Inter- and intra-specific genomic divergence in *Drosophila montana* shows evidence for cold adaptation

D. J. Parker[1,2,3], R. A. W. Wiberg[2], U. Trivedi[4], V. I. Tyukmaeva[1], K. Gharbi[4], R. K. Butlin[5], A. Hoikkala[1], M. Kankare[1*], and M. G. Ritchie[2*]

1. Department of Biological and Environmental Science, University of Jyväskylä, P.O. Box 35, FI-40014, Jyväskylä, Finland
2. Centre for Biological Diversity, School of Biology, University of St Andrews, Fife, KY16 9TH, UK.
3. Department of Ecology and Evolution, University of Lausanne, Biophore, Lausanne 1015, Switzerland
4. Edinburgh Genomics, School of Biological Sciences, University of Edinburgh, Edinburgh, UK
5. Department of Animal and Plant Sciences, The University of Sheffield, Sheffield, U.K; and Department of Marine Sciences, University of Gothenburg, 405 30 Göteborg, Sweden

*Corresponding authors: Maaria Kankare: s.maaria.kankare@jyu.fi, Michael Ritchie: mgr@st-andrews.ac.uk

Abstract

The genomes of species that are ecological specialists will likely contain signatures of genomic adaptation to their niche. However, distinguishing genes related to their ecological specialism from other sources of selection and more random changes is a challenge. Here we describe the genome of *Drosophila montana*, the most extremely cold-adapted Drosophila species. We describe the genome, which is similar in size and gene content to most Drosophila species. We look for evidence of accelerated divergence from a previously sequenced relative, and do not find strong evidence for divergent selection on coding sequence variation. We use branch tests to identify genes showing accelerated divergence in contrasts between cold- and warm adapted species and identify about 250 genes that show differences, possibly driven by a lower synonymous substitution rate in cold-adapted species. Divergent genes are involved in a variety of functions, including cuticular and olfactory processes. We also re-sequenced three populations of *D. montana* representing its ecological and geographic range. Outlier loci were more likely to be found on the X chromosome and there was a greater than expected overlap between population outliers and those genes implicated in cold adaptation between Drosophila species, implying some continuity of selective process at these different evolutionary scales.

INTRODUCTION

Comparative genomic analyses provide new insights into our understanding of evolutionary processes by helping to identify genes contributing to adaptive divergence. If strong divergent selection due to environmental adaptation or social interactions, such as sexual selection, act as "barrier loci" by influencing species isolation, then identifying them can help to understand the process of speciation [1,2]. However, accurately identifying such genes is a considerable challenge [3–6].

Genomic analysis has often been hampered by poor understanding of the sources of selection contributing to species divergence. However, the increasing availability of genomes from species with distinct ecological specialisations has allowed improved identification of genes and genomic features involved in adaptations to specific ecologies. Examples of this are the cactophilic *Drosophila* [7,8], Asian longhorn beetles with specialized feeding habits [9], climate-mediated adaptations in honey bees [10] and adaptation to high altitude in humans [11]. Here we describe the genome of *Drosophila montana*, a widely-distributed northern member of the virilis group of *Drosophila*, which shows unique adaptations to seasonally varying environmental conditions prevailing at high latitudes and altitudes. *D. montana* is the most cold-tolerant *Drosophila* species known [12,13], and this cold tolerance, as well as a robust photoperiodic diapause [14] and inducible cold acclimation [15], contribute to its ability to survive through cold and dark winters. The daily and seasonal activity patterns of *D. montana*, and the interactions and neurochemistry of the core circadian clock genes behind these patterns, differs from those of more temperate species such as *D. melanogaster* [16,17]. These features have likely played an important role in allowing *D. montana* to colonize and persist in high-latitude environments [17–19].

D. montana belongs to the *virilis* group of *Drosophila*, which comprises 13 species or subspecies divided into two clades, the virilis and montana phylads, the latter being further split into three lineages [20]. These phylads are thought to have diverged in South Asia during the Early Miocene, after which both of them entered the New World by way of Beringia [21]. The virilis phylad is constrained mostly within the temperate zone, while the montana phylad has expanded into a variety of habitats and spread to higher latitudes [21]. Divergence of the two phylads has been estimated to have occurred 7 [22] to 11 [20] million years ago, while the North American, European and Asian *D. montana* populations have diverged within the last 450,000 to 900,000 years [23]. Interestingly, conspecific *D. montana* populations have been shown to diverge in traits that play a role in ecological adaptation (e.g. Lankinen et al. [24] and Tyukmaeva et al. [25]), male sexual cues and female preferences (e.g. Klappert et al. [26]), and also to show sexual and post-mating pre-zygotic reproductive barriers [27]. Information on potential candidate genomic regions and genes for traits involved in adaptation and sexual selection has been accumulated through QTL analyses [25,28], microarray [29,30] and transcriptome [31–33] studies.

Here we compare the genome of *D. montana* to that of a more temperate species of the same group, *D. virilis*, to identify genes showing evidence of divergent selection. We also compare the divergence of these species at candidate gene sets suspected to evolve rapidly,

85 including genes known to be involved in cold tolerance. We then extend these analyses to
86 multiple species and identify genes showing elevated nonsynonymous substitution counts
87 across cold-tolerant species of *Drosophila* to recover broader patterns in the evolution of
88 insect cold-tolerance genes. In order to examine the continuity of divergence at the
89 intraspecific level, we also use pooled genomic sequencing to detect outliers between three
90 *D. montana* populations known to be divergent in ecology and behaviour [25,34] and to
91 show evidence of sexual isolation and gametic incompatibilities [27]. We compare
92 divergence seen at these different evolutionary levels. Our results thus give a novel insight
93 into genomic patterns of selection-driven divergence at different evolutionary scales, in
94 addition to providing a well-annotated genome for a uniquely cold adapted insect species.
95
96
97

MATERIALS AND METHODS

Samples and Sequencing

Genomic DNA for the *D. montana* reference genome was extracted from an inbred isofemale line originating from Vancouver, Canada (Can3F9) in summer 2003. This line was inbred via full-sib matings for 37 generations, relaxed for 9 generations and maintained on malt food [35] at 19°C in constant light. Quality checked DNA extracted from 210 males using a Gentra Puregene Tissue Kit (Qiagen) was used to produce 3 libraries with different insert sizes: 200bp, 400bp and 3,000bp. The 200bp and 400bp libraries were sequenced using an Illumina HiSeq 2000 at Edinburgh Genomics to produce paired-end reads (101 + 101bp). The 3,000bp library was sequenced using an Illumina MiSeq at The Centre for Genomic Research, University of Liverpool to produce mate-pair reads (101+101bp). This strategy produced 65107854 paired-end reads for the 200bp library, 25618163 paired end reads for the 400bp library and 19020110 mate-pair reads for the 3000bp library. Reads from the 200bp and 400bp libraries were trimmed using scythe [36] to remove adaptors and sickle [37] to quality trim reads (bases with phred quality of <20 were trimmed from the tail end of each read). Reads from the 3,000bp library were trimmed in the same manner, with the addition of a linker sequence removal step.

An initial assembly using reads from the 200bp and 400bp libraries was made using CLC assembly cell (4.0.12). Contigs from this were then blasted (blastN) to two subsets of NCBI's nt database (arthropod and bacteria) with a e-value threshold of 1×10^{-40}. Bit scores of blast hits from the arthropod and bacterial databases were compared for each contig, and any with a higher bit score for bacteria than arthropods were considered to be contaminants (**Supplementary Figure 1**). Reads were mapped to contigs identified as contaminants using BWA (v. 0.7.12) [38] and then the unmapped reads were assembled using CLC assembly cell (4.0.12) (default options, minimum contig length = 200bp). Contigs were then scaffolded using the 3,000bp mate pair library using SSPACE-BASIC-2.0. This assembly contained 68950 scaffolds (N50 = 39341). This assembly was then further screened for contaminants using DeconSeq (v. 0.4.3) [39]. Bacterial (2786) and viral (4359) genomes were downloaded from NCBI on January 20th 2016 and used as the contamination databases in DeconSeq along with the human genome (hg38). The *D. melanogaster* (r6.09) and *D. virilis* (r1.05) genomes were used as retention databases. DeconSeq identified 5208 scaffolds as contaminants, which were removed from our assembly. We then used this assembly for all subsequent analyses. To assess the completeness of our genome assembly we used the CEGMA analysis pipeline (v. 2.4) [40,41] which identifies the presence of 248 conserved eukaryote genes, and the BUSCO pipeline (v.1.22) [42] which identifies the presence of 2675 conserved arthropod genes.

Genome Annotation

Full details of the genome annotation are given in the supplementary methods. Briefly, we used the Maker2 pipeline [43] to first mask putative repeats within the genome, and then used *ab initio* gene predictors SNAP and AUGUSTUS, and gene evidence (from proteins homology and RNA-seq data) to generate gene predictions. Gene predictions from Maker2 were reciprocally blasted to proteins from *D. virilis* (r1.2) with the following cutoffs: e-value < 3 x10^{-13}, query cover > 60% to give reciprocal best blast hits (RBBH). Orthologs for *D. melanogaster*, *D. sechellia*, *D. simulans*, *D. erecta*, *D. yakuba*, *D. ananassae*, *D. persimilis*, *D. pseudoobscura*, *D. willistoni*, *D. mojavensis*, and *D. grimshawi* were then obtained from FlyBase using *D. virilis* FlyBase numbers. Genes without a single ortholog for each species were discarded from multi-species selection analyses (below).

Linkage map construction

For the genetic map construction, we selected 192 samples from a previous QTL study [25], which consisted of two families (four parent individuals and their F2 progeny, females only). We used RAPiD Genomics facilities to develop a set of oligonucleotide probes for 13,975 selected regions in the largest scaffolds of the *Drosophila montana* genome. These probes were used to capture sequence these target loci with 100bp single end reads using HiSeq 2000. A resulting SNP dataset was cleaned with Genotype Checker to eliminate possible errors in pedigree/genotyping [44]. The R/qtl package [45] was used to construct a genetic linkage map after discarding any polymorphic loci that were heterozygous for both parents, duplicated markers, markers showing segregation distortion, and individuals with fewer than 2000 markers. Reads from the 200bp and 400bp genome reference libraries were mapped back to anchored scaffolds using BWA (v. 0.7.12) [38]. Multi-mapping reads were discarded. Since the genome reference libraries were produced from males, X linked regions should have half the coverage of autosomal regions, we used the coverage of these scaffolds to validate our linkage map.

Selection Analyses

Pairwise analysis

To identify protein-coding genes with elevated signatures of selection we estimated pairwise ω (dN/dS) for each gene that had a reciprocal best blast hit (RBBH) to a *D. virilis* gene. The longest coding sequence of each gene and its RBBH ortholog were codon-aligned using PRANK (v.140110) [46], before estimating ω using codeml in PAML (v. 4.8) [47,48]. To determine if any genes showed ω > 1, we compared genes using a Bayesian estimation of ω in codeml (runmode = -3, model = 0, NSsites = 0) [49] with default priors. The p-values were corrected for multiple testing using a strict Bonferroni correction. We further filtered to exclude any genes where estimates of *dN, dS* or ω were greater than 10.

We then compared mean ω values in several candidate gene sets (genes involved in immune function, reproduction, and cold tolerance) against the genomic background. Genes were

184 classified into two 'immune' classes firstly using the GO term 'immune response' from
185 FlyBase (version 6.05) and secondly using orthologs of genes identified as being involved in
186 immune function by Sackton et al. [50]. Next, genes connected to reproduction were
187 classified into several reproductive classes following Haerty et al. [51]: sex and reproduction
188 related genes (SRR), female reproductive tract (FRTP) and seminal fluid proteins (SFP).
189 Finally, cold tolerance genes were classified into two classes with genes differentially
190 expressed in response to cold in *D. montana* and in *D. virilis* [31]. Parker et al. (2015) found
191 that from the differentially expressed genes, 42 were the same in both species but 550 were
192 different, allowing genes to be classified into 'cold tolerance same' and 'cold tolerance
193 different' groups.
194
195 **Multispecies analysis**
196
197 13 species with fully annotated genome sequences available were divided into cold-tolerant
198 and non-cold-tolerant ones; six species with a knockdown temperature <3°C [12] were
199 classified as cold-tolerant, the remainder as non-cold-tolerant **Fig. 1**). To identify genes
200 showing elevated signatures of selection in these species we extracted the longest CDS (N =
201 5,619) for each ortholog and codon-aligned them using PRANK (v.140110) [46]. Sequences
202 were then analysed in codeml from the PAML (v4.8) package [47,48]. Two models were
203 compared; the "null" model (clock = 0; fix_omega = 0, model = 0, NSSites = 0) which assumes
204 a single common value for ω with an alternative model (clock = 0; fix_omega = 0, model = 2,
205 NSSites = 0) which assumes one value of omega for all the cold-tolerant species and a
206 separate value of omega for the non-cold-tolerant species. Nested models were compared
207 using a likelihood ratio test and p-values corrected for multiple testing using a Bonferroni
208 correction. Additionally, results were filtered to exclude sequences with dN, dS or ω > 10.
209 This comparison tests whether there is a different rate of molecular evolution in cold-
210 tolerant species compared to non-cold-tolerant species.
211
212 **Population resequencing**
213
214 For population comparisons we used *D. montana* flies from 3 populations: Oulanka (Finland;
215 66°N), Crested Butte, Colorado (USA; 39°N) and Vancouver (Canada; 49°N). These
216 populations were established from the progenies of fertilized females collected in the
217 summer of 2008 in Oulanka and Vancouver, and in the summer of 2009 in Colorado.
218 Population cages were set up using 20 F3 generation individuals from approximately 20
219 isofemale lines for each population. Population cages were maintained at 19°C in constant
220 light (for more details see Jennings et al. [52]). In March 2013 Genomic DNA was extracted
221 from a pool of 50 females for each population and sequenced at Beijing Genomics Institute
222 using an Illumina HiSeq 2000 to produce paired-end reads (90 + 90bp, insert size = 500bp).
223
224 Sequencing produced 84938118 paired-end reads for Colorado and 82663801 for Oulanka.
225 Two runs for Vancouver resulted in 303365095 reads. Reads were quality trimmed (leading
226 or trailing bases with a phred score of <20, or if two consecutive bases had an average phred
227 score of <32 the read was trimmed at this point) and screened for adaptor sequence using
228 trimmomatic (v. 0.30) [53]. Reads containing adaptor sequence or that had a length of less

than 85 bp after quality trimming, were discarded. Since coverage depth can influence the estimation of allele frequency [54], reads for Vancouver were randomly sampled prior to mapping to the mean number of reads from Colorado and Oulanka. Reads were mapped to the genome assembly using BWA (v. 0.7.12) [38]. Reads with a mapping quality of <20 were then removed, and an mpileup file was produced using samtools (v. 0.1.19) [55]. From this, a sync file was produced using PoPoolation2 pipeline (v 1.201) [56]. Outlier detection was performed on the raw read count data with BayeScan v. 2.1 [57–59], which performs comparably alongside other outlier methods in several simulation studies [60–62]. SNPs were filtered to include only sites with a minimum coverage of 25 and a maximum coverage of 93 (corresponding to the median 10th and 90th percentiles of the population coverage distributions). At the same time, SNPs were only considered if the minor allele had a read count > 4 across all populations. BayeScan was run with 5 pilot runs of 1,000 iterations each followed by a main run of 2,000 iterations, a thinning interval of 10 and a burn in of 1,500. Additionally, three pairwise runs of BayeScan were performed with the same parameters as above. The three pairwise analyses compared Colorado to Vancouver, Vancouver to Oulanka, and Colorado to Oulanka populations, respectively.

Functional Enrichment

To examine functional enrichment of genes for the species level selection analyses and population level F_{ST} scans, we used GOrilla [63]. For the pairwise selection analyses genes were ranked by ω (from high to low and low to high). For the multispecies selection analyses, we ranked genes by p-value and direction so that genes with the lowest p-values and a higher ω in cold-tolerant species were at the top, and genes with lowest p-values and have higher ω in non-cold-tolerant species were at the bottom, allowing us to identify enriched GO terms for genes showing elevated ω in cold-tolerant species. To examine GO terms for genes showing elevated ω in non-cold-tolerant species the list order was simply reversed. For population level analyses genes were ranked by the most significantly differentiated SNP occurring within 1kb, 10kb, or 100kb of a gene for each population. Results from GOrilla were then visualised using ReviGO [64], using the January 2017 version of Gene Ontology.

We used DAVID (v6.8) [65,66] to identify enriched functional groups of genes. A functional group was considered to be significantly enriched if its enrichment score (the geometric mean (in -log scale) of the p-values of the GO terms in the group) was >1 ($p < 0.1$). For the pairwise selection analyses we identified functional clusters for genes occurring in the top and bottom 10% of genes for ω estimates. For the multispecies selection analyses we identified functional clusters for genes that showed a significantly (FDR < 0.1) higher omega in cold-tolerant species or in non-cold-tolerant species separately. For population level analyses we identified functional clusters for genes containing (within 1kb) significantly differentiated SNPs for each population.

To take advantage of the superior annotation of *D. melanogaster* [67], we used *D. melanogaster* orthologs for all of the above function enrichment analyses. For the DAVID

273 analyses the 'background' list used was the subset of *D. melanogaster* genes available for
274 each analysis.
275
276

RESULTS

Genome Sequencing and Assembly

The assembled *D. montana* genome (**Table 1**) has a total length of 183.6 Mb, which falls within the range seen for *Drosophila* species (111-187 Mb), and is similar to that of *D. virilis* (172 Mb), a close relative of *D. montana* with a sequenced genome. CEGMA identified 238 complete orthologs (96%) and 244 partial orthologs (98%) of the 248 CEGMA proteins and BUSCO identified 2457 genes as complete (92%), and failed to identify only 46 (1.7%). RepeatMasker identified that 14.4% of the assembly was composed of repeat elements, the major classes of which were: Simple repeats (4.5%), LTR elements (4.3%), Unclassified (2.9%), and LINEs (1.9%) (**Supplementary Figure 2**). The total percentage of repeat elements identified was around half of that found for related *Drosophila* species (*D. virilis* = 25.9%, *D. mojavensis* = 23.8%, and *D. grimshawi* = 26.1%) likely reflecting the problem of assembling repetitive regions with short reads.

For the genetic map construction, the final dataset contained 5,858 polymorphic SNPs. The median depth of the SNPs in the final dataset was 52.4 and the average missing data rate was 0.003. The initial anaylsis formed five major linkage groups (as expected since *D. montana* has five chromosomes in total). Chromosome number was assigned by blasting genes assigned to the linkage groups to the *D. virilis* genome, which have been localised to chromosomes and is largely syntenic with *D. montana* [28]. While the analysis showed clear linkage groups, the order of markers was not totally resolved, likely due to lack of recombination events among F2 progeny (**Supplementary Figure 3**). The tentative scaffold order and position are given in **Supplementary Table 1**. Using this map we were able to anchor approximately one third of the genome assembly to chromosomes. To validate our linkage map we examined coverage of anchored scaffolds. X-linked regions were found to have approximately half the coverage of autosomal regions, as expected since the reference genome was produced from male-only samples (**Supplementary Figure 4**).

Selection analyses

We obtained 13,683 predicted gene models with a mean coding region length of 1,461 bp. RBBH identified 10898 orthologs with *D. virilis*. Using the FlyBase gene numbers for *D. virilis* for these we then obtained a set of 5619 one-to-one orthologs between *D. montana*, *D. virilis*, *D. melanogaster*, *D. sechellia*, *D. simulans*, *D. erecta*, *D. yakuba*, *D. ananassae*, *D. persimilis*, *D. pseudoobscura*, *D. willistoni*, *D. mojavensis*, and *D. grimshawi*.

We estimated ω (dN/dS) for each of the one-to-one orthologs between *D. montana* and *D. virilis* (**Supplemental Table 20**). No genes had a ω significantly greater than 1 after filtering and multiple-test correction. Comparison of mean ω for several candidate gene sets (genes involved in cold tolerance, immune function, and reproduction) found that none of the candidate genes sets differed significantly from the genomic background (**Fig. 2**). By ranking genes by ω we identified GO terms enriched in genes with relatively high and low ω. For those with high ω we identified 23 enriched GO terms (Biological Processes: Molecular

Functions: Cellular Components = 10:4:9) (FDR < 0.1) (**Supplementary Table 2**). Semantic clustering of these GO terms shows that they fall into the following categories: Reproduction, detection of chemical binding / olfaction, amino sugar metabolism, and chitin binding (**Fig. 3**). DAVID identified 9 functional group clusters (**Supplementary Table 3**) including 2 related to chitin production and 2 related to olfactory functions, congruent with the findings from the single GO term enrichment analysis (above). In addition DAVID also identified 2 clusters involved in: immune defence (C-type lectin domain carrying genes, and Fibrinogen related genes), Transcription factor binding, and a cluster containing genes with either a CAP (cysteine-rich secretory protein) or SCP (Sperm-coating protein) domain. We identified 662 enriched GO terms for genes with low ω between *D. montana* and *D. virilis* (Biological Processes: Molecular Functions: Cellular Components = 485:80:97) (FDR < 0.1). As expected for genes with very low ω the enriched GO terms are consistant with housekeeping roles in the cell (cell cycle control, cell communication, cell developmental process etc.), which are expected to be under strong purifying selection (**Supplementary Table 4, Supplementary Figures 5-7**).

Across the 13 *Drosophila* species we found 250 genes that had significantly different rates of evolution (ω) in cold- and non-cold-tolerant species (**Fig. 1, Supplemental Table 20**). dS was on average lower for cold-tolerant species than for non-cold-tolerant species while dN was very similar (**Fig. 4**, **Supplementary Table 5**). ω was on average greater for cold-tolerant species, probably driven by generally lower values of dS in these species (**Supplementary Table 5**). 203 and 47 genes showed higher values of ω for cold-tolerant and for non-cold-tolerant species, respectively (**Figure 5**). Genes with elevated ω in cold-tolerant species were enriched for 23 GO terms (Biological Processes: Molecular Functions: Cellular Components = 6:10:7) (FDR < 0.1) (**Supplementary Table 6**), which semantically cluster into the following categories: response to drug, male courtship behaviour, olfaction, ion-channel activity, and developmental processes (**Fig. 6**). Of genes with elevated ω in non-cold-tolerant species we identified 50 enriched GO terms (Biological Processes: Molecular Functions: Cellular Components = 34:3:13) (FDR < 0.1) (**Supplementary Table 7**), which semantically cluster into the following categories: proteasome-mediated ubiquitin-dependent protein catabolic process, reproductive processes, response to fungus, animal organ morphogenesis, regulation of biological and cellular processes (**Fig. 6**). Moreover, DAVID identified 11 functional group clusters for genes with significantly higher ω in cold-tolerant species (**Supplementary Table 8**) including: Nucleotide-binding, Olfaction, Transmembrane proteins, Neural development, Leucine-rich repeat containing proteins, GTPase / GTP binding, Cytoskeleton / Microtubule, and Ion Transport. Finally, DAVID identified 3 functional group clusters for genes with significantly higher ω in cold-tolerant species (**Supplementary Table 9**) including: calcium-binding EGF domain containing proteins, Transmembrane proteins, and Cytoskeleton / Microtubule.

Population resequencing

A total of 4,802,266 SNPs were analysed. Of these, 1,669,212 (35%) could be placed on the main linkage groups. Using an FDR threshold of $q < 0.05$, a total of 29,556 SNPs were identified as outliers (10,113 could be placed on linkage groups, **Supplementary Figures 8-10**). The outlier SNPs that could be placed on linkage groups were not randomly distributed throughout the genome (**Supplementary Figures 9-10**) (Chi-squared = 9087.7, df = 4, $p < 0.001$). When the proportion of the total genome length of each chromosome was used to calculate the expected numbers of SNPs, there was still a significant deviation (Chi-squared = 9541.1, df = 4, $p < 0.001$). There were more observed outlier SNPs on the X-chromosome (5,840) than expected (1,977). However, these results should be taken with caution because the SNP counts on each chromosome are not completely independent due to LD between SNPs (**Supplementary Figure 10**). In the three pairwise analyses, results were qualitatively similar. The numbers of SNPs in each comparison along with the number of outlier SNPs are presented in (**Supplementary Table 10)**.

There were fewer polymorphic SNPs between the Colorado and Vancouver populations and there were also fewer outlier SNPs in this comparison (**Supplementary Table 10**, **Supplementary Figure 11**). As above, the outlier SNPs are not randomly distributed throughout the genome (**Fig. 7 and Supplementary Figure 12**). There was a significant excess of outlier SNPs on the X-chromosome in all pairwise comparisons (Colorado: Oulanka – Chi-squared = 3,029.4, d.f. = 4, $p < 0.01$; Colorado: Vancouver - Chi-squared = 31.9,d.f. = 4, $p < 0.01$; Oulanka: Vancouver - Chi-squared = 2477.7, d.f. = 4, $p < 0.01$). These results held when the proportion of the total genome length of each chromosome was taken to calculate the expected numbers of SNPs. There were no shared outlier SNPs between all three pairwise comparisons but there were some shared outlier SNPs between pairs of population comparisons and the majority of outlier SNPs were private to each comparison (**Supplementary Figure 13A**).

To examine patterns of divergence at the gene level we considered a gene to be differentiated if it contained a significantly differentiated SNP (with 1kb) (**Supplemental Table 20**). Significant SNPs were found in 1801, 155, and 1387 genes (from pairwise comparisons between Colorado: Oulanka, Colorado: Vancouver, and Oulanka: Vancouver respectively). 10 genes overlapped between all the three pairwise comparisons (**Supplementary Figure 13B, Supplemental Table 11**). Although this is a relatively small number of genes, it is significantly greater than expected by chance ($p = 0.00013$). By ranking genes by q-value we could identify GO terms enriched in genes with high divergence for each population comparison (Colorado: Oulanka = 74 (Biological Processes: Molecular Functions: Cellular Components = 27:29:18) (**Supplemental Table 12**), Colorado: Vancouver = 66 (Biological Processes: Molecular Functions: Cellular Components = 19:28:19) (**Supplemental Table 13**), Oulanka: Vancouver = 91 (Biological Processes: Molecular Functions: Cellular Components = 37:39:14) (**Supplemental Table 14**). As with genes, there was a significant overlap of enriched GO-terms between population comparisons (N = 22, $p = 1.74 \times 10^{-79}$, **Supplementary Figure 13B, Supplemental Table 15**). Semantic clustering of GO terms (**Fig. 8**) and functional clustering (**Supplemental Table 16-18**) showed that the

411 dominant terms include: membrane components, ion transport, small molecule binding, and
412 neuron / synaptic associated terms.
413
414 **Intraspecific and Interspecific divergence**
415
416 We examined whether genes showing significant divergence between populations were the
417 same as those showing higher rates of evolution between cold-tolerant and non-cold-
418 tolerant species. We found 68 genes that had both an elevated rate of evolution between
419 species and significant divergence in at least one population comparison (**Supplemental**
420 **Table 19**). This is significantly greater than we would expect by chance (Fisher's exact test =
421 1.447, p = 0.0006) and implies that genes under divergent selection within species are also
422 more likely to diverge between species.
423

Discussion

Ecological studies with *Drosophila montana* have shown that it is able to thrive at high latitudes due to a number of adaptations including the evolution of increased cold tolerance and reproductive diapause. By sequencing the genome of this species we were able to use comparative genomics to identify genes and functional processes that differ between *D. montana* and its less cold-adapted relatives. We find evidence for selection acting on neuronal, membrane-transport and ion-transport related genes at both the inter- and intra-specific levels. These findings likely result from selection for an ability to overwinter under harsh environmental conditions, as these processes have clear links to both increased cold tolerance and reproductive diapause.

Genome assembly and features

We assembled the *D. montana* genome using a combination of Illumina paired-end reads and mate-pair reads. We annotated 13,683 genes, which is comparable to other Drosophila species that have been sequenced [68]. 10,898 of these genes (80 %) were then assigned to a *D. virilis* ortholog, comparable to the number of orthologs identified between *D. melanogaster* and *D. simulans*. Together with the high BUSCO and CEGMA scores, this suggests that the genic component of the assembled genome is largely complete and successfully annotated.

Inter- and intra- specific comparisons show evidence for cold adaptation

We used three separate comparisons to identify genes potentially involved in adaptive divergence: between cold-tolerant and non-cold-tolerant *Drosophila* species, between *D. montana* and its sister species *D. virilis*, and between three populations of *D. montana* from different biogeographic localities.

Firstly, in the comparison between multiple *Drosophila* species, we identified 250 genes with an elevated rate of evolution between cold-tolerant and non-cold-tolerant *Drosophila* species. Interestingly, the increased rate of evolution was biased toward cold-tolerant species, with 77% of these genes showing a higher rate of evolution in these species. Secondly, we compared *D. montana* to its sequenced relative *D. virilis*. Although *D. montana* and *D. virilis* are both relatively cold-tolerant species, *D. montana* is significantly more cold-tolerant than *D. virilis* [29], and *D. montana* is also more desiccation tolerant (Kellerman et al. 2012). In addition, unlike *D. virilis*, *D. montana* females enter reproductive diapause in late summer, which further increases their chances to survive over the cold season and produce progeny in spring [69]. However, genes in the *D. montana* genome showed little evidence for divergent selection when compared to *D. virilis*, most genes showing evidence for purifying selection. Finally, we compared *D. montana* populations from Oulanka, Colorado and Vancouver. These populations face quite different abiotic and biotic conditions throughout the year, and hence can be expected to vary in several traits affecting flies' life-cycle and stress tolerances. We identified many SNPs that show significant divergence between the three populations; the number of divergent SNPs was smallest between

Colorado and Vancouver populations reflecting the later divergence times of these populations. Although no divergent SNPs were shared between all population comparisons, when SNPs were grouped by gene, we found evidence for a significant number of overlapping genes. Divergent SNPs were overrepresented in the X chromosome which often shows elevated rates of evolution due to a combination of effects including a smaller effective population size, increased efficacy of selection in hemizygous males, and sexual antagonism. However, as some of the populations are known to differ in sexual behaviour and post-mating pre-zygotic reproductive barriers [26,27], as well as ecological adaptations, it is not possible to distinguish among the multiple possible sources of any divergent selection on X-linked SNPs.

In most of these comparisons the genes with elevated dN/dS or F_{ST} were enriched for functional processes previously demonstrated as important in cold adaptation (see below). In particular changes to membrane components and ion transport, as well as in the neurological system were heavily represented in our enrichment analyses in each comparison. In addition, we also found enrichment of many small-molecule binding terms, but these tended to be more varied across the different comparisons. Finally several comparisons were also enriched for many reproduction-associated terms, which are unlikely to be linked to cold adaptation per se. We discuss each of these functional groups below.

Functional processes enriched in inter- and intra- specific comparisons

Cellular membranes

The composition of the cell membrane is critical for maintenance of cellular function in sub-optimal temperatures [70,71]. We found enrichment of many terms associated with membrane structure (e.g. intrinsic component of membrane, integral component of membrane, plasma membrane, transmembrane region, etc) across all our comparisons, providing further evidence for the importance of adjusting cell membrane structure to better survive in cold environments. In addition to these terms we also found enrichment of other key processes that likely contribute to the functioning of cell membranes at low temperatures. The most important of these are functions associated with cellular ionic balance (e.g. ion channel activity, transmembrane transporter activity, calcium transport, ion binding). Many of the mechanisms involved in the maintenance of cellular ion balance are known to be temperature specific [72,73]. Failure to maintain the ionic balance of cells leads to metabolic perturbations which can cause a wide range of negative consequences, including cellular damage and even death [74,75]. One class of cells particularly affected by low temperature are neurons [76–78] which are particularly susceptible to cold injury [79]. In line with this we also found enrichment of several terms related to neuron function (cell projection membrane, dendrite membrane, signal transducer activity, etc.). Finally, we observed that membrane, ion transport, and neuronal terms often functionally clustered together, showing that changes to each of these functions are in fact interrelated. Taken together this suggests that the adjustment of cell membranes for increased cold tolerance is complex, requiring changes to many genes to improve cellular functioning at low temperatures.

Small-molecule binding

We observed enrichment of many small-molecule binding terms (small-molecule binding, ATP-binding, kinase, nucleotide-binding, nucleotide phosphate-binding, carbohydrate derivative binding, ribonucleotide binding, anion binding, etc.), both in the population and in the multi-species comparisons. At low temperatures the activity levels of many reactions are reduced meaning that during cold adaptation there is selection to adjust chemical reactions to work better in cold environments [80]. In particular ATP-binding and associated terms were enriched in most of our comparisons suggesting that adjustments to ATP-binding may be particularly important for cold adaptation. This finding is supported by the fact that low temperatures adversely affect ATP metabolism across a broad range of taxa [81,82].

Reproduction

Genes involved in reproduction typically show faster rates of divergence than other genes [83,84]. Consistent with this we find reproductive-associated terms (male courtship behaviour, single organism reproductive process, reproductive process) are enriched at each comparison level. Different species of Drosophila (including *D. montana* and *D. virilis*) are known to vary for a number of reproductive traits and so this finding is not too surprising. Interestingly, the only pairwise population comparison that shows enrichment for reproductive-associated terms (reproductive structure development, gonad development) is between Colorado and Vancouver. Although all populations show some evidence of reproductive isolation, crosses between Colorado and Vancouver showed the highest proportion of non-developing eggs [27]. Moreover, although the exact cause of non-developing eggs is unknown, one possibility is that it could be due to a negative interaction between sperm and the female reproductive tract. Some support for this idea comes from examining the top differentiated genes between Colorado and Vancouver which include the transcription factor *ken and barbie* which has a major role in the development of genitalia of *D. melanogaster* [85].

Olfaction

Drosophila flies have various kinds of olfaction-driven behaviors including the location of food and mates [86,87] and the genomic repertoire of olfactory loci is correlated with environmental variation [88]. A cold environment may affect the perception of olfactory signals as the detection of odorants at low temperatures is more difficult due to the reduced concentration of olfactory cues in the air. Previous work in *D. melanogaster* has shown that the sensitivity of the olfactory system increases in response to cold temperature (e. g. Dalton [89]), and that this change is accompanied by a change in expression in olfactory genes [90,91]. Since both sexual and non-sexual olfactory signals are likely to be affected by colder temperatures we hypothesize that the changes in olfaction-related genes we observe in the present study are a product of adaptation to living in a colder environment as well as of sexual selection to distinguish conspecific flies from the heterospecific ones.

559 **Intraspecific and interspecific divergence**
560
561 Phenotypic variation in traits between and within species may or may not arise from the
562 same genes [92]. Here we find that genes showing divergence between populations were
563 also more likely to show differences between species. The functions of these genes mirror
564 those enriched in each of the separate comparisons (transmembrane transport / ion
565 transport (9/68), sexual reproduction (16/68), and neurological system
566 process/neurogenesis (15/68)), implicating these genes' involvement in similar differences in
567 cold adaptation and reproduction between populations and species. Although any of these
568 genes may be important in cold adaptation, one gene in particular, *Task6*, stands out as an
569 interesting candidate. *Task6* encodes a subunit of two-pore domain potassium (K2P)
570 channels, which are important in setting the membrane potential and input resistance of
571 neurons in Drosophila [93]. Temperature impacts a cells ability to maintain ionic balance,
572 and in particular loss of potassium ion balance has been shown to cause membrane
573 depolarization, induction of chill-coma, and cell death [13,94]. As such the changes we
574 observe in *Task6* may be involved in thermal adaptation of species and populations.
575
576 **Conclusion**
577
578 *D. montana* is an exceptional species of Drosophila in terms of cold adaptation, as well as a
579 species used for studies of behavioural variation and reproductive isolation. Here we report
580 the first description of its genome. Although there are few strong signals of divergent
581 selection on coding sequence variation, contrasts between cold-adapted species and
582 intraspecific population sequencing suggest that the genome contains a clear signal of
583 selection for cold tolerance. We identify many genes potential important in adaptation and
584 speciation in this ecological specialist species.
585
586

DATA AVAILABILITY

This project has been deposited at NCBI under the BioProject accession PRJNA312336. The accession number for the assembly is LUVX00000000. Raw reads were deposited in the SRA under the following accession numbers: mate-pair reads: SRX1604922, paired-end reads: SRX1602883, SRX1602879, population resequencing reads: SRX1625831, SRX1625832, SRX1625834.

ACKNOWLEDGEMENTS

We would like to thank Paris Veltsos for help with DNA extractions, and RAPiD Genomics (FL, USA) for help with probe design, library preparation, sequencing and initial analysis of the SNP data for the linkage map. The work has been supported by the Academy of Finland to AH (projects 132619 and 267244) and to MK. (projects 268214 and 272927) and NERC (UK) funding to MGR (grants NE/E015255/1 and NE/J020818/1) and PhD studentship to DJP (NE/I528634/1).

607 **REFERENCES**

608

609 1. Nosil P, Funk DJ, Ortiz-Barrientos D. Divergent selection and heterogeneous genomic
610 divergence. Mol Ecol. 2009;18:375–402.

611 2. Smadja CM, Butlin RK. A framework for comparing processes of speciation in the presence
612 of gene flow. Mol Ecol. 2011;20:5123–40.

613 3. Noor MA, Bennett SM. Islands of speciation or mirages in the desert? Examining the role
614 of restricted recombination in maintaining species. Heredity. 2009;103:439.

615 4. Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of speciation are due
616 to reduced diversity, not reduced gene flow. Mol Ecol. 2014;23:3133–3157.

617 5. Ravinet M, Faria R, Butlin RK, Galindo J, Bierne N, Rafajlović M, et al. Interpreting the
618 genomic landscape of speciation: a road map for finding barriers to gene flow. J Evol Biol.
619 2017;30:1450–1477.

620 6. Wolf JBW, Ellegren H. Making sense of genomic islands of differentiation in light of
621 speciation. Nat Rev Genet. 2017;18:87–100.

622 7. Matzkin LM, Watts TD, Bitler BG, Machado CA, Markow TA. Functional genomics of cactus
623 host shifts in *Drosophila mojavensis*. Mol Ecol. 2006;15:4635–4643.

624 8. Smith G, Fang Y, Liu X, Kenny J, Cossins AR, Oliveira CC, et al. Transcriptome-wide
625 expression variation associated with environmental plasticity and mating success in
626 cactophilic *Drosophila mojavensis*. Evolution. 2013;67:1950–1963.

627 9. McKenna DD, Scully ED, Pauchet Y, Hoover K, Kirsch R, Geib SM, et al. Genome of the
628 Asian longhorned beetle (*Anoplophora glabripennis*), a globally significant invasive species,
629 reveals key functional and evolutionary innovations at the beetle–plant interface. Genome
630 Biol. 2016;17:227.

631 10. Chen C, Liu Z, Pan Q, Chen X, Wang H, Guo H, et al. Genomic analyses reveal
632 demographic history and temperate adaptation of the newly discovered honey bee
633 subspecies *Apis mellifera sinisxinyuan* n. ssp. Mol Biol Evol. 2016;33:1337–1348.

634 11. Foll M, Gaggiotti OE, Daub JT, Vatsiou A, Excoffier L. Widespread signals of convergent
635 adaptation to high altitude in Asia and America. Am J Hum Genet. 2014;95:394–407.

636 12. Kellermann V, Loeschcke V, Hoffmann AA, Kristensen TN, Flojgaard C, David JR, et al.
637 Phylogenetic constraints in key functional traits behind species' climate niches: patterns of
638 desiccation and cold resistance across 95 *Drosophila* species. Evolution. 2012;66:3377–3389.

639 13. Andersen JL, Manenti T, Sørensen JG, MacMillan HA, Loeschcke V, Overgaard J. How to
640 assess Drosophila cold tolerance: chill coma temperature and lower lethal temperature are
641 the best predictors of cold distribution limits. Funct Ecol. 2015;29:55–65.

642 14. Tyukmaeva VI, Salminen TS, Kankare M, Knott KE, Hoikkala A. Adaptation to a seasonally
643 varying environment: a strong latitudinal cline in reproductive diapause combined with high
644 gene flow in *Drosophila montana*. Ecol Evol. 2011;1:160–168.

15. Vesala L, Hoikkala A. Effects of photoperiodically induced reproductive diapause and cold hardening on the cold tolerance of *Drosophila montana*. J Insect Physiol. 2011;57:46–51.

16. Kauranen H, Menegazzi P, Costa R, Helfrich-Förster C, Kankainen A, Hoikkala A. Flies in the north: locomotor behavior and clock neuron organization of *Drosophila montana*. J Biol Rhythms. 2012;27:377–387.

17. Kauranen H, Ala-Honkola O, Kankare M, Hoikkala A. Circadian clock of *Drosophila montana* is adapted to high variation in summer day lengths and temperatures prevailing at high latitudes. J Insect Physiol. 2016;89:9–18.

18. Terhzaz S, Teets NM, Cabrero P, Henderson L, Ritchie MG, Nachman RJ, et al. Insect capa neuropeptides impact desiccation and cold tolerance. Proc Natl Acad Sci U S A [Internet]. 2015 [cited 2015 Feb 21]; Available from: http://www.pnas.org/content/early/2015/02/12/1501518112.short

19. Menegazzi P, Dalla Benetta E, Beauchamp M, Schlichting M, Steffan-Dewenter I, Helfrich-Förster C. Adaptation of circadian neuronal network to photoperiod in high-latitude European drosophilids. Curr Biol. 2017;27:833–839.

20. Spicer GS, Bell CD. Molecular phylogeny of the *Drosophila virilis* species group (Diptera: Drosophilidae) inferred from mitochondrial 12S and 16S ribosomal RNA genes. Ann Entomol Soc Am. 2002;95:156–161.

21. Throckmorton LH. The virilis species group. In: Ashburner M, Carson HL, Thompson JN, editors. Genet Biol Drosoph. London, UK: Academic Press; 1982. p. 227–296.

22. Ostrega MS. Restriction endonuclease analysis of the relatedness of *D. montana* and *D. virilis* lines. Drosoph Inf Serv. 1985;61:132–133.

23. Mirol PM, Schäfer MA, Orsini L, Routtu J, Schlötterer C, Hoikkala A, et al. Phylogeographic patterns in *Drosophila montana*. Mol Ecol. 2007;16:1085–1097.

24. Lankinen P, Tyukmaeva VI, Hoikkala A. Northern *Drosophila montana* flies show variation both within and between cline populations in the critical day length evoking reproductive diapause. J Insect Physiol. 2013;59:745–751.

25. Tyukmaeva VI, Veltsos P, Slate J, Gregson E, Kauranen H, Kankare M, et al. Localization of quantitative trait loci for diapause and other photoperiodically regulated life history traits important in adaptation to seasonally varying environments. Mol Ecol. 2015;24:2809–19.

26. Klappert K, Mazzi D, Hoikkala A, Ritchie MG. Male courtship song and female preference variation between phylogeographically distinct populations of *Drosophila montana*. Evolution. 2007;61:1481–1488.

27. Jennings JH, Snook RR, Hoikkala A. Reproductive isolation among allopatric *Drosophila montana* populations. Evolution. 2014;68:3095–3108.

28. Schäfer MA, Mazzi D, Klappert K, Kauranen H, Vieira J, Hoikkala A, et al. A microsatellite linkage map for *Drosophila montana* shows large variation in recombination rates, and a courtship song trait maps to an area of low recombination. J Evol Biol. 2010;23:518–527.

683 29. Vesala L, Salminen TS, Laiho A, Hoikkala A, Kankare M. Cold tolerance and cold-induced
684 modulation of gene expression in two *Drosophila* virilis group species with different
685 distributions. Insect Mol Biol. 2012;21:107–118.

686 30. Salminen TS, Vesala L, Laiho A, Merisalo M, Hoikkala A, Kankare M. Seasonal gene
687 expression kinetics between diapause phases in *Drosophila virilis* group species and
688 overwintering differences between diapausing and non-diapausing females. Sci Rep. 2015;5.

689 31. Parker DJ, Vesala L, Ritchie MG, Laiho A, Hoikkala A, Kankare M. How consistent are the
690 transcriptome changes associated with cold acclimation in two species of the *Drosophila*
691 *virilis* group? Heredity. 2015;

692 32. Kankare M, Parker DJ, Merisalo M, Salminen TS, Hoikkala A. Transcriptional differences
693 between diapausing and non-diapausing *D. montana* females reared under the same
694 photoperiod and temperature. PloS One. 2016;11:e0161852.

695 33. Parker DJ, Ritchie MG, Kankare M. Preparing for Winter: the Transcriptomic Response
696 Associated with Different Day Lengths in *Drosophila montana*. G3 GenesGenomesGenetics.
697 2016;g3.116.027870.

698 34. Routtu J, Mazzi D, Van Der Linde K, Mirol P, Butlin RK, Hoikkala A. The extent of variation
699 in male song, wing and genital characters among allopatric *Drosophila montana* populations.
700 J Evol Biol. 2007;20:1591–1601.

701 35. Lakovaara S. Malt as a culture medium for Drosophila species. Drosoph Inf Serv.
702 1969;44:128.

703 36. Buffalo V. Scythe. Available at https://github.com/vsbuffalo/scythe; 2014.

704 37. Joshi NA, Fass JN. Sickle: A sliding-window, adaptive, quality-based trimming tool for
705 FastQ files. Available at https://github.com/najoshi/sickle.; 2011.

706 38. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform.
707 Bioinformatics. 2009;25:1754–1760.

708 39. Schmieder R, Edwards R. Fast Identification and Removal of Sequence Contamination
709 from Genomic and Metagenomic Datasets. PLOS ONE. 2011;6:e17288.

710 40. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in
711 eukaryotic genomes. Bioinformatics. 2007;23:1061–1067.

712 41. Parra G, Bradnam K, Ning ZM, Keane T, Korf I. Assessing the gene space in draft
713 genomes. Nucleic Acids Res. 2009;37:289–297.

714 42. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing
715 genome assembly and annotation completeness with single-copy orthologs. Bioinforma Oxf
716 Engl. 2015;31:3210–2.

717 43. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management
718 tool for second-generation genome projects. BMC Bioinformatics. 2011;12:491.

719 44. Paterson T, Law A. Genotypechecker: an interactive tool for checking the inheritance
720 consistency of genotyped pedigrees. Anim Genet. 2011;42:560–2.

721 45. Broman KW, Wu H, Sen Ś, Churchill GA. R/qtl: QTL mapping in experimental crosses.
722 Bioinformatics. 2003;19:889–90.

723 46. Löytynoja A, Goldman N. An algorithm for progressive multiple alignment of sequences
724 with insertions. Proc Natl Acad Sci U S A. 2005;102:10557–10562.

725 47. Yang ZH. PAML: a program package for phylogenetic analysis by maximum likelihood.
726 Comput Appl Biosci. 1997;13:555–556.

727 48. Yang ZH, Bielawski JP. Statistical methods for detecting molecular adaptation. Trends
728 Ecol Evol. 2000;15:496–503.

729 49. Angelis K, Dos Reis M, Yang Z. Bayesian estimation of nonsynonymous/synonymous rate
730 ratios for pairwise sequence comparisons. Mol Biol Evol. 2014;31:1902–13.

731 50. Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG. Dynamic evolution
732 of the innate immune system in Drosophila. Nat Genet. 2007;39:1461–1468.

733 51. Haerty W, Jagadeeshan S, Kulathinal RJ, Wong A, Ravi Ram K, Sirot LK, et al. Evolution in
734 the fast lane: rapidly evolving sex-related genes in *Drosophila*. Genetics. 2007;177:1321–
735 1335.

736 52. Jennings JH, Mazzi D, Ritchie MG, Hoikkala A. Sexual and postmating reproductive
737 isolation between allopatric *Drosophila montana* populations suggest speciation potential.
738 BMC Evol Biol. 2011;11:68.

739 53. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence
740 data. Bioinformatics. 2014;30:2114–2120.

741 54. Zhu Y, Bergland AO, González J, Petrov DA. Empirical validation of pooled whole genome
742 population re-sequencing in *Drosophila melanogaster*. PloS One. 2012;7:e41901.

743 55. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence
744 alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–2079.

745 56. Kofler R, Pandey RV, Schlötterer C. PoPoolation2: identifying differentiation between
746 populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics.
747 2011;27:3435–6.

748 57. Foll M, Gaggiotti O. A genome-scan method to identify selected loci appropriate for both
749 dominant and codominant markers: a bayesian perspective. Genetics. 2008;180:977–993.

750 58. Foll M, Fischer MC, Heckel G, Excoffier L. Estimating population structure from AFLP
751 amplification intensity. Mol Ecol. 2010;19:4638–4647.

752 59. Fischer MC, Foll M, Excoffier L, Heckel G. Enhanced AFLP genome scans detect local
753 adaptation in high-altitude populations of a small rodent (*Microtus arvalis*). Mol Ecol.
754 2011;20:1450–1462.

755 60. Pérez-Figueroa A, García-Pereira MJ, Saura M, Rolán-Alvarez E, Caballero A. Comparing
756 three different methods to detect selective loci using dominant markers. J Evol Biol.
757 2010;23:2267–2276.

758 61. Vilas A, Pérez-Figueroa A, Caballero A. A simulation study on the performance of
759 differentiation-based methods to detect selected loci using linked neutral markers. J Evol
760 Biol. 2012;25:1364–1376.

761 62. Villemereuil P, Frichot É, Bazin É, François O, Gaggiotti OE. Genome scan methods
762 against more complex models: when and how much should we trust them? Mol Ecol.
763 2014;23:2006–2019.

764 63. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and
765 visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009;10:48.

766 64. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO Summarizes and Visualizes Long Lists of
767 Gene Ontology Terms. PLOS ONE. 2011;6:e21800.

768 65. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the
769 comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37:1–13.

770 66. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene
771 lists using DAVID bioinformatics resources. Nat Protoc. 2009;4:44–57.

772 67. Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, et al. FlyBase:
773 enhancing *Drosophila* Gene Ontology annotations. Nucleic Acids Res. 2009;37:D555–D559.

774 68. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, et al. Evolution of
775 genes and genomes on the Drosophila phylogeny. Nature. 2007;450:203–218.

776 69. Watabe H. Photoperiodic responses in the *Drosophila virilis* species group (Diptera,
777 Drosophilidae) from Japan. Kontyo. 1983;51:628–634.

778 70. Hazel JR. Thermal adaptation in biological membranes - is homeoviscous adaptation the
779 explanation? Annu Rev Physiol. 1995;57:19–42.

780 71. Koštál V, Berková P, Šimek P. Remodelling of membrane phospholipids during transition
781 to diapause and cold-acclimation in the larvae of *Chymomyza costata* (Drosophilidae). Comp
782 Biochem Physiol B Biochem Mol Biol. 2003;135:407–419.

783 72. Heitler WJ, Goodman CS, Fraserrowell CH. Effects of temperature on threshold of
784 identified neurons in locust. J Comp Physiol. 1977;117:163–182.

785 73. Kivivuori L, Lehti S, Lagerspetz KYH. Effect of temperature-acclimation on thermal-
786 dependence and hysteresis of the resting membrane-potential of the stretch-receptor
787 neuron in crayfish *Astacus astacus* (L). J Therm Biol. 1990;15:9–14.

788 74. Hochachka PW. Defense strategies against hypoxia and hypothermia. Science.
789 1986;231:234–241.

790 75. Koštál V, Vambera J, Bastl J. On the nature of pre-freeze mortality in insects: water
791 balance, ion homeostasis and energy charge in the adults of *Pyrrhocoris apterus*. J Exp Biol.
792 2004;207:1509–1521.

793 76. Montgomery JC, Macdonald JA. Effects of temperature on nervous system: implications
794 for behavioral performance. Am J Physiol - Regul Integr Comp Physiol. 1990;259:R191–6.

77. Janssen R. Thermal influences on nervous system function. Neurosci Biobehav Rev. 1992;16:399–413.

78. Robertson RM, Money TG. Temperature and neuronal circuit function: compensation, tuning and tolerance. Curr Opin Neurobiol. 2012;22:724–34.

79. Hochachka PW, Somero GN. Temperature. Biochem Adapt Mech Process Physiol Evol. New York: Oxford University Press; 2002. p. 290–449.

80. Margesin R. Psychrophiles: From Biodiversity to Biotechnology. London, UK: Springer; 2017.

81. Napolitano MJ, Shain DH. Four kingdoms on glacier ice: convergent energetic processes boost energy levels as temperatures fall. Proc R Soc Lond B Biol Sci. 2004;271:S273–6.

82. Morrison BA, Shain DH. An AMP nucleosidase gene knockout in *Escherichia coli* elevates intracellular ATP levels and increases cold tolerance. Biol Lett. 2008;4:53–6.

83. Swanson WJ, Vacquier VD. Reproductive Protein Evolution. Annu Rev Ecol Syst. 2002;33:161–79.

84. Clark NL, Aagaard JE, Swanson WJ. Evolution of reproductive proteins from animals and plants. Reproduction. 2006;131:11–22.

85. Lukacsovich T, Yuge K, Awano W, Asztalos Z, Kondo S, Juni N, et al. The *ken and barbie* gene encoding a putative transcription factor with a BTB domain and three zinc finger motifs functions in terminalia development of Drosophila. Arch Insect Biochem Physiol. 2003;54:77–94.

86. Amrein H. Pheromone perception and behavior in Drosophila. Curr Opin Neurobiol. 2004;14:435–442.

87. Libert S, Zwiener J, Chu X, VanVoorhies W, Roman G, Pletcher SD. Regulation of Drosophila life span by olfaction and food-derived odors. Science. 2007;315:1133–1137.

88. Gardiner A, Barker D, Butlin RK, Jordan WC, Ritchie MG. Drosophila chemoreceptor gene evolution: selection, specialization and genome size. Mol Ecol. 2008;17:1648–1657.

89. Dalton P. Psychophysical and behavioral characteristics of olfactory adaptation. Chem Senses. 2000;25:487–492.

90. Riveron J, Boto T, Alcorta E. The effect of environmental temperature on olfactory perception in *Drosophila melanogaster*. J Insect Physiol. 2009;55:943–51.

91. Riveron J, Boto T, Alcorta E. Transcriptional basis of the acclimation to high environmental temperature at the olfactory receptor organs of *Drosophila melanogaster*. BMC Genomics. 2013;14:259.

92. Wittkopp PJ, Stewart EE, Arnold LL, Neidert AH, Haerum BK, Thompson EM, et al. Intraspecific polymorphism to interspecific divergence: genetics of pigmentation in Drosophila. Science. 2009;326:540–544.

831 93. Döring F, Scholz H, Kühnlein RP, Karschin A, Wischmeyer E. Novel Drosophila two-pore
832 domain K+ channels: rescue of channel function by heteromeric assembly. Eur J Neurosci.
833 2006;24:2264–2274.

834 94. Andersen MK, Jensen SO, Overgaard J. Physiological correlates of chill susceptibility in
835 Lepidoptera. J Insect Physiol. 2017;98:317–326.

836
837

838 **Table 1 | Summary statistics of *D. montana* genome assembly**

839

Metric	Value
Total assembled length (bp)	183585048
Scaffolds (n)	63742
Scaffold N50 (bp)	40647
Largest scaffold (bp)	515352
GC content (%)	40.57
Number of predicted gene models	13683
Number of predicted gene models with RBBH to *D. virilis* genome	10898
CEGMA pipeline analysis (% complete/partial)	95.97/98.39
BUSCO (% complete/missing)	91.85/1.72

840

841

842

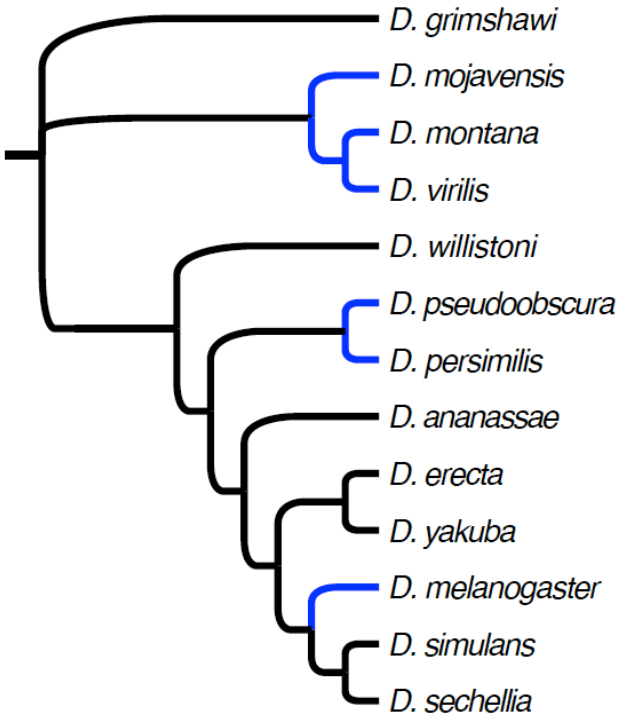


Figure 1 | Tree used for multi-species PAML analyses. Cold-tolerant species (species that have a knockdown temperature of <3°C) are shown in blue (data from Kellermann et al. (2012)).

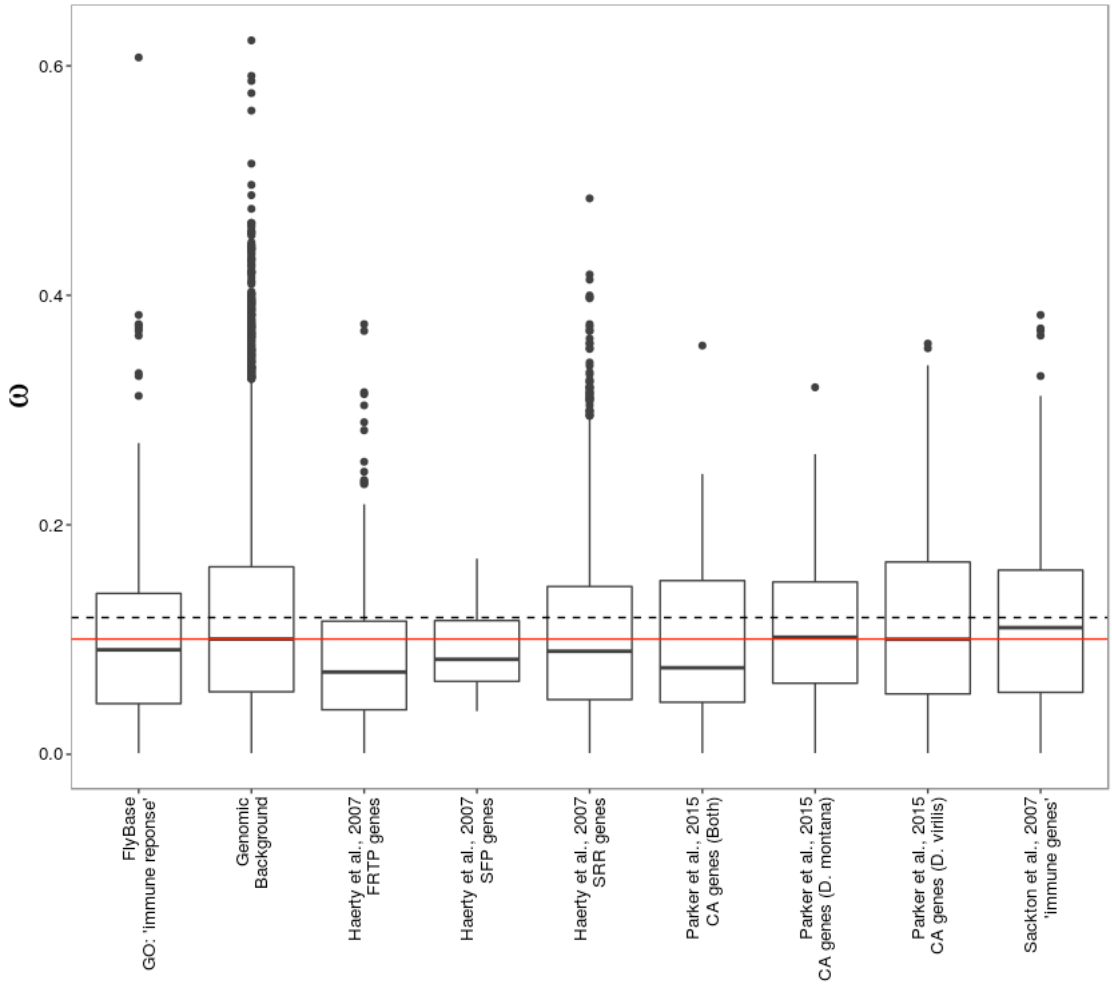


Figure 2 | Average values of ω between *D. montana* and *D. virilis* for candidate gene sets. FRTP = female reproductive tract SFP = seminal fluid proteins SRR = sex and reproduction related genes CA = cold acclimation genes. The red and dashed lines indicate the median and mean ω of the genomic background respectively.

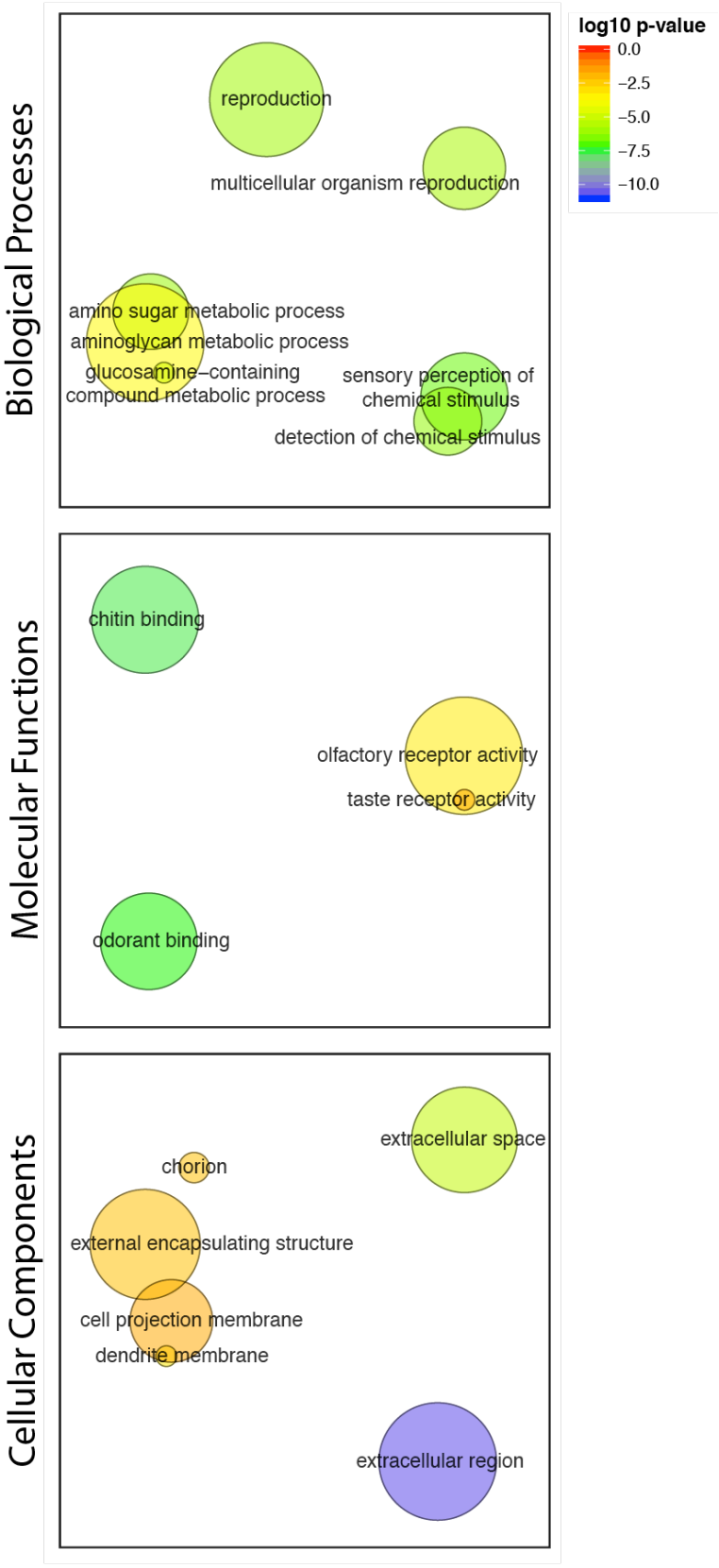
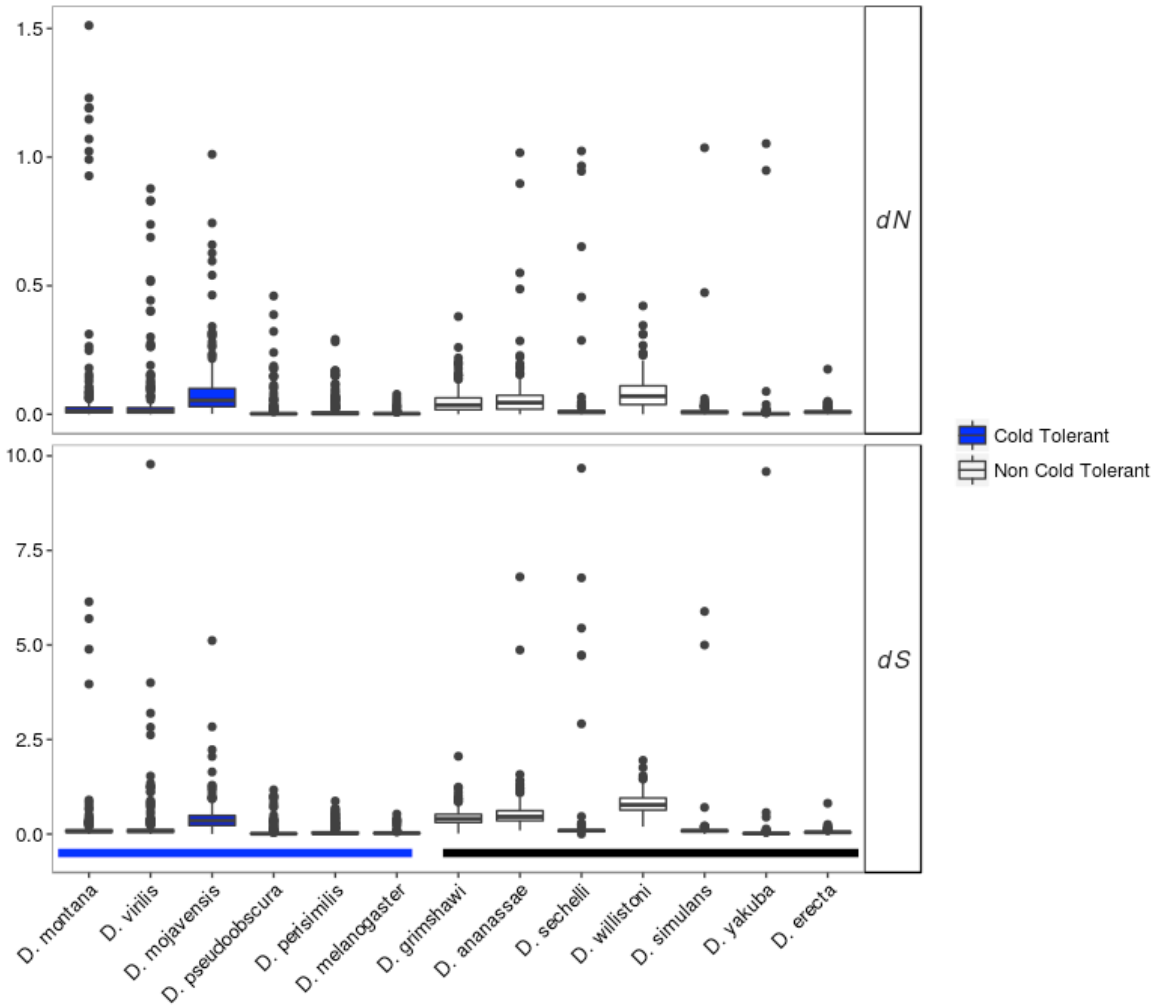


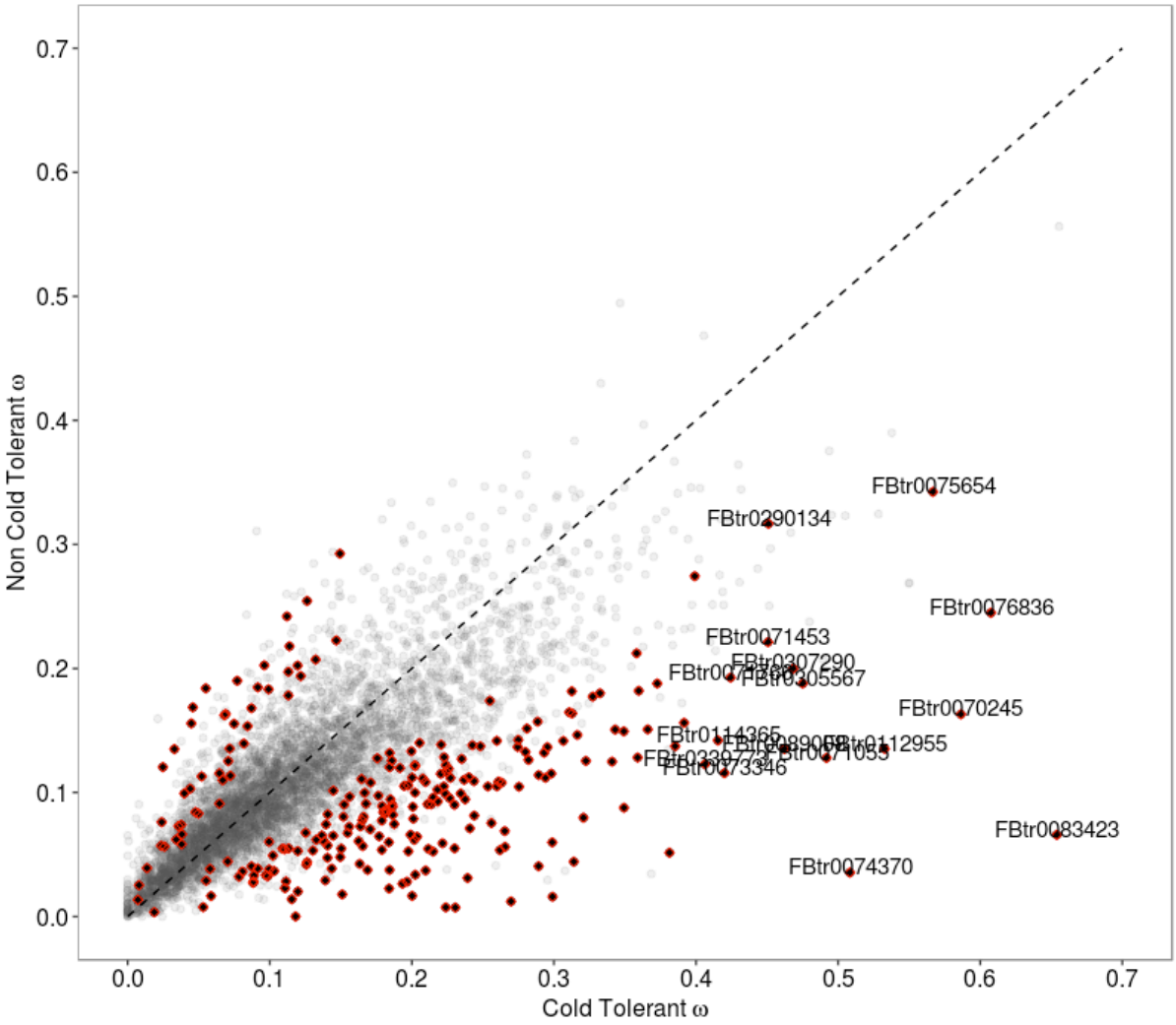
Figure 3 | Semantic clustering of significantly (FDR < 0.1) enriched GO-terms for genes showing high dN/dS between *D. montana* and *D. virilis*. Circle size corresponds to the number of genes annotated to the term in the reference database. Circle colour indicates $\log_{10}$ FDR of the GO term.

863



864

865 **Figure 4 |** Distributions of dN and dS estimates for each of the 250 genes from 13 *Drosophila*
866 species with significant differences in omega between cold-tolerant and non-cold-tolerant
867 species.



868

869 **Figure 5 |** The relationship between values of omega estimated for 5,619 genes in cold-
870 tolerant and non-cold-tolerant species of Drosophila. 250 genes with significantly different
871 estimates of omega are shown in black with red outline. Diagonal line indicates the 1-1
872 diagonal, points below the diagonal line show elevated levels of omega in cold-tolerant
873 species compared to non-cold-tolerant species, while points above the diagonal show
874 elevated levels of omega in non-cold-tolerant species.

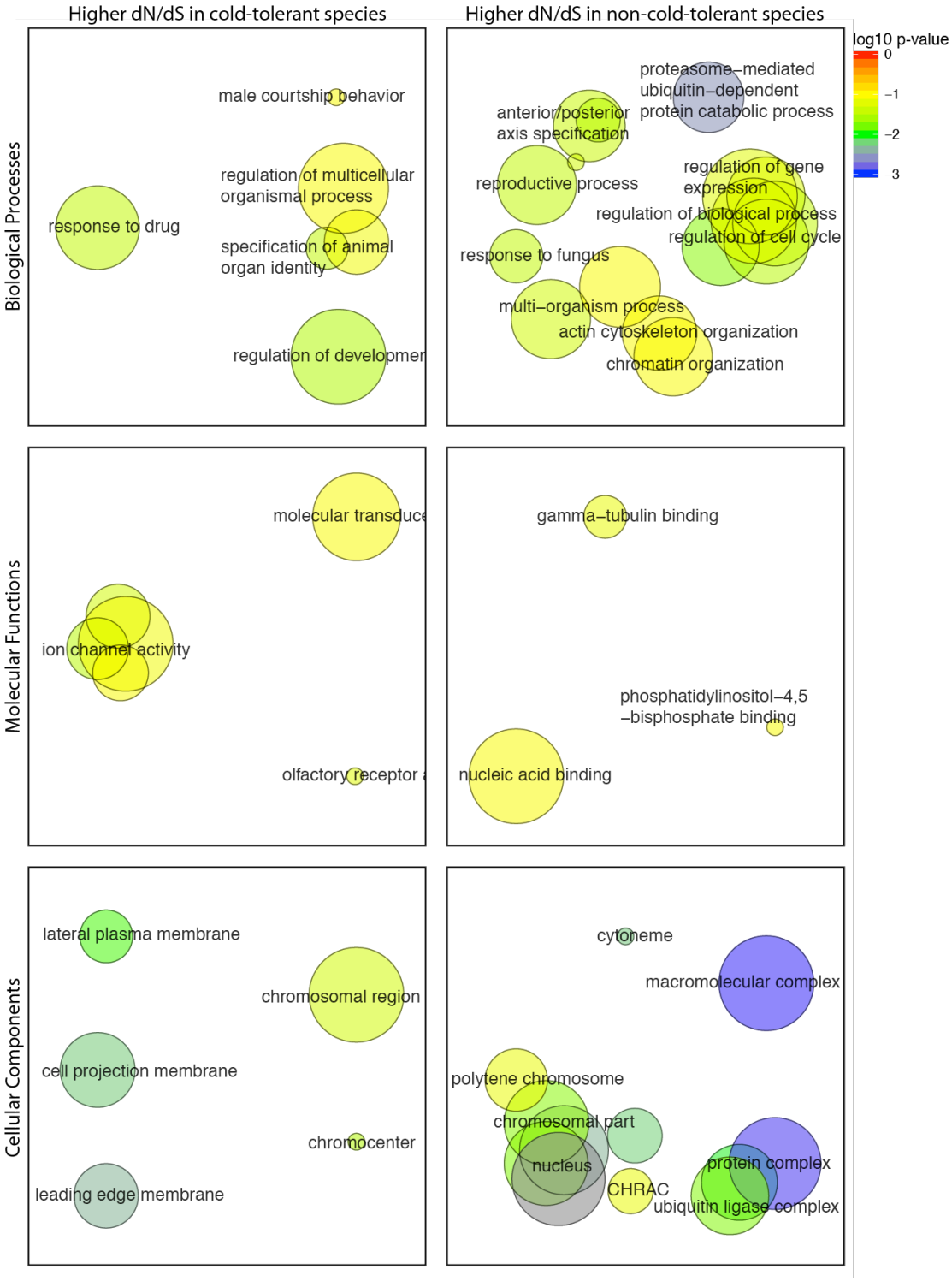


Figure 6 | Semantic clustering of significantly (FDR < 0.1) enriched GO-terms for genes showing significantly elevated dN/dS in cold-tolerant or non-cold-tolerant species. Circle size corresponds to the number of genes annotated to the term in the reference database. Circle colour indicates $\log_{10}$ FDR of the GO term

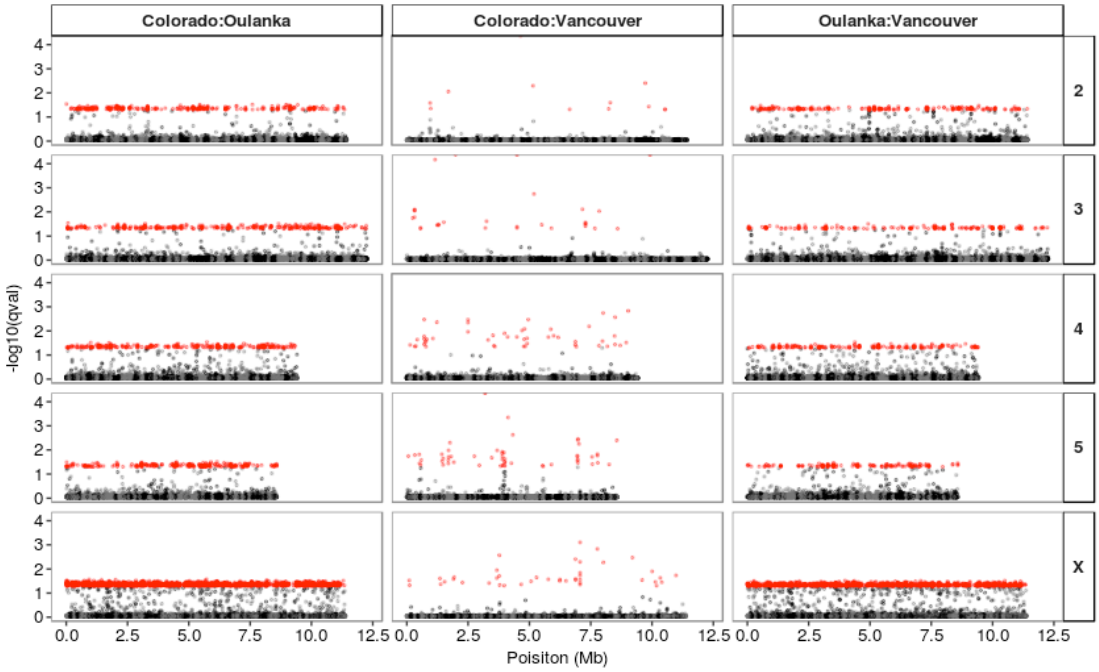


Figure 7 | Manhattan plot of q-values from the three pairwise BayeScan analyses for the SNPs on the mapped chromosomes. Red points denote SNPs which passed the 0.05 q-value FDR threshold. Alternating grey and black points denote different scaffolds that have been anchored to the chromosomes. The order of the mapped scaffolds is established but not their orientation.

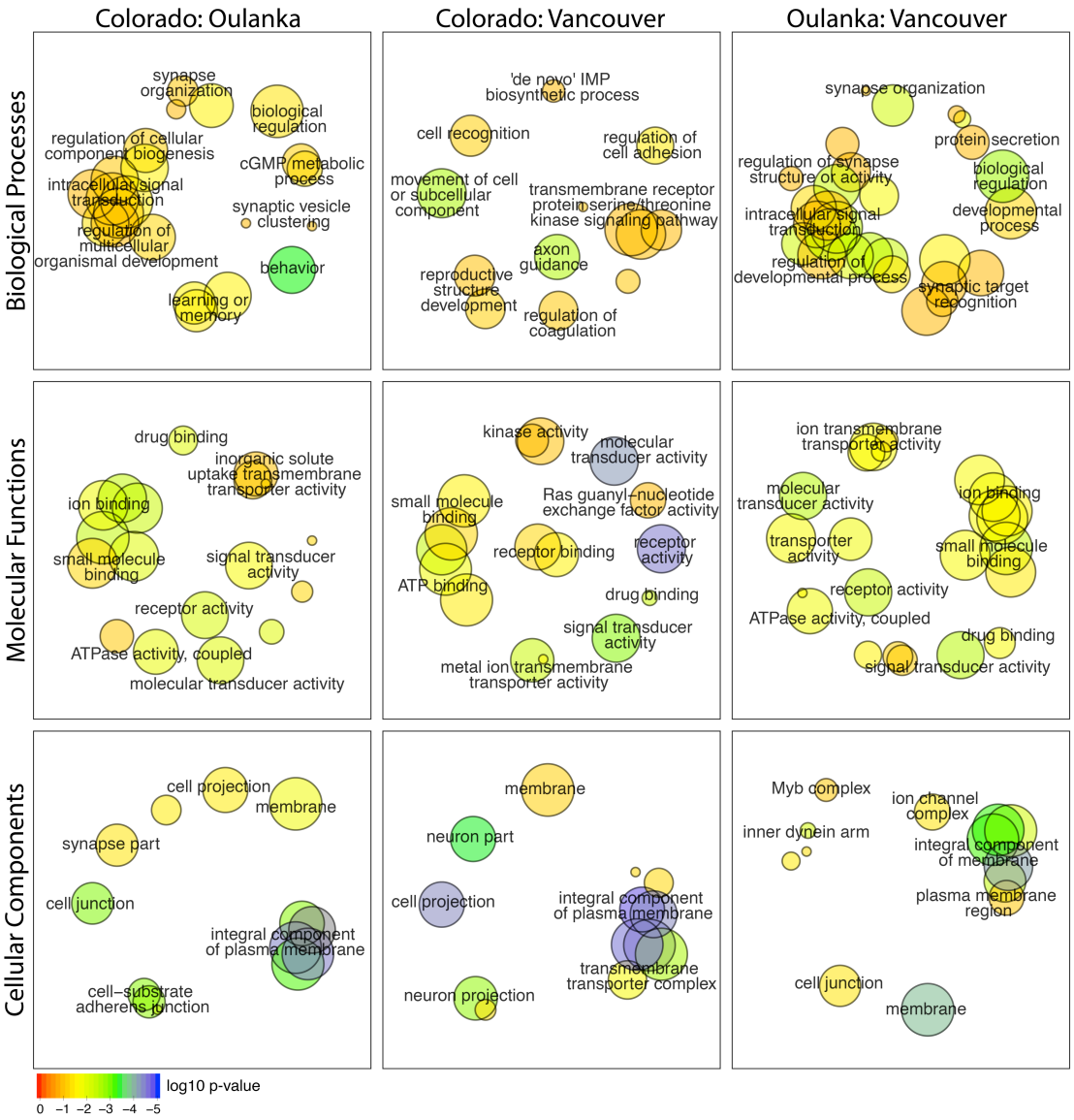


Figure 8 | Semantic clustering of significantly (FDR < 0.1) enriched GO-terms for genes showing significant divergence between populations of *D. montana*. Circle size corresponds to the number of genes annotated to the term in the reference database. Circle colour indicates $\log_{10}$ FDR of the GO term