A comprehensive collection of signal artifact blacklist regions in the human genome

Anshul Kundaje
akundaje@stanford.edu
Stanford University

Summary

We aim to identify a comprehensive set of regions in the human genome that have anomalous, unstructured, high signal/read counts in next gen sequencing experiments independent of cell line and type of experiment. The breadth of cell-lines covered by the ENCODE datasets allows us to accomplish this in a systematic manner. We also explore the relationship of these empirical signal artifact regions to sequence mappability and known repeat annotations. The complete list of regions and comparisons to repeat annotations can be found in this google spreadsheet . It is important to note that RNA based sequencing experiments were not used to identify these artifact regions and it is unclear how these regions affect mapping and quantification in RNA-seq experiments.

Relevant files and datasets

All the lists as a google doc with links to genome browser shots	https://spreadsheets.google.com/ccc?key=0Am6FxqAtrFDwdE5LYWh2MkVscmtCWEdtNUN2eEVEYmc&hl=en
Merged Consensus blacklist (BED file)	ftp://encodeftp.cse.ucsc.edu/users/akundaje/rawdata/blacklists/hg19/wgEncodeHg19ConsensusSignalArtifactRegions.bed.gz
Ultra-high signal artifacts identified by this pipeline (BED file)	ftp://encodeftp.cse.ucsc.edu/users/akundaje/rawdata/blacklists/hg19/Anshul_Hg19UltraHighSignalArtifactRegions.bed.gz
Terry's blacklist (based on repeat annotations) (BED file)	ftp://encodeftp.cse.ucsc.edu/users/akundaje/rawdata/blacklists/hg19/Duke_Hg19SignalRepeatArtifactRegions.bed.gz

For the FTP site username: encode, password: human

How do we identify these regions?

We first use an automated procedure to identify seed suspect regions in the genome and follow this up with manual curation to reliably collect artifact regions.

We use 80 open chromatin tracks (DNase and FAIRE datasets) and 12 ChIP-seq input/control tracks spanning ~60 cell lines in total.

We use the signal tracks generated by the align2rawsignal tool. The tool computes tag extended **uniquely mapping** read counts and normalizes the signal to obtain a fold change w.r.t. an equivalent uniform distribution of reads in the genome. **It also includes a mappability and stack filter**. So the signal tracks are already partially filtered for artifacts. Thus, our goal is to find artifacts that manage to escape these simple filters. We visualize these signal tracks in the UCSC genome browser (This custom track just shows the relevant Tier 1 tracks. These tracks are infact sufficient to identify the artifacts)

We also use the DNAnexus browser (username: encode, password: encodehuman) to get a different perspective of the tracks. The DNAnexus mapper allows for probabilistics mapping of multimapping reads and the browser includes a **visualization of multimapping reads**. This allows use to compare the relative number of multimapping reads to unique mapping reads in a region. A region that has a very high number of multimapping reads assigned to it is suspect since some the uniquely mappable coordinates might be just outliers and a large number of unique mapping reads might just creep through the mappability filters. The browser also allows for uniformly scaled and compact viewing of a large number of signal tracks that is useful for quick manual curation.

We scan the genome for ~2000 bp windows where

- The maximum signal in the open chromatin tracks (uniquely mapping reads ONLY) > 200 fold
- The maximum signal in the input/control tracks (uniquely mapping reads ONLY) > 30 fold.
- The ration of multimapping to unique mapping reads > 20 fold
- The tag-extended sequence mappability in these regions is NOT uniformly high > 0.4

We call these ***ultra-high signal suspect regions***

We then manually filter and extend, merge and adjust the boundaries of these suspect regions. We also manually scan 1 Mbp regions surrounding these suspect regions for artifacts that might have been missed by the automated procedure. We flag regions as artifacts if they adhere to the following criteria

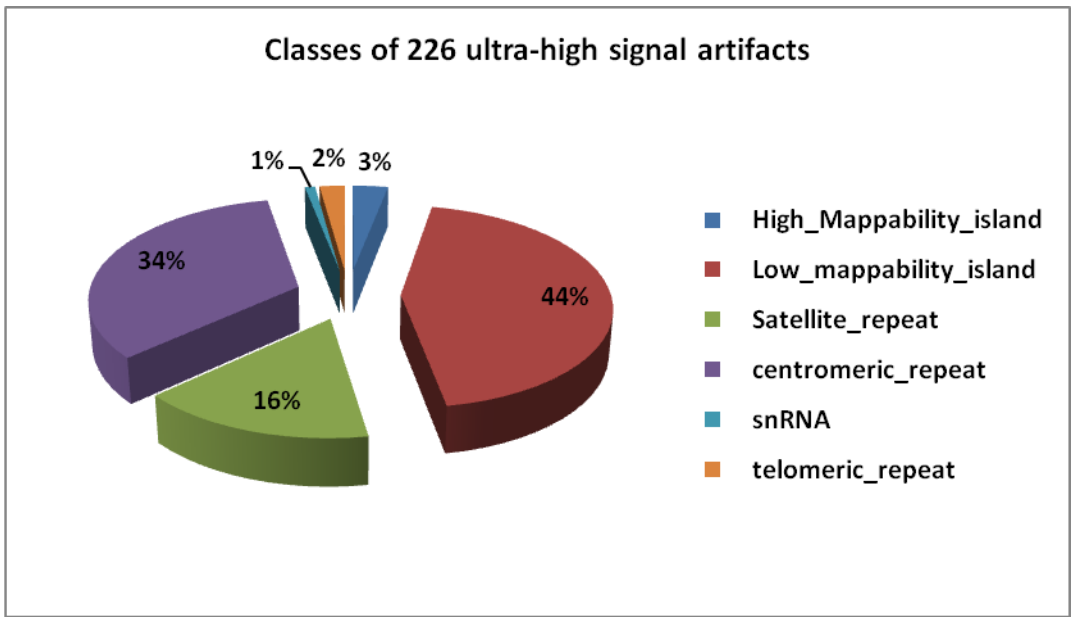
- If the signal artifact **(uniquely mapping)** is extremely severe (> 1000 fold) we flag the region.
- If the signal artifact is present in most of the tracks **independent of cell-line or experiment type** we flag the region.
- If the region has **dispersed high mappability** and low mappability coordinates then it is more likely to be an artifact region.
- If the region has a **known repeat element** then it is more likely to be an artifact
- We check if the stranded **read counts/signal is structured** in the UwDNase, UncFAIRE and input/control datasets i.e. do we see offset mirror peaks on the + and - strand that is typical observed in real, functional peaks. If so we remove these regions from the artifact list.
- If the region exactly overlaps a **known gene's TSS**, or is in the vicinity or within a known gene, it is more likely to be removed from the artifact list. Our intention is to give such regions the benefit of doubt of being real peaks.

We call these ***ultra-high signal artifact regions***

How many regions do we obtain and what are their characteristics?

We converge on a set of **226 regions** in the genome that we call ***ultra-high signal artifact regions***. The list of regions can be found in this google spreadsheet (Sheet: HS-list), with links to the UCSC genome browser custrom track view (you will need to manually paste the coordinates from the google spreadsheet into the browser) and DNAnexus bookmarks (these take you directly to the artifact regions).

We annotate these regions based on overlapping known repeat annotations and/or mappability characteristics. The classes of regions and the number of regions in each class are shown in the figure below.

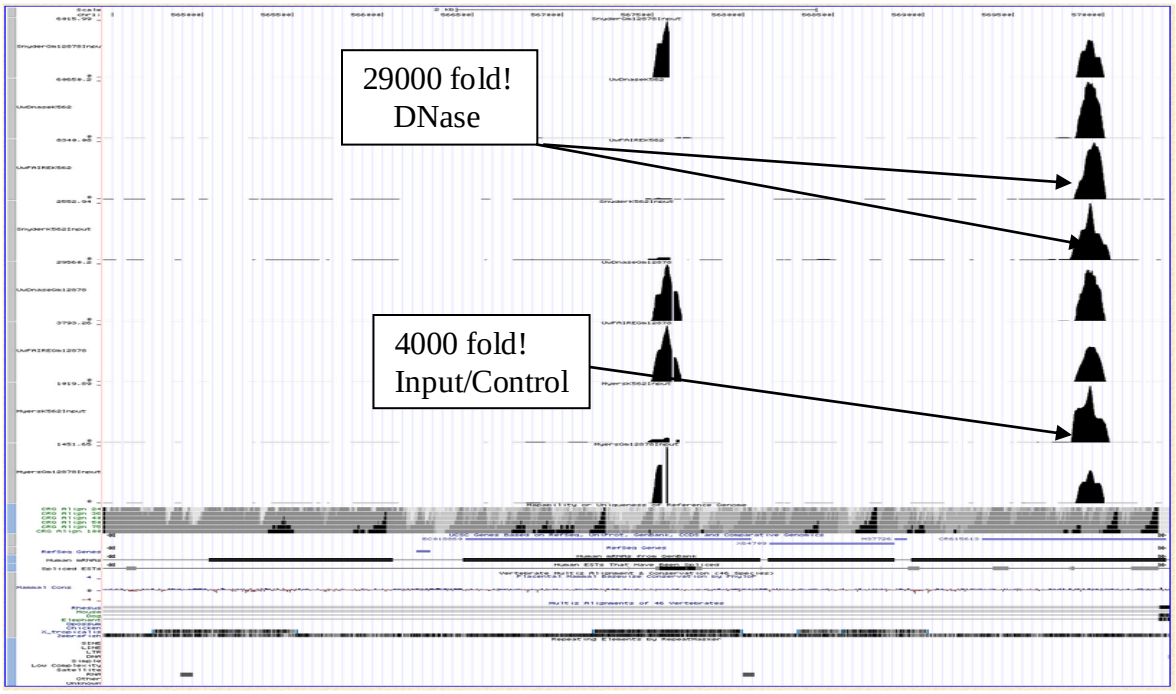


The definitions of the classes are below:

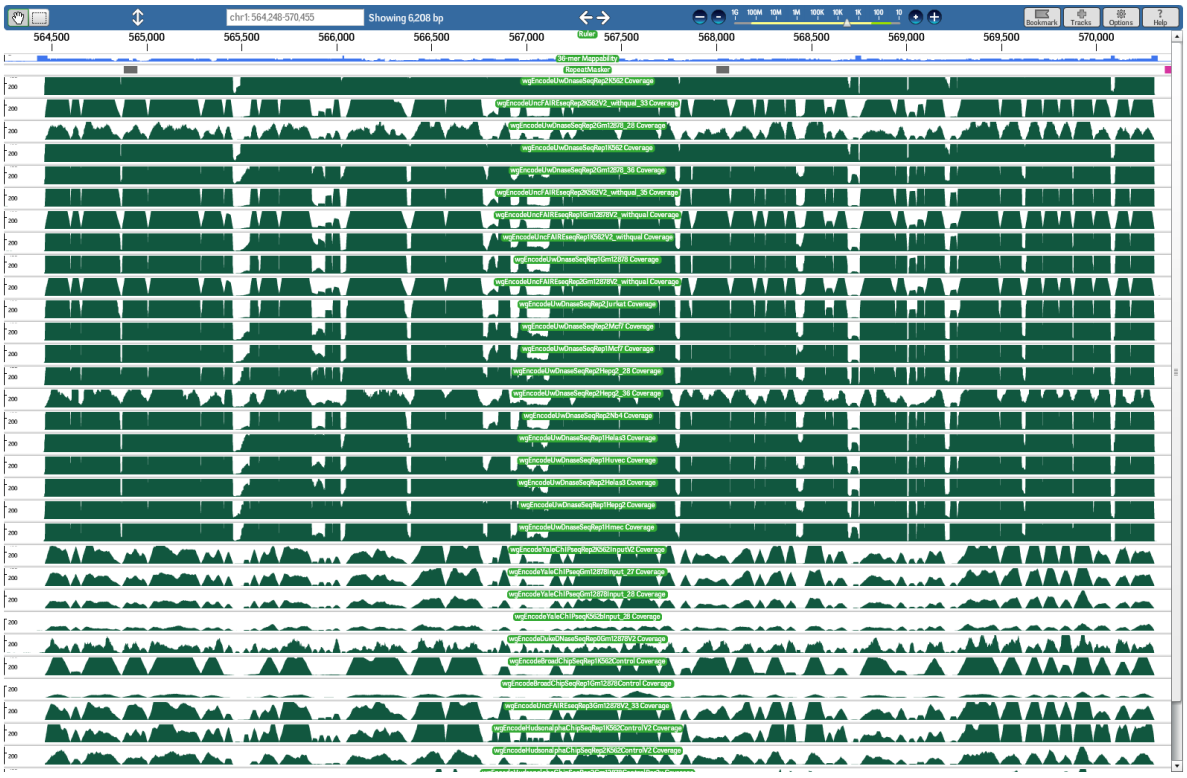
Artifact Type	Definition	#artifacts
High_Mappability_island	Regions that have overall high mappability > 0.5 but with interspersed low mappability locations	7
Low_mappability_island	Regions that have overall low mappability > 0.5 but with interspersed high mappability locations	100
Satellite_repeat	Overlap known satellite repeats	36
centromeric_repeat	Overlap known centromeric repeats	76
snRNA	Overlap known snRNA repeat annotation	2
telomeric_repeat	Overlap known telomeric repeats	5
Total		**226**

Examples of ultra-high signal artifact regions

High-mappability Island



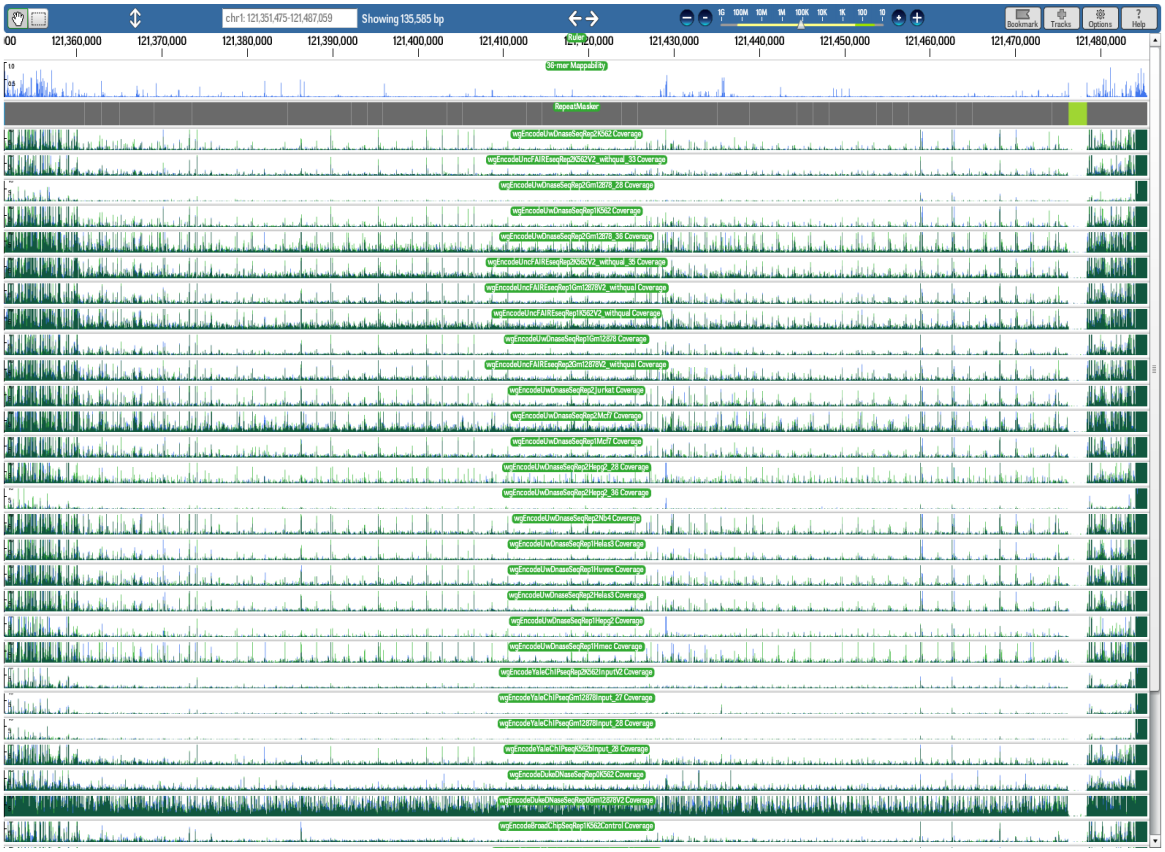
Signal due to unique mapping reads



Signal due to multimapping reads

Centromeric repeat

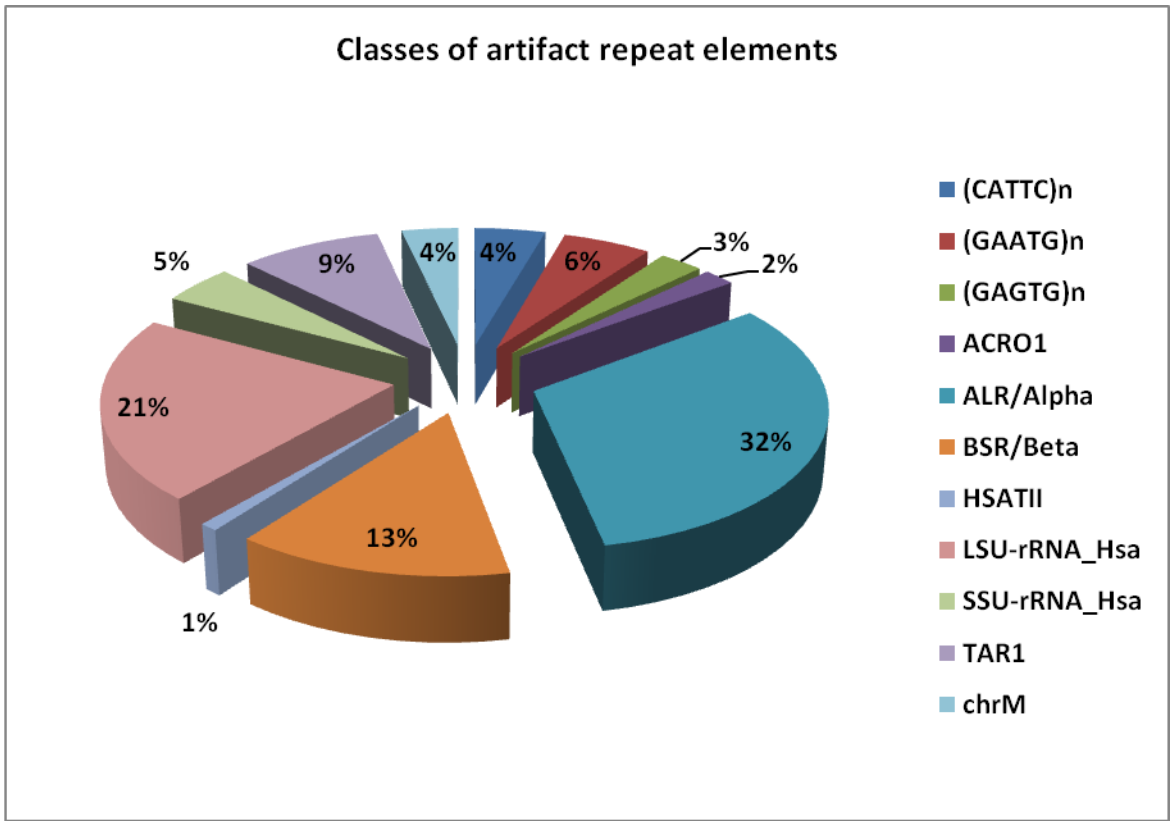
Signal due to unique mapping reads



Signal due to multimapping reads

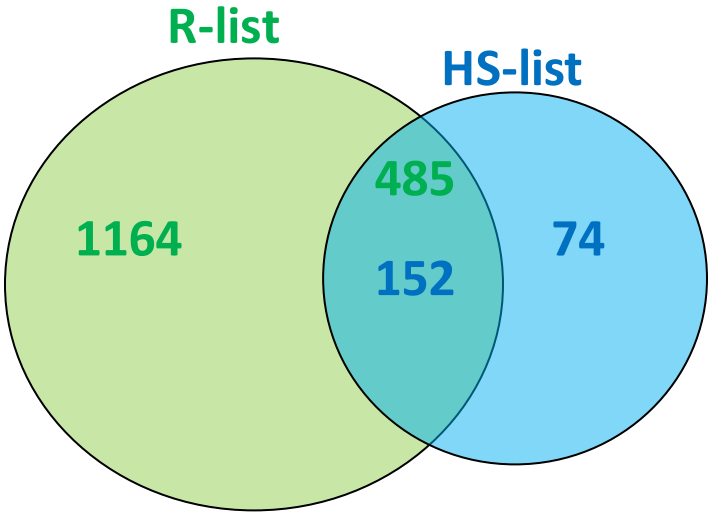
Comparison to known repeat annotations and mappability artifacts

Terry Furey et al. have created a blacklist primarily based on repeat annotations. The BED file can be downloaded from the ENCODE wiki . The classes of repeat elements and the number of regions in each class are listed below.



Artifact Type	#artifacts
(CATTC)n	74
(GAATG)n	92
(GAGTG)n	46
ACRO1	33
ALR/Alpha	529
BSR/Beta	223
HSATII	15
LSU-rRNA_Hsa	354
SSU-rRNA_Hsa	76
TAR1	148
chrM	59
Total	1649

We compared our set of signal artifact regions **(HS-list)** with Terry's list of artifact repeat regions **(R-list).**



Common Regions

152 of the 226 HS-list regions are in R-list
485 of the 1649 R-list regions are in HS-list

Regions specific to each list

74 regions in HS-list are NOT found in R-list
1164 regions in R-list are NOT found in HS-list

5 of the 7 High-mappability islands are exclusive to the HS-list. Of the remaining 2, one maps to a chrM repeat annotation and the other maps to the rRNA repeat annotation.

We reanalyzed the signal/read count properties of the 1164 regions that were exclusive to Terry's repeat-based blacklist.

Of the 1164 regions unique to T-list

- **1087 regions are completely benign** in terms of signal artifacts i.e. they have no noticeable signal (< 2 fold)
- **77 regions have medium-scale signal (> 40 fold for open chromatin and > 15 fold for control/input)**
    - 41 of the 77 regions are annotated as rRNA repeats
    - These regions were not identified by our pipeline mainly because reads mapping to ribosomal RNA were being filtered out by the DNAnexus mapper and so the ratio of the number of multimapping to unique mapping reads for these regions was not exceeding the artifact threshold.
- **None have ultra-high signal** (> 200 fold for open chromatin and > 30 fold for control/input) in all tracks
- All the regions have **low mappability** (since they are repeats).

Consensus artifact regions

We merged the 77 medium-scale signal artifact regions that were exclusive to Terry's list with the ultra-high signal artifact regions identified by our pipeline.

One may also wish to combine all of Terry's repeat based artifact regions with the ultra-high signal artifact regions.

Analyses that benefit from this blacklist

- Pearson correlation is highly affected by signal artifacts. Hence any analyses involving correlation between signal tracks should pre-filter the signal tracks using this artifact list.
- Mean and max measures are not robust to artifacts. Hence, aggregation plot analyses should prefilter peaks that lie in these artifact regions
- It might be wise to eliminate any kind of hypotheses testing (e.g. enhancer predictions) for regions that lie in these artifact regions.
- Genome segmentation/large-scale analyses might want to nullify the signal in these artifact regions before using the signal tracks for learning.
- Supervised learning methods should remove any training examples that lie in these artifact regions.