DNA sequence information was generated from the libraries using Illumina Genome Analyser (GA) and HiSeq2000 platforms. For the male *E. lutescens*, PE libraries had insert sizes of 175 bp, 300 bp, and 500 bp, and the MP libraries had insert sizes of 2.5 kbp, 5 kbp, and 10 kbp. The libraries were sequenced to a total of 321 Gbp.
For the female *E. lutescens* and *E. talpinus*, PE libraries with an insert size of 400 bp were prepared, and sequenced with a lower coverage, to a total of 84 Gbp and 129 Gbp, respectively (Supplemental Table S1).
Reads were trimmed using their Phred quality score, and reads with a probability error higher than 0.05 (quality score Q lower than 13) were excluded. Moreover, reads shorter than 15 bp and reads with N's were removed. Together, this guarantees a high quality of the sequence reads used for the assembly.

Genome assembly using ABySS

De novo sequence assembly was performed, based on the *de Bruijn* graph algorithm, using ABySS 1.2.5 software (Simpson et al. 2009) on the *Huygens* supercomputer equipped with 512 Gb RAM and 64 dual core processors (SARA, Amsterdam, The Netherlands). First, all possible substrings of length *k* (aka *k*-mers) were generated from the sequence reads, and the *k*-mer data set was processed to remove read errors and to build contigs. To this end, different *k*-mers (29, 35, 45, 51, 55, 59, 61, 63 bp) of the PE reads were tested, and it was found that a *k*-mer of 55 bp gave the highest N50, the longest contigs, and the lowest number of contigs. Using unique *k*-mers of the PE reads, the male *Ellobius lutescens* genome was reconstructed to a length of 2.35-2.45 Gbp. The *k*-mers (55 bp) were then loaded into the distributed *de Bruijn* graph, and further processed to assemble 9.1 million contigs. Following this, the *k*-mers were aligned to the contigs using KAligner. More than 90% of the *k*-mers were aligned to the contigs, with the correct PE read insert size estimated before sequencing and orientation. This was followed by further processing to resolve ambiguities between contigs and the merging of contigs using PE information. The distance information from the PE reads was then used to build the final contigs. The final contigs obtained as described above were further processed to generate scaffolds using SSPACE premium version 1.0 (Boetzer et al. 2011). Distance information available from both the PE and MP reads was applied, sequentially from the shortest to the longest inserts, to generate scaffolds. However, distance information of the 10 kbp reads were excluded from this scaffolding process, in view of errors in sequencing MP with long inserts.

Genome assembly using CLCbio

CLCbio (https://www.qiagenbioinformatics.com/), is less demanding regarding computing power, compared to ABySS, and we used a computer with 4 dual core processors and 48 Gb of RAM. Scaffolds were generated using SSPACE. Different sequence read combinations were tested to reach the optimum assembly, and the genome assembly statistics are presented in Supplemental Tables S4-S6.

Construction of super-scaffolds

To obtain a single molecule restriction map of *E. lutescens,* 97 high density MapCards were collected using *Kpn*I enzyme. On average, ~83,000 molecules were marked up on each card, and a total of 2.38 million molecules larger than 250kb were used in the Genome-BuilderTM analysis. Data summary is shown in Supplemental Table S2. The assembled *E. lutescens*

genome was compared to the single-molecule restriction maps, and Genome-BuilderTM was run for 4 and 8 iterations respectively. Four iterations is the standard which was used in benchmarking the software. Eight iterations is thought to provide more true joints, but may also increase the number of false joints.

Genome quality and gene prediction

A number of known *E. lutescens* genes, sequenced and banked before, were found to be present in the assembly with 99-100% sequence alignment (Supplemental Table S7). The validity of the assembled genome was also confirmed by perfect alignment with *E. lutescens* BACs, which we sequenced separately, and by different PCR primer sets yielding amplified DNA fragments of the expected size (data not shown). The GC content of the *E. lutescens* genome is 42%, identical to that of the mouse genome (Mouse Genome Sequencing et al. 2002). Interspersed repeats and low complexity DNA of the assembled genomes were screened using RepeatMasker open-3.0 (Smith et al. 1996-2010), using mouse, rat and human repeat databases as a reference. Tandem repeats in the genomes were assigned using Tandem Repeat Finder (Benson 1999). In reference to the mouse, rat and human databases, respectively, 31.6%, 31.5% and 13.5% of *Ellobius lutescens* sequences were masked (Supplemental Table S8). For *E. talpinus* genome, these values are 32.8%, 32.7% and 13.9%, respectively. (Supplemental Table S8). The mouse genome consists for 38% of repeats (Mouse Genome Sequencing et al. 2002), and using the mouse repeats as a reference we detected 32% of repeats in the *E. lutescens* draft genome (Supplemental Table S8). The number of genes in the repeat masked genome of *E. lutescens* was predicted using the *de novo* gene prediction software Augustus (Stanke et al. 2008) with model parameters trained for the human genome. The *E.lutescens* genome may contain 21,864 protein-coding genes, compared to 22,011 in mouse. However, a small proportion of repeats and genes likely is not represented in the present *E. lutescens* draft genome.

Gene hunting and sequence alignment

We searched Y chromosomal genes by combining several methods.

1) Mouse Y chromosomal genes were obtained from Ensembl biomart (genome assembly GRCm38). Homologs of mouse Y genes were searched in *E.lutescens* (male and female) and *E. talpinus* genomes using blastN.

2) Mouse Y chromosomal transcript sequences (CDS) were obtained from Ensembl, and BLAT/blastN was used to search in *E.lutescens* (male and female) and *E. talpinus* genomes.

3) We aligned *E.lutescens* (male and female) and *E. talpinus* genomes to the mouse Y-Chromosome using LASTZ.

4) *E.lutescens* (male and female), *E. talpinus* scaffolds founds in the above procedures were reverse searched in using BLAT and blastN against the mouse genome, and when necessary other mammalian genomes. The final scaffolds where manually inspected to distinguish X and Y homologues.

To search for selected genes, sequences of selected genes and genomic regions from mouse, rat and human were imported from NCBI and Ensembl Genome Browser and were blasted against the *Ellobius sp.* genomes using BLAST. The resulting hits, scaffolds or contigs were reverse searched against the NCBI non redundant (nr/nr) database and the Ensembl database using BLAST and BLAT (BLAST-like Alignment Tool) algorithms, respectively, for confirmation. Sequence alignments of genes or genomic regions were performed using ClustalW 2.1 (Larkin et al. 2007) and T-Coffee (Notredame et al. 2000), and phylogenetic trees were built using software MEGA5 (Tamura et al. 2011).

RNA sequencing, assembly

RNA quality was quantified using Nanodrop and Bioanalyzer. RNA sequencing was performed using Illumina HiSeq 2000. ~ 100 million paired end read (200 million reads) were generated. The RNA sequence quality control was performed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Trimmomatic (Bolger et al. 2014) was used to remove adapter sequences, sequences with bad and trim sequence reads from 5' and 3' (several parameters were tested). Transcriptome assembly was performed using the trinity short-read *de novo* assembly software (Grabherr et al. 2011). Several rounds of *de novo* assembly were performed to find the optimum assembly (long contigs). Transcript contigs generated by *de novo* assembly were annotated using mouse cDNA (mm10). The abundance of each transcript and gene was generated using RNA-Seq by expectation-maximization (RSEM). Further data was processed in R to plot and analyse the expression of genes or expression of genes per chromosome.

Global analysis of d_N and d_S

The genome analysis tool gKaKs (version 1.2.1) (Zhang et al. 2013) can be used for genome-wide computation of nucleotide substitution rates between the genes of a well-annotated genome, such as that of the mouse, and their orthologous sequences in a non-annotated genome, such as that of *E. lutescens*. The best match between mouse coding sequences (CDSs, mm10) and the *E. lutescens* genome was found using BLAT, and bl2seq was used to align every CDS to the BLAT-identified target genome region. After merging the aligned sequences and removing gaps according to the reference CDS codons, PAML (CodeML) (Yang 1997; Yang 2007; Xu and Yang 2013) was used to compute d_N, dS, and d_N/d_S between mouse CDSs and orthologous sequences in the *E.*
lutescens genome. In total 90,956 mouse CDSs were used for the analysis, and d_N, d_S, and d_N/d_S were computed for 40,020 mouse and *E. lutescens* orthologous sequences. Transcripts from the same gene were represented by the longest transcript, and mouse chromosomal annotations of CDSs were used to calculate the mean d_N, d_S, and d_N/d_S for *E. lutescens* orthologous sequences, per mouse chromosome. To remove the effect of pseudogenes and retrotransposed sequences, the *E. lutescens* sequences that map to two or more different mouse chromosomes were removed. In the end, d_N, d_S, and d_N/d_S were calculated for 16,844 autosomal and 456 X-linked mouse-*E. lutescens* orthologs.

SNV analysis

We selected the SNVs with a high genotype quality (GQ above 90) that were found for the assembled male genome compared to all female *E. lutescens* Illumina reads. The numbers of SNVs per chromosome are represented using mouse chromosome annotation, for the *E. lutescens* orthologs.

Immunostaining of meiotic nuclei

Spread nuclei of E. Lutescens spermatocytes were obtained from frozen testis material essentially as described (Peters et al. 1997), except that the thawed tissue was homogenized manually in PBS, followed by direct spreading of 10 microliter aliquots, without hypotonic treatment. Slides were immunostained with a polyclonal rabbit anti-SYCP3 (gift from dr. C. Heyting), and mouse monoclonal anti–γH2AFX (Upstate), Anti-RNA polymerase II CTD repeat YSPTSPS antibody [8WG16] (Abcam), and anti-MLH1 (BD Pharmingen), followed by washings and incubation with the appropriate fluorescent labeled secondary antibody. After final washings in PBS, slides were embedded in Vectashield containing DAPI (4',6-diamidino-2-phenylindole).

Quantitative PCR analyses

For quantitative RT–PCR (RT–qPCR) cDNA was synthesized using standard methods and all samples were analyzed in duplicate in a 25 μl final reaction volume using the BioRad CFX 98 Real-time System. The reaction mixture contained Platinum*taq* DNA polymerase, 10x PCR-buffer, $MgCl_2$, dNTP's (Invitrogen), SYBR Green I (Sigma-Aldrich S9430), primers (for *E. lutescens Actin*, *Eif2s3x*, *Eif2s3y, Zfx, Zfy, Usp9x, Usp9y, or Ssty*) and 1.0 μl of cDNA. (The primer sequences are in Supplemental Table S13).
After incubation at 95°C for 3 minutes, reaction mixtures underwent 40 cycles of 30s at 95°C, 30s at 57°C, and 30s at 72°C. Results were expressed as Cycle threshold (Ct) values. Gene expression levels were normalized over *Actin* gene expression, according to the 2^{-DCT} method (Livak and Schmittgen 2001).

DNA FISH
The DNA was labeled with digoxygenin-dUTP. In addition, we labeled an X-chromosomal BAC with biotin-dUTP. FISH was carried out on slides that were first immunostained with anti-SYCP3 as described above, but not embedded. Subsequently, slides were washed in 2x SSC, at 55°C (10 minutes), 2xSSC at room temperature (4 minutes), dehydrated in ethanol series and air dried. Slides were denatured at 78°C in 50% formamide/10% dextrane/ in 2x SSC pH 7,5 containing the pooled PCR probe and the BAC probe and hybridization was carried out overnight at 37°C. After washings, hybridized PCR probes were detected with fluorescent labeled anti-digoxygenin, and the BAC probe was detected with fluorescent labeled streptavidin. Pachytene nuclei were selected based on the SYCP3 immunostaining and images were processed in Adobe Photoshop.

References:

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**: 573-580.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**: 578-579.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**: 644-652.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947-2948.

Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**: 402-408.

Mouse Genome Sequencing C Waterston RH Lindblad-Toh K Birney E Rogers J Abril JF Agarwal P Agarwala R Ainscough R Alexandersson M et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520-562.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**: 205-217.

Peters AH, Plug AW, van Vugt MJ, de Boer P. 1997. A drying-down technique for the spreading of mammalian meiocytes from the male and female germline. *Chromosome Res* **5**: 66-68.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Res* **19**: 1117-1123.

Smith A, Hubley R, Green P. 1996-2010. Repeat Masker Open-3.0.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**: 637-644.
Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution* **28**: 2731-2739.
Xu B, Yang Z. 2013. PAMLX: a graphical user interface for PAML. *Molecular biology and evolution* **30**: 2723-2724.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences : CABIOS* **13**: 555-556.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**: 1586-1591.
Zhang C, Wang J, Long M, Fan C. 2013. gKaKs: the pipeline for genome-level Ka/Ks calculation. *Bioinformatics* **29**: 645-646.