SUPPLEMENTARY MATERIALS

Details of the record collection procedure

The process of collecting records is shown in Figure S1, consisting of 6 steps. Here we explain each step.

The first step is to download a complete list consisting of all of the record identifiers associated to an organism. Searching the specific organism in the NCBI taxonomy database shows the links of the records under that organism and the taxonomies below the species level (if any) across different databases; for example, the links for *Homo sapiens* and its records are shown in a single search.[1] In this study we select *Nucleotide* database. This links to GenBank and displays its *CoreNucleotide* division results by default. *CoreNucleotide* (also called *Nuccore*) is the main collection of GenBank. Its data are also exchanged with EMBL and DDBJ. Hence we use it in this study. We access *Direct links* of *Nucleotide* database shown from the search to get all the records directly belonging to the organism.[2] The complete identifier list of those records then can be downloaded using the *Send to* function in that web page. It is worthwhile to mention that using E-Utilities (5), the first step can be replaced with a single internal query. Note that we used `gi` number as record identifier. However it was phased out recently (after Sept, 2016). So please use `accession.version` instead. We used `gi` in the study since both options were available.

The records can then be downloaded in batch using E-Utilities based on the downloaded list (the second step). An analysis of the downloaded records can then be performed. In particular, we look for records that have more than one accession number in the *ACCESSION* field. This is because when a record replaces other records, their accession numbers will be recorded in the primary record.

After tracking records having multiple accession numbers, the third step is to search for their revision history using the online NCBI tools. The revision history documents when a record replaces another record. It records the identifiers of the two related records and the relevant replacement or merge date. Figure S2 shows an example of revision history, for record `AC098870.3`. Using the revision history, we can find out which specific version of the record replaced other a record. In this example, a previous version (`AC098870.1`) has replaced another record (`AC040969.2`). For our study, we tracked the exact versions of both replaced records and the records replacing them; these are the versions that were current at the time of the replacement or merge. Hence for this example, record `AC098870.1` and record `AC040969.2` were tracked as a duplicate pair. Record `AC098870.2` and record `AC098870.3` are not tracked because they are not the exact version of the record which replaced record `AC040969.2`. Notably record `AC098870.2` and record `AC098870.3` are not duplicates either. These are just standard updates. Standard updates refer to record is updated on itself, not replacing or merging other records. When only meta-data of a record is updated, `accession` number and `version` number remain the same. In contrast, When the sequence of a record is updated, `version` number will increase by 1. Therefore they refer to a single base record, whereas the replacement involves two completely separate records. The revision history can be used to identify the precise version of a record where a replacement was processed.

Nevertheless, in some situations the exact versions cannot be traced. For instance, when a record is only updated with meta data (any attributes except the sequence), the identifiers will remain the same. Under such conditions, it is not possible to find out the exact version of the record that has replaced others. Given the current available resources, it is still the best way to find the exact versions of records involving in the merge wherever possible.

In order to trace the revision history as above, we used a simple program to access the urls of the revision history of the records tracked in Step 3. The url of the record revision history is constructed as http://www.ncbi.nlm.nih.gov/nuccore/record_id?report=girevhist and the *record_id* refers to `accession.version`. For example, the revision history of record `AB968079.1` is found at http://www.ncbi.nlm.nih.gov/nuccore/AB968079.1?report=girevhist. The program documents the identifiers of the records and possible merge date if the revision history shows that a merge occurred.

Using the tracked record identifiers, we download records using E-Utilities (5) and cleanse accordingly.

We downloaded records in both *FASTA* and *GenBank* format. This is because we experienced the issue that downloading *GenBank* format records that have dynamic links to other records using E-Utilities will result in incorrect sequences. The sequences have the correct lengths but are filled with "N" characters. On the other hand, the *FASTA* format for the sequence is correct. Therefore we downloaded the two formats to ensure the correct results. We used *FASTA* format records to do sequence-based analysis, such as running alignment and calculating GC content and melting temperature because they have the correct sequences. We used *GenBank* format records to do annotation-based analysis, like measuring submitter similarity and categorizing based on annotations because *GenBank* format contains this information.

In terms of cleansing, we removed the repeated merged groups. This is because we observed inconsistencies occurred in some records' revision histories. For instance, record `NM_001168605.1` replaced record `XM_001251235.3` twice [3]. One was on 29 June 2010 whereas another was on 13 Dec 2009. This seems to be inconsistent as the same record has been replaced at different times. The record will be marked as "obsolete" if it has been replaced or merged so that it should not involve any further updates. Here a record was replaced in 2009 but it was replaced again in 2010, which causes ambiguity. Therefore we only leave one of them. It

[1] http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Undef&name=Homo+sapiens&lvl=0&srchmode=1

[2] Continuing the example, http://www.ncbi.nlm.nih.gov/nuccore/?term=txid9606[Organism:noexp] shows all the records in the nucleotide database for *Homo sapiens*.

[3] http://www.ncbi.nlm.nih.gov/nuccore/NM_001168605.1?report=girevhist

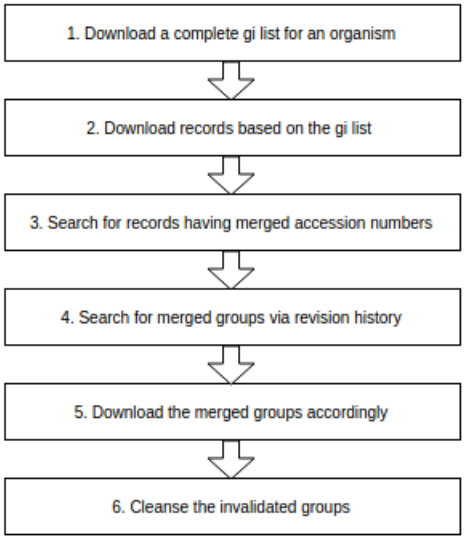

Figure S1. A complete data collection process. Note that `accession.version` format is now preferred since `gi` number is phased out. So please use `accession.version` instead. We used gi number as both options were available that time.

Figure S2. An example of revision history from GenBank

does not matter which we remove because both links point to the same record.

We also removed the records that are segments, that is, a record actually contains sub-records and has been labelled explicitly as a "segmented set". For instance, record GenBank accession and version number AH009070.1 has three segments.[4] In this case, it is not a single record. We speculate that submitters or database staff found the three records are actually segments so that they grouped them into one record and clearly marked it as "segmented set". This is a reasonable way to avoid duplicates, but we only found 18 such cases during the data collection.

We have also made the collection available.[5] Notice that some of these records are from RefSeq, but they are all accessible from INSDC databases.

Details of measuring submitter similarity

In order to estimate the proportion of duplicate pairs that are merged by different groups, we measured the submitter similarity between each pair. NCBI provides an annotated sample record to provide definitive guidance on the record fields in GenBank flat file format.[6] It defines that *Direct Submission* field (under *REFERENCE* field) contains submitter details. Hence we extracted these values and computed the similarity. We label a pair *Same* as long as it shares one common submitter. We label it *Different* if there are no submitters in common. If a record does not include the *Direct Submission* field, we label the pair *N/A*. Note that since typically only the original submitter can initiate a record merge, we would expect an overlap in submitters to occur frequently. However, from the data in Supplementary Table S1 we can see that this is not consistently the case.

From the submitter similarity results, many pairs were labelled *N/A*. The above definitive guidance specifies that "Some older records do not contain the 'Direct Submission' reference. However, it is required in all new records", so probably those records were published some time ago. (We do not know when the documentation was established, but the documentation on the web page was revised on 23 Oct 2006). We observed that some records do not have submitter details at all whereas some records placed such information in different places like the *COMMENT* field and other places under *REFERENCE* field and they are not as detailed as the ones in *Direct Submission* as requested. To avoid incorrect interpretation, we strictly followed the documentation and used *Direct submission* to measure the submitter similarity, and we cautiously labelled "N/A" for pairs not having such field.

Details of measuring sequence similarities

Measurement of sequence identity is based on the BLAST software provided by NCBI (1). We used the stand alone version 2.2.30 and its *bl2seq* application that aligns sequence pairwise such that sequence identity of all the pairs can be reported. NCBI BLAST staff gave valuable advice on the recommended parameters for running BLAST pairwise alignment in general. In particular, we disabled the dusting parameter which automatically filters non-complexity regions and selected the smallest word size (4) aiming to achieve the highest accuracy as possible. With this setting, we can reasonably conclude that a pair has low sequence identity if there are *"no hits"* found or its *expected value* score is above 0.001 in the output. The specific command employed for producing alignments was:

./ncbi-blast-2.2.30+/bin/blastn -task blastn -query query_file_path -subject subject_file_path -word_size 4 -dust no -out output_file_path

In some cases although sequences have high local alignment identity, the alignment length is actually very small. For instance, a duplicate with over 10,000 base pairs may have over 90% local sequence identity with its replacement but the actual alignment length might be less than 100 base pairs. It is therefore necessary to use an additional metric, which we refer to as *local alignment proportion*. This can also estimate the coverage of the pair globally instead of running global alignment. Getting explicit global sequence identity for each pair requires running the full global alignment pairwise without applying heuristics to skip pairs that have estimated lower identity than the threshold. Thus it is computational intensive and is also the major reason that why non-redundant database in NCBI is no longer non-redundant as mentioned before. Hence we use *local alignment proportion* to estimate

[4]http://www.ncbi.nlm.nih.gov/nuccore/AH009070.1?report=genbank
[5]https://cloudstor.aarnet.edu.au/plus/index.php/s/Xef2fvsebBEAv9w
[6]http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

the likelihood of global identity between each pair. It is computed by the length of the identical bases dividing by the larger of the two sequence lengths, that is:

$$L = \frac{len(I)}{max(len(D), len(R))}$$

where L is the local alignment proportion, I is the locally aligned identical bases, D is the duplicate sequence, R is the replacement sequence, and $len(S)$ is the length of a sequence S.

Details of formulas in the case study

The formulas used in the case study in section 5.1 are presented as below.

For GC content, the equation appears in **1**. For this calculation, we removed special characters in sequences, like gaps, beforehand. The results would have larger differences if special characters are included.

$$GC = \frac{yG + zC}{(wA + xT + yG + zC)} \quad (\mathbf{1})$$

For melting temperature, we adopted three formulas: T_{basic} or T_b (Formula **2**): base line or "rule of thumb" (3); T_{salt} or T_s (Formula **3**): measuring the temperature using Na^+ concentration approach (3); $T_{advanced}$ or T_a (Formula **4**): a more advanced approach which has the best fit constant based on the previous study (2). We do not use the state of art Nearest Neighbour approach (originally published in (4)) since it has more external parameters (Mg_2^+, dNTPs, and DMSO) to control. Three formulas are sufficient to represent the melting temperature in different aspects.

$$T_{basic} = \begin{cases} (wA + xT) \times 2 + (yG + zC) \times 4 & \text{if } length < 14 \\ \frac{64.9 + 41 \times (yG + zC - 16.4)}{(wA + xT + yG + zC)} & \text{if } length >= 14 \end{cases} \quad (\mathbf{2})$$

$$T_{salt} = \begin{cases} (wA + xT) \times 2 + (yG + zC) \times 4 - 16.6 \times \log_{10}(0.050) \\ +16.6 \times \log_{10}([Na^+]) & \text{if } length < 14 \\ 100.5 + \frac{41 \times (yG + zC)}{(wA + xT + yG + zC)} - \frac{820}{(wA + xT + yG + zC)} \\ +16.6 \times \log_{10}([Na^+]) & \text{if } length >= 14 \end{cases} \quad (\mathbf{3})$$

$$T_{advanced} = 77.1 + 11.7 \times \log_{10}[Na^+] + 0.41 \times GC \times 100 - \frac{528}{length} \quad (\mathbf{4})$$

Full results of the case study for the organisms in the dataset with a minimum of 100 merged groups or duplicate pairs appear in Supplementary Tables S4 and S5 respectively.

REFERENCES

1. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T.L. Madden. Blast+: architecture and applications. *BMC bioinformatics*, 10(1):421, 2009.
2. S. Chavali, A. Mahajan, R. Tabassum, S. Maiti, and D. Bharadwaj. Oligonucleotide properties determination and primer designing: a critical examination of predictions. *Bioinformatics*, 21(20):3918–3925, 2005.
3. W.A. Kibbe. Oligocalc: an online oligonucleotide properties calculator. *Nucleic acids research*, 35(suppl 2):W43–W46, 2007.
4. W. Rychlik, W.J. Spencer, and R.E. Rhoads. Optimization of the annealing temperature for dna amplification in vitro. *Nucleic acids research*, 18(21):6409–6412, 1990.
5. E. Sayers. The e-utilities in-depth: parameters, syntax and more, 2014.

Table S1. Duplicate data collected across 21 organisms from the INSDC databases

Organism	Total records	Merged accessions	Available merged groups	Duplicate pairs	Submitters (%)
Saccharomyces cerevisiae	68,236	3,517	165	191	5.8, 1.6, 92.6
Plasmodium falciparum	43,375	51	18	26	34.6, 3.8, 61.6
Zea mays	613,768	1,077	454	471	0.0, 7.6, 92.4
Mycoplasma pneumoniae	1,009	4	2	3	66.7, 0.0, 33.3
Bos taurus	245,188	38,557	12,822	20,945	0.0, 0.0, 100.0
Drosophila melanogaster	211,143	734	431	3,039	87.8, 0.1, 12.1
Homo sapiens	12,506,281	113,576	16,545	30,336	0.1, 17.9, 82.0
Escherichia coli	512,541	398,924	201	231	13.0, 2.2, 84.8
Xenopus laevis	35,544	1,690	1,620	1,660	0.0, 0.7, 99.3
Pneumocystis carinii	528	8	1	1	0.0, 0.0, 100.0
Oryza sativa	108,395	13	6	6	16.7, 0.0, 83.3
Chlamydomonas reinhardtii	24,891	1,601	10	17	5.9, 5.9, 88.2
Caenorhabditis elegans	74,404	2,029	1,881	1,904	5.6, 92.5, 1.9
Rattus norvegicus	318,577	20,180	12,411	19,295	0.0, 0.0, 100.0
Danio rerio	153,360	9,350	7,895	9,227	0.0, 0.0, 100.0
Mus musculus	1,730,943	291,842	13,222	23,733	1.3, 9.8, 88.9
Hepatitis C virus	130,456	91	32	48	93.8, 0.0, 6.2
Schizosaccharomyces pombe	4,086	54	39	545	10.8, 81.8, 7.4
Arabidopsis thaliana	337,640	6,058	47	50	26.0, 8.0, 66.0
Takifugu rubripes	51,654	14,294	64	72	0.0, 4.2, 95.8
Dictyostelium discoideum	7943	64	25	26	11.5, 0.0,88.5

Total records: Number of records in total directly belong to the organism (derived from NCBI taxonomy database); Merged accessions Number of records having more than one accession number defined in *ACCESSION* fields; Available merged groups: out of those records having merged accessions, the number of groups that are currently tracked in record revision history; Duplicates: the number of duplicates in total; Submitters: the proportion of the records were processed by the same submitter, different submitters, and submitters that cannot be traced respectively

Table S2. The duplicate data collection, broken down by the different types of duplicates categorized at both the sequence level and the annotation level

Organism	Sequence-based					Annotation-based			Others	
	ES	SS	EF	SF	LI	WD	SP	PR	LS	UC
Saccharomyces cerevisiae	100	27	49	11	4	0	0	0	0	0
Plasmodium falciparum	6	5	11	3	1	0	0	0	0	0
Zea mays	163	188	85	30	1	0	1	357	4	4
Mycoplasma pneumoniae	0	0	3	0	0	0	0	0	0	0
Bos taurus	2923	3633	5167	6984	149	0	0	18120	2089	2089
Drosophila melanogaster	106	35	1816	503	57	0	1586	0	522	522
Homo sapiens	2844	7139	11325	6890	642	2951	316	17243	1496	930
Escherichia coli	86	33	70	40	2	0	0	1	0	0
Xenopus laevis	1601	7	37	13	2	0	0	0	0	0
Pneumocystis carinii	0	0	1	0	0	0	0	0	0	0
Oryza sativa	3	1	1	1	0	0	0	0	0	0
Chlamydomonas reinhardtii	4	5	8	0	0	0	0	0	0	0
Caenorhabditis elegans	1736	7	109	44	5	0	121	0	0	0
Rattus norvegicus	2511	5302	7556	3817	107	0	0	15832	2	2
Danio rerio	721	2740	1662	3504	75	1	34	7684	525	491
Mus musculus	2597	4689	6678	7379	379	1926	1305	16510	2011	2011
Hepatitis c	0	15	31	2	0	0	0	0	0	0
Schizosaccharomyces pombe	18	7	5	4	1	0	0	0	510	510
Arabidopsis thaliana	27	9	7	4	3	0	2	0	0	0
Takifugu rubripes	24	24	12	12	0	0	0	56	0	0
Dictyostelium discoideum	12	5	6	3	0	0	0	0	0	0

ES: exact sequences; SS: similar sequences; EF: exact fragments; SF: similar fragments; LI: low-identity sequences; WD: working draft; SP: sequencing-in-progress record; PR: predicted sequence; LS: long sequence (we did not run BLAST to compute pariwise local identity if at least one record of the pair has sequence greater or equal to 1 million bases); UC: unclassified pairs (for pairs whose label for sequence level is LS and no keywords found for annotation level)

Table S3. Merged record examples

Category	Sample records	Comments
ES	U88068.1 and Y10925*	Both of them have exactly same sequences
SS	NM_001163193.1 and XM_002688000.1	Local identity 100%; the alignment proportion over 90%
EF	AH011371.2 and AY044907.1	Local identity 100%; the alignment proportion below 30%
SF	AB011371.1 and D76439*	Local identity 98%; the alignment proportion below 60%
LI	X66742.1 and S45098.1	Local identity about 83%
WD	AC156562.1 and AC120385.3	Annotated as "WORKING DRAFT"
SP	AP000014.2 and AP000015.1	Annotated as "SEQUENCING IN PROGRESS"
PR	NM_001177453.1 and XM_002667401.1	Annotated as "PREDICTED"

ES: exact sequences; SS: similar sequences; EF: exact fragments; SF: similar fragments; LI: low-identity sequences; WD: working draft; SP: sequencing-in-progress record; PR: predicted sequence. The first five are sequence-based whereas the others are annotation-based. *: the version number is not provided. Also note that the example records having NM and XM prefix are RefSeq records, which are searchable and browsable in INSDC

Table S4. Case study results, comparing GC content and melting temperature differences for each pair (Exemplar vs. Original merged group). The study includes organisms whose categories have at least 100 groups

Organism	Category	Size	GC (%)		Melting temperature					
			mdiff	std	T_b mdiff	std	T_s mdiff	std	T_a mdiff	std
Saccharomyces cerevisiae	ES	100	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	ALL	165	0.23	0.57	0.30	**1.64**	0.35	**2.05**	0.31	**1.67**
Zea mays	ES	162	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.01
	SS	185	0.29	0.27	0.12	0.11	0.12	0.11	0.15	0.14
	PR	357	0.34	0.58	0.14	0.22	0.13	0.22	0.17	0.29
	ALL	454	0.38	0.71	0.17	0.37	0.17	0.39	0.21	0.42
Bos taurus	ES	2866	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	SS	2788	0.53	0.54	0.22	0.22	0.22	0.22	0.28	0.28
	EF	3530	**1.85**	**1.83**	0.74	0.76	0.74	0.78	0.94	0.94
	SF	4441	**1.61**	**1.61**	0.64	0.64	0.64	0.64	0.82	0.81
	LI	101	**2.80**	**3.10**	**1.14**	**1.40**	**1.15**	**1.46**	**1.45**	**1.69**
	PR	11382	**1.15**	**1.60**	0.46	0.63	0.45	0.63	0.58	0.81
	ALL	12822	**1.11**	**1.54**	0.44	0.63	0.44	0.63	0.57	0.79
Drosophila melanogaster	ES	105	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.01
	EF	202	0.83	**1.11**	0.42	0.64	0.45	0.77	0.46	0.60
	SF	194	0.86	0.89	0.36	0.36	0.36	0.37	0.45	0.46
	SP	185	0.74	0.77	0.30	0.32	0.30	0.33	0.37	0.41
	ALL	431	0.56	0.90	0.26	0.49	0.28	0.56	0.30	0.50
Homo sapiens	ES	2454	0.01	0.03	0.00	0.01	0.00	0.01	0.01	0.01
	SS	4887	0.28	0.34	0.11	0.14	0.11	0.14	0.14	0.18
	EF	5360	**1.51**	**2.04**	0.92	**1.28**	**1.01**	**1.50**	**1.01**	**1.28**
	SF	5003	**1.01**	**1.60**	0.41	0.68	0.41	0.71	0.52	0.84
	LI	369	**3.47**	**3.28**	**1.56**	**2.11**	**1.60**	**2.42**	**1.93**	**2.43**
	WD	2432	0.19	0.56	0.08	0.23	0.08	0.23	0.10	0.29
	SP	257	0.52	**1.18**	0.22	0.47	0.22	0.47	0.27	0.61
	PR	9214	**1.11**	**1.89**	0.44	0.76	0.45	0.77	0.56	0.96
	ALL	16545	0.87	**1.65**	0.46	0.92	0.48	**1.04**	0.52	0.99
Xenopus laevis	ES	1589	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ALL	1620	0.02	0.24	0.02	0.22	0.02	0.26	0.02	0.20
Caenorhabditis elegans	ES	1736	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	EF	100	**1.01**	**1.38**	0.41	0.46	0.41	0.51	0.52	0.54
	SP	121	0.88	0.87	0.35	0.31	0.35	0.32	0.44	0.41
	ALL	1878	0.07	0.41	0.03	0.17	0.03	0.18	0.04	0.20
Rattus norvegicus	ES	2113	0.01	0.02	0.00	0.01	0.00	0.01	0.00	0.01
	SS	3696	0.36	0.35	0.15	0.15	0.15	0.15	0.19	0.18
	EF	4880	**1.47**	**1.48**	0.58	0.60	0.58	0.62	0.74	0.74
	SF	2846	**1.21**	**1.25**	0.47	0.48	0.47	0.48	0.61	0.62
	PR	9286	0.97	**1.31**	0.38	0.50	0.37	0.50	0.49	0.65
	ALL	12411	0.91	**1.25**	0.36	0.50	0.36	0.51	0.46	0.63
Danio rerio	ES	720	0.01	0.02	0.00	0.01	0.00	0.01	0.00	0.01
	SS	2554	0.30	0.30	0.12	0.12	0.12	0.12	0.15	0.16
	EF	1496	**1.59**	**1.54**	0.59	0.57	0.58	0.57	0.77	0.75
	SF	3142	**1.55**	**1.44**	0.59	0.55	0.58	0.55	0.76	0.71
	PR	6761	**1.06**	**1.35**	0.40	0.51	0.39	0.50	0.52	0.66
	ALL	7895	**1.01**	**1.32**	0.38	0.50	0.38	0.49	0.50	0.65
Mus musculus	ES	2187	0.01	0.03	0.00	0.01	0.00	0.01	0.01	0.01
	SS	3402	0.29	0.34	0.12	0.14	0.12	0.14	0.15	0.18
	EF	3809	**1.47**	**1.71**	0.64	0.80	0.65	0.87	0.78	0.92
	SF	5179	0.96	**1.39**	0.38	0.56	0.38	0.57	0.48	0.71
	LI	235	**2.92**	**2.80**	**1.31**	**1.86**	**1.35**	**2.23**	**1.62**	**2.02**
	WD	1926	0.16	0.28	0.07	0.12	0.07	0.12	0.08	0.15
	SP	1305	0.15	0.22	0.06	0.09	0.06	0.09	0.08	0.12
	PR	8844	0.96	**1.51**	0.38	0.60	0.38	0.60	0.49	0.76
	ALL	13222	0.83	**1.39**	0.34	0.64	0.35	0.68	0.43	0.76

ES: exact sequences; SS: similar sequences; EF: exact fragments; SF: similar fragments; LI: low-identity sequences; WD: working draft; SP: sequencing-in-progress record; PR: predicted sequence; mdiff: the mean of absolute value of each pair difference; T_b, T_s, T_a refer to melting temperature calculated using basic (Formula **1**), salted (Formula **2**), and advanced (Formula **3**) respectively;std: standard deviations

Table S5. Case study results, comparing GC content and melting temperature differences for each pair (Exemplar vs. Duplicate). The study includes organisms whose categories have at least 100 duplicates

Organism	Category	Size	GC (%)		Melting temperature					
			mdiff	std	T_b mdiff	std	T_s mdiff	std	T_a mdiff	std
Saccharomyces cerevisiae	ES	100	0.00	0.02	0.00	0.01	0.00	0.01	0.00	0.01
	ALL	191	0.85	**2.26**	0.95	**5.40**	**1.12**	**6.81**	**1.00**	**5.50**
Zea mays	ES	163	0.00	0.03	0.00	0.01	0.00	0.01	0.00	0.01
	SS	188	0.59	0.55	0.24	0.22	0.24	0.22	0.31	0.28
	PR	357	0.69	**1.16**	0.27	0.45	0.27	0.44	0.35	0.58
	ALL	471	0.83	**1.52**	0.37	0.83	0.38	0.90	0.45	0.94
Bos taurus	ES	2923	0.00	0.02	0.00	0.01	0.00	0.01	0.00	0.01
	SS	3633	0.98	0.97	0.40	0.39	0.40	0.39	0.51	0.50
	EF	5167	**3.44**	**3.41**	**1.40**	**1.58**	**1.41**	**1.69**	**1.77**	**1.85**
	SF	6984	**2.86**	**2.86**	**1.14**	**1.13**	**1.13**	**1.13**	**1.46**	**1.45**
	LI	149	**5.47**	**5.41**	**2.22**	**2.42**	**2.22**	**2.50**	**2.83**	**2.93**
	PR	18120	**2.24**	**2.89**	0.89	**1.14**	0.88	**1.14**	**1.14**	**1.46**
	ALL	20945	**2.18**	**2.80**	0.88	**1.19**	0.88	**1.23**	**1.12**	**1.46**
Drosophila melanogaster	ES	106	0.00	0.03	0.00	0.01	0.00	0.01	0.00	0.01
	EF	1816	**2.51**	**2.20**	**1.06**	**1.02**	**1.07**	**1.09**	**1.32**	**1.19**
	SF	503	**2.19**	**2.22**	0.91	0.91	0.91	0.91	**1.15**	**1.15**
	SP	1586	**2.56**	**2.28**	**1.06**	0.93	**1.07**	0.93	**1.34**	**1.18**
	ALL	3039	**2.18**	**2.10**	0.91	0.94	0.92	0.99	**1.14**	**1.12**
Homo sapiens	ES	2844	0.02	0.05	0.01	0.02	0.01	0.02	0.01	0.03
	SS	7139	0.52	0.59	0.21	0.24	0.21	0.24	0.27	0.31
	EF	11325	**3.38**	**3.79**	**1.99**	**2.85**	**2.20**	**3.35**	**2.14**	**2.73**
	SF	6890	**2.19**	**3.02**	0.89	**1.27**	0.89	**1.31**	**1.13**	**1.57**
	LI	642	**5.67**	**5.40**	**2.49**	**3.32**	**2.54**	**3.78**	**3.09**	**3.86**
	WD	2951	0.80	**1.70**	0.33	0.70	0.33	0.70	0.42	0.89
	SP	316	**1.41**	**2.59**	0.60	**1.07**	0.60	**1.07**	0.75	**1.35**
	PR	17243	**2.05**	**3.20**	0.83	**1.29**	0.83	**1.30**	**1.05**	**1.63**
	ALL	30336	**2.15**	**3.24**	**1.11**	**2.09**	**1.19**	**2.40**	**1.26**	**2.13**
Xenopus laevis	ES	1601	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ALL	1660	0.14	**1.13**	0.09	0.77	0.10	0.87	0.09	0.78
Caenorhabditis elegans	ES	1736	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	EF	109	**2.77**	**3.96**	**1.02**	**1.24**	**1.06**	**1.33**	**1.30**	**1.56**
	SP	121	**1.76**	**1.73**	0.69	0.63	0.69	0.63	0.89	0.81
	ALL	1901	0.21	**1.23**	0.08	0.44	0.08	0.46	0.10	0.54
Rattus norvegicus	ES	2511	0.01	0.04	0.01	0.01	0.01	0.01	0.01	0.02
	SS	5302	0.61	0.62	0.25	0.26	0.25	0.26	0.32	0.32
	EF	7556	**2.58**	**2.59**	**1.03**	**1.14**	**1.04**	**1.20**	**1.31**	**1.36**
	SF	3817	**2.19**	**2.27**	0.85	0.88	0.85	0.88	**1.10**	**1.13**
	LI	107	**3.73**	**3.43**	**1.58**	**1.48**	**1.59**	**1.53**	**1.98**	**1.81**
	PR	15832	**1.63**	**2.19**	0.63	0.84	0.63	0.83	0.82	**1.09**
	ALL	19295	**1.63**	**2.21**	0.65	0.93	0.65	0.96	0.83	**1.14**
Danio rerio	ES	721	0.01	0.03	0.01	0.01	0.01	0.01	0.01	0.02
	SS	2740	0.59	0.59	0.24	0.24	0.24	0.23	0.30	0.30
	EF	1662	**3.06**	**3.00**	**1.14**	**1.11**	**1.12**	**1.10**	**1.49**	**1.45**
	SF	3504	**3.03**	**2.81**	**1.15**	**1.07**	**1.14**	**1.07**	**1.49**	**1.39**
	PR	7684	**2.06**	**2.62**	0.78	0.98	0.77	0.98	**1.01**	**1.28**
	ALL	9227	**1.95**	**2.55**	0.74	0.96	0.73	0.95	0.96	**1.25**
Mus musculus	ES	2597	0.02	0.05	0.01	0.02	0.01	0.02	0.01	0.03
	SS	4689	0.53	0.62	0.22	0.25	0.22	0.25	0.28	0.32
	EF	6678	**2.88**	**3.22**	**1.33**	**1.78**	**1.39**	**1.99**	**1.58**	**1.92**
	SF	7379	**2.01**	**2.58**	0.80	**1.04**	0.80	**1.05**	**1.02**	**1.32**
	LI	379	**4.79**	**4.80**	**2.07**	**3.05**	**2.12**	**3.62**	**2.59**	**3.35**
	WD	1926	0.32	0.57	0.13	0.23	0.13	0.23	0.17	0.29
	SP	1305	0.29	0.45	0.12	0.19	0.12	0.19	0.15	0.23
	PR	16510	**1.83**	**2.66**	0.72	**1.05**	0.72	**1.05**	0.92	**1.34**
	ALL	23733	**1.87**	**2.68**	0.80	**1.31**	0.82	**1.43**	0.99	**1.51**

ES: exact sequences; SS: similar sequences; EF: exact fragments; SF: similar fragments; LI: low-identity sequences; WD: working draft; SP: sequencing-in-progress record; PR: predicted sequence; mdiff: the mean of absolute value of each pair difference; T_b, T_s, T_a refer to melting temperature calculated using basic (Formula **1**), salted (Formula **2**), and advanced (Formula **3**) respectively;std: standard deviations