

1 **The population genomics of structural variation in a songbird genus**

2

3

4 Matthias H. Weissensteiner<sup>1,2</sup>, Ignas Bunikis<sup>3</sup>, Ana Catalán<sup>2</sup>, Kees-Jan Francoijs<sup>6</sup>,  
5 Ulrich Knief<sup>2</sup>, Wieland Heim<sup>5</sup>, Valentina Peona<sup>1</sup>, Saurabh D. Pophaly<sup>2</sup>, Fritz J.  
6 Sedlazeck<sup>4</sup>, Alexander Suh<sup>1</sup>, Vera M. Warmuth<sup>2</sup>, Jochen B.W. Wolf<sup>1,2</sup>

7

8 <sup>1</sup>Department of Evolutionary Biology and Science for Life Laboratory, Uppsala  
9 University, SE-752 36, Uppsala, Sweden.

10 <sup>2</sup>Division of Evolutionary Biology, Faculty of Biology, LMU Munich, Grosshaderner  
11 Str. 2, 82152 Planegg-Martinsried, Germany

12 <sup>3</sup>SciLife Lab Uppsala, Uppsala University SE-751 85 Uppsala, Sweden

13 <sup>4</sup>Human Genome Sequencing Center at Baylor College of Medicine, 1 Baylor Plaza,  
14 Houston, TX 77030, USA

15 <sup>5</sup>Institute of Landscape Ecology, University of Münster, Heisenbergstrasse 2, 48149,  
16 Münster, Germany

17 <sup>6</sup>BioNanoGenomics, San Diego, CA 92121, USA

18

19 Correspondence should be addressed to M.H.W. (email:  
20 matthias.weissensteiner@ebc.uu.se) or J.B.W.W. (email: j.wolf@biologie.uni-  
21 muenchen.de).

22

23 **Abstract**

24

25 **Structural variation (SV) accounts for a substantial part of genetic mutations**  
26 **segregating across eukaryotic genomes with important medical and evolutionary**  
27 **implications. Here, we characterized SV across evolutionary time scales in the**  
28 **songbird genus *Corvus* using *de novo* assembly and read mapping approaches.**  
29 **Combining information from short-read ( $N = 127$ ) and long-read re-sequencing**  
30 **data ( $N = 31$ ) as well as from optical maps ( $N = 16$ ) revealed a total of 201,738**  
31 **insertions, deletions and inversions. Population genetic analysis of SV in the**  
32 **Eurasian crow speciation model revealed an evolutionary young (~530,000 years)**  
33 ***cis*-acting 2.25-kb retrotransposon insertion reducing expression of the *NDP* gene**  
34 **with consequences for premating isolation. Our results attest to the wealth of SV**  
35 **segregating in natural populations and demonstrate its evolutionary significance.**

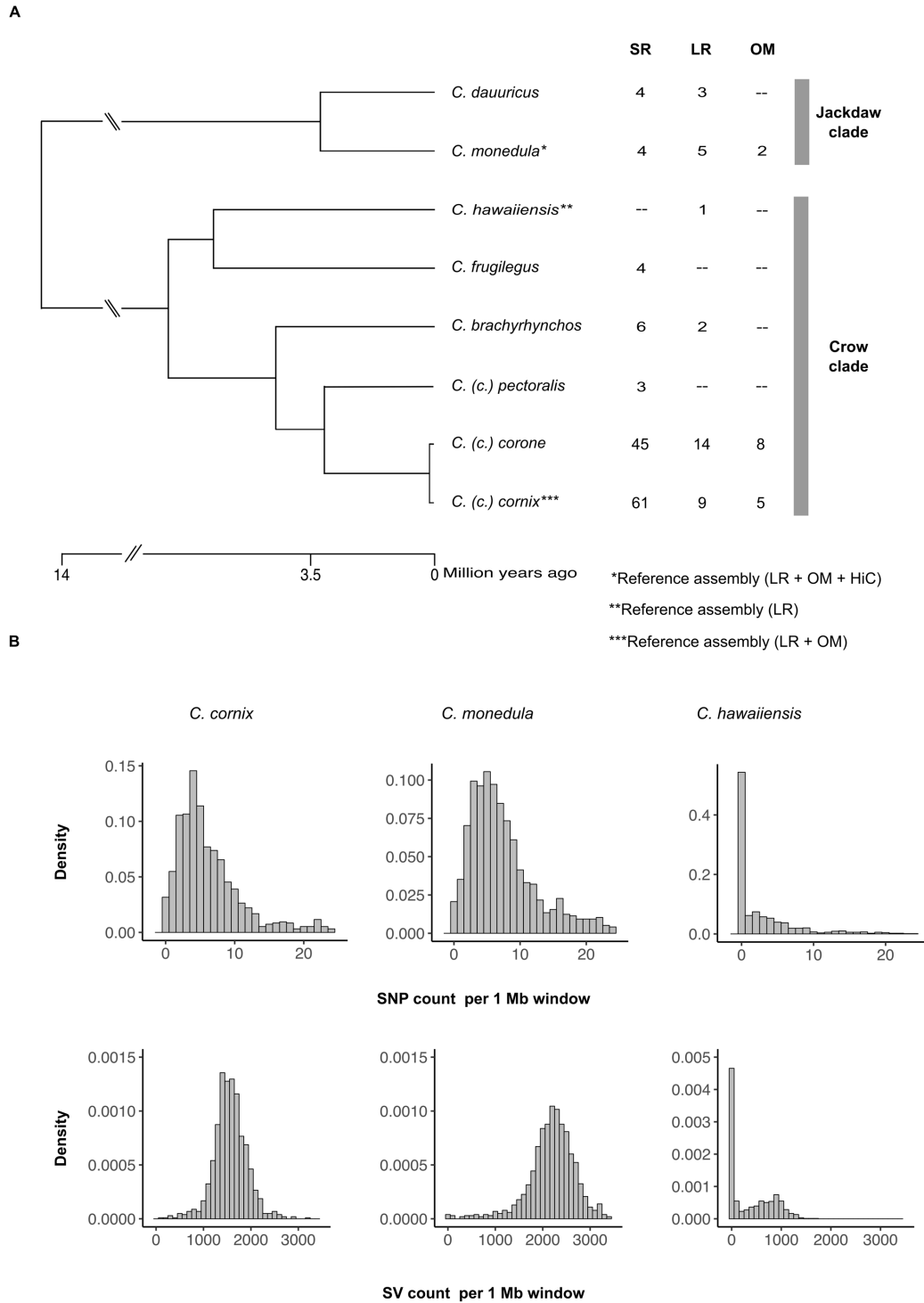
36

37 Mutations altering the structure of DNA have the potential to drastically change  
38 phenotypes with medical and evolutionary implications (1–3). Yet, technological  
39 constraints have long impeded genome-wide characterization of (4). The detection of  
40 SV requires highly contiguous genome assemblies accurately representing the  
41 repetitive fraction of genomes which is known to be a vibrant source and catalyst of SV  
42 (5). Moreover, SV likely remains hidden unless sequence reads traverse it completely  
43 (6, 7). As a consequence, despite the rapidly increasing number of short-read (SR)  
44 based genome assemblies (8) and associated population genomic investigations (9), SV  
45 generally remains unexplored. Even in genetic model organisms, population-level  
46 analysis of SV has been restricted to pedigrees (10) or organisms with smaller, less  
47 complex genomes (11, 12), and few studies have provided a comprehensive account of  
48 SV segregating in natural populations (12, 13).

49

50 To investigate the dynamics of SV and uncover its role in causing phenotypic  
51 differences, we first generated high-quality phased *de novo* genome assemblies  
52 combining long-read (LR) data from single-molecule, real-time (SMRT, PacBio)  
53 sequencing and nanochannel optical mapping (OM) for the hooded crow (*Corvus*  
54 (*corone*) *cornix*; data from (14)), and the European jackdaw (*Corvus monedula*). For  
55 the former, we also generated chromatin interaction mapping data (Hi-C) to obtain a  
56 chromosome-level reference genome (**Fig. 1A**, see **Supplementary Table S1** for  
57 assembly statistics). In addition, we included a previously published LR assembly of  
58 the Hawaiian crow (*Corvus hawaiiensis*) in the analyses (15). All assemblies were  
59 generated with the diploid-aware FALCON-UNZIP assembler (16), facilitating the  
60 comparison of haplotypes within species to identify heterozygous variants and  
61 determine genetic diversity at the level of single individuals. After aligning the two  
62 haplotypes of each assembly, we identified single-nucleotide polymorphisms (SNPs),  
63 insertions and deletions in all three species (**Table 1**). Genome-wide numbers of SV  
64 and SNPs per 1 Mb window were highest in jackdaw and lowest in the highly inbred  
65 Hawaiian crow (**Fig. 1B**), consistent with a positive correlation between census  
66 population size and genetic diversity (15, 17).

67



68  
 69  
 70  
 71  
 72  
 73  
 74  
 75  
 76  
 77  
 78  
 79

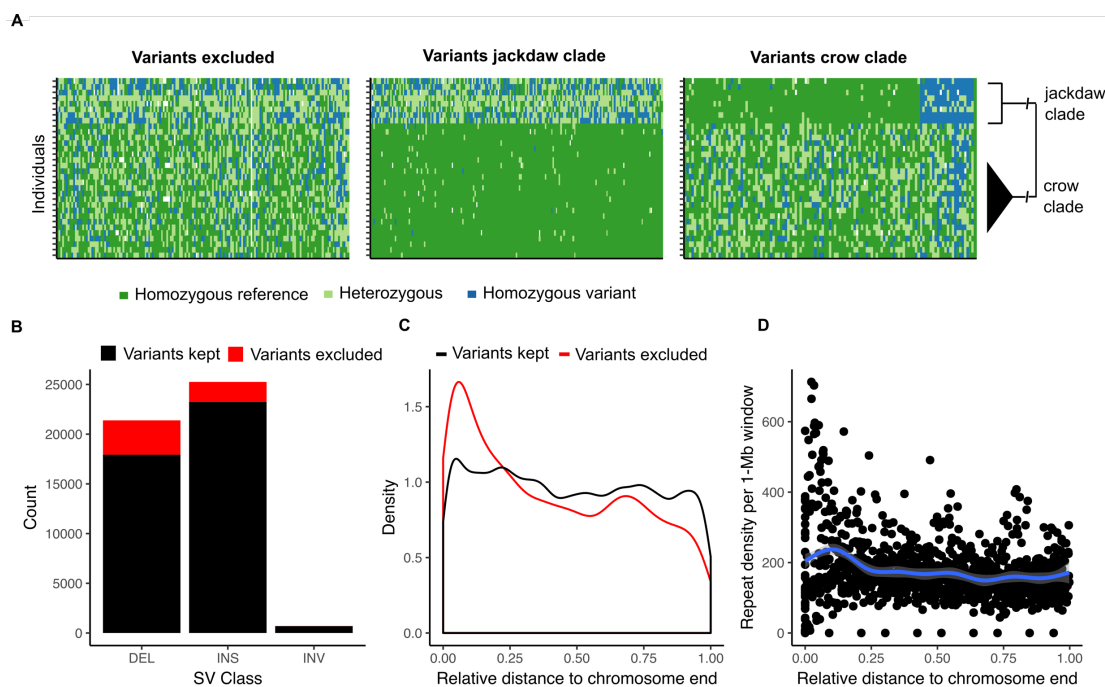
**Fig. 1 | Sampling setup and assembly-based structural and single-nucleotide variation. (A),** Phylogeny of sampled species in the genus *Corvus* (after (50)). Numbers in columns represent individual numbers for short-read sequencing (SR), long-read sequencing (LR) and optical mapping (OM). **(B),** Density histogram showing the abundance of genetic variation within single individuals. Counts of variants per 1 Mb windows are based on comparing the two haplotypes of each assembly. The upper panel reflects structural variation (SV) densities, the lower panel reflects densities for single-nucleotide polymorphisms (SNP).

80 **Table 1 | Assembly-based structural variation and single-nucleotide polymorphism detection.**  
81

Species	Total number	Mean density per 1 Mb	Median density per 1 Mb	Total number	Mean density per 1 Mb	Median density per 1 Mb
Hooded crow	1637609	1568	1558	9916	9.19	5
Jackdaw	2262079	2189	2228	9903	9.29	7
Hawaiian crow	414229	366	0	4841	3.82	0

82  
83

84 Next, to uncover SV segregating within and between natural populations, we generated  
85 LR re-sequencing data for 31 individuals. Spanning the phylogeny of the genus, this  
86 dataset included samples from the European and Daurian jackdaw (*C. monedula*, *C.*  
87 *dauuricus*), the American crow (*C. brachyrhynchos*) and the Eurasian crow complex  
88 (*C. (corone)* spp.). The latter comprised individuals from the phenotypically divergent  
89 hooded crow (Sweden and Poland), and carrion crow populations (Spain and Germany)  
90 (18) (Fig. 1A). Individuals were sequenced to a mean sequence coverage of 15 (range:  
91 8.47 – 27.91) with a mean read length of 7,535 bp (range: 5,219 - 10,034 bp;  
92 **Supplementary Table S2**). Mapping reads to the hooded crow reference allowed us to  
93 identify variants and genotypes for each diploid individual, which resulted in a set of  
94 47,346 variants. SV genotyping is nontrivial and associated with high uncertainty (7).  
95 Thus, we utilized the sampling scheme to filter for variants complying with basic  
96 population genetic assumptions (Fig. 2A)(19). Variants that were excluded according  
97 to these criteria were enriched for deletions and clustered near the end of chromosomes  
98 (linear model,  $p = 10^{-16}$ , Fig. 2B, C). Increased densities of repetitive elements (Fig.  
99 2D), particularly tandem repeats, in these regions are conducive to erroneous genotype  
100 calling, though it is possible that a subset of these phylogenetically recurring variants  
101 indeed represent true positive, hypermutable sites.  
102



103  
104  
105  
106

**Fig. 2 | Phylogenetic filtering of read mapping-based structural variants.** (A), Example genotype plots of LR-based variants according to phylogenetically informed filtering. Given the large divergence time of 13 million years (50) between the crow and jackdaw lineage, the proportion of polymorphisms

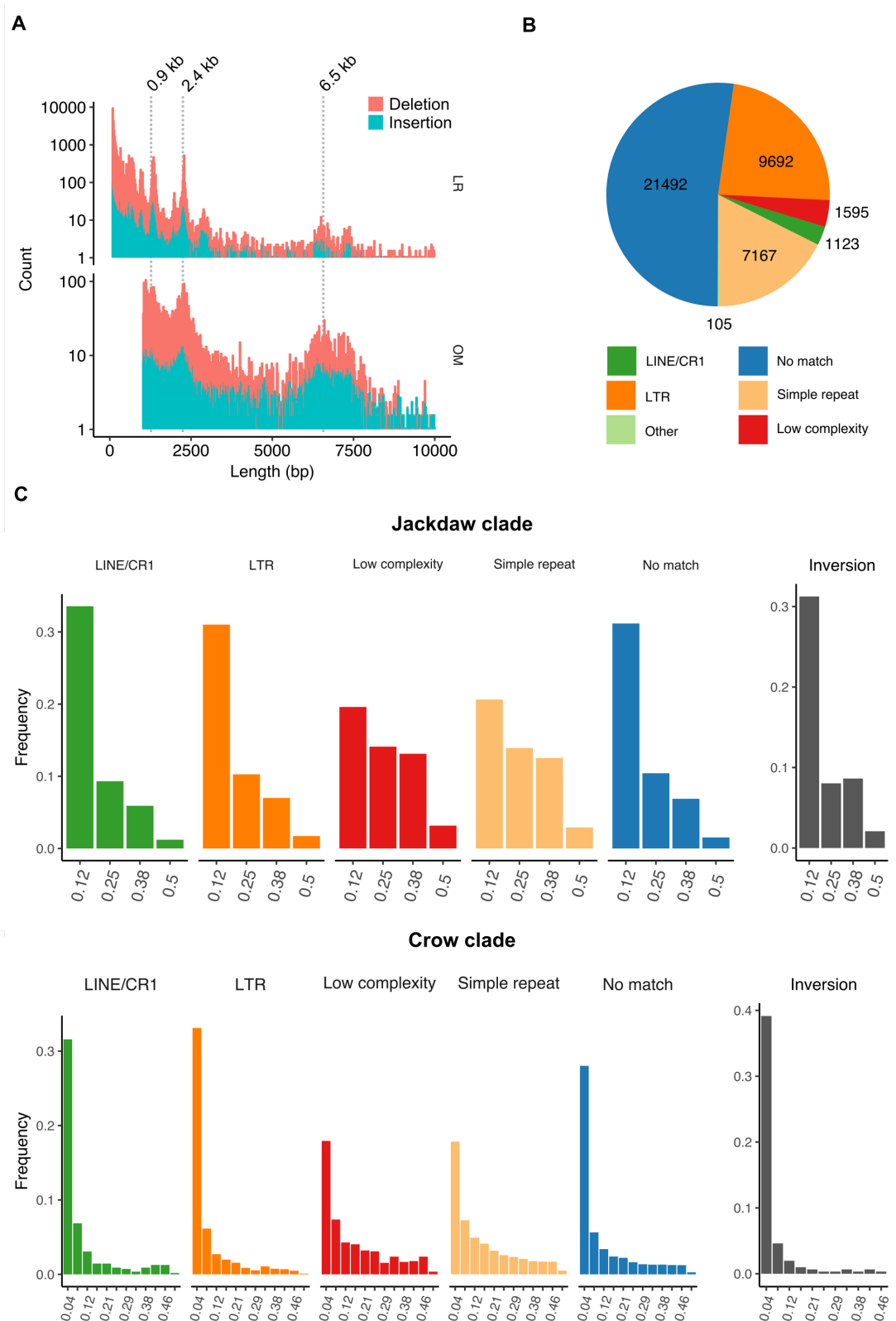
107 shared by descent is negligible (51) and therefore likely constitutes false positives or hypermutable sites  
108 (left panel). Variants segregating exclusively in the jackdaw or crow clade (middle and right panel),  
109 however, comply with the infinite sites model and were retained accordingly. Plotted are genotypes of  
110 one representative chromosome (chromosome 18), with genotypes of variants in different colors, where  
111 each row corresponds to one individual ( $N = 8$  individuals jackdaw clade and  $N = 24$  individuals crow  
112 clade). Note that, due to the tolerance of a certain number of mis-genotyped variants per clade, some  
113 variants are present in both clades. **(B)**, Excluded versus retained variants in relation to SV class and  
114 chromosomal distribution. Excluded variants are enriched for deletions (LMM,  $p < 10^{-16}$ ) and **c**, are most  
115 abundant at chromosome ends, coinciding with **(D)**, an increased repeat density.

116

117

118 After the phylogenetically informed filtering step, we retained a final set of 41,868  
119 variants (88.43 % of the initial, unfiltered set) segregating within and between species.  
120 Of these, a small proportion was classified as inversions (694, 1.657 %), whereas the  
121 vast majority was attributed to insertions (23,235, 55.495 %) and deletions (17,939 ,  
122 42.846 %) relative to the hooded crow reference. Variant sizes were largest for  
123 inversions, with a median size of 980 bp (range: 51 – 99,824 bp), followed by insertions  
124 (248 bp, range: 51 – 45,373 bp) and deletions (154 bp, range: 51 – 94,167 bp). The  
125 latter showed noticeable peaks in the size distribution at around 900, 2,400 and  
126 6,500 bp (**Fig. 3A**, for inversions see **Supplementary Fig. S1**), which likely stem from  
127 an overrepresentation of paralogous repeat elements. The five most common repeat  
128 motifs in insertions and deletions belonged to endogenous retrovirus-like LTR  
129 retrotransposon families and accounted for 22.78 % of all matches to a manually  
130 curated repeat library (**Supplementary Table S3**). This suggests recent activity of this  
131 transposable element group, as has been previously reported in other songbird species  
132 (20). More than half of all insertions and deletions could not be associated with any  
133 known repeat motif (52.19 %). The remainder was distributed approximately equally  
134 between tandem repeats (e.g. simple and low complexity repeats) and interspersed  
135 repeats. The latter category was dominated by LTR and LINE/CR retrotransposons with  
136 only a small number of SINE retrotransposons (**Fig. 3B, Table 2**). These different types  
137 of repeat elements exhibit fundamentally different mutation mechanisms (21) and  
138 effects on neighboring genes (22), such that repeat annotations are of crucial  
139 importance for the downstream population genetic analysis of SV.

140



141  
142  
143  
144  
145  
146  
147  
148

**Fig. 3 | Characterization and allele frequencies of SV. (A)** Length distributions of deletions and insertions shorter than 10 kb identified with LR (upper panel) and OM (lower panel) data. Pronounced peaks at 0.9, 2.2 kb in the LR and at 2.3 and 6.5 kb in the OM variants likely stem from an overrepresentation of specific repeats. Indeed, among the five most common repeats found in insertions and deletions are LTR retrotransposons with a consensus sequence length of 670, 1,315, 6,022 bp, respectively. **(B)** Content of insertion and deletion sequences. About half of all variants were assigned to a known repeat family, of which transposable elements from the LTR retrotransposon subclass were

149 most common, followed by simple repeats (including microsatellites) and low complexity repeats. (C)  
150 Folded allele frequency spectra of structural variants. Upper and lower panels correspond to the jackdaw  
151 and crow clade, respectively. The five left panels depict the minor allele frequencies of insertions and  
152 deletions, and the rightmost panel that of inversions.

153  
154

**Table 2 | Characterization of LR insertions and deletions.**

<b>Classification</b>	<b>Number</b>	<b>Percentage</b>
<b>Tandem repeat total</b>	8847	21.48
Simple repeat	7167	17.4
Low complexity repeat	1595	3.87
Satellite	75	0.18
rRNA	5	< 0.05
tRNA	3	< 0.05
Macrosatellite	1	< 0.05
<b>Interspersed repeat total</b>	10828	26.3
LTR retrotransposon	9692	23.53
LINE / CR1 retrotransposon	1123	2.27
SINE retrotransposon	11	< 0.05
D/hAT-Charlie element	2	< 0.05
<b>No match</b>	21492	52.19

155  
156

157 We then scrutinized structural variation segregating within clades sharing recent  
158 common ancestry. A total of 35,723 and 29,555 variants remained after filtering in the  
159 jackdaw ( $N = 8$  individuals; *C. monedula*, *C. dauuricus*) and crow clade ( $N = 24$ ; *C.*  
160 *corone*) spp., *C. brachyrhynchos*), respectively. Using the full data set across all  
161 populations within each clade allowed us to compare folded allele frequency spectra  
162 between SV classes and repeat types with high resolution (for population specific  
163 spectra unbiased by population structure see **Supplementary Fig. S2**). Consistent with  
164 recent studies in grapevine and *Drosophila* SV (12, 23), the distribution of allele  
165 frequencies was skewed towards rare alleles (**Fig. 3C**). However, allele frequency  
166 spectra of different SV classes differed in shape. While insertions and deletions  
167 associated with LTR elements, LINE/CR1 elements or without any known match as  
168 well as inversions exhibited the typical pattern of a strongly right-skewed frequency  
169 distribution, allele frequencies of simple and low complexity repeats were shifted  
170 towards intermediate frequencies. Besides a potential technical bias due to the more  
171 difficult genotyping and variant discovery of these classes (24), this pattern is consistent  
172 with convergence to intermediate allele frequencies due to high mutation rates (21).  
173 These results illustrate how different underlying mutation dynamics potentially impact  
174 the analysis of population genetic parameters for SV.

175

176 To improve our ability to detect larger SV and to provide an independent orthogonal  
177 approach for SV discovery, we generated an additional 14 optical maps (**Fig. 1A**) and  
178 compared them to the hooded crow reference assembly. Following that approach, we  
179 identified 12,807 insertions, 8,799 deletions and 293 inversions. As expected from the

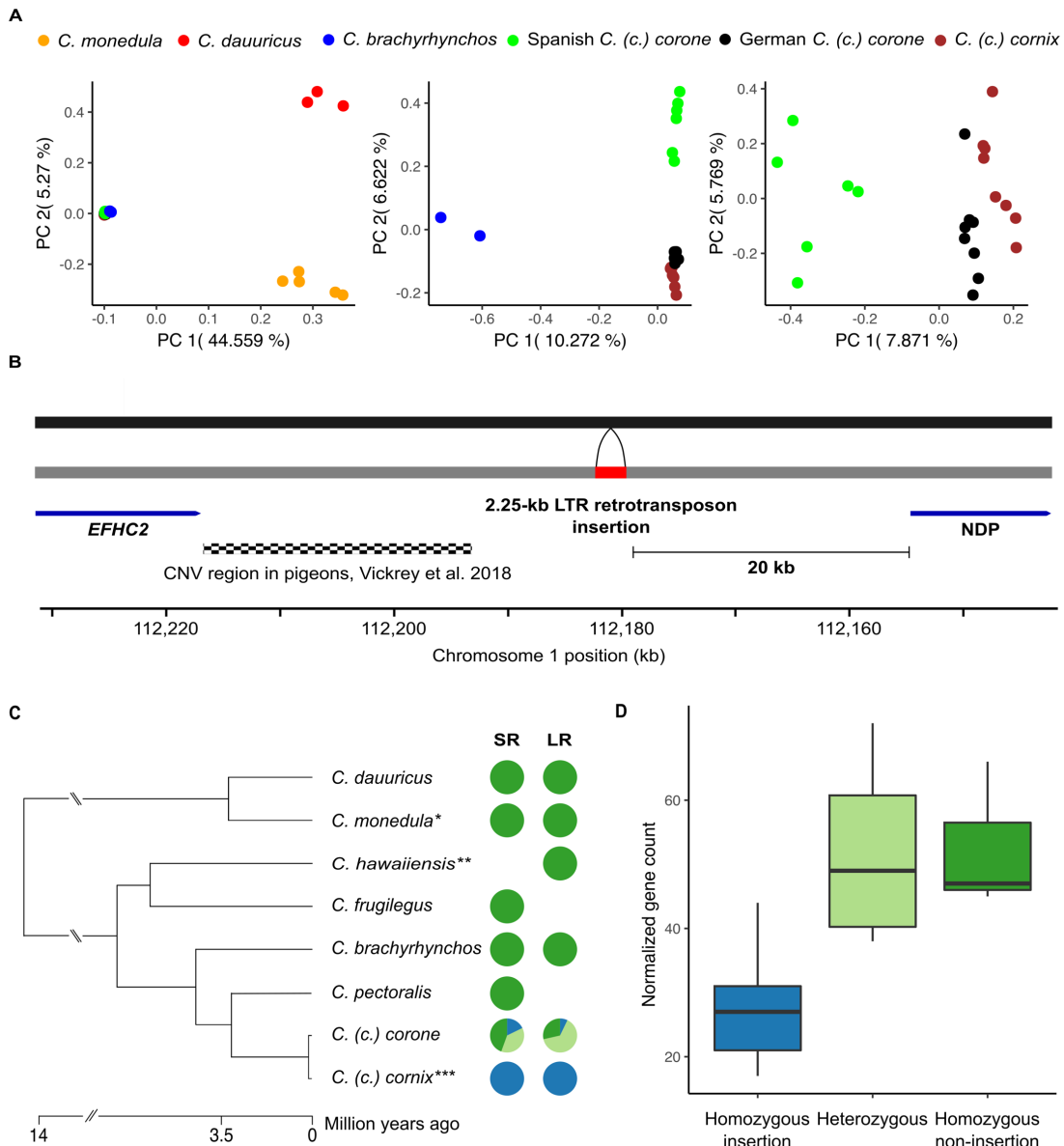
180 increased size of individually assessed DNA molecules (mean molecule N50 = 223.38  
181 kb), variants identified with this approach exhibited a different size range (**Fig. 3A**)  
182 after applying the same upper limit (100 kb) as for the LR SV calls and a lower limit of  
183 resolution (1 kb) (25). Interestingly, insertion and deletions were not only enriched at  
184 lengths around 0.9 and 2.4 kb as seen in the LR-based SV calling, but also at ~ 6.5 kb,  
185 indicating an influence of the TguERV1-Ld\_I\_corCor LTR retrotransposon, which was  
186 the third most common single repeat in the LR variant set with a consensus sequence  
187 length of 6,022 bp (**Supplementary Table S3**). Thus, independent approaches  
188 targeting different size ranges of SV are vital to increase sensitivity in detecting hidden  
189 genetic variation.

190  
191 To increase our sample size and expand our analysis to further populations and species  
192 (**Fig. 1A**), we applied a combination of three different short-read (SR) based SV  
193 detection approaches on previously published data of 127 individuals (18, 26). In total,  
194 we identified 132,025 variants of which 97,524 (73.87%) were unique to single  
195 individuals. In total, only 11,951 variants overlapped with the final set of variants  
196 identified in the long-read data set (corresponding to 9.05 % of SR and 28.54 % LR  
197 calls). This disconnect cannot be explained solely by differences in sample size. More  
198 likely, it indicates a high number of false-positives and false-negatives in the SR-based  
199 approach known for its sensitivity to the calling method (27) and disparity to LR-based  
200 calls (7). Therefore, we focused on the LR-based SV calls in the subsequent analysis  
201 and considered SR calls only for specific mutations.

202  
203 Next, we investigated population structure using principal component analyses (PCA).  
204 The pattern in **Fig. 4A** (based on LR data) recapitulates the pattern of population  
205 stratification found in Vijay et al. based on 16.6 million SNPs (18), and thus supports  
206 the general suitability of SV genotypes for population genetic analyses (for SR data see  
207 **Supplementary Fig. S3**). In order to identify SV associated with prezygotic  
208 reproductive isolation, we calculated genetic differentiation between phenotypically  
209 divergent populations connected by gene flow (18, 26) and allopatric populations  
210 within the same phenotype (18). Mean  $F_{ST}$  was low overall with values ranging from  
211 0.03 in the hooded versus carrion crow comparison to 0.156 in the hooded versus  
212 American crow comparison.

213





**Fig. 4 | SV-based population structure and LTR retrotransposon insertion upstream of the *NDP* gene.** (A) Principal component analysis based on SV genotypes. The first two principal components separate the crow and jackdaw clade, while principal components 3 to 5 separate lineages within the crow clade. (B) A 2.25-kb LTR retrotransposon insertion into the crow lineage (black bar: ancestral state, grey bar: derived, reference allele) belongs to the endogenous retrovirus-like family ERVK and the subfamily TguERV1-Ld-I and is located 20 kb upstream of the *NDP* gene. In close proximity, variation in copy number is associated with plumage pattern variation in pigeons. (C) Genotypes of the LTR element in short-read (SR) and long-read (LR) data. In both datasets, the LTR element insertion (blue) is fixed in all hooded crow populations. Species and populations with a black plumage are either polymorphic (light green) or fixed non-insertion (green). (D) Gene expression of *NDP*. Normalized gene counts of 18 individuals are significantly associated with the insertion genotypes (LMM,  $p = 0.002$ ).

214  
 215  
 216  
 217  
 218  
 219  
 220  
 221  
 222  
 223  
 224  
 225  
 226  
 227  
 228  
 229  
 230  
 231  
 232  
 233

A total of 103 variants fell into the 99<sup>th</sup> percentile of  $F_{ST}$  in the gray-coated hooded versus all-black carrion crow population comparison in central Europe. These variants, located on in total 23 chromosomes, were considered as *ad hoc* candidate outlier loci subject to divergent selection (9), and were found at a median distance of 14.32 kb to

234 adjacent genes (range: 0 - 695.84 kb). (**Supplementary Table S4**). Ten of these outliers  
235 (10.31 %) were placed on chromosome 18, which only represents 1.22 % of the entire  
236 assembly, corresponding to an ~8.5-fold enrichment. Given that outliers are located in  
237 the proximity of previously identified genes presumably under divergent selection  
238 (such as *AXIN2* and *RGS9*, Supplementary Table S5), this supports a crucial role of  
239 chromosome 18 in maintaining plumage divergence (26, 28).

240  
241 The three highest  $F_{ST}$  outliers included an 86 bp indel on chromosome 18 inside of a  
242 tandem repeat array, a 1.56 kb indel on chromosome 3 and a 2.25 kb indel on  
243 chromosome 1 (**Supplementary Table S5**). The latter, an LTR retrotransposon  
244 insertion, was located 20 kb upstream of the *NDP* gene on chromosome 1 (**Fig. 4B**), a  
245 gene known to contribute to the maintenance of color divergence across the European  
246 crow hybrid zone (28). Molecular dating based on the LTR region suggest an insertion  
247 event at <534,000 years ago upon diversification of the European crow lineage (**Fig.**  
248 **4C**) (18). In current day populations, the insertion still segregates in all-black crows  
249 including *C. (c.) corone* in Europe and *C. (c.) orientalis* in Russia ( $N$  individuals with  
250 LR = 14 and with SR = 45 genotypes) (**Fig. 4C**). All hooded crow *C. (c.) cornix*  
251 individuals, however, genotyped with LR ( $N = 9$ ) and SR data ( $N = 61$ ) were  
252 homozygous for the insertion regardless of their population of origin. This finding is  
253 consistent with a selective sweep in proximity to the *NDP* gene that has previously been  
254 suggested for hooded crow populations (26, 28). Recent work has also shown that the  
255 *NDP* gene exhibits decreased gene expression in grey feather follicles of hooded crows,  
256 suggesting a role in modulating overall plumage color patterning (29). Following re-  
257 analysis of normalized gene expression data for 8 carrion and 10 hooded crows (29),  
258 we found a significant association between the homozygous insertion genotype and  
259 decreased *NDP* gene expression levels (linear model,  $p = 0.002$ ) (**Fig. 4D**), consistent  
260 with reduced pigmentation in hooded crows (29).

261  
262 To further investigate the relationship between the abovementioned insertion and  
263 phenotypic differences between all-black *C. (c.) corone* and gray-coated *C. (c.) cornix*  
264 populations, we genotyped 120 individuals from the European hybrid zone using PCR  
265 (Methods, (28)). Including data of adjacent SNPs for the same individuals, we tested  
266 the association between genotype and pigmentation phenotype. A statistical model  
267 including the insertion fit best to the observed phenotypes ( $\Delta AICc = 2.33$ , but  $\Delta BIC =$   
268  $-0.12$ ) explaining an additional 10.32% of the variance of the phenotype-derived PC1  
269 relative to the adjacent SNPs. The insertion lies upstream of *NDP* in close proximity to  
270 an orthologous region in pigeons containing a copy number variation shown to  
271 modulate plumage patterning (**Fig. 4B**) (30). Reminiscent of the wing color altering TE  
272 insertion in the peppered moth (3), this insertion thus constitutes a prime candidate  
273 causal mutation modulating gene expression with phenotypic consequences;  
274 reminiscent of the TE insertion in the peppered moth altering wing coloration (3). While  
275 such insertions have usually been associated with increased expression of the affected  
276 gene (31), there are also examples of TE insertions repressing gene activity, as observed  
277 here (32).

278  
279 In conclusion, this study provides the first comprehensive population-level SV  
280 catalogue in a non-model organism, further elucidating the role of SV on modulating  
281 expression of evolutionary important genes with phenotypic consequences. Given that  
282 the majority of SV is likely still uncovered in most organisms (33), these results mark

283 an important hallmark for the field highlighting the evolutionary importance of SV in  
284 natural populations and the need for rigorous methodological approaches.

285

## 286 **Material and Methods**

287

### 288 **Short-read sequencing data**

289 We compiled raw short-read sequencing data from Poelstra et al. 2014 and Vijay et al.  
290 2016 (18, 26) for *Corvus (corone) spp.*, *C. frugilegus*, *C. dauuricus*, *C. monedula* and  
291 *C. brachyrhynchos* (for more information on the origin of samples and accession  
292 numbers of the data see **Supplementary Table S5**). Overall, 127 individuals had an  
293 average 12.6-fold sequencing coverage using paired-end libraries (primarily)  
294 sequenced on an Illumina HiSeq2000 machine.

295

### 296 **DNA extraction and long-read sequencing**

297 First, we extracted high-molecular weight DNA from a total of 32 samples using either  
298 a modified phenol-chloroform extraction protocol (14), or the Qiagen Genomic-tip kit  
299 (following manufacturer's instructions) from frozen blood samples. For sampling  
300 details, see **Supplementary Table S5**. Extracted DNA was eluted in 10 mM Tris buffer  
301 and stored at -80 °C. The quality and concentration of the DNA was assessed using a  
302 0.5 % agarose gel (run for >8 h at 25 V) and a Nanodrop spectrophotometer  
303 (ThermoFisherScientific). Long-read sequencing DNA libraries were prepared using  
304 the SMRTbell Template Prep Kit 1.0 (Pacific Biosciences). For each library, 10 µg  
305 genomic DNA was sheared into 20-kb fragments with the Hydroshear  
306 (ThermoFisherScientific) instrument. SMRTbell libraries for circular consensus  
307 sequencing were generated after an Exo VII treatment, DNA damage repair and end-  
308 repair before ligation of hairpin adaptors. Following an exonuclease treatment and PB  
309 AMPure bead wash, libraries were size-selected using the BluePippin system with a  
310 minimum cutoff value of 8,500 bp. All libraries were then sequenced on either the RSII  
311 or Sequel instrument from Pacific Biosciences, totaling 324 RSII and 76 Sequel SMRT  
312 cells, respectively, resulting in 754 Gbp of raw data.

313

### 314 **Genome assembly**

315 In birds, females are the heterogametic sex (ZW). For this study, we were interested in  
316 a high-quality assembly of all autosomes and the shared sex chromosome (Z) and  
317 accordingly chose male individuals for the genome assemblies. Note, however, that this  
318 choice excludes the female-specific W chromosome *a priori*. Diploid genome assembly  
319 was performed for both a hooded crow and a jackdaw individual. For the former a long-  
320 read based genome assembly has previously been published (14) and is available under  
321 the accession number GCA\_002023255.2 at the repository of the National Center for  
322 Biotechnology Information (NCBI, [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)). Here, we (re)assembled  
323 raw reads using updated filtering and assembly software. First, all SMRT cells for the  
324 respective individuals (102 for the hooded crow individual S\_Up\_H32, 70 for the  
325 jackdaw individual S\_Up\_J01) were imported into the SMRT Analysis software suite  
326 (v2.3.0). Subreads shorter than 500 bp or with a quality (QV) <80 were filtered out.  
327 The resulting data sets were used for *de novo* assembly with FALCON UNZIP v0.4.0  
328 (16). Initial FALCON UNZIP assemblies of hooded crow and jackdaw consisted of  
329 primary and associated contigs with a total length of 1,053.37 Mb and 965.95 Mb for  
330 the hooded crow and 1,073.84 and 1,092.55 Mb for the jackdaw, presumably  
331 corresponding to the two chromosomal haplotypes (for assembly statistics see  
332 **Supplementary Table S1**). To further improve the assembly, we performed consensus

333 calling of individual bases using ARROW (16). In addition, we obtained the genome  
334 of the Hawaiian crow (*Corvus hawaiiensis*) from the repository of NCBI with accession  
335 number GCA\_003402825.1. This genome had been likewise derived from long-reads  
336 generated with the SMRT technology and assembled using FALCON UNZIP (16). To  
337 assess the completeness of the newly assembled genomes we used BUSCO v2.0.1 (34).  
338 The aves and the vertebrate databases were used to indentify ultra conserved  
339 orthologous gene sets (**Supplementary Table S1**).

340

#### 341 **Optical mapping data and assembly**

342 We generated additional optical map assemblies for two jackdaw individuals, 8 carrion  
343 crow individuals and 4 additional hooded crow individuals, following the same  
344 approach used for the optical map assembly of the hooded crow individual (see  
345 Weissensteiner et al.(14)). In brief, we extracted nuclei of red blood cells and captured  
346 them in low-melting point agarose plugs. DNA extraction was followed by melting and  
347 digesting of the agarose resulting in a high-molecular weight DNA solution. After  
348 digestion with a nicking endonuclease (Nt.BspQI) which inserts a fluorescently labelled  
349 nick strand, the sample was loaded onto an IrysChip, which was followed by  
350 fluorescent label detection on the Irys instrument. The assembled consensus maps were  
351 then used to perform SV calling as part of the Bionano Access 1.3.1 Bionano Solve  
352 pipeline 3.3.1 (pipeline version 7841). As reference an *in-silico* map of the hooded crow  
353 reference assembly was used. Molecule and assembly statistics of optical maps can be  
354 found in **Supplementary Table S6**. For details regarding the hybrid scaffolding see  
355 Weissensteiner et al. 2017 (14).

356

357

#### 358 **Hi-C chromatin interaction mapping and scaffolding**

359 One Dovetail Hi-C library was prepared from a hooded crow sample following  
360 Lieberman-Aiden et al. (2009) (35). In brief, chromatin was fixed in place with  
361 formaldehyde in the nucleus and extracted thereafter. Fixed chromatin was digested  
362 with DpnII, the 5' overhangs filled with biotinylated nucleotides and free blunt ends  
363 were ligated. After ligation, crosslinks were reversed and the DNA purified from the  
364 protein. Purified DNA was treated such that all biotin was removed that was not internal  
365 to ligated fragments. The DNA was then sheared to ~350 bp mean fragment size and  
366 sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-  
367 compatible adapters. Biotin-containing fragments were isolated using streptavidin  
368 beads before PCR enrichment of each library. The library was then sequenced on an  
369 Illumina HiSeq X (rapid run mode). The Dovetail Hi-C library reads and the contigs of  
370 the primary FALCON UNZIP assembly were used as input data for HiRise, a software  
371 pipeline designed specifically for using proximity ligation data to scaffold genome  
372 assemblies (36). An iterative analysis was conducted. First, Hi-C library sequences  
373 were aligned to the draft input assembly using a modified SNAP read mapper  
374 (<http://snap.cs.berkeley.edu>). The separation of read pairs mapped within draft  
375 scaffolds were analyzed by HiRise to produce a likelihood model for genomic distance  
376 between read pairs, and the model was used to identify and break putative misjoins, to  
377 score prospective joins and make joins above a threshold. The resulting 48 super-  
378 scaffolds were assigned to 27 chromosomes based on synteny to the flycatcher genome  
379 version (NCBI accession GCA\_000247815.2) (37) using LASTZ (38). The final Hi-C  
380 scaffolded hooded crow assembly is available as a Dryad repository, file XX.

381

382

### 383 **Assembly-based SV and SNP detection**

384 We aligned the associated contigs of all three assemblies (hooded crow, jackdaw and  
385 Hawaiian crow) to the primary contigs (super-scaffolded to chromosome level for  
386 hooded crow) using MUMmer (39). SNPs were then identified using *show-snps* with  
387 the options `-Clr` and `-T` following a filtering step with `delta-filter -r` and `-q`. We only  
388 considered single-nucleotide differences in this analysis.

389 Structural variants between the two haplotypes of each assembly were identified using  
390 two independent approaches. First, we used the alignments produced with MUMmer to  
391 identify variants using the Assemblytics tool (40). We then converted the output to a  
392 *vcf* file using SURVIVOR (v1.0.3) (27). Independently, we used the smartie-sv pipeline  
393 to identify structural variants (41), and then converted and merged the output with the  
394 Assemblytics-based variant set with SURVIVOR. This final unified variant set was  
395 then used to calculate SV-density in non-overlapping 1-Mb windows.

396

### 397 **Repeat annotation and characterization of insertions and deletions**

398

399 To characterize the repeat content of the hooded crow assembly, we used the repeat  
400 library from Vijay et al. (18). Raw consensus sequences were manually curated  
401 following the method used in Suh et al. (2018) (20). Every consensus sequence was  
402 aligned back to the reference genome, then the best 20 BLASTN (42) hits were  
403 collected, extended by 2 kb and aligned to one another using MAFFT (v6; (43)). The  
404 alignments were manually curated applying a majority-rule and the superfamily of each  
405 repeat assessed following Wicker et al. (44). We then masked the new consensus  
406 sequences in CENSOR (<http://www.girinst.org/censor/index.php>) and named them  
407 according to homology to known repeats present in Repbase(45). Repeats with high  
408 sequence similarity to known repeats were given the name of the known repeat + suffix  
409 "\_corCor"; repeats with partial homology were named with the suffix "-L\_corCor"  
410 where "L" stands for "like" (20). Repeats with no homology to other known repeats  
411 were considered as new families and named with the prefix "corCor" followed by the  
412 name of their superfamilies. Using this fully curated repeat library (**Supplementary**  
413 **file S1**), we performed a RepeatMasker (46) search on all sequences reported for  
414 insertion and deletion variants. In case of multiple different matches per variant or  
415 individual, we took the match with the highest overlap with the query sequence to yield  
416 a single match for each variant. We also performed a RepeatMasker search with the  
417 curated library to estimate repeat density per 1 Mb window in the hooded crow  
418 reference assembly.

419

### 420 **Read-mapping based SV and SNP detection**

421 We aligned PacBio long-read data of all re-sequenced individuals to both the hooded  
422 crow and jackdaw reference assembly using NGM-LR (47) (v0.2.2) with the `-pacbio`  
423 option and sorted and indexed resulting alignments with samtools (48)(v1.9). Initial SV  
424 calling per individual was then performed using Sniffles (47) (v1.0.8) with parameters  
425 set to a minimum support of 5 reads per variant (`--min_support 5`) and enabled `-`  
426 `genotype`, `-cluster` and `-report_seq` options. We removed abundant translocation calls  
427 indicative of an excess of false positives and filtered remaining variants for a maximum  
428 length of 100 kb and a maximum read support of 60 with bcftools (49). Both of these  
429 filtering steps have been shown to be necessary to remove erroneously called variants.  
430 Next, we generated a merged multi-sample *vcf* file consisting of all individuals from  
431 both the crow and the jackdaw clades with SURVIVOR merge and options set to 1000  
432 1 1 0 0 50. This merged *vcf* file was then used as an input to reiterate SV calling with

433 Sniffles for each individual with the `-Ivcf` option enabled, effectively genotyping each  
434 variant per individual. Resulting single individual *vcf* files were again merged with the  
435 SURVIVOR command described above and variants overlapping with assembly gaps  
436 were removed. We converted the *vcf* file into a genotype file with *vcftools* (49)  
437 (v0.1.15) for downstream analysis.

438 To account for the high amount of genotyping errors and false positives after initial  
439 filtering, we employed a ‘phylogenetic’ filtering strategy. The crow and jackdaw clades  
440 diverged roughly 13 million years ago (50), such that the proportion of polymorphisms  
441 shared by descent is near negligible (51). Moreover, under the infinite sites model,  
442 recurrent mutations are not expected, such that polymorphisms segregating in both  
443 lineages most likely constitute false positives. For population genetic analyses of the  
444 jackdaw clade, we therefore considered only variants which were homozygous for the  
445 reference in crow clade individuals, allowing for a maximum of four genotyping errors.  
446 In the crow clade analyses, we only retained variants which were either fixed for the  
447 reference or the variant allele in the jackdaw clade, allowing for 2 genotyping errors. It  
448 is likely that this conservative approach excludes variants with a high mutation rate  
449 (52). However, since it is difficult to differentiate such variants from genotyping errors,  
450 we deemed this filter necessary to yield a set of more reliable variants. Due to the  
451 tolerance of genotyping errors, there is a number of variants present in both clades,  
452 most of them fixed or almost fixed in both clades. Extensive manual curation would be  
453 necessary to differentiate between genotyping errors and variants truly polymorphic  
454 between clades. To find common features in filtered versus kept variants, we applied a  
455 generalized linear mixed-effects model with a binomial error structure, in which we  
456 coded the dependent variable as 1 for a retained variant and as 0 for a filtered variant.  
457 As covariates we included the distance to the chromosome end and variant class as a  
458 factor (insertion, deletion or inversion). We further fitted chromosome identity as a  
459 random intercept term. All models were run in R (v3.2.3, R Core Team) using the *lme4*  
460 package (53) (v1.1-19).

461 The short-read data were mapped using BWA-MEM with the `-M` option to the hooded  
462 crow reference assembly (54). We used LUMPY (55), DELLY (56) and Manta (57) to  
463 obtain SV calls for each sample using their respective default parameters. Subsequently  
464 the individual SV calls per sample were merged using SURVIVOR (27) merge with  
465 the parameters: “1000 2 1 0 0 0”. This filtering step retained only SV calls for which 2  
466 out of the 3 callers had reported a call within 1 kbp. Next, we computed the coverage  
467 of low mapping quality reads (MQ<5) for each sample independently and recorded  
468 regions where the low MQ coverage exceeded 10. SV calls which overlapped these  
469 regions were filtered out.

470

### 471 **Optical mapping-based SV detection**

472 The assembled optical maps were used to identify SV compared to the provided  
473 reference, which is part of the assembly pipeline or can be run manually. SV calling  
474 was based on the alignment between an individual assembled consensus *cmap* and the  
475 *in-silico* generated map of the reference using a multiple local alignment algorithm and  
476 detecting SV signatures. The detection algorithm identifies insertions, deletions,  
477 translocation breakpoints, inversion breakpoints and duplications. The results are in a  
478 generated file in the Bionano specific format *smap* in which the SVs are classified as  
479 homozygous or heterozygous. This resulting *smap* file was converted to *vcf* format  
480 (version 4.2) for further downstream processing.

481

482

## 483 **Population genetic analysis of structural variants**

484 To investigate population structure, we performed principal component analyses (PCA)  
485 with both the long-read and short-read variant sets using the R packages SNPrelate  
486 (v1.4.2.) and gdsfmt (v1.6.2) (58). We further calculated the folded allele frequency  
487 spectrum using minor allele frequencies of variants for all populations and clades.

488 To estimate genetic differentiation of structural variations, we calculated  $F_{ST}$  for each  
489 variant using vcftools (59). We employed the Weir and Cockerham estimator for  $F_{ST}$   
490 (60), variants with an  $F_{ST}$  exceeding the 99<sup>th</sup> percentile were considered as outliers.

491

## 492 **Analyses of SV in the vicinity of the *NDP* gene**

493 The LTR retrotransposon insertion identified upstream of the *NDP* gene on  
494 chromosome 1 - an ERV1 element belonging to the subfamily TguERV1-Ld-I - has  
495 initially been called as a deletion relative to the reference (hooded crow) assembly. To  
496 estimate its age, we assumed that the two long terminal repeats of the full-length LTR  
497 retrotransposon were identical at the time of insertion (61). Thus, we quantified the  
498 number of substitutions and 1-bp indels between the left and right LTR of the insertion  
499 at position 112,179,329 on chromosome 1 of the hooded crow reference. The LTRs  
500 showed 5 differences which we then divided by the length of the LTR (296 bp) and by  
501 twice the neutral substitution rate per site and million years (0.0158 (18)). Assuming  
502 that all differences between the left and right LTR of this insertion are fixed, this  
503 estimate yields an upper bound of the insertion age. However, overlap with SNPs  
504 segregating in the hooded crow population suggests that all 5 differences were not fixed  
505 and the insertion could thus be considerably younger.

506 To investigate a potential link between the LTR insertion and differences in plumage  
507 coloration, we re-analyzed gene expression data from 10 black-and-grey hooded crows  
508 and 8 all-black carrion crows raised under common garden conditions (29). Expression  
509 was measured for messenger RNA derived from feather buds at the torso, where carrion  
510 crows have black feathers and hooded crows are grey. We inferred the insertion  
511 genotype for each individual using short-read sequencing data via visual inspection of  
512 the alignments to the hooded crow reference. We then fitted a linear model with  
513 normalized *NDP* expression data as the dependent variable and *NDP* indel genotype as  
514 the predictor. We decomposed the effect of the insertion genotype into an additive  
515 component (the number of non-inserted minor allele copies - 0, 1, or 2 - as a covariate)  
516 and a dominance component (homozygous = 0, heterozygous = 1).

517 To further establish a link between the LTR retrotransposon insertion and phenotypic  
518 differences, we made use of a hybrid admixture data set from the European hybrid zone  
519 (28). We designed three sets of PCR primers to genotype the insertion for 120  
520 phenotyped individuals from the European hybrid zone of all-black *C. (c.) corone* and  
521 black-and-grey *C. (c.) cornix* crows. For absence of the insertion, a pair of primers  
522 located in the sequence flanking the insertion was used (A\_F\_3  
523 ‘AGTAACTGTCCTCTGTAGTGCAGG’ and A\_R\_3  
524 ‘CCTGGGTAAGATCACAGTGTTC’) resulting in a 197 bp fragment. For presence  
525 of the insertion, a pair of primers with one in the flanking and one in either left or right  
526 LTR region of the insertion (P\_L\_F\_1 ‘TCCTCTGTAGTGCAGGACTGG’ and  
527 P\_L\_R\_2 ‘CACCCATGGTTTCCCTCACA’, as well as P\_R\_F\_1  
528 ‘GGATCGGGGATCGTTCTGCT’ and P\_R\_R\_1  
529 ‘CACAGCCCCAGAAGATGTGC’), resulting in fragments of 659 and 564 bp,  
530 respectively. A representative gel picture used for genotyping can be found in the  
531 **Supplementary Fig. S4**. Phenotypic data was taken from Knief et al. (28) who  
532 summarized 11 plumage color measures on the dorsal and ventral body into a principal

533 component (PC1), explaining 78% of the phenotypic variation. We then tested whether  
534 the interaction between chromosome 18 and the insertion genotype explained more  
535 variation in plumage color than the interaction between chromosome 18 and the most  
536 significant SNP near the *NDP* gene (28). We fitted two linear regression models on the  
537 same subset of the data that contained no missing genotypes ( $N = 120$  individuals). In  
538 both models, we used color PC1 as our dependent variable. In the first model, we fitted  
539 the interaction between chromosome 18 and the insertion genotype, and in the second  
540 model the interaction between chromosome 18 and the SNP genotype as our  
541 independent variables. Both variables were coded as 0, 1, 2 copies of the derived allele  
542 and fitted as factors. We selected the model with the better fit to the data by estimating  
543 the AICc and BIC and deemed a  $\Delta\text{AICc} \geq 2$  as significant.

544

#### 545 **Competing interests**

546 Kees-Jan Francoijs is an employee of BioNano Genomics (San Diego, CA).

547

#### 548 **Acknowledgements**

549

550 We thank John Marzluff, Vittorio Bagglione and Kristaps Sokolovskis for providing  
551 sample material. Reto Burri and Sergio Tusso Gomez provided helpful input for the  
552 downstream analysis. We are thankful for being able to use the UPPMAX Next-  
553 Generation Sequencing Cluster and Storage (UPPNEX) project, funded by the Knut  
554 and Alice Wallenberg Foundation and the Swedish National Infrastructure for  
555 Computing. This work was supported by the Swedish Research Council (grant number  
556 621-2010-5553 to J.B.W.W. and grant number 2016-05139 to A.S.), the European  
557 Research Council (grant number ERCStG-336536 to J.B.W.W.) and the National  
558 Institutes of Health (grant number UM1 HG008898 to F.J.S).

559

#### 560 **Author contributions**

561 M.W. and J.W. conceived of the study, conducted field work and wrote the manuscript  
562 with input from all other authors. M.W. conducted lab work and all bioinformatic  
563 analyses with help from V.P., V.W., S.D.P., A.S. (repeat annotation) and U.K.  
564 (statistical analyses). W.H. conducted field work. I.B. generated genome assemblies  
565 and F.J.S. performed short-read based SV calling.

566

#### 567 **References and Notes**

- 568 1. L. Feuk, A. R. Carson, S. W. Scherer, Structural variation in the human  
569 genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
- 570 2. C. Küpper, M. Stocks, J. E. Risse, N. dos Remedios, L. L. Farrell, S. B. McRae, T.  
571 C. Morgan, N. Karlionova, P. Pinchuk, Y. I. Verkuil, A. S. Kitaysky, J. C.  
572 Wingfield, T. Piersma, K. Zeng, J. Slate, M. Blaxter, D. B. Lank, T. Burke, A  
573 supergene determines highly divergent male reproductive morphs in the  
574 ruff. *Nat. Genet.* **48**, 79–83 (2016).
- 575 3. A. E. van't Hof, P. Campagne, D. J. Rigden, C. J. Yung, J. Lingley, M. A. Quail, N.  
576 Hall, A. C. Darby, I. J. Saccheri, The industrial melanism mutation in British  
577 peppered moths is a transposable element. *Nature.* **534**, 102–105 (2016).
- 578 4. J. Huddleston, E. E. Eichler, An Incomplete Understanding of Human Genetic  
579 Variation. *Genetics.* **202**, 1251–1254 (2016).



- 580 5. B. Weckselblatt, M. K. Rudd, Human Structural Variation: Mechanisms of  
581 Chromosome Rearrangements. *Trends Genet.* **31**, 587–599 (2015).
- 582 6. M. J. P. Chaisson, R. K. Wilson, E. E. Eichler, Genetic variation and the de novo  
583 assembly of human genomes. *Nat. Rev. Genet.* **16**, 627–640 (2015).
- 584 7. F. J. Sedlazeck, H. Lee, C. A. Darby, M. C. Schatz, Piercing the dark matter:  
585 bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* **19**,  
586 329–346 (2018).
- 587 8. S. Goodwin, J. D. McPherson, W. R. McCombie, Coming of age: ten years of  
588 next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351  
589 (2016).
- 590 9. J. B. W. Wolf, H. Ellegren, Making sense of genomic islands of differentiation  
591 in light of speciation. *Nat. Rev. Genet.* **18**, 87–100 (2017).
- 592 10. M. J. P. Chaisson, A. D. Sanders, X. Zhao, A. Malhotra, D. Porubsky, T. Rausch, E.  
593 J. Gardner, O. Rodriguez, L. Guo, R. L. Collins, X. Fan, J. Wen, R. E. Handsaker,  
594 S. Fairley, Z. N. Kronenberg, X. Kong, F. Hormozdiari, D. Lee, A. M. Wenger, A.  
595 Hastie, D. Antaki, P. Audano, H. Brand, S. Cantsilieris, H. Cao, E. Cerveira, C.  
596 Chen, X. Chen, C.-S. Chin, Z. Chong, N. T. Chuang, C. C. Lambert, D. M. Church,  
597 L. Clarke, A. Farrell, J. Flores, T. Galeev, D. Gorkin, M. Gujral, V. Guryev, W.  
598 Haynes Heaton, J. Korlach, S. Kumar, J. Y. Kwon, J. E. Lee, J. Lee, W.-P. Lee, S.  
599 P. Lee, S. Li, P. Marks, K. Viaud-Martinez, S. Meiers, K. M. Munson, F. Navarro,  
600 B. J. Nelson, C. Nodzak, A. Noor, S. Kyriazopoulou-Panagiotopoulou, A. Pang,  
601 Y. Qiu, G. Rosanio, M. Ryan, A. Stutz, D. C. J. Spierings, A. Ward, A. E. Welch,  
602 M. Xiao, W. Xu, C. Zhang, Q. Zhu, X. Zheng-Bradley, E. Lowy, S. Yakneen, S.  
603 McCarroll, G. Jun, L. Ding, C. L. Koh, B. Ren, P. Flicek, K. Chen, M. B. Gerstein,  
604 P.-Y. Kwok, P. M. Lansdorp, G. Marth, J. Sebat, X. Shi, A. Bashir, K. Ye, S. E.  
605 Devine, M. Talkowski, R. E. Mills, T. Marschall, J. O. Korbel, E. E. Eichler, C.  
606 Lee, Multi-platform discovery of haplotype-resolved structural variation in  
607 human genomes (2018), doi:10.1101/193144.
- 608 11. S. Tusso, B. P. S. Nieuwenhuis, F. J. Sedlazeck, J. W. Davey, D. C. Jeffares, J. B. W.  
609 Wolf, Ancestral Admixture Is the Main Determinant of Global Biodiversity in  
610 Fission Yeast. *Mol. Biol. Evol.* **36**, 1975–1989 (2019).
- 611 12. M. Chakraborty, J. J. Emerson, S. J. Macdonald, A. D. Long, Structural variants  
612 exhibit widespread allelic heterogeneity and shape variation in complex  
613 traits. *Nat. Commun.* **10** (2019), doi:10.1038/s41467-019-12884-1.
- 614 13. L. E. Flagel, J. H. Willis, T. J. Vision, The Standing Pool of Genomic Structural  
615 Variation in a Natural Population of *Mimulus guttatus*. *Genome Biol. Evol.* **6**,  
616 53–64 (2014).
- 617 14. M. H. Weissensteiner, A. W. C. Pang, I. Bunikis, I. H?ijer, O. Vinnere-Petterson,  
618 A. Suh, J. B. W. Wolf, Combination of short-read, long-read, and optical  
619 mapping assemblies reveals large-scale tandem repeat arrays with  
620 population genetic implications. *Genome Res.* **27**, 697–708 (2017).

- 621 15. J. T. Sutton, M. Helmkampf, C. C. Steiner, M. R. Bellinger, J. Korlach, R. Hall, P.  
622 Baybayan, J. Muehling, J. Gu, S. Kingan, B. M. Masuda, O. A. Ryder, A High-  
623 Quality, Long-Read De Novo Genome Assembly to Aid Conservation of  
624 Hawaii's Last Remaining Crow Species. *Genes*. **9**, 393 (2018).
- 625 16. C.-S. Chin, P. Peluso, F. J. Sedlazeck, M. Nattestad, G. T. Concepcion, A. Clum, C.  
626 Dunn, R. O'Malley, R. Figueroa-Balderas, A. Morales-Cruz, G. R. Cramer, M.  
627 Delledonne, C. Luo, J. R. Ecker, D. Cantu, D. R. Rank, M. C. Schatz, Phased  
628 diploid genome assembly with single-molecule real-time sequencing. *Nat.*  
629 *Methods*. **13**, 1050 (2016).
- 630 17. R. B. Corbett-Detig, D. L. Hartl, T. B. Sackton, Natural Selection Constrains  
631 Neutral Diversity across A Wide Range of Species. *PLOS Biol.* **13**, e1002112  
632 (2015).
- 633 18. N. Vijay, C. M. Bossu, J. W. Poelstra, M. H. Weissensteiner, A. Suh, A. P.  
634 Kryukov, J. B. W. Wolf, Evolution of heterogeneous genome differentiation  
635 across multiple contact zones in a crow species complex. *Nat. Commun.* **7**,  
636 13195 (2016).
- 637 19. See Supplementary Materials. *Online*.
- 638 20. A. Suh, L. Smeds, H. Ellegren, Abundant recent activity of retrovirus-like  
639 retrotransposons within and among flycatcher species implies a rich source  
640 of structural variation in songbird genomes. *Mol. Ecol.* **27**, 99–111 (2018).
- 641 21. B. Charlesworth, P. Sniegowski, W. Stephan, The evolutionary dynamics of  
642 repetitive DNA in eukaryotes. *Nature*. **371**, 215–220 (1994).
- 643 22. E. B. Chuong, N. C. Elde, C. Feschotte, Regulatory activities of transposable  
644 elements: from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).
- 645 23. Y. Zhou, A. Minio, M. Massonnet, E. Solares, Y. Lv, T. Beridze, D. Cantu, B. S.  
646 Gaut, The population genetics of structural variants in grapevine  
647 domestication. *Nat. Plants*. **5**, 965–979 (2019).
- 648 24. M. Gymrek, A genomic view of short tandem repeats. *Curr. Opin. Genet. Dev.*  
649 **44**, 9–16 (2017).
- 650 25. M. Levy-Sakin, S. Pastor, Y. Mostovoy, L. Li, A. K. Y. Leung, J. McCaffrey, E.  
651 Young, E. T. Lam, A. R. Hastie, K. H. Y. Wong, C. Y. L. Chung, W. Ma, J. Sibert, R.  
652 Rajagopalan, N. Jin, E. Y. C. Chow, C. Chu, A. Poon, C. Lin, A. Naguib, W.-P.  
653 Wang, H. Cao, T.-F. Chan, K. Y. Yip, M. Xiao, P.-Y. Kwok, Genome maps across  
654 26 human populations reveal population-specific patterns of structural  
655 variation. *Nat. Commun.* **10** (2019), doi:10.1038/s41467-019-08992-7.
- 656 26. J. W. Poelstra, N. Vijay, C. M. Bossu, H. Lantz, B. Ryll, I. Müller, V. Baglione, P.  
657 Unneberg, M. Wikelski, M. G. Grabherr, J. B. W. Wolf, The genomic landscape  
658 underlying phenotypic integrity in the face of gene flow in crows. *Science*.  
659 **344**, 1410–1415 (2014).

- 660 27. D. C. Jeffares, C. Jolly, M. Hoti, D. Speed, L. Shaw, C. Rallis, F. Balloux, C.  
661 Dessimoz, J. Bähler, F. J. Sedlazeck, Transient structural variations have  
662 strong effects on quantitative traits and reproductive isolation in fission  
663 yeast. *Nat. Commun.* **8**, 14061 (2017).
- 664 28. U. Knief, C. M. Bossu, N. Saino, B. Hansson, J. Poelstra, N. Vijay, M.  
665 Weissensteiner, J. B. W. Wolf, Epistatic mutations under divergent selection  
666 govern phenotypic variation in the crow hybrid zone. *Nat. Ecol. Evol.* (2019),  
667 doi:10.1038/s41559-019-0847-9.
- 668 29. J. W. Poelstra, N. Vijay, M. P. Hoepfner, J. B. W. Wolf, Transcriptomics of  
669 colour patterning and coloration shifts in crows. *Mol. Ecol.* **24**, 4617–4628  
670 (2015).
- 671 30. A. I. Vickrey, R. Bruders, Z. Kronenberg, E. Mackey, R. J. Bohlender, E. T.  
672 Maclary, R. Maynez, E. J. Osborne, K. P. Johnson, C. D. Huff, M. Yandell, M. D.  
673 Shapiro, Introgression of regulatory alleles and a missense coding mutation  
674 drive plumage pattern diversity in the rock pigeon. *eLife.* **7** (2018),  
675 doi:10.7554/eLife.34803.
- 676 31. E. B. Chuong, N. C. Elde, C. Feschotte, Regulatory activities of transposable  
677 elements: from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).
- 678 32. M. F. L. Derks, K. M. Schachtschneider, O. Madsen, E. Schijlen, K. J. F.  
679 Verhoeven, K. van Oers, Gene and transposable element methylation in  
680 great tit (*Parus major*) brain and blood. *BMC Genomics.* **17** (2016),  
681 doi:10.1186/s12864-016-2653-y.
- 682 33. M. Chakraborty, N. W. VanKuren, R. Zhao, X. Zhang, S. Kalsow, J. J. Emerson,  
683 Hidden genetic variation shapes the structure of functional elements in  
684 *Drosophila*. *Nat. Genet.* **50**, 20–25 (2018).
- 685 34. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov,  
686 BUSCO: assessing genome assembly and annotation completeness with  
687 single-copy orthologs. *Bioinformatics.* **31**, 3210–3212 (2015).
- 688 35. E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A.  
689 Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B.  
690 Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A.  
691 Mirny, E. S. Lander, J. Dekker, Comprehensive Mapping of Long-Range  
692 Interactions Reveals Folding Principles of the Human Genome. *Science.* **326**,  
693 289–293 (2009).
- 694 36. N. H. Putnam, B. L. O’Connell, J. C. Stites, B. J. Rice, M. Blanchette, R. Calef, C. J.  
695 Troll, A. Fields, P. D. Hartley, C. W. Sugnet, D. Haussler, D. S. Rokhsar, R. E.  
696 Green, Chromosome-scale shotgun assembly using an in vitro method for  
697 long-range linkage. *Genome Res.* **26**, 342–350 (2016).
- 698 37. T. Kawakami, L. Smeds, N. Backström, A. Husby, A. Qvarnström, C. F. Mugal, P.  
699 Olason, H. Ellegren, A high-density linkage map enables a second-generation  
700 collared flycatcher genome assembly and reveals the patterns of avian

- 701 recombination rate variation and chromosomal evolution. *Mol. Ecol.* **23**,  
702 4035–4058 (2014).
- 703 38. R. S. Harris, *Improved Pairwise Alignment of Genomic DNA* (Pennsylvania  
704 State Univ., 2007), *PhD thesis*.
- 705 39. S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, S. L.  
706 Salzberg, Versatile and open software for comparing large genomes.  
707 *Genome Biol.* **5**, R12 (2004).
- 708 40. M. Nattestad, M. C. Schatz, Assemblytics: a web analytics tool for the  
709 detection of variants from an assembly. *Bioinformatics.* **32**, 3021–3023  
710 (2016).
- 711 41. Z. N. Kronenberg, I. T. Fiddes, D. Gordon, S. Murali, S. Cantsilieris, O. S.  
712 Meyerson, J. G. Underwood, B. J. Nelson, M. J. P. Chaisson, M. L. Dougherty, K.  
713 M. Munson, A. R. Hastie, M. Diekhans, F. Hormozdiari, N. Lorusso, K.  
714 Hoekzema, R. Qiu, K. Clark, A. Raja, A. E. Welch, M. Sorensen, C. Baker, R. S.  
715 Fulton, J. Armstrong, T. A. Graves-Lindsay, A. M. Denli, E. R. Hoppe, P. Hsieh,  
716 C. M. Hill, A. W. C. Pang, J. Lee, E. T. Lam, S. K. Dutcher, F. H. Gage, W. C.  
717 Warren, J. Shendure, D. Haussler, V. A. Schneider, H. Cao, M. Ventura, R. K.  
718 Wilson, B. Paten, A. Pollen, E. E. Eichler, High-resolution comparative  
719 analysis of great ape genomes. *Science.* **360**, eaar6343 (2018).
- 720 42. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local  
721 alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
- 722 43. K. Katoh, H. Toh, Recent developments in the MAFFT multiple sequence  
723 alignment program. *Brief. Bioinform.* **9**, 286–298 (2008).
- 724 44. T. Wicker, F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A.  
725 Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, A. H.  
726 Schulman, A unified classification system for eukaryotic transposable  
727 elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
- 728 45. J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, J. Walichiewicz,  
729 Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet.*  
730 *Genome Res.* **110**, 462–467 (2005).
- 731 46. A. F. Smit, R. Hubley, P. Green, RepeatMasker. Open-3.0 (1996).
- 732 47. F. J. Sedlazeck, P. Rescheneder, M. Smolka, H. Fang, M. Nattestad, A. von  
733 Haeseler, M. C. Schatz, Accurate detection of complex structural variations  
734 using single-molecule sequencing. *Nat. Methods.* **15**, 461–468 (2018).
- 735 48. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G.  
736 Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The  
737 Sequence Alignment/Map format and SAMtools. *Bioinformatics.* **25**, 2078–  
738 2079 (2009).

- 739 49. P. Danecek, S. A. McCarthy, BCFtools/csq: haplotype-aware variant  
740 consequences. *Bioinformatics*. **33**, 2037–2039 (2017).
- 741 50. K. A. Jønsson, P.-H. Fabre, J. D. Kennedy, B. G. Holt, M. K. Borregaard, C.  
742 Rahbek, J. Fjeldså, A supermatrix phylogeny of corvid passerine birds  
743 (Aves: Corvides). *Mol. Phylogenet. Evol.* **94**, 87–94 (2016).
- 744 51. C. F. Mugal, V. E. Kutschera, F. Botero-Castro, J. B. W. Wolf, I. Kaj,  
745 Polymorphism Data Assist Estimation of the Nonsynonymous over  
746 Synonymous Fixation Rate Ratio  $\omega$  for Closely Related Species. *Mol. Biol.*  
747 *Evol.* (2019), doi:10.1093/molbev/msz203.
- 748 52. H. Ellegren, Microsatellite mutations in the germline: Implications for  
749 evolutionary inference. *Trends Genet.* **16**, 551–558 (2000).
- 750 53. D. Bates, M. Maechler, B. Bolker, S. Walker, *lme4: Linear mixed-effects models*  
751 *using Eigen and S4. R package version 1.1–7. 2014* (2015).
- 752 54. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows–  
753 Wheeler transform. *Bioinformatics*. **26**, 589–595 (2010).
- 754 55. R. M. Layer, C. Chiang, A. R. Quinlan, I. M. Hall, LUMPY: a probabilistic  
755 framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
- 756 56. T. Rausch, T. Zichner, A. Schlattl, A. M. Stutz, V. Benes, J. O. Korbel, DELLY:  
757 structural variant discovery by integrated paired-end and split-read  
758 analysis. *Bioinformatics*. **28**, i333–i339 (2012).
- 759 57. X. Chen, O. Schulz-Trieglaff, R. Shaw, B. Barnes, F. Schlesinger, M. Källberg, A.  
760 J. Cox, S. Kruglyak, C. T. Saunders, Manta: rapid detection of structural  
761 variants and indels for germline and cancer sequencing applications.  
762 *Bioinformatics*. **32**, 1220–1222 (2016).
- 763 58. X. Zheng, D. Levine, J. Shen, S. M. Gogarten, C. Laurie, B. S. Weir, A high-  
764 performance computing toolset for relatedness and principal component  
765 analysis of SNP data. *Bioinformatics*. **28**, 3326–3328 (2012).
- 766 59. P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E.  
767 Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, 1000  
768 Genomes Project Analysis Group, The variant call format and VCFtools.  
769 *Bioinformatics*. **27**, 2156–2158 (2011).
- 770 60. B. S. Weir, C. C. Cockerham, Estimating F-Statistics for the Analysis of  
771 Population Structure. *Evolution*. **38**, 1358 (1984).
- 772 61. T. E. Kijima, H. Innan, On the Estimation of the Insertion Time of LTR  
773 Retrotransposable Elements. *Mol. Biol. Evol.* **27**, 896–904 (2010).
- 774