Myrmecological News

ISSN 1997-3500
myrmecologicalnews.org

Myrmecol. News 30: 151-159 doi: 10.25849/myrmecol.news_030:151 24 June 2020

Original Article

Revisiting museum collections in the genomic era: potential of MIG-seq for retrieving phylogenetic information from aged minute dry specimens of ants (Hymenoptera: Formicidae) and other small organisms

Katsuyuki EGUCHI, Emiko OGURI, Takanori SASAKI, Ayumi MATSUO, Dai Dac NGUYEN, Weeyawat JAITRONG, Bakhtiar Effendi YAHYA, Zhilin CHEN, Rijal SATRIA, Wendy Yanling WANG & Yoshihisa SUYAMA

Abstract

Multiplexed inter-simple sequence repeat genotyping by sequencing – MIG-seq – is an effective PCR-based method for genome-wide SNP detection using the Next-Generation Sequencing platform, and it provides a potential solution to a central problem in museomics – the difficulties of obtaining useful sequence data from aged specimens with often degraded and / or low yields of DNA. We demonstrate and validate the cost effectiveness and utility of the MIG-seq workflow in obtaining useful and robust sequence data from aged museum specimens. We applied the MIG-seq approach to 55 aged (10 - 23 years old) millimeter-sized dry-mounted specimens of the hyper-diverse ant genus *Pheidole*. A total of 50,782,736 reads were generated from the 55 samples (259,902 - 3,693,375 reads per sample). The reads corresponded to 36,862 SNPs from 4,849 polymorphic loci; the SNP dataset was then used to construct a Bayesian phylogenetic tree. The topology of the phylogenetic tree was highly compatible with existing knowledge of phylogenetic relationships among species of *Pheidole*. Therefore, we recommend the MIG-seq method as a cost-effective and highly applicable pipeline for conducting phylogenetic and population genetic studies on aged museum specimens, potentially enhancing the relevance of specimen repositories in general towards modern biodiversity science and conservation biology.

Key words: Insecta, museomics, Next-Generation Sequencing, phylogenetics, SNP detection.

Received 24 February 2020; revision received 14 April 2020; accepted 19 April 2020
Subject Editor: Alexander S. Mikheyev

Katsuyuki Eguchi (contact author), Tokyo Metropolitan University, Graduate School of Science, Tokyo, 1920397, Japan. E-mail: antist2007@gmail.com

Emiko Oguri, Tokyo Metropolitan University, Graduate School of Science, Tokyo, 1920397, Japan; Tokyo Gakugei University, Faculty of Education, Tokyo, 1848501, Japan.

Takanori Sasaki, Ayumi Matsuo, Yoshihisa Suyama, Tohoku University, Graduate School of Agricultural Science, Miyagi, 9896711, Japan.

Dai Dac Nguyen, Tokyo Metropolitan University, Graduate School of Science, Tokyo, 1920397, Japan; Institute of Ecology and Biological Resources, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet Road, Nghia Do, Cau Giay, Hanoi, Vietnam.

Weeyawat Jaitrong, National Science Museum, Thailand, Natural History Museum, Pathum Thani, 12120, Thailand.

Bakhtiar Effendi Yahya, Universiti Malaysia Sabah, Institute for Tropical Biology and Conservation, Sabah, 88400, Malaysia.

Zhilin Chen, Ministry of Education, Key Laboratory of Ecology of Rare and Endangered Species and Environmental Protection (Guangxi Normal University), Guilin, 541004, China.

Rijal Satria, Universitas Negeri Padang, Faculty of Mathematics and Natural Sciences, West Sumatra, 25173, Indonesia.

Wendy Yanling Wang, National University of Singapore, Lee Kong Chian Natural History Museum, 117377, Singapore.

Introduction

Biological collections preserved in museums are archives of the past and present life and environment on the earth. These might provide valuable information on temporal and spatial distributions of wild organisms, diversification and evolutionary changes, and also baselines for the practices of conservation of endangered species and habitats, management of invasive species, pests and pathogens, and exploration of novel biological resources (Graham & al. 2004, Suarez & Tsutsui 2004, Lister & al. 2011, Ward 2012). Since the mid-2000s, novel genomic methods based on Next Generation Sequencing (NGS), such as RAD-seq (Baird & al. 2008) and targeted enrichment of ultra-conserved elements (Faircloth & al. 2012), have been rapidly developed and generalized, and then increasingly used in basic and practical research involving wild organisms and their conservation (e.g., snails – Razkin & al. 2016, birds – Battery & Klicka 2017, weevils – Chen & al. 2017). Consequently, the term "Museomics" was coined to describe studies focused on extracting and analyzing genomic data from museum material. A key aim in museomics is to link genomic data from museum resources to existing biodiversity information, for stronger support of meaningful biological inferences (Buerki & Baker 2016, Nakazato 2018).

In recent times, species identification and phylogenetic analyses with standard DNA barcodes, generated via conventional Sanger methods or NGS, have been increasingly performed on aged museum specimens, including types (Miller & al. 2013, Prosser & al. 2016, Schäffer & al. 2017, Wachi & al. 2018, Velasco-Cuervo & al. 2019). In addition, genome-wide SNP detection using NGS has also been performed on aged specimens (Tin & al. 2014, Blaimer & al. 2016, Suchan & al. 2016, Battery & Klicka 2017, Linck & al. 2017, Wood & al. 2018, Ewart & al. 2019). Suyama & Matsuki (2015) proposed "multiplexed ISSR genotyping by sequencing" (hereafter referred to as MIG-seq), an effective PCR-based method for genome-wide de novo identification of single-nucleotide polymorphisms using the NGS platform. As the MIG-Seq process is PCR-based, a major advantage is that it does not require substantial amounts of absolutely pure DNA in order to be successfully performed. Ample useful sequence data for phylogenetic and population genetic analyses can be obtained from a sample with only a very small volume of input DNA (Suyama & Matsuki 2015) – this critical feature of MIG-seq makes the workflow particularly applicable to and useful for museomics. That is to say that the MIG-seq process provides a potential solution to a central problem in museomics – the difficulties of obtaining useful sequence data from aged specimens with often degraded and / or low yields of DNA.

Thus, in the present study, we demonstrate and validate the utility of the MIG-Seq workflow in obtaining useful and robust sequence data from aged millimeter-sized specimens in museum collections.

Materials and methods

Taxon sampling: Southeast and East Asian species of the ant genus *Pheidole* Westwood, 1839 (Insecta: Hymenoptera: Formicidae: Myrmicinae) were chosen as the target taxa in the present study, based on four reasons. First, minor workers (smaller subcaste of the worker caste) of *Pheidole* species are millimeter-sized, usually within a range from 1.5 to 6 millimeters in the approximate body length (when excluding appendages), and thus they are good representatives of "small" invertebrates in a typical museum collection. Second, the morphology-based species delimitation of Southeast and East Asian *Pheidole* species has been well-established by K. Eguchi, the first author (Eguchi 2001, 2008). Third, phylogenetic relationships among major lineages of the genus known from Southeast and East Asia have been resolved with seeming high reliability in a recent study (Economo & al. 2015a, b), based on a sequence dataset of nine loci (8,820 bp in total). Finally, a huge collection is available of dry-mounted specimens of Southeast and East Asian *Pheidole* species, including paratypes and specimens collected from their type localities, that is managed by the first author.

A total of 55 specimens comprising 46 *Pheidole* species were used for the present study (for details see Tab. 1). Of these, 22 species were chosen sensu Economo & al. (2015b), including all of the named species of the "*P. quadricuspis* Emery, 1900 clade", "*P. nodus* F. Smith, 1874 clade", "*P. quinata* Eguchi, 2000 clade" and "*P. smythiesii* Forel, 1902 clade". Two specimens of *P. sexspinosa* Mayr, 1870 were designated as the outgroup OTUs, following Economo & al. (2015b).

The identification of *Pheidole* spp. was done by K. Eguchi, a specialist on alpha-taxonomy of Asian *Pheidole* (Eguchi 2001, 2008, Eguchi & al. 2016, Wang & al. 2018). The voucher specimens are housed in the collection of the Systematic Zoology Laboratory, Tokyo Metropolitan University, and managed by K. Eguchi.

Condition of specimens: Of the 55 specimens, two (OTU-2 and OTU-3) had been preserved in 100% ethanol and three (OTU-52, OTU-53 and OTU-54) in 75%, and the others were dry-mounted and had been preserved for approximately 10 - 23 years (Tab. 1). Nineteen are paratypes and nine were collected from the type localities of the species. The approximate body length of the specimens varied from 1.5 mm (OTU-51: *Pheidole parvicorpus* Eguchi, 2001) to 6.0 mm (OTU-48: *P. singaporensis* Özdikmen, 2010).

DNA extraction: For 53 specimens, total DNA was extracted from two legs (OTU-39 and OTU-48) or whole body by using the "Chelex-TE-ProK method" with 105 µL of elution buffer (Satria & al. 2015). For the other two specimens (OTU-4 and OTU-5; Tab. 1) DNA was extracted from a whole body using QIAGEN DNeasy Blood & Tissue Kit (QIAGEN, Hilden, Germany) by following the manufacturer's instructions (hereafter referred to as "DNeasy method"). Lysis time was extended to approximately 48 hours to maximize DNA yield, and the purified DNA bound

Tab. 1: Specimens analyzed in the present study. PT, paratype; TT, specimen from the type locality. Abbreviations of the names of clades, groups, and grade recognized by Economo & al. (2015b): ARS, *Pheidole aristotelis* grade; FER, *P. fervens* clade; KNW, *P. knowlesi* group; MEG, *P. megacephala* group; NOD, *P. noda* clade; PAR, *P. parva* clade; QDC, *P. quadricuspis* clade; QNT, *P. quinata* clade; SAU, *P. sauberi* clade; SMY, *P. smythiesii* clade; SXS, *P. sexspinosa* clade. OTU-2 and 3 were preserved in 100% ethanol, and OTU-52, 53 and 54 were preserved in 70% ethanol.

OTU ID	Colony ID	Identification and treatment in Economo & al. (2015b)	Status	Locality	Collecting year	Period after dry-mounting (year)	DNA concentration (ng/µL)	DNA purity ($OD_{260/280}$)	No. of reads	Removal rate (%) after quality filtering using FASTX-Toolkit	Statistics of the dataset assembled by Stacks (m = 3, M = 8, n = 9, N = 10, r = 0.1)			
											No. of loci	No. of polymorphic loci	No. of SNPs	% of missing SNP sites
OTU-1	Eg01-TH-122	*P. plagiaria* F. Smith, 1860; FER	-	N. Thailand	2001	16	77.8	0.95	564,125	0.4	833	776	6,939	81.2
OTU-2	WW-SG18-Phei3	*P. sexspinosa* Mayr, 1870; SXS	-	Singapore	2018	n/a	41.8	0.76	2,050,772	0.5	2,033	1,931	13,039	64.6
OTU-3	Eg16iii13-24	*P. leloi* Eguchi & Bui, 2016	PT	C. Vietnam	2013	n/a	122.4	0.71	2,062,897	0.5	2,364	2,202	15,242	58.7
OTU-4	Eg00-BOR-102	*P. lokitae* Forel, 1913	-	Sabah, Borneo	2000	12	3.3	1.59	3,693,375	0.3	2,361	2,243	15,778	57.2
OTU-5	No. 10/16a	*P. sperata* Forel, 1915	-	W. Java	1999	12	2.7	1.42	2,748,261	0.2	2,371	2254	15,322	58.4
OTU-6	Eg98-BOR-880	*P. spinicornis* Eguchi, 2001	PT	Sabah, Borneo	1998	12	71.3	0.98	899,513	0.4	1,155	1,044	8,597	76.7
OTU-7	Eg00-BOR-101	*P. comata* F. Smith, 1858; SMY	-	Sabah, Borneo	2000	12	173.7	1.10	2,015,635	0.3	1,812	1,725	13,579	63.2
OTU-8	No. 10/1a	*P. tjibodana* Forel, 1905; ARS	TT	W. Java	1999	18	66.9	0.94	589,091	1.1	412	367	3,705	89.9
OTU-9	No. 10/4b	*P. sauberi* Forel, 1905; SAU	TT	W. Java	1999	18	56.5	0.83	635,305	0.8	573	521	4,740	87.1
OTU-10	Eg98-BOR-850	*P. sabahna* Eguchi, 2000; QNT	PT	Sabah, Borneo	1998	19	174.9	1.05	1,121,208	0.4	1,037	945	8,218	77.7
OTU-11	Eg98-BOR-865	*P. retivertex* Eguchi, 2001	PT	Sabah, Borneo	1998	18	92.8	0.81	347,478	1.2	391	336	3,285	91.1
OTU-12	AU01-SKY-12	*P. megacephala* (Fabricius, 1793); MEG	-	NE. Australia	2001	17	76.5	0.74	606,224	0.9	623	574	5,151	86.0
OTU-13	Eg98-BOR-821	*P. aristotelis* Forel, 1911; ARS	-	Sarawak, Borneo	1998	18	89.0	0.83	443,654	1.0	599	482	4,759	87.1
OTU-14	FI92-242	*P. aristotelis* Forel, 1911; ARS	-	W. Sumatra	1992	18	57.7	0.77	277,418	1.4	143	112	1,201	96.7
OTU-15	Eg97-BOR-394	*P. tenebricosa* Eguchi, 2001	PT	Sabah, Borneo	1997	18	55.9	0.73	323,390	1.3	183	157	1,631	95.6
OTU-16	Eg98-BOR-836	*P. rugifera* Eguchi, 2001	PT	Sarawak, Borneo	1998	18	59.7	0.69	393,863	1.0	267	237	2,241	93.9
OTU-17	Eg96-BOR-035	*P. clypeocornis* Eguchi, 2001	PT	Sabah, Borneo	1996	18	43.3	0.79	283,737	1.1	571	460	4,202	88.6
OTU-18	FI93-253	*P. clypeocornis* Eguchi, 2001	-	W. Sumatra	1993	18	39.8	0.78	278,659	1.2	320	204	1,893	94.9
OTU-19	09Q21S4	*P. kikutai* Eguchi, 2001	TT	Sabah, Borneo	1997	18	38.7	0.83	570,930	0.9	510	435	4,423	88.0
OTU-20	FI93-256	*P. hortensis* Forel, 1913	-	W. Sumatra	1993	18	34.4	0.74	372,451	1.3	102	65	648	98.2
OTU-21	No. 10/5a	*P. hortensis* Forel, 1913	TT	W. Java	1999	18	34.1	0.80	511,394	0.9	730	627	5,664	84.6
OTU-22	SU02-SKY-96	*P. lokitae* Forel, 1913	-	Sumatra	2002	10	209.6	1.00	831,051	0.5	890	841	7,390	80.0
OTU-23	Eg96-BOR-292	*P. lucioccipitalis* Eguchi, 2001; FER	PT	Sabah, Borneo	1996	18	47.4	0.81	364,901	1.5	425	339	3,449	90.6
OTU-24	Eg01-VN-155	*P. vulgaris* Eguchi, 2006; FER	PT	N. Vietnam	2001	13	57.4	0.84	490,433	1.3	324	185	1,534	95.8
OTU-25	23July1997#168	*P. knowlesi* Mann, 1921; KNW	-	Viti Levu, Fiji	1997	unknown	64.2	0.90	268,045	1.2	139	103	1,332	96.4
OTU-26	FI97-440	*P. quadrensis* Forel, 1900; QDC	-	W. Sumatra	1997	18	127.1	0.98	259,902	1.5	321	284	3,065	91.7
OTU-27	SU02-SKY-134	*P. quadricuspis* Emery, 1900 QDC	-	Nias I., N. Sumatra	2002	16	114.5	0.95	563,030	0.7	766	723	6,441	82.5

OTU ID	Colony ID	Identification and treatment in Economo & al. (2015b)	Status	Locality	Collecting year	Period after dry-mounting (year)	DNA concentration (ng/µL)	DNA purity ($OD_{260/280}$)	No. of reads	Removal rate (%) after quality filtering using FASTX-Toolkit	Statistics of the dataset assembled by Stacks (m = 3, M = 8, n = 9, N = 10, r = 0.1)			
											No. of loci	No. of polymorphic loci	No. of SNPs	% of missing SNP sites
OTU-28	Eg26iii06-09	*P. rugithorax* Eguchi, 2008; NOD	PT	C. Vietnam	2006	11	137.9	0.52	734,475	0.6	784	668	5,649	84.7
OTU-29	Eg01-TH-084	*P. protea* Forel, 1912; NOD	-	N. Thailand	2001	15	73.7	0.92	401,388	0.9	387	358	3,682	90.0
OTU-30	FI97-551	*P. nodgii* Forel, 1905	TT	W. Java	1997	18	68.2	0.87	398,608	1.4	348	241	2,073	94.4
OTU-31	Eg02-JPN-02	*P. nodus* F. Smith, 1874; NOD	-	S. Kyushu, Japan	2002	17	86.7	0.96	1,524,337	0.4	849	795	7,278	80.3
OTU-32	SU02-SKY-77	*P. bluntschlii* Forel, 1911; QNT	-	N. Sumatra	2002	16	194.9	1.01	341,396	1.0	751	698	6,388	82.7
OTU-33	FI99-109	*P. singaporensis* Özdikmen, 2010; SMY	-	S. Malay Peninsula	1999	18	322.4	1.08	586,056	0.7	788	732	7,300	80.2
OTU-34	Eg98-BOR-847	*P. montana* Eguchi, 1999; SMY	PT	Sabah, Borneo	1998	20	222.6	1.04	956,331	0.3	1,458	1,387	13,032	64.6
OTU-35	Eg00-BOR-100	*P. angulicollis* Eguchi, 2001	PT	Sabah, Borneo	2000	18	73.3	0.90	984,303	0.5	1,067	927	8,213	77.7
OTU-36	Eg97-BOR-404	*P. acantha* Eguchi, 2001	PT	Sabah, Borneo	1997	18	132.8	0.98	541,067	0.8	648	601	5,647	84.7
OTU-37	HD-105	*P. quinata* Eguchi, 2000	PT	Sabah, Borneo	1996	22	91.1	0.92	608,236	0.6	1,023	900	5,647	84.7
OTU-38	Eg04-VN-800	*P. aspidata* Eguchi & Bui, 2005	PT	S. Vietnam	2004	14	60.5	0.85	1,289,075	0.3	1,514	1,375	10,438	71.7
OTU-39	Bottle-Eg-A	*P. upeneci* Forel, 1913	-	Sabah, Borneo	2000	18	30.3	0.68	636,007	0.5	845	730	6,610	82.1
OTU-40	Eg01-VN-200	*P. dugasi* Forel, 1911; SMY	-	N. Vietnam	2001	17	184.1	0.99	656,452	0.4	631	596	5,680	84.6
OTU-41	Eg01-VN-176	*P. tumida* Eguchi, 2008	PT	N. Vietnam	2001	11	83.8	0.90	323,091	1.20	276	233	2,529	93.1
OTU-42	Eg01-TH-116	*P. gatesi* (Wheeler, 1927); SMY	-	N. Thailand	2001	17	205.2	1.03	1,561,389	0.2	873	822	7,664	79.2
OTU-43	Eg01-VN-222	*P. colpigaleata* Eguchi, 2006	PT	N. Vietnam	2001	13	45.8	0.83	861,631	0.7	884	793	7,107	80.7
OTU-44	Eg01-TH-070	*P. smythiesii* Forel, 1902	-	N. Thailand	2001	17	115.7	0.98	1,186,060	0.4	1,066	964	8,598	76.7
OTU-45	Eg02-VN-210	*P. foveolata* Eguchi, 2006	PT	N. Vietnam	2002	13	40.3	0.77	469,894	1.0	565	536	5,072	86.2
OTU-46	Eg00-HK-25	*P. taipoana* Wheeler, 1928	TT	Hongkong	2000	18	68.7	0.67	1,178,770	0.3	1,488	1,415	11,195	69.6
OTU-47	Eg99-HK-34	*P. hongkongensis* Wheeler, 1928	TT	Hongkong	1999	18	63.5	0.78	1,107,415	0.4	1,111	1,022	8,269	77.6
OTU-48	4xii1995	*P. singaporensis* Özdikmen, 2010; SMY	TT	Singapore	1995	23	89.2	0.75	369,313	1.1	215	145	1,567	95.7
OTU-49	10Aug1997#272	*P. sexspinosa* Mayr, 1870; SXS	-	Efate, Vanuatu	1997	unknown	108.9	0.83	334,920	0.9	200	151	1,691	95.4
OTU-50	No. 4186	*P. schoedli* Eguchi, Hashimoto & Malsch, 2006	PT	Sabah, Borneo	1998	13	42.7	0.77	1,227,049	0.4	1,103	989	8,270	77.6
OTU-51	Eg97-BOR-584	*P. parvicepus* Eguchi, 2001	PT	Sabah, Borneo	1997	18	34.3	0.73	913,285	0.6	962	907	7,713	79.1
OTU-52	Q625	*P. parva* Mayr, 1865; PAR	-	Marine vessel	2011	n/a	182.5	0.54	2,065,739	0.8	2,061	1,929	13,151	64.3
OTU-53	Q677	*P. parva* Mayr, 1865; PAR	-	Marine vessel	2012	n/a	90.3	0.87	2,166,809	0.6	2,177	2,023	13,471	63.5
OTU-54	Q2146	*P. parva* Mayr, 1865; PAR	-	Marine vessel	2011	n/a	56.2	0.70	1,702,973	0.5	2,307	2,176	14,207	61.5
OTU-55	Eg01-JPN-002	*P. parva* Mayr, 1865; PAR	-	Okinawa I., Japan	2001	11	129.8	0.92	2,089,925	0.3	2,106	1,979	13,586	63.1
Min					1992	10	2.7	0.52	259,902		102	65	648	57.2
Max					2018	23	322.4	1.59	3,693,375		2,371	2,254	15,778	98.2
Total									50,782,736					

Tab. 2: The numbers of SNPs and polymorphic loci under different values of the m, M, n, and r parameters of Stacks.

	Number of SNPs					Number of polymorphic loci				
	r=0.8	r=0.7	r=0.6	r=0.5	r=0.1	r=0.8	r=0.7	r=0.6	r=0.5	r=0.1
m=1, M=2, n=1	0	0	0	0	0	0	0	0	0	0
m=2, M=2, n=1	0	4	23	60	8,461	0	1	6	16	2,700
m=3, M=2, n=1	1	5	34	111	8,574	1	2	8	26	2,767
m=4, M=2, n=1	1	1	37	128	7,813	1	1	8	27	2,500
m=5, M=2, n=1	1	1	38	113	7,254	1	1	9	23	2,271
m=6, M=2, n=1	1	1	27	107	6,606	1	1	7	24	2,040
m=7, M=2, n=1	1	8	39	102	6,231	1	2	8	21	1,901
m=3, M=1, n=1	0	5	39	107	8,029	0	2	10	28	2,781
m=3, M=2, n=1	1	5	34	111	8,574	1	2	8	26	2,767
m=3, M=3, n=1	1	5	19	111	8,974	1	2	7	26	2,770
m=3, M=4, n=1	1	5	19	108	9,231	1	2	5	24	2,777
m=3, M=5, n=1	1	5	19	108	9,297	1	2	5	24	2,769
m=3, M=6, n=1	1	5	26	110	9,574	1	2	7	24	2,774
m=3, M=7, n=1	10	15	33	121	9,868	1	2	6	24	2,793
m=3, M=8, n=1	8	13	32	123	10,024	1	2	7	27	2,798
m=3, M=8, n=7	0	11	144	422	33,548	0	2	15	48	4,716
m=3, M=8, n=8	0	19	164	425	35,702	0	3	16	47	4,800
m=3, M=8, n=9	0	19	164	421	36,862	0	3	17	47	4,849

to the silica membrane column was eluted with 50 µL of elution buffer. In the cases where the whole bodies of ants were used for DNA extraction, the exoskeletons were preserved as vouchers in 80% ethanol.

MIG-seq analysis: The quality and quantity of the total DNA were assessed using NanoDrop Lite (Thermo Scientific, Waltham, MA, USA). Approximately 1 ng of DNA was used for the 1st PCR as template DNA. The first PCR step was performed to amplify inter-simple-sequence repeats (ISSR) from genomic DNA with MIG-seq primer set-1 (Suyama & Matsuki 2015). The fragments were amplified with Multiplex PCR Assay Kit Ver. 2 (TaKaRa Bio Inc., Otsu, Shiga, Japan) using 7 µL reaction volumes in a thermal cycler with the following profile: 94 °C for 1 min followed by 28 cycles of 94 °C for 30 s, 35 °C for 1 min, 72 °C for 1 min, and finally 72 °C for 10 min. The fragments were purified and normalized using AMPure XP beads (Beckman Coulter, Brea, CA, USA), and then size selected in the range of > 250 bp on AMPure XP beads. The second PCR step was conducted independently to add individual indices to each sample using eight forward primers and each reverse primer (Suyama & Matsuki 2015). The fragments were amplified with PrimeSTAR GXL DNA polymerase (TaKaRa Bio Inc., Otsu, Shiga, Japan) using 6 µL reaction volumes in a thermal cycler with the following profile: 12 cycles of 98 °C for 10 s, 54 °C for 15 s, 68 °C for 1 min. Then, products from the second PCR were pooled as a single mixture library with 1 µL of each product. Size selection (350 - 800 bp) and product purification were performed using AMPure XP beads. Library concentration was measured with a Qubit fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) and a SYBR green quantitative PCR assay (Library Quantification Kit, Clontech, Mountain View, CA, USA) with primers specific to the Illumina system. Approximately 4 nM of concentration-adjusted library was denatured using 0.2 N NaOH and mixed with Illumina-generated PhiX Control libraries (PhiX Control v3, Illumina, San Diego, CA, USA) according to Illumina's protocol, and finally, approximately 12 pM of library was used for sequencing on an Illumina MiSeq Sequencer (Illumina, San Diego, CA, USA) using a MiSeq Reagent Kit v3 (150 cycles; illumina), 80 bp of sequences were determined for Read 1 and Read 2, respectively. DarkCycle option was changed "Amplicon-dark 17-3" to "Amplicon-dark 17-17" on the "Chemistry" line (see also Suyama & Matsuki 2015).

SNP detection: The raw reads from each indexed sample were grouped together using the index reads option of the sequencer. Removal of the adapter and anchor sequences and quality filtering were performed by FASTX-Toolkit (http://hannonlab.cshl.edu./fastx_toolkit/) according to the procedure of Suyama & Matsuki (2015). The quality-filtered reads were analyzed with Stacks v. 1.47 (Catchen & al. 2013) to detect SNPs.

By referring to Paris & al. (2017), Díaz-Arce & Rodríguez-Ezpeleta (2019), and Campbell & al. (2020), the optimal m, M and n parameter values for the Stacks were determined under r = 0.8, 0.7, 0.6, 0.5 and 0.1, respectively. N, p and max_obs_het were set as default (Paris & al. 2017); all samples (OTUs) were regarded to belong to one population; min_maf was set as 0.02 (1 / 49 OTUs ≈ 0.02). The optimal values were determined when the number of polymorphic loci and the number of SNPs appeared to be maximized or stabilized around the maximum: (i) the

optimal m value being determined (among values ranging from 3 to 7 at 1 interval) under M = 2, n = 1, N = M + 2 (default), p = 1 (default) and max_obs_het = 1 (default); (ii) the optimal M value being determined (among values ranging from 1 to 8) under the m value optimized previously, and the other parameters set as above; and (iii) determining the optimal n value (among M - 1, M and M + 1) under the m and M value optimized previously, and the other parameters set as above (Tab. 2).

Phylogenomic analyses: Bayesian inference analysis was performed using ExaBayes 1.4 (Aberrer & al. 2014). The analysis was performed with 10,000,000 generations and default parameter setting (GTR model, sampling every 500 generation and tuning parameters every 100 generation, a burn-in of 25%), and a consensus tree and parameter summary files were produced using the "consense" and "postProcParam" post-processing tools in ExaBayes. Effective sampling size (ESS) was checked using the Tracer v.1.7.1 (http://tree.bio.ed.ac.uk/software/tracer/). The consensus tree was visualized using Figtree v.1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/) and saved as a pdf, then edited using Inkscape 0.92 (https://inkscape.org/ja/).

Results

DNA quantity and quality: The concentration and purity ($OD_{260/280}$) of total DNA extracted from the specimens varied in the range of 30.3 - 322.4 ng / µL and 0.52 - 1.10 in the Chelex-TE-ProK method, and 2.7 - 3.3 ng / µL and 1.42 - 1.59 in the DNeasy method, respectively. Preservation period after dry-mounting the specimen did not correlate with both the total DNA yield (R = 0.02; Fig. 1A) and DNA purity (R = 0.12; Fig. 1B). The degree of DNA fragmentation was not measured in this study.

Numbers of reads, SNPs and polymorphic loci: The number of reads per sample varied from 263,820 to 3,703,022 in the raw data, and from 259,902 to 3,693,375 after quality filtering using FASTX-Toolkit; the removal rate varied from 0.2 to 1.5% (Tab. 1). The number of reads after quality filtering correlated negatively with the preservation period after dry-mounting the specimen (R = 0.61; Fig. 1C).

Stacks with the parameter combination of m = 3, M = 8, n = 9 and N = 10 yielded 4,849 polymorphic loci with 36,862 SNPs under r = 0.1 (Tab. 2). This dataset (% of missing data ≈ 81.4) was used for phylogenetic analysis. The number of SNPs per sample varied from 648 to 15,778, and negatively correlated with the preservation period after dry-mounting the specimen (R = 0.63; Tab. 1; Fig. 1D).

Phylogenetic tree of Southeast and East Asian *Pheidole* species: A consensus Bayesian tree was obtained (ESS > 800 for all parameters), and rooted by designating two OTUs of *Pheidole sexspinosa* as outgroups (Fig. 2). The clade "A" involving "*P. quadricuspis* clade", "*P. nodus* clade", "*P. quinata* clade" and "*P. smythiesii* clade" was highly supported to be monophyletic (posterior probability (pp) = 1), and the most of the internal nodes were also supported with higher posterior probability values.

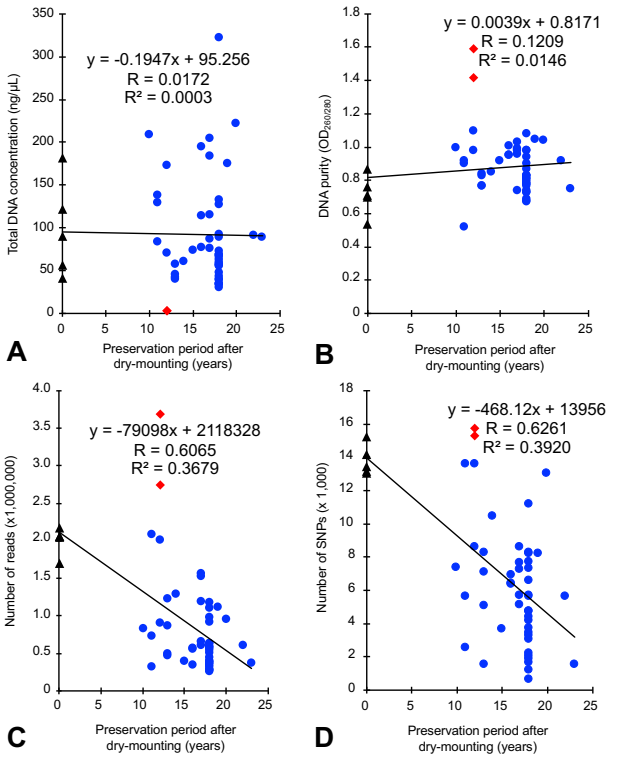

Fig. 1: Performance of DNA extraction and MIG-seq. (A) Total DNA concentration (ng / µL); (B) DNA purity (OD260/280); (C) Number of reads; (D) Number of SNPs. Blue circles: DNA templates from dry-mounted specimens extracted by Chelex TE ProK method; black triangles: from ethanol-preserved specimens by Chelex TE ProK method; red diamond: from dry-mounted specimens by DNeasy method.

The "*Pheidole smythiesii* clade" was nested in the clade "B" (pp = 1), and then subdivided into two monophyletic subclades, namely "*P. smythiesii* group (pp = 1; B1 in Fig. 2)" and "*P. singaporensis* subclade (pp = 1; B2 in Fig. 2)". This subdivision is in agreement with the phylogenetic tree inferred by Economo & al. (2015b, 2019) as well as morphology-based grouping (Eguchi 2001, 2008). The "*P. quinata* clade" was nested in the clade "C" (pp = 1). The internal topology of the clade C agreed with the morphology-based grouping (Eguchi 2001) in which the *P. quinata* group (*P. quinata* + *P. sabahna* Eguchi, 2000) is expected to be the closest with *P. bluntschlii* Forel, 1911. The close relationship between *P. sabahna* and *P. bluntschlii* was also inferred by Economo & al. (2019). The "*P. quadricuspis* clade" was nested in the clade "D" (pp = 1) which was compatible with the "*P. quadricuspis* group" sensu Eguchi & al. (2016), that is, the species group consisting of *P. leloi* Eguchi & Bui, 2016, *P. lokitae* Forel, 1913, *P. acantha* Eguchi, 2001, *P. quadrensis* Forel, 1900, *P. spinicornis* Eguchi, 2001, *P. sperata* Forel, 1915 and *P. quadricuspis*. The clade "D" is also compatible with the topology of Economo & al. (2019).

In contrast to the other clades, "*Pheidole nodus* clade" was not supported to be monophyletic in the present analysis. *Pheidole nodus* was sister to *P. tumida* Eguchi, 2008,

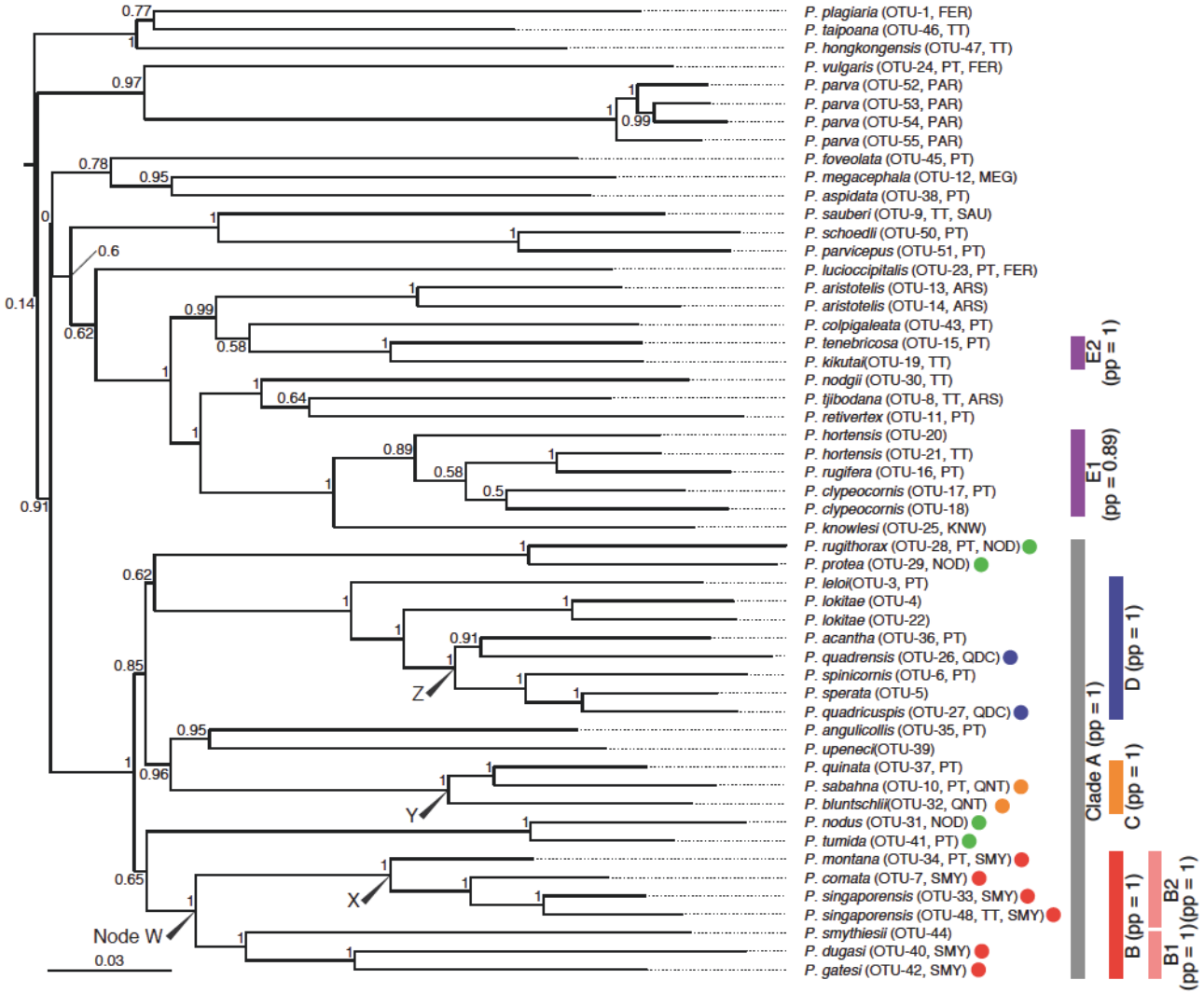

Fig. 2: A consensus Bayesian phylogenetic tree of 46 Southeast and East Asian species of the genus *Pheidole* (Insecta: Hymenoptera: Formicidae: Myrmicinae), rooted by designating two OTUs of *P. sexspinosa* as outgroups. The dataset consisting of 36,862 SNPs were obtained under the following optimized Stacks parameters: $m = 3$, $M = 8$, n = 9, N = 10, and r = 0.1. All of the named species assigned to the "*P. quadricuspis* clade", "*P. nodus* clade", "*P. quinata* clade" and "*P. smythiesii* clade" in Economo & al. (2015b) were marked with blue dots, green dots, orange dots and red dots, respectively.

a representative of the "*P. spathifera* Forel, 1902 group" sensu Eguchi (2008). On the other hand, *P. protea* and *P. rugithorax* were sister to each other, and formed an independent lineage which was deeply divergent from the other lineages. Interestingly the subdivision of "*P. nodus* clade" is compatible with the topology of Economo & al. (2019) where *P. rugithorax* appears far from the clade involving *P. nodus* and *P. tumida*.

However, basal relationships outside the clade "A" were often less reliable (Fig. 2). A remarkable obscurity was observed in the "*Pheidole hortensis* Forel, 1913 group" sensu Eguchi (2001), that is, a morphologically well-defined group consisting of *P. hortensis*, *P. clypeocornis* Eguchi, 2001, *P. kikutai* Eguchi, 2001, *P. rugifera* Eguchi, 2001 and *P. tenebricosa* Eguchi, 2001. The group appeared to be polyphyletic in the present tree (Fig. 2). The clade E1 (including the specimen of *P. hortensis* collected from the type locality; pp = 0.89) was located far from the clade E2 (pp = 1) which was sister to *P. colpigaleata* Eguchi, 2006.

Discussion

Reliability of the phylogenetic analyses: The monophyly of the larger clade "A" and the major internal nodes were highly supported, and topology of the Bayesian tree obtained in this study is highly compatible with Economo & al. (2015b). According to Economo & al. (2015b), the ages of the nodes "W", "X", "Y", and "Z" inside the clade A were estimated to be 14 Mya (million years ago), 9 Mya, 8 Mya and 8 Mya, respectively. With that in mind, if used in combination with time-calibration data, SNP data obtained via MIG-seq also has the potential for similar use in revealing prehistorical divergence events (i.e., from mid Miocene to present).

Basal relationships outside the clade "A" were often less reliable (Fig. 2). Because MIG-seq relies on the conservation of the SSR sites to which the universal multiplex PCR primers are attached, the number of polymorphic loci shared among the majority of OTUs decreases drastically and the level of missing sites in the assembled dataset increases as the OTUs are deeply divergent from each other. Therefore, we set a very low value of the Stacks parameter "r" (r = 0.1 in which 10% of OTUs that must possess a particular locus) for enlarging the otherwise much smaller dataset (4,849 polymorphic loci with 36,862 SNPs under r = 0.1, and 47 polymorphic loci with 421 SNPs under "r = 0.5"). Phylogenetic studies based on genome-wide detection usually struggle with a dilemma between "a larger dataset with a larger level of missing data" and "a smaller dataset with a smaller level of missing data". The size of dataset and the level of missing data can affect the resolution and reliability of the phylogenetic tree. Impacts of missing data on phylogenetic inference have been actively studied and discussed in bioinformatics (Roure & al. 2012).

The polyphyly of the *Pheidole hortensi*s group (as seen in the clades E1 and E2) might be an artifact caused by larger level of missing sites, and can be resolved by a "hierarchic approach", that is, MIG-seq and downstream phylogenetic analyses based on further taxon sampling of ingroup OTUs which might comprehensively cover the full range of genetic diversity of the *P. hortensis* group and several outgroup OTUs located "nearby" the *P. hortensis* group, that is, *P. aristotelis* Forel, 1911, *P. nodgii* Forel, 1905, *P. tjibodana* Forel, 1905, *P. retivertex* Eguchi, 2001, *P. colpigaleata* and *P. knowlesi* Mann, 1921 as inferred from the present tree (Fig. 2). We will deal with the *P. hortensi*s group in a separate paper with a new dataset.

Cost efficiency of the MIG-seq workflow used in the present study: The detailed estimations of reagent cost and operation times for 96 samples loaded into one sequencing run are as given in Suyama & Matsuki (2015). Because in the present study a total of 55 samples were processed in one sequencing run (reads per sample after quality filtering: 259,902 - 3,693,375), therefore the reagent cost is approximately 26 US dollars per sample (mainly cost for the MiSeq Kit and qPCR reagents, excluding the initial cost for the 1st and 2nd PCR primers). On the other hand, in our standard sequencing operation, a total of 384 samples (4 × 96-hole PCR plates; expected number of reads per sample ≈ 130,000) can be loaded into one sequence run, thereby keeping reagent costs down to approximately 4 USD per sample.

High applicability of the MIG-seq to millimeter-sized dry-mounted specimens: In this study, using the workflow of MIG-seq modified from Suyama & Matsuki (2015), phylogenetically-informative SNPs were successfully detected from millimeter-sized specimens of ants preserved for more than 15 years after dry-mounting. The phylogenetic tree constructed using MIG-seq SNP data was also overall reliable and in agreement with established existing knowledge. In conclusion, as well as targeted enrichment of ultra-conserved elements (Blaimer & al. 2016, Wood & al. 2018) and hybridization capture using RAD probes (Suchan & al. 2016), MIG-seq is highly applicable for use in phylogenetic and population genetic studies based on aged minute specimens in museum collections. Museomics using the MIG-seq approach can effectively expend the potential of museum material as critical DNA resources, contributing to studies involving the revelation of various phenomena relevant to applications in biodiversity sciences and conservation. For example, MIG-seq-generated data may enable studies of changes in genetic structure over historical timeframes, genetic degradation of endemic and / or endangered species and local populations caused by habitat loss and / or fragmentation, genetic pollution caused by introduced species or lineages, and origin and expansion routes of invasive species. Museum specimens, especially types and vouchers of endemic and / or endangered or even extinct species, are extremely valuable and sometimes indispensable for such research. In a world undergoing environmental changes at unprecedented rates, the intrinsic value of specimen repositories and the importance of drawing upon museum collections as data resources cannot be more understated (Schlick-Steiner & al. 2003).

Acknowledgements

We thank Seiki Yamane and Fuminori Ito for providing us valuable specimens. We also thank staff of the protected areas in Vietnam, Malaysia, Indonesia and China for their kind help and assistance during field surveys. KE's research activities were partly supported by Asahi Glass Foundation (Leader: Katsuyuki Eguchi; FY2017-FY2020) and Tokyo Metropolitan University Fund for TMU Strategic Research (Leader: Noriaki Murakami; FY2020-FY2022). Oguri's research activities were supported by Grants-in-Aid for Scientific Research, Japan Society for the Promotion of Science (No. 17J40194).

To the memory of our dear friend Bakhtiar Effendi Yahya, who passed away before this manuscript was accepted for publication.

References

Aberer, A.J., Kobert, K. & Stamatakis, A. 2014: ExaBayes: massively parallel bayesian tree inference for the whole-genome era. – Molecular Biology and Evolution 31: 2553-2556.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A. & Johnson, E.A. 2008: Rapid SNP discovery and genetic mapping using sequenced RAD markers. – Public Library of Science One 3: art. e3376.

Battery, C.J. & Klicka, J. 2017: Cryptic speciation and gene flow in a migratory songbird species complex: insights from the red-eyed vireo (*Vireo olivaceus*). – Molecular Phylogenetics and Evolution 113: 67-75.

Blaimer, B.B., Lloyd, M.W., Guillory, W.X. & Brady, S.G. 2016: Sequence capture and phylogenetic utility of genomic ultraconserved elements obtained from pinned insect specimens. – Public Library of Science One 11: art. e0161531.

Buerki, S. & Baker, W.J. 2016: Collections-based research in the genomic era. – Biological Journal of the Linnean Society 117: 5-7.

Campbell, E.O., Gage, E.V., Gage, R.V. & Sperling, F.A.H. 2020: Single nucleotide polymorphism-based species phylogeny of greater fritillary butterflies (Lepidoptera: Nymphalidae: *Speyeria*) demonstrates widespread mitonuclear discordance. – Systematic Entomology 45: 269-280.

Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A. & Cresko, W.A. 2013: Stacks: an analysis tool set for population genomics. – Molecular Ecology 22: 3124-3140.

Chen, Y.-T., Tseng, H.-Y., Jeng, M.-L., Su, Y.-C., Huang, W.-S. & Lin, C.-P. 2017: Integrated species delimitation and conservation implications of an endangered weevil *Pachyrhynchus sonani* (Coleoptera: Curculionidae) in Green and Orchid Islands of Taiwan. – Systematic Entomology 42: 796-813.

Díaz-Arce, N. & Rodríguez-Ezpeleta, N. 2019: Selecting RAD-seq data analysis parameters for population genetics: the more the better? – Frontiers in Genetics 10: art. 533.

Economo, E.P., Huang, J.P., Fischer, G., Sarnat, E.M., Narula, N., Janda, M., Guénard, B., Longino, J.T. & Knowles, L.L. 2019: Evolution of the latitudinal diversity gradient in the hyperdiverse ant genus *Pheidole*. – Global Ecology and Biogeography 28: 456-470.

Economo, E.P., Klimov, P., Sarnat, E.M., Guénard, B., Weiser, M.D., Lecroq, B. & Knowles, L.L. 2015a: Global phylogenetic structure of the hyperdiverse ant genus *Pheidole* reveals the repeated evolution of macroecological patterns. – Proceedings of the Royal Society B-Biological Sciences 282: art. 20141416.

Economo, E.P., Sarnat, E.M., Janda, M., Clouse, R., Klimov, P.B., Fischer, G., Blanchard, B.D., Ramirez, L.N., Andersen, A.N., Berman, M., Guénard, B., Lucky, A., Rabeling, C., Wilson, E.O. & Knowles, L.L. 2015b: Breaking out of biogeographical modules: range expansion and taxon cycles in the hyperdiverse ant genus *Pheidole*. – Journal of Biogeography 42: 2289-2301.

Eguchi, K. 2001: A revision of the Bornean species of the ant genus *Pheidole* (Insecta: Hymenoptera: Formicidae: Myrmicinae). – Tropics Monograph Series 2: 1-154.

Eguchi, K. 2008: A revision of Northern Vietnamese species of the ant genus *Pheidole* (Insecta: Hymenoptera: Formicidae: Myrmicinae). – Zootaxa 1902: 1-118.

Eguchi, K., Bui, T.V., Oguri, E. & Yamane, S. 2016: The first discovery of the "*Pheidole quadricuspis* group" in the Indo-Chinese Peninsula (Insecta: Hymenoptera: Formicidae: Myrmicinae). – Revue Suisse de Zoologie 123: 45-55.

Ewart, K.M., Johnson, R.N., Ogden, R., Joseph, L., Frankham, G.J. & Lo, N. 2019: Museum specimens provide reliable SNP data for population genomic analysis of a widely distributed but threatened cockatoo species. – Molecular Ecology Resources 19: 1578-1592.

Faircloth, B.C., McCormack, J.E., Grawford, N.G., Harvey, M.G., Brumfield, R.T. & Glenn, T.C. 2012: Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. – Systematic Biology 61: 717-726.

Graham, H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A.T. 2004: New developments in museum-based informatics and applications in biodiversity analysis. – Trends in Ecology & Evolution 19: 497-503.

Linck, E.B., Hanna, Z.R., Sellas, A. & Dumbacher, J.P. 2017: Evaluating hybridization capture with RAD probes as a tool for museum genomics with historical bird specimens. – Ecology and Evolution 7: 4755-4767.

Lister, A.M. & Climate Change Research Group 2011: Natural history collections as sources of long-term datasets. – Trends in Ecology & Evolution 26: 153-154.

Miller, J.A., Beentjes, K.K., van Helsdingen, P. & Ijland, S. 2013: Which specimens from a museum collection will yield DNA barcodes? A time series study of spiders in alcohol. – ZooKeys 365: 245-261.

Nakazato, T. 2018: A challenge to integrate bioinformatics and biodiversity informatics data as museomics. – Biodiversity Information Science and Standards 2: art. e26102.

Paris, J.R., Stevens, J.R. & Catchen, J.M. 2017: Lost in parameter space: a road map for STACKS. – Methods in Ecology and Evolution 8: 1360-1373.

Prosser, S.W., Dewaard, J.R., Miller, S.E. & Hebert, P.D.N. 2016: DNA barcodes from century-old type specimens using next-generation sequencing. – Molecular Ecology Resources 16: 487-497.

Razkin, O., Sonet, G., Breugelmans, K., Madeira, M.J., Gómez-Molier, B.J. & Backeljau, T. 2016: Species limits, interspecific hybridization and phylogeny in the cryptic land snail complex *Pyramidula*: the power of RADseq data. – Molecular Phylogenetics and Evolution 101: 267-278.

Roure, B., Baurain, D. & Philippe, H. 2012: Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. – Molecular Biology and Evolution 30: 197-214.

Satria, R., Kurushima, H., Herwina, H., Yamane, S. & Eguchi, K. 2015: The trap-jaw ant genus *Odontomachus* Latreille (Hymenoptera: Formicidae) from Sumatra, with a new species description. – Zootaxa 4048: 1-36.

Schäffer, S., Zachos, F. & Koblmüller, S. 2017: Opening the treasure chest: a DNA-barcoding primer set for most higher taxa of Central European birds and mammals from museum collections. – Public Library of Science One 12: art. e0174449.

Suarez, A.V. & Tsutsui, N.D. 2004: The value of museum collections for research and society. – BioScience 54: 66-74.

Suchan, T., Pitteloud, C., Gerasimova, N.S., Kostikova, A., Schmid, S., Arrigo, N., Pajkovic, M., Ronikier, M. & Alvarez, N. 2016: Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on collection specimens. – Public Library of Science One 11: art. e0151651.

Suyama, Y. & Matsuki, Y. 2015: MIG-seq: an effective PCR-based method for genome-wide single-nucleotide polymorphism genotyping using the next-generation sequencing platform. – Scientific Reports 5: art. 16963.

Tin, M.M.Y., Economo, E.P. & Mikheyev, A.S. 2014: Sequencing degraded DNA from non-destructively sampled museum specimens for RAD-tagging and low-coverage shotgun phylogenetics. – Public Library of Science One 9: art. e96793.

Velasco-Cuervo, S.M., Aguirre-Ramirez, E., Gallo-Franco, J.J., Obando, R.G., Carrejo, N. & Toro-Perea, N. 2019: Saving DNA from museum specimens: the success of DNA mini-barcodes in haplotype reconstruction in the genus *Anastrepha* (Diptera: Tephritidae). – Journal of Advanced Research 16: 123-134.

Wang, W.Y., Yamada, A. & Eguchi, K. 2018: First discovery of the mangrove ant *Pheidole sexspinosa* Mayr, 1870 (Formicidae Myrmicinae) from the Oriental region, with redescriptions of the worker, queen and male. – Raffles Bulletin of Zoology 66: 652-663.

Wachi, N., Matsubayashi, K.W. & Maeto, K. 2018: Application of next-generation sequencing to the study of non-model insects. – Entomological Science 21: 3-11.

Ward, D.F. 2012: More than just records: analyzing natural history collections for biodiversity planning. – Public Library of Science One 7: art. e50346.

Wood, H.M., González, V.L., Lloyd, M., Coddington, J. & Scharff, N. 2018: Next-generation museum genomics: phylogenetic relationships among palpimanoid spiders using sequence capture techniques (Araneae: Palimanoidea). – Molecular Phylogenetics and Evolution 127: 907-918.