excluderanges: exclusion sets for T2T-CHM13, GRCm39, and other genome assemblies

Jonathan D. Ogata[1], Wancen Mu[3], Eric S. Davis[5], Bingjie Xue[10], J. Chuck Harrell[2,11], Nathan C. Sheffield[10], Douglas H. Phanstiel[5,6,7,8,9], Michael I. Love[3,4], Mikhail G. Dozmorov[1,2*]

[1]Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, 23298, USA.

[2]Department of Pathology, Virginia Commonwealth University, Richmond, VA, 23284, USA.

[3]Department of Biostatistics, University of North Carolina-Chapel Hill, Chapel Hill, NC 27514, USA

[4]Department of Genetics, University of North Carolina-Chapel Hill, Chapel Hill, NC 27514, USA

[5]Curriculum in Bioinformatics and Computational Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[6]Thurston Arthritis Research Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[7]Department of Cell Biology and Physiology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[8]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[9]Curriculum in Genetics and Molecular Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[10]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, 22908, USA

[11]Massey Cancer Center, Virginia Commonwealth University, Richmond, VA 23220, USA

Abstract

Summary: Exclusion regions are sections of reference genomes with abnormal pileups of short sequencing reads. Removing reads overlapping them improves biological signal, and these benefits are most pronounced in differential analysis settings. Several labs created exclusion region sets, available primarily through ENCODE and Github. However, the variety of exclusion sets creates uncertainty which sets to use. Furthermore, gap regions (e.g., centromeres, telomeres, short arms) create additional considerations in generating exclusion sets. We generated exclusion sets for the latest human T2T-CHM13 and mouse GRCm39 genomes and systematically assembled and annotated these and other sets in the *excluderanges* R/Bioconductor data package, also accessible via the BEDbase.org API. The package provides unified access to 82 GenomicRanges objects covering six organisms, multiple genome assemblies and types of exclusion regions. For human hg38 genome assembly, we recommend *hg38.Kundaje.GRCh38_unified_blacklist* as the most well-curated and annotated, and sets generated by the Blacklist tool for other organisms.

Availability and implementation: https://bioconductor.org/packages/excluderanges/

Contact: Mikhail G. Dozmorov (mdozmorov@vcu.edu)

Supplementary information: Package website: https://dozmorovlab.github.io/excluderanges/

Introduction

Up to 87% of sequencing reads generated by chromatin targeting technologies (e.g., ChIP-seq) can map to a reference genome in distinct clusters (aka high-signal pileups)[1,2] (*1*). These pileups frequently occur in regions near assembly gaps, copy number-high regions, and in low-complexity regions (*2*, *3*). Removing reads overlapping those regions, referred hereafter as exclusion sets, improves normalization of the signal between samples, correlation between replicates, and increases accuracy of both peak calling and differential ChIP-seq analysis (*4*–*6*). Therefore, standardized availability of those exclusion sets is critical for improving reproducibility and quality of bioinformatics analyses.

Finding and choosing an exclusion set can be a non-trivial task. The ENCODE project returns 94 hits using the "exclusion" search term (as of 11/08/2022)[3], most of them having minimal annotation and unknown curation methods. These sets are available for human and mouse genome assemblies; however, the ENCODE project lacks exclusion sets for the latest Telomere-to-Telomere (T2T-CHM13) human and Genome Reference Consortium Mouse Build 39 (GRCm39/mm39) mouse assemblies. Converting exclusion set coordinates between genomic assemblies using liftOver is not advisable since new artifact-prone regions are added and others are lost due to closed gaps (*1*); therefore, exclusion sets should be generated and used for their respective genome assemblies. Furthermore, exclusion regions have been observed in genomes of other species and many exclusion sets for model organisms remain unpublished and scattered across GitHub repositories. We curated a collection of exclusion sets for six model organisms and 12 genome assemblies, including the newly generated T2T and mm39 exclusion sets. We included two other types of potentially problematic regions: University of California Santa Cruz (UCSC)-annotated gap sets, e.g., centromere, telomere, short arm, and Nuclear mitochondrial (NUMT) sets containing mitochondrial sequences present in the nuclear genome (*7*). We assemble a total of 82 uniformly processed and annotated exclusion sets in the *excluderanges* R/Bioconductor data package and provide API access via BEDbase.org.

---

1 https://docs.google.com/spreadsheets/d/1G4SkqUMiGcUlvR6homc7RW33nSOf4mS9QYJifsd4qo0

2 https://sites.google.com/site/anshulkundaje/projects/blacklists

3 https://www.encodeproject.org/search/?searchTerm=exclusion

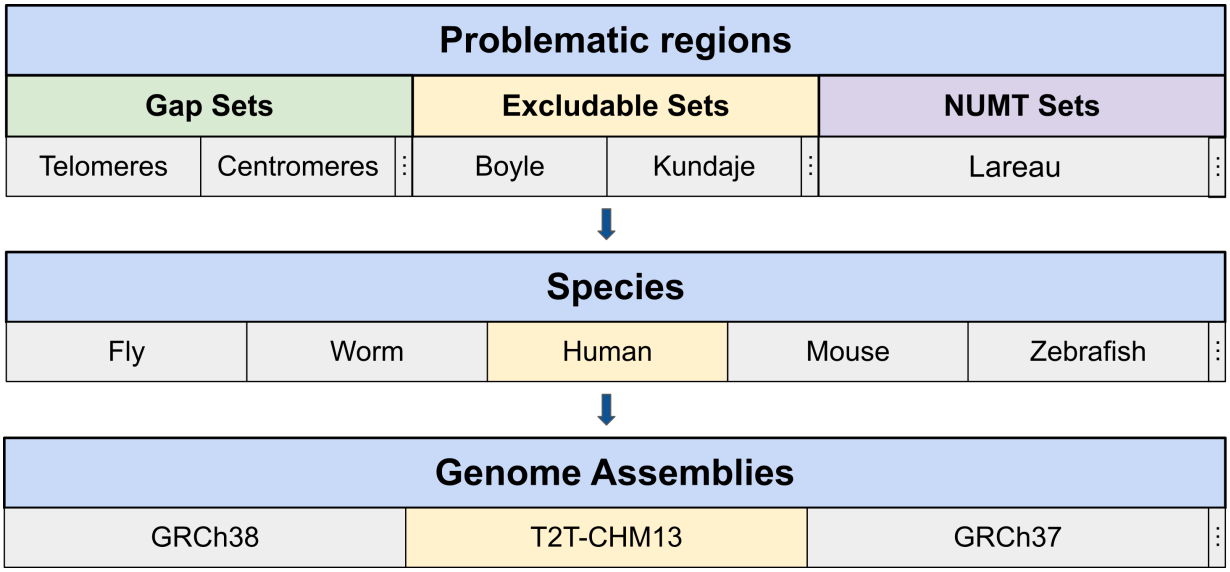


Figure 1. Schematic overview of the *excluderanges* package. Data for each type of problematic region (exclusion sets, gaps, Nuclear Mitochondrial (NUMT) sets) were obtained from public sources for each model organism and the corresponding genome assemblies. Exclusion sets for T2T-CHM13 and GRCm39 genome assemblies were *de novo* generated. Three vertical dots indicate more categories in the corresponding section.

Implementation

An overview of the *excluderanges* data is shown in Figure 1. To create this resource, we performed a systematic internet and literature search. The ENCODE project was the largest source of exclusion sets for human (11 sets) and mouse (6 sets) organisms, covering hg19, hg38, mm9, and mm10 genome assemblies. We also obtained exclusion sets generated by the Blacklist (*1*) and PeakPass (*5*) software. Additionally, we obtained exclusion sets for *C. elegans* (ce10 and ce11 genome assemblies), *D. melanogaster* (dm3 and dm6), *D. rerio* (danRer10), and *A. thaliana* (TAIR10). Using the Blacklist software, we generated exclusion sets for the latest Telomere-to-Telomere (T2T-CHM13) human and Genome Reference Consortium Mouse Build 39 (GRCm39/mm39) mouse assemblies (Table 1, Supplementary Table S1).

Mitochondrial DNA sequences (mtDNA, 100-600K mitochondria per human cell) transferred to the nucleus give rise to the so-called mitochondrial DNA sequences in the nuclear genome (NUMTs). These sequences are found in genomes of various species (*7*), suggesting NUMTs may be a pervasive phenomenon. In the settings of DNA/chromatin sequencing (e.g., ATAC-seq), up to 80% of mitochondrial sequencing reads (*8*) may pile up in the NUMT sequences. Similar to exclusion sets, genomic regions highly homologous to mtDNA can be masked to improve biological signal. The reference human nuclear mitochondrial sequences have been available in the UCSC genome browser for hg18 (RHNumtS.2 database (*9*)) and lifted over to hg19 human genome assembly. Similarly, mouse NUMTs (RMNumtS database (*10*)) are available for the mm9 mouse genome assembly. However, recent human, mouse, and other organism genome assemblies lack NUMTs annotations in the UCSC database. We collected NUMT sets for more recent human and mouse genome assemblies, including hg38, T2T-CHM13, mm10, generated by Caleb Lareau in the mitoblacklist GitHub repository[4].

[4] https://github.com/caleblareau/mitoblacklist

Gaps in the genome represent another type of problematic regions. These include centromere and telomere sequences, short arms, gaps from large heterochromatin blocks, etc. While some are present in genome assemblies of most organisms (centromeres, telomeres, short arms, covering 2.47% $\pm$ 1.64, 0.01% $\pm$ 0.01, and 15.39% $\pm$ 3.66 of hg38 chromosomes, respectively), many are assembly-specific (e.g., gaps between clones, contigs, scaffolds in hg19 and hg38 assemblies). Gap data are available from the UCSC Genome Browser database or UCSC-hosted data hubs. The T2T-CHM13 assembly lacks assembly-specific gaps by the definition of telomere-to-telomere sequencing (*11*); however, coordinates of centromeres and telomeres are available from the CHM13 GitHub repository[5]. Additionally, we obtained T2T peri/centromeric satellite annotations, known to be associated with constitutive heterochromatin and span sites involved in kinetochore assembly or sequences epigenetically marked as centromeres (*12*). We also included the rDNA gap regions and regions unique to T2T-CHM13 v2.0 as compared with GRCh38/hg38 and GRCh37/hg19 assemblies under the rationale that alignments within these previously problematic regions might warrant extra attention. We characterized hg38 exclusion sets for overlap with gap regions and found that *hg38.Kundaje.GRCh38_unified_Excludable*, *hg38.Boyle.hg38-Excludable.v2*, and *hg28.Wimberley.peakPass60Perc_sorted* cover 99.40%, 99.08%, and 59.60% of centromeric regions, respectively. Notably, relatively few large regions were responsible for these overlaps (e.g., 27 out of 910 in *hg38.Kundaje.GRCh38_unified_Excludable*). In contrast, over 60% of the *hg38.Nordin.CandRblacklist_hg38* exclusion set for the CUT&RUN technology overlapped centromeres on chromosomes 1 and 13. Only sets generated by the Blacklist software overlapped centromeres, telomeres, and short arms, and there results were consistent across organisms and genome assemblies (Supplementary Table S2). Given the distinct properties of gap regions and inconsistency of their presence in exclusion sets, the aforementioned NUMTs and gap sets may be combined with other exclusion sets.

The large number of exclusion sets (e.g., nine for hg38 human genome assemblies) creates uncertainty in which set to use for a given genome assembly. We annotated exclusion sets by their creation methods, date of last update, width distribution, percent of the genome covered, and other properties (Supplementary Table S1, BEDbase.org[6]). Only sets generated by the Boyle's lab Blacklist (*1*) or PeakPass by Eric Wimberley (*5*) software had published methods. While methods for some sets may be inferred (e.g., the hg38 *Yeo.eCLIP_Excludableregions.hg38liftover* set may have been lifted over from hg19), we advise against using poorly annotated sets. We also characterized hg38 exclusion sets and found they vary dramatically in terms of number (12,052 - 38) and width (median 10,151 - 30bp) (Supplementary Figure S1A, B). We calculated Jaccard overlap between each pair of hg38 exclusion sets, $J(A,B) = \frac{width(\cap_{A,B})}{width(\cup_{A,B})}$. We found that *hg38.Kundaje.GRCh38_unified_Excludable* had the best Jaccard overlap with other sets, followed by_hg38.Wimberley.peakPass60Perc_sorted_ and *hg38.Boyle.hg38-Excludable.v2* sets (Supplementary Figure S1C). We additionally calculated overlap coefficient $C(A,B) = \frac{width(\cap_{A,B})}{Min(width(A),width(B))}$ to minimize the effect of set size differences. We similarly found Kindaje-generated sets showing the best overlap with other sets, followed by *hg38.Boyle.hg38-*

[5] https://github.com/marbl/CHM13

[6] Example of BEDbase overview screen for hg38.Kundaje.GRCh38_unified_blacklist: http://bedbase.org/#/bedsplash/1a561729234c2844303a051b16f66656

Excludable.v2. We also observed *hg38.Wold.hg38mitoExcludable* and *hg38.Lareau.hg38.full.Excludable* sets overlapping *hg38.Kundaje.GRCh38_unified_Excludable*, suggesting it contains NUMTs (Supplementary Figure S1D). Because of its agreement with other sets, we recommend *hg38.Kundaje.GRCh38_unified_Excludable* set and list other recommended sets Table 1.

Table 1. Characteristics of recommended exclusion sets for human and mouse genome assemblies. Unless specified otherwise, exclusion sets were defined by the Boyle-Lab/Blacklist software. The complete list is provided in Supplementary Table S1.

Name	Assembly	Number of regions	Width, min/median/max, bp	Percent of the genome, %	Year last updated
T2T.excluderanges	T2T	2066	1001/9701/25738901	8.358	2022
hg38.Kundaje.GRCh38_unified_Excludable[7]	hg38	910	19/384/5407756	2.317	2020
hg38.Boyle.hg38-Excludable.v2	hg38	636	1200/10150/30590100	7.355	2018
hg38.Wimberley.peakPass60Perc_sorted[8]	hg38	5078	1000/2000/1852000	2.387	2021
hg19.Boyle.hg19-Excludable.v2	hg19	834	1100/9350/30590100	8.882	2018
mm39.excluderanges	mm39	3147	1100/12500/5487000	6.272	2022
mm10.Boyle.mm10-Excludable.v2	mm10	3435	1000/8100/50585400	8.768	2018

Discussion

Limited annotation remains the main problem when selecting exclusion sets as it remains unclear which method and/or data were used. Examples include Wold's lab-generated "mitoblack" sets for mm9 and mm10 assemblies. Their curation method is unknown, and the exact number (123 regions), width distribution, and other characteristics suggest that one may be a liftOver version of the other. Similarly, it remains unknown why Bernstein's lab-generated "Mint_Blacklist" hg19 and hg38 exclusion sets have a very large number of regions (9,035 and 12,052, respectively) as compared with under 1,000 regions for other exclusion sets. Additionally, hg19 and hg38 "full.blacklist" sets were generated by Caleb Lareau as a combination of NUMTs and unknown ENCODE exclusion sets, the source of which we were unable to infer. Given annotation shortcomings, we recommend using assembly-specific

[7] Defined as a combination of *hg38.Lareau.hg38_peaks*, *hg38.Boyle.hg38-Excludable.v2*, and *hg38.Wimberley.peakPass60Perc_sorted*, followed by manual curation, https://www.encodeproject.org/files/ENCFF356LFX/

[8] Defined by the PeakPass software, https://github.com/ewimberley/peakPass/raw/main/excludedlists/

exclusion sets generated by a published method and, if relevant, combining them with other problematic region sets.

Most annotated exclusion sets were created via Blacklist, a tool for detecting regions with abnormally high signal and/or low mappability (*1*). These genomic properties are commonly accepted as problematic; however, they may not be exhaustive. The Peakpass algorithm was developed to learn genomic properties associated with problematic regions using a random forest model (*5*). It reported distance to nearest assembly gap or gene, and frequency of unique 4-mers or softmasked base pairs, as the most predictive of problematic regions. A limitation of Peakpass is that its extensive collection of Python, R, and bash scripts is poorly documented. A limitation of Blacklist, on the other hand, is computational resource requirements (64+ GB; CPU: 24+ cores, 3.4+ GHz/core) and disk storage (~ 1TB) due to a large number of required BAM files (hundreds). A recent preprint introduced the Greenscreen pipeline, a promising tool for identifying exclusion sets using as few as three ChIP-seq data. It reports a 99.9% overlap with a Blacklist-generated exclusion set, identical performance on ChIP-seq quality metrics but a smaller genome footprint (*13*). We utilized Blacklist as the most well-known tool to generate exclusion sets for the T2T-CHM13 and GRCm39 genome assemblies. The aforementioned tools detect problematic regions in ChIP-seq data; however, they may be different in data generated by other technologies due to different biochemical procedures (*14*). Additional collaborative efforts are needed to develop a consensus approach for defining well-documented exclusion sets.

Acknowledgements

We thank Tim Triche and Stuart Lee for the helpful feedback and suggestions.

Funding

This work was supported in part by the George and Lavinia Blick Research Scholarship to M.G.D., the Essential Open Source Software (EOSS) award from the Chan Zuckerberg Initiative (CZI) to M.I.L., the National Institutes of Health [R35-GM128645 to D.H.P.].

References

1. H. M. Amemiya, A. Kundaje, A. P. Boyle, The ENCODE blacklist: Identification of problematic regions of the genome. *Sci Rep*. **9**, 9354 (2019).

2. K. H. Miga, C. Eisenhart, W. J. Kent, Utilizing mapping targets of sequences underrepresented in the reference assembly to reduce false positive alignments. *Nucleic Acids Res*. **43**, e133 (2015).

3. J. K. Pickrell, D. J. Gaffney, Y. Gilad, J. K. Pritchard, False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics*. **27**, 2144–6 (2011).

4. P. V. Kharchenko, M. Y. Tolstorukov, P. J. Park, Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol*. **26**, 1351–9 (2008).

5. C. E. Wimberley, S. Heber, PeakPass: Automating ChIP-seq blacklist creation. *J Comput Biol* (2019), doi:10.1089/cmb.2019.0295.

6. T. S. Carroll, Z. Liang, R. Salama, R. Stark, I. de Santiago, Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. *Front Genet*. **5**, 75 (2014).

7. H. Qu, F. Ma, Q. Li, Comparative analysis of mitochondrial fragments transferred to the nucleus in vertebrate. *J Genet Genomics*. **35**, 485–90 (2008).

8. L. Montefiori, L. Hernandez, Z. Zhang, Y. Gilad, C. Ober, G. Crawford, M. Nobrega, N. Jo Sakabe, Reducing mitochondrial reads in ATAC-seq using CRISPR/Cas9. *Sci Rep*. **7**, 2451 (2017).

9. D. Simone, F. M. Calabrese, M. Lang, G. Gasparre, M. Attimonelli, The reference human nuclear mitochondrial sequences compilation validated and implemented on the UCSC genome browser. *BMC Genomics*. **12**, 517 (2011).

10. F. M. Calabrese, D. Simone, M. Attimonelli, Primates and mouse NumtS in the UCSC genome browser. *BMC Bioinformatics*. **13 Suppl 4**, S15 (2012).

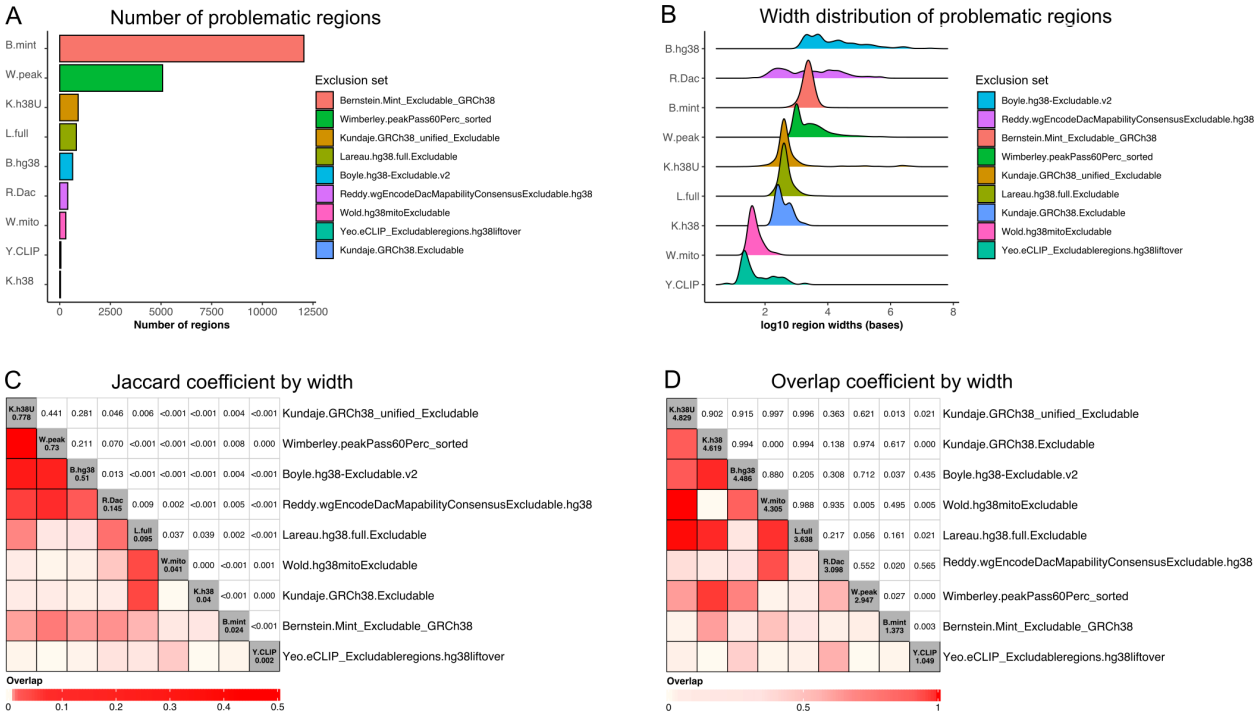
11. S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, M. R. Vollger, N. Altemose, L. Uralsky, A. Gershman, S. Aganezov, S. J. Hoyt, M. Diekhans, G. A. Logsdon, M. Alonge, S. E. Antonarakis, M. Borchers, G. G. Bouffard, S. Y. Brooks, G. V. Caldas, N.-C. Chen, H. Cheng, C.-S. Chin, W. Chow, L. G. de Lima, P. C. Dishuck, R. Durbin, T. Dvorkina, I. T. Fiddes, G. Formenti, R. S. Fulton, A. Fungtammasan, E. Garrison, P. G. S. Grady, T. A. Graves-Lindsay, I. M. Hall, N. F. Hansen, G. A. Hartley, M. Haukness, K. Howe, M. W. Hunkapiller, C. Jain, M. Jain, E. D. Jarvis, P. Kerpedjiev, M. Kirsche, M. Kolmogorov, J. Korlach, M. Kremitzki, H. Li, V. V. Maduro, T. Marschall, A. M. McCartney, J. McDaniel, D. E. Miller, J. C. Mullikin, E. W. Myers, N. D. Olson, B. Paten, P. Peluso, P. A. Pevzner, D. Porubsky, T. Potapova, E. I. Rogaev, J. A. Rosenfeld, S. L. Salzberg, V. A. Schneider, F. J. Sedlazeck, K. Shafin, C. J. Shew, A. Shumate, Y. Sims, A. F. A. Smit, D. C. Soto, I. Sović, J. M. Storer, A. Streets, B. A. Sullivan, F. Thibaud-Nissen, J. Torrance, J. Wagner, B. P. Walenz, A. Wenger, J. M. D. Wood, C. Xiao, S. M. Yan, A. C. Young, S. Zarate, U. Surti, R. C. McCoy, M. Y. Dennis, I. A. Alexandrov, J. L. Gerton, R. J. O'Neill, W. Timp, J. M. Zook, M. C. Schatz, E. E. Eichler, K. H. Miga, A. M. Phillippy, The complete sequence of a human genome. *Science*. **376**, 44–53 (2022).

12. J. J. Yunis, W. G. Yasmineh, Heterochromatin, satellite DNA, and cell function. Structural DNA of eucaryotes may support and protect genes and aid in speciation. *Science*. **174**, 1200–9 (1971).

13. S. Klasfeld, D. Wagner, Greenscreen decreases type I errors and increases true peak detection in genomic datasets including ChIP-seq. *bioRxiv* (2022).

14. A. Nordin, G. Zambanini, P. Pagella, C. Cantu, The CUT&RUN blacklist of problematic regions of the genome. *bioRxiv* (2022).

Supplementary Figure S1. Characteristics of hg38 exclusion sets. (A) Number and (B) width distribution of problematic regions in hg38-specific exclusion sets. (C) Jaccard overlap $J(A,B) = \frac{width(\cap_{A,B})}{width(\cup_{A,B})}$ and (D) overlap coefficient $C(A,B) = \frac{width(\cap_{A,B})}{Min(width(A),width(B))}$ among hg38 exclusion sets by width. Diagonal counts represent sum of overlap coefficients of a list with all others.



Supplementary Table S1. Characteristics of exclusion sets. "AHub IDs" - AnnotationHub IDs for objects in Bioconductor version 3.16 and above; "Original/Filtered regions" - the number of regions in the original set and in the subset to the assembled (autosomal) chromosomes; "ID/URL" - ENCODE ID or URL for data download. "BEDbase ID" - unique identifies for BEDbase.org API access, "Ahub IDs BioC 3.15 and 3.14" - AnnotationHub IDs for objects in Bioconductor version 3.14 and 3.15.

Name	Assembly	Description	AHub IDs BioC 3.16 and above	Original Region count	Filtered Region count	Missing Chromosomes	Width, min/median/max, bp	Percent of the genome, %	Year last updated	Source	ID/URL	BEDbase ID	AHub IDs BioC 3.15 and 3.14
T2T.excluderanges	T2T-CHM13	Defined by the Boyle-Lab/Blacklist software, High Signal and L	AH107304	2066	2066	chrY, chrMT	1001/9701/25738901	8.3582	2022	excluderange	excluderanges	8329d8c624880308ab51ba05149a737d	NA
hg38.Kundaje.GRCh38_unified_Excludable	hg38	Defined as a combination of hg38.Lareau.hg38_peaks, hg38.Bc	AH107305	910	910	chrM	20/385/5407757	2.3175	2020	ENCODE	ENCFF356LFX	1a561729234c2844303a051b16f66656	AH95917
hg38.Bernstein.Mint_Excludable_GRCh38	hg38	Defined from Mint-ChIP (low input, multiplexed ChIP-seq) data	AH107306	12052	12052	chrM	502/2365/46435	0.9786	2019	ENCODE	ENCFF023CZC	80e335903b77b597b8245f9817fcd9cd	AH95915
hg38.Boyle.hg38-Excludable.v2	hg38	Defined by the Boyle-Lab/Blacklist software, High Signal and L	AH107307	636	636	chrM	1201/10151/30590101	7.3557	2018	GitHub	https://github.com/	ac58962c9ec98fe9258c12092a0c8832	NA
hg38.Kundaje.GRCh38.Excludable	hg38	Defined by Anshul Kundaje as a part of ENCODE and modENCO	AH107308	38	38	chr6, chr7, ch	221/301/1761	0.0012	2016	ENCODE	ENCFF419RSJ	cb701496bde7eeb18add96fdbc3b8b11	AH95916
hg38.Lareau.hg38.full.Excludable	hg38	ENCODE excludable regions combined with regions of high hor	AH107309	820	820	chrY, chrM	201/384/9421	0.0144	2017	GitHub	https://github.com/	5a12c1de138ace1a73a45e6faf9ba669	NA
hg38.Reddy.wgEncodeDacMapabilityConsensusExclu	hg38	Defined by the ENCODE consortium, includes satellite repeats	AH107310	401	396	NA	42/2520/618655	0.3182	2016	ENCODE	ENCFF220FIN	148622e896f6798f7c4abf448bab67c4	AH95918
hg38.Wimberley.peakPass60Perc_sorted	hg38	Defined by the ewimberley/peakPass software	AH107311	5078	5078	chrM	1001/2001/1852001	2.3875	2021	GitHub	https://github.com/	f4a9bb19ed29e993592813e970e7dd90	NA
hg38.Wold.hg38mitoExcludable	hg38	Definition method unknown	AH107312	299	299	chr10, chr15,	31/40/295	6.00E-04	2016	ENCODE	ENCFF940NTE	a714dcba99821801b5c426fba9c80988	AH95919
hg38.Yeo.eCLIP_Excludableregions.hg38liftover.bed.f	hg38	Defined from eCLIP data	AH107313	56	56	chr18, chr21,	5/30/1850	3.00E-04	2019	ENCODE	ENCFF269URO	1a02a65fafefefd65ff4a060273304ed	AH95920
hg38.Nordin.CandRblacklist_hg38	hg38	Defined from CUT&RUN negative controls as 0.1% top significa	NA	1049	885	NA	3/2880/93435	0.1451	2022	Publication	https://www.biorxi	NA	NA
hg19.Boyle.hg19-Excludable.v2	hg19	Defined by the Boyle-Lab/Blacklist software, High Signal and L	AH107314	834	834	chrM	1101/9351/30590101	8.8824	2018	GitHub	https://github.com/	6eb180d456f2f3b71b419e5fab107fc9	NA
hg19.Bernstein.Mint_Excludable_hg19	hg19	Defined from Mint-ChIP (low input, multiplexed ChIP-seq) data	AH107315	9035	9035	chrX, chrY, ch	502/2418/49368	0.8111	2019	ENCODE	ENCFF200UUD	d1a6047ed5bec84acefe9c52cf63b593	AH95910
hg19.Birney.wgEncodeDacMapabilityConsensusExclu	hg19	Defined by the ENCODE consortium, includes satellite repeats	AH107316	411	411	NA	42/2567/1400396	0.3743	2011	ENCODE	ENCFF001TDO	5b6b19dea85a8bc6007ef07a0960267b	AH95911
hg19.Crawford.wgEncodeDukeMapabilityRegionsExc	hg19	Defined by the ENCODE consortium, includes satellite repeats	AH107317	1649	1566	NA	21/553/160603	0.3269	2011	ENCODE	ENCFF001THR	dac2eda4e8687eb039611ac6cd595821	AH95912
hg19.Lareau.hg19.full.Excludable	hg19	ENCODE excludable regions combined with regions of high hor	AH107318	902	902	chrM	91/388/1400396	0.3424	2017	GitHub	https://github.com/	d934d47e8035da9c5a1767c8153db4cc	NA
hg19.Wold.hg19mitoExcludable	hg19	Definition method unknown	AH107319	295	295	chr10, chr15,	31/41/301	6.00E-04	2016	ENCODE	ENCFF055QTV	182046a0f055b0176178241a95cbd637	AH95913
hg19.Yeo.eCLIP_Excludableregions.hg19	hg19	Defined from eCLIP data, includes skyscraper, rRNA pseudogen	AH107320	57	57	chr18, chr21,	5/30/1850	3.00E-04	2019	ENCODE	ENCFF039QTN	350f49dc47e5307109e1e17d60223a31	AH95914
mm39.excluderanges	mm39	Defined by the Boyle-Lab/Blacklist software, High Signal and L	AH107321	3147	3147	chrY, chrM	1101/12501/5487001	6.2721	2022	excluderange	excluderanges	edc716833d4b5ee75c34a0692fc353d5	NA
mm10.Boyle.mm10-Excludable.v2	mm10	Defined by the Boyle-Lab/Blacklist software, High Signal and L	AH107322	3435	3435	chrM	1001/8101/50585401	8.7683	2018	GitHub	https://github.com/	a5311e39fe1590de66c1df6a5881a942	NA
mm10.Hardison.Excludable.full	mm10	Definition method unknown	AH107323	7865	7865	NA	10/1301/220008	0.9546	2016	ENCODE	ENCFF790DJT	087541f51cf8c7d7078995d1bd95fd27	AH95921
mm10.Hardison.psuExcludable.mm10	mm10	Definition method unknown	AH107324	5552	5552	NA	3/529/220008	0.7337	2016	ENCODE	ENCFF226BDM	fc6b88f936c5cd880545943708e4c2af	AH95922
mm10.Kundaje.anshul.Excludable.mm10	mm10	Defined by Anshul Kundaje as a part of ENCODE and modENCO	AH107325	3010	3010	chrM	1001/1501/121601	0.3125	2016	ENCODE	ENCFF999QPV	e6a89a8432f4a69bae41f60ed0c7e704	AH95923
mm10.Kundaje.mm10.Excludable	mm10	Defined by Anshul Kundaje as a part of ENCODE and modENCO	AH107326	164	164	chrX, chrY, ch	161/241/4331	0.0033	2016	ENCODE	ENCFF547MET	76c03b6c831f8fecdf4fee7adf2def6a	AH95924
mm10.Lareau.mm10.full.Excludable	mm10	ENCODE excludable regions combined with regions of high hor	AH107327	523	523	chrY, chrM	161/381/13031	0.0095	2017	GitHub	https://github.com/	1bd30517be79d4d051308c693b822798	NA
mm10.Wold.mm10mitoExcludable	mm10	Definition method unknown	AH107328	123	123	chr7, chr14,	31/40/3068	6.00E-04	2016	ENCODE	ENCFF759PJK	830f1ffd31689e3e7c22ff856f0ba02c	AH95925
mm10.Nordin.CandRblacklist_mm10	mm10	Defined from CUT&RUN negative controls as 0.1% top significa	NA	559	559	NA	5/2648/82820	0.1025	2022	Publication	https://www.biorxi	NA	NA
mm9.Lareau.mm9.full.Excludable	mm9	ENCODE excludable regions combined with regions of high hor	AH107329	3415	3415	chrM	201/1401/121601	0.3272	2017	GitHub	https://github.com/	e903b285baefce8167367ce57a8c3d48	NA
mm9.Wold.mm9mitoExcludable	mm9	Definition method unknown	AH107330	123	123	chr7, chr14,	31/40/3068	6.00E-04	2016	ENCODE	ENCFF299EZH	9b4389a6a4b937df8abd62dad30fa3a3	AH95926
ce11.Boyle.ce11-Excludable.v2	ce11	Defined by the Boyle-Lab/Blacklist software, High Signal and L	AH107331	97	97	chrM	1301/5001/47501	0.7266	2018	GitHub	https://github.com/	7235114a78b1709be96f0d6a82b4ea36	NA
ce10.Boyle.ce10-Excludable.v2	ce10	Defined by the Boyle-Lab/Blacklist software, High Signal and L	AH107332	100	100	chrM	1301/5401/1130801	2.1993	2018	GitHub	https://github.com/	6de11bb5f50ee015b23ac96f433f00bb	NA
ce10.Kundaje.ce10-Excludable	ce10	Defined by Anshul Kundaje, superseded by ce10.Boyle.ce10-Exc	AH107333	122	122	chrM	1001/2201/25801	0.3937	2012	Stanford.edu	http://mitra.stanfor	32b59590fa83161687cec4cabfa2bb2b	AH95908
danRer10.Domingues.Excludableed	danRer10	Defined manually using total RNA-seq.	AH107334	62	57	chr3, chr6, ch	37/481/82628	0.0731	2020	GitHub	https://github.com/	a0a94af275f858d63550005627d260b7	NA
danRer10.Yang.Supplemental_Table_19.ChIP-seq_bl	danRer10	Defined via MACS2 peak calling using ChIP-seq (PMID: 332397	AH107335	853	853	chrM	410/1170/6033	0.0774	2020	Publication	https://www.ncbi.n	78f5eb585019a4d795ef80159a597b15	NA
dm6.Boyle.dm6-Excludable.v2	dm6	Defined by the Boyle-Lab/Blacklist software, High Signal and L	AH107336	182	182	chrM	1201/7401/236601	2.7194	2018	GitHub	https://github.com/	24186dc2aac492074d3de9caede730a0	NA
dm3.Boyle.dm3-Excludable.v2	dm3	Defined by the Boyle-Lab/Blacklist software, High Signal and L	AH107337	271	248	chrM	1401/5701/127701	1.7485	2018	GitHub	https://github.com/	7427399e18d9c01e423b2f4963b409ea	NA
dm3.Kundaje.dm3-Excludable	dm3	Defined by Anshul Kundaje. Contains heterochromatin chromos	AH107338	492	306	chrM	1001/1851/24301	0.6889	2012	Stanford.edu	http://mitra.stanfor	0801a522159f7ebf2f669d8cade4aa8f	AH95909
TAIR10.Wimberley.predicted_excluded_list_sorted_(	TAIR10	Defined by the ewimberley/peakPass software	AH107339	887	887	chrMT, chrPl	501/1001/60001	2.0944	2021	GitHub	https://github.com/	6f3a3ae3ee878b88a92093eb8e3fe982	NA
TAIR10.Klasfeld.arabidopsis_Excludable_20inputs	TAIR10	Defined by the Boyle-Lab/Blacklist software, High Signal and L	AH107340	83	83	chrMT, chrPl	1301/14601/308301	2.3959	2021	GitHub	https://github.com/	aa1c99c2dd2aef874486b1c0c3bf6b92	NA
TAIR10.Klasfeld.arabidopsis_greenscreen_20inputs	TAIR10	Defined by the green screen pipeline (DOI: 10.1101/2022.02.27	AH107341	36	36	chrMT, chrPl	121/7506/80842	0.4069	2021	GitHub	https://github.com/	e5d66ee787a8cb0c76438bba768c2331	NA
T2T.Lareau.chm13v2.0_peaks	T2T-CHM13	Regions of high homology to mtDNA (NUMT regions) defined l	AH107342	817	817	chrMT	201/384/9422	0.0138	2022	GitHub	https://github.com/	354dfced295f54f70ae9656ca8f9b141	NA
hg38.Lareau.hg38_peaks	hg38	Regions of high homology to mtDNA (NUMT regions) defined l	AH107343	784	784	chrY, chrM	201/385/9421	0.0139	2017	GitHub	https://github.com/	9fa55701a3bd3e7a598d1d2815e3390f	NA
hg19.Lareau.hg19_peaks	hg19	Regions of high homology to mtDNA (NUMT regions) defined l	AH107344	779	779	chrY, chrM	201/384/9422	0.0137	2017	GitHub	https://github.com/	79e924141251afbd4cde0c38456913fd	NA
mm10.Lareau.mm10_peaks	mm10	Regions of high homology to mtDNA (NUMT regions) defined l	AH107345	387	387	chrY, chrM	201/381/5011	0.0064	2017	GitHub	https://github.com/	1b76ab775549e116da5e1a89aad7019b	NA
mm9.Lareau.mm9_peaks	mm9	Regions of high homology to mtDNA (NUMT regions) defined l	AH107346	395	395	chrY, chrM	201/381/5011	0.0065	2017	GitHub	https://github.com/	5c4b1cb28175b72bc56adb0bd7384dfd	NA
hg19.UCSC.numtS	hg19	Human NumtS mitochondrial sequence	AH107347	766	766	chrM, chrMT	12/212/14835	0.0175	2011	UCSC	numtS	cc4fd05fdfe015e4acd5111dac5b372f	NA
mm9.UCSC.numtS	mm9	Mouse NumtS mitochondrial sequence	AH107348	172	172	chr19, chrY,	33/196/4654	0.0023	2011	UCSC	numtS	29dc50750f0535b6b9c746ee8371c211	NA
T2T.CHM13.chm13.draft_v2.0.cen_mask	T2T-CHM13	Centromeric satellite masking bed file (v2.0)	AH107349	23	23	chrY, chrM	2081535/5479655/3175	6.4944	2022	CHM13	https://s3-us-west-	44138ebb0d3340e70164d12649a47dc8	NA
T2T.CHM13.chm13.draft_v1.1.telomere	T2T-CHM13	Telomere identified by the VGP pipeline (v1.1)	AH107350	48	48	chrM	1001/2964/4749	0.0045	2022	CHM13	https://s3-us-west-	b72dd2fa5f8a916cc36960b93169c743	NA
T2T.UCSC.censat	T2T-CHM13	T2T peri/centromeric satellite annotation (v2.0, 20220329, CH	AH107351	2523	2523	chrM	2/17108/27638497	14.4957	2022	UCSChub	https://hgdownloac	f28798df2c4d72810e7c4626b5a62106	NA
T2T.UCSC.gap	T2T-CHM13	Locations of assembly gaps, as determine by strings of 'N' cha	AH107352	5	5	chr1, chr2, ch	675001/2700001/40500	0.3754	2021	UCSChub	http://t2t.gi.ucsc.ec	0747aae5f4cac92367a16c3eb1c7f3f1	NA
T2T.UCSC.hgUnique.hg38	T2T-CHM13	Regions unique to the T2T-CHM13 v2.0 assembly compared to	AH107353	615	615	chrM	2/15829/29694330	8.0625	2022	UCSChub	https://hgdownloac	c3839f43c53a3c47733388528b853690	NA
hg38.UCSC.centromere	hg38	Gaps from centromeres	AH107354	109	109	chrM	341/76959/4763585	1.9282	2014	UCSC	centromeres	0b1f161675fa0f52ac6d0d4f54b1efb9	NA
hg38.UCSC.telomere	hg38	Gaps from telomeres	AH107355	48	48	chrM	10000/10000/10000	0.0155	2018	UCSC	gap	79f964e68d5daa1462c52ca54855b06a	AH95938
hg38.UCSC.short_arm	hg38	Gaps on the short arm of the chromosome	AH107356	5	5	chr1, chr2, ch	5000000/15990000/169	2.0876	2018	UCSC	gap	92fc8f64f92d525c6b92c9aab5e2c711	AH95937
hg38.UCSC.heterochromatin	hg38	Gaps from large blocks of heterochromatin	AH107357	11	11	chr3, chr4, ch	20000/207000/3000000	2.3452	2018	UCSC	gap	8af7b48ab48183229d3bc72005040dc1	AH95935
hg38.UCSC.contig	hg38	Gaps between contigs in scaffolds	AH107358	285	285	chrM	100/50000/400000	0.3309	2018	UCSC	gap	2dd1b22f2add15bc7508580d18bc9495	AH95934
hg38.UCSC.scaffold	hg38	Gaps between scaffolds in chromosome assemblies. Has extra	AH107359	478	254	chr8, chrM	10/796/180000	0.0976	2018	UCSC	gap	de0c7f42f29fb83ac393e86a2ec28374	AH95936
hg19.UCSC.centromere	hg19	Gaps from centromeres	AH107360	24	24	chrM, chrMT	3000000/3000000/3000	2.3258	2020	UCSC	gap	26ecf1381b6323791656f800ad39b69c	AH95927
hg19.UCSC.telomere	hg19	Gaps from telomeres	AH107361	46	46	chr17, chrM,	10000/10000/10000	0.0149	2020	UCSC	gap	2bcad8794847411e9b3f52ff39c4f377	AH95933
hg19.UCSC.short_arm	hg19	Gaps on the short arm of the chromosome	AH107362	5	5	chr1, chr2, ch	5201193/15990000/169	2.1695	2020	UCSC	gap	e09fac8aedf1230ab77ac4194fd75784	AH95932
hg19.UCSC.heterochromatin	hg19	Gaps from large blocks of heterochromatin	AH107363	12	12	chr3, chr4, ch	20000/128500/3000000	2.3412	2020	UCSC	gap	8ea9b6cdfe68a4b4111e5b03157af371	AH95930
hg19.UCSC.clone	hg19	Gaps between clones in the same map contig. Has extra chron	AH107364	207	107	chr13, chr14,	40442/50000/486181	0.1979	2020	UCSC	gap	4f3b1098a0f4ea5e81747f4414a8d294	AH95928
hg19.UCSC.contig	hg19	Gaps between contigs in scaffolds	AH107365	163	163	chrM, chrMT	700/50000/4200000	0.5207	2020	UCSC	gap	a4da41916b0b213d4e3b89f5ab20e1e8	AH95929
mm39.UCSC.centromere	mm39	Gaps from centromeres	AH107366	20	20	chrY, chrM	2890000/2890000/2890	2.1223	2020	UCSC	gap	1aa3f73ffa8e6d498f0f3f22e0302472	NA
mm39.UCSC.telomere	mm39	Gaps from telomeres	AH107367	42	42	chrM	100000/100000/100000	0.1542	2020	UCSC	gap	883bdae38244c6f0e0facfbd4fcc601b	NA
mm39.UCSC.short_arm	mm39	Gaps on the short arm of the chromosome	AH107368	21	21	chrM	10000/10000/10000	0.0077	2020	UCSC	gap	f61de62eae0898943e6c9d163b0a3989	NA
mm39.UCSC.contig	mm39	Gaps between contigs in scaffolds	AH107369	60	60	chr3, chr18,	8000/50000/500000	0.1567	2020	UCSC	gap	732437d9fcb8992b2e5c6513bfed2586	NA
mm39.UCSC.scaffold	mm39	Gaps between scaffolds in chromosome assemblies	AH107370	181	115	chr1, chr15,	27/50000/522000	0.2575	2020	UCSC	gap	97e738326c0681f5ebe94b0d28d058c5	AH95939
mm10.UCSC.centromere	mm10	Gaps from centromeres	AH107371	20	20	chrY, chrM	2890000/2890000/2890	2.1207	2021	UCSC	gap	b0f9aa3cc8a4a43f59b463891b5d12c8	AH95945
mm10.UCSC.telomere	mm10	Gaps from telomeres	AH107372	42	42	chrM	100000/100000/100000	0.1541	2021	UCSC	gap	051090e82c227bcc55dba3e953bc6daa	AH95944
mm10.UCSC.short_arm	mm10	Gaps on the short arm of the chromosome	AH107373	21	21	chrM	10000/10000/10000	0.0077	2021	UCSC	gap	f4d2d6fe334deca5800ca9ae39ce95ce	AH95940
mm10.UCSC.clone	mm10	Gaps between clones in the same map contig. Has extra chron	AH107374	114	4	chr1, chr3, ch	50000/50000/50000	0.0073	2021	UCSC	gap	9ecfce46335e4d5b3a1230b69690a25a	AH95941
mm10.UCSC.contig	mm10	Gaps between contigs in scaffolds	AH107375	104	104	chrM	717/55000/800000	0.3483	2021	UCSC	gap	82d2374cf5524a2b13dcf9c3dc487d6f	AH95943
mm10.UCSC.other	mm10	Sequence of Ns in the assembly that were not marked as gaps	AH107376	384	383	chrM	1/100/300000	0.2236	2021	UCSC	gap	75662812e5eb228b25c9ae5a28fbb402	AH95948
mm9.UCSC.centromere	mm9	Gaps from centromeres	AH107377	21	21	chrM	3000000/3000000/3000	2.373	2007	UCSC	gap	99e4d2c9a794d321bfcf01709787caac	NA
mm9.UCSC.fragment	mm9	Gaps between the contigs of a draft clone. (In this context, a c	AH107378	709	436	chr6, chr15,	100/100/222253	0.127	2007	UCSC	gap	ecf3f802759c5dc93f1446b6942c58b3	NA
mm9.UCSC.contig	mm9	Gaps between contigs in scaffolds. Has extra chromosomes, us	AH107379	281	105	chrM	1700/50000/10000000	1.1305	2007	UCSC	gap	4dd8bb54f6432144c045619337d8212e	NA
danRer10.UCSC.contig	danRer10	Gaps between contigs in scaffolds	AH107380	2338	2338	chrM	100/100/100	0.0174	2015	UCSC	gap	5d41a9fa328769b63734e11e6ae4252b	NA
danRer10.UCSC.scaffold	danRer10	Gaps between scaffolds in chromosome assemblies	AH107381	18955	16496	chrM	10/100/100	0.1161	2015	UCSC	gap	a5feefb2d573d265f1085043c208c2ed	NA
dm6.UCSC.other	dm6	Sequence of Ns in the assembly that were not marked as gaps	AH107382	572	268	chrM	13/100/53860	0.3565	2014	UCSC	gap	11a80264dcdad6c0868ea48637a799df	NA
dm3.UCSC.contig	dm3	Gaps between contigs in scaffolds	AH107383	37665	7	chr3R, chrM	100/100/18000	0.0154	2006	UCSC	gap	f9822a0f88047e92cf92824fe025b2f2	NA
dm3.UCSC.scaffold	dm3	Gaps between scaffolds in chromosome assemblies	AH107384	8	1	chr2L, chr2R,	72000/72000/72000	0.0598	2006	UCSC	gap	d4e31a2c488de8ff335b0cb779c9cef5	NA
TAIR10.UCSC.araTha1.gap	TAIR10	Gaps in the May 2011 Arabidopsis thaliana genome assembly	AH107385	357	357	chrMT, chrPl	1/2/53000	0.1556	2013	UCSChub	https://genome-tes	74585119b9b90d3b4ad077b10b487d39	NA

Supplementary Table S2. Gap overlap statistics for human and mouse exclusion sets. "% centromeres/short arms/telomeres covered" - proportion of gap regions covered by the corresponding exclusion set. "% regions intersecting centromeres/short arms/telomeres" - proportion of exclusion regions from a set covering gaps (number of overlapping regions over total).

Name	% centromeres covered	% regions intersecting centromeres	% short arms covered	% regions intersecting short arms	% telomeres covered	% regions intersecting telomeres
T2T.excluderanges	93.58	2.27% (47/2066)	74.13	1.79% (37/2066)	94.59	2.18% (45/2066)
hg38.Bernstein.Mint_Excludable_GRCh38	< 1.0	0.04% (5/12052)	0	0% (0/12052)	0	0% (0/12052)
hg38.Boyle.hg38-Excludable.v2	99.08	4.56% (29/636)	58.91	0.47% (3/636)	72.72	5.5% (35/636)
hg38.Kundaje.GRCh38_unified_Excludable	99.4	2.97% (27/910)	0	0% (0/910)	< 1.0	0.11% (1/910)
hg38.Kundaje.GRCh38.Excludable	< 1.0	7.89% (3/38)	0	0% (0/38)	0	0% (0/38)
hg38.Lareau.hg38.full.Excludable	< 1.0	0.37% (3/820)	0	0% (0/820)	0	0% (0/820)
hg38.Reddy.wgEncodeDacMapabilityConsensusExcludable.hg38	0	0% (0/396)	0	0% (0/396)	< 1.0	0.25% (1/396)
hg38.Wimberley.peakPass60Perc_sorted	59.6	26.98% (1370/5078)	< 1.0	0.08% (4/5078)	< 1.0	0.41% (21/5078)
hg38.Wold.hg38mitoExcludable	0	0% (0/299)	0	0% (0/299)	0	0% (0/299)
hg38.Yeo.eCLIP_Excludableregions.hg38liftover	0	0% (0/56)	0	0% (0/56)	0	0% (0/56)
hg38.Nordin.CandRblacklist_hg38	4.13	60.34% (534/885)	0	0% (0/885)	1.42	0.79% (7/885)
hg19.Bernstein.Mint_Excludable_hg19	0	0% (0/9035)	0	0% (0/9035)	0	0% (0/9035)
hg19.Birney.wgEncodeDacMapabilityConsensusExcludable	< 1.0	3.89% (16/411)	0	0% (0/411)	< 1.0	0.24% (1/411)
hg19.Boyle.hg19-Excludable.v2	100	2.88% (24/834)	92.26	0.48% (4/834)	62.85	3.48% (29/834)
hg19.Crawford.wgEncodeDukeMapabilityRegionsExcludable	< 1.0	0.45% (7/1566)	0	0% (0/1566)	0	0% (0/1566)
hg19.Lareau.hg19.full.Excludable	< 1.0	1.77% (16/902)	0	0% (0/902)	< 1.0	0.11% (1/902)
hg19.Wold.hg19mitoExcludable	0	0% (0/295)	0	0% (0/295)	0	0% (0/295)
hg19.Yeo.eCLIP_Excludableregions.hg19	0	0% (0/57)	0	0% (0/57)	0	0% (0/57)
mm39.excluderanges	83.3	1.11% (35/3147)	43.66	0.89% (28/3147)	76.19	0.51% (16/3147)
mm10.Boyle.mm10-Excludable.v2	88.07	1.08% (37/3435)	100	0.12% (4/3435)	76.19	0.47% (16/3435)
mm10.Hardison.Excludable.full	0	0% (0/7865)	0	0% (0/7865)	0	0% (0/7865)
mm10.Hardison.psuExcludable.mm10	0	0% (0/5552)	0	0% (0/5552)	0	0% (0/5552)
mm10.Kundaje.anshul.Excludable.mm10	0	0% (0/3010)	0	0% (0/3010)	0	0% (0/3010)
mm10.Kundaje.mm10.Excludable	0	0% (0/164)	0	0% (0/164)	0	0% (0/164)
mm10.Lareau.mm10.full.Excludable	0	0% (0/523)	0	0% (0/523)	0	0% (0/523)
mm10.Wold.mm10mitoExcludable	0	0% (0/123)	0	0% (0/123)	0	0% (0/123)
mm10.Nordin.CandRblacklist_mm10	< 1.0	1.97% (11/559)	< 1.0	0.36% (2/559)	0	0% (0/559)