Supplemental Material: sample quality

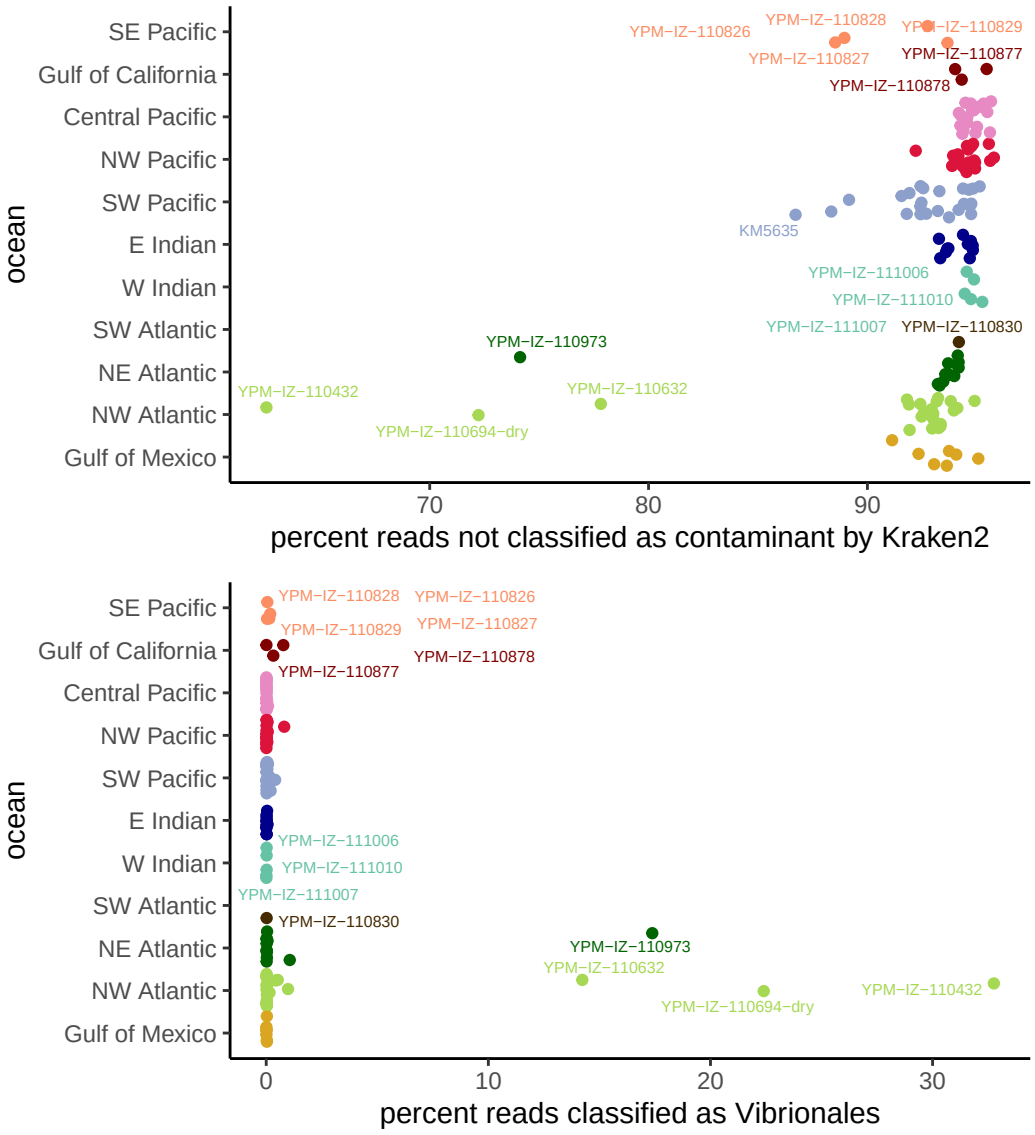
Sample quality

`Kraken2`: cross-species contamination

`Kraken2` attempts to classify reads based on k-mers, against a standard database of known contaminants. `Kraken2` identifies several samples as having a large percentage of "classified" reads. These reads are largely classified as Vibrionales bacteria. Based on these, we will exclude the following samples as potentially highly contaminated:

- YPM-IZ-110432
- YPM-IZ-110973
- YPM-IZ-110632

Sample YPM-IZ-110694-dry was a resequenced sample from dried tissue. For main analyses we will use the other sample, YPM-IZ-110694, and not the dried specimen.



Cross-contamination and duplicated samples

In addition to YPM-IZ-110694-dry, we sequenced a second biological replicat (YPM-IZ-110474-2 of YPM-IZ-110474) and we generated two technical replicates (YPM-IZ-110269-A and -B of YPM-110269). We used these replicates as quality control for library preparation and sequencing, and as a reference for identifying duplicated samples.

We assesses similarity by calculating relatedness scores using the King-robust metric in `PLINK2`. Using a cutoff of 0.35 to detect potential duplicate samples, we identify the following pairs:

- YPM-IZ-106941 and YPM-IZ-106944
- YPM-IZ-110878 and YPM-IZ-110879
- YPM-IZ-111013 and YPM-IZ-111015
- YPM-IZ-111018 and YPM-IZ-111019

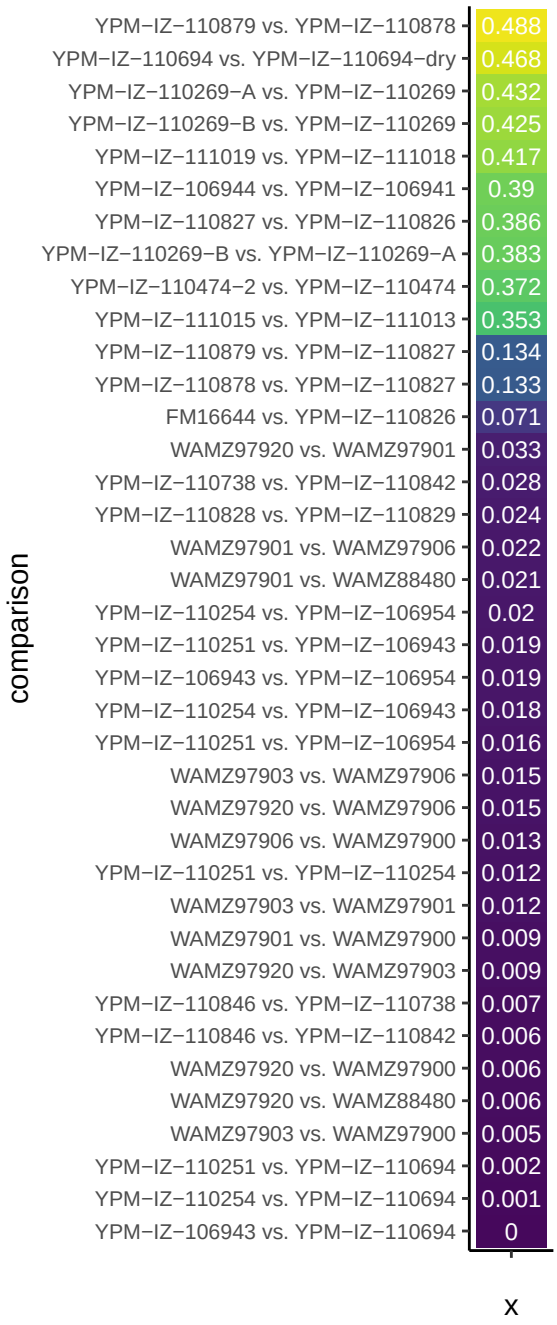
The latter sample in each pair was subsequently excluded.

We also identified two pairs of samples that had moderately high relatedness (0.07 and 0.13, respectively) with samples from improbably long distances. In both cases, DNA was extracted in the same batch, suggested cross contamination.

- YPM-IZ-110826 (Chile) with sample FM16644 (Canary Islands)
- YPM-IZ-110827 (Chile) with samples YPM-IZ-110878 and YPM-IZ-110879 (Gulf of California).

The former in each pair was subsequently excluded.

No other samples indicated a relatedness above the predicted score for a close relative.



Mapping statistics

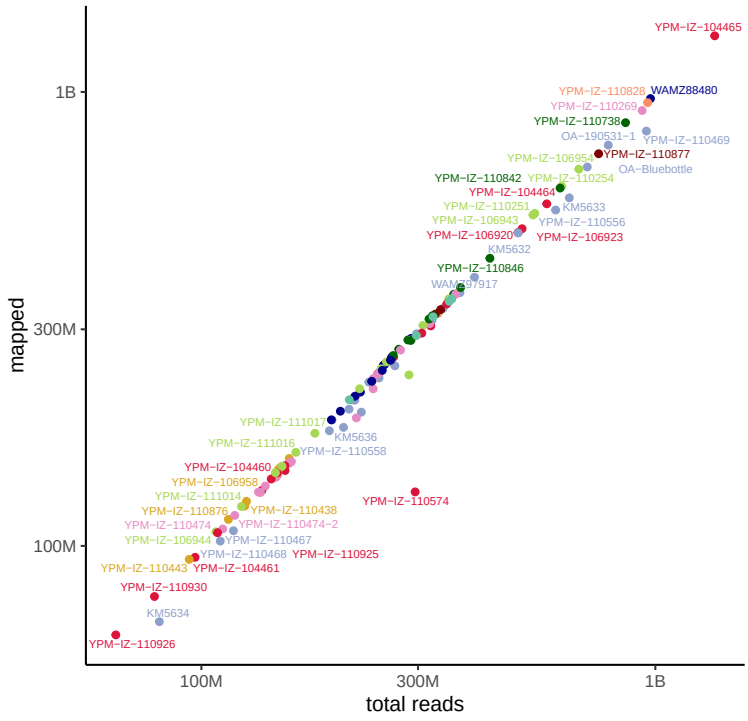
We examine sample quality by mapping to our assembled genome. Samples with low mapping percentages indicate potential contamination or low-quality. We can count the number of total

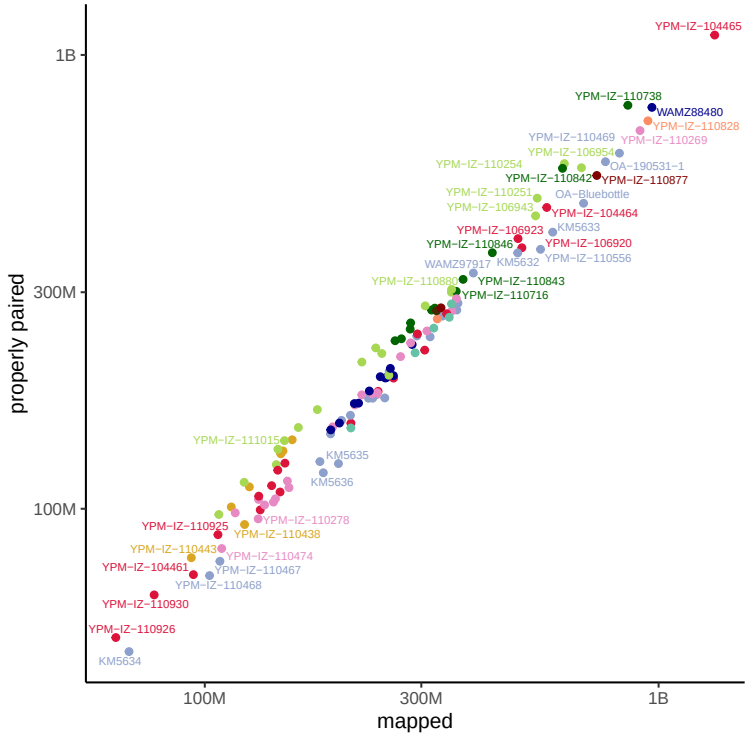
reads, reads that map to the reference genome, and reads that are properly paired. From this we can detect a significantly lower mapping percentage for the sample:

- YPM-IZ-110574

Tests using *in silico* PCR to extract 18S from this sample further indicated contamination from arthropod DNA. This sample will likewise be excluded.

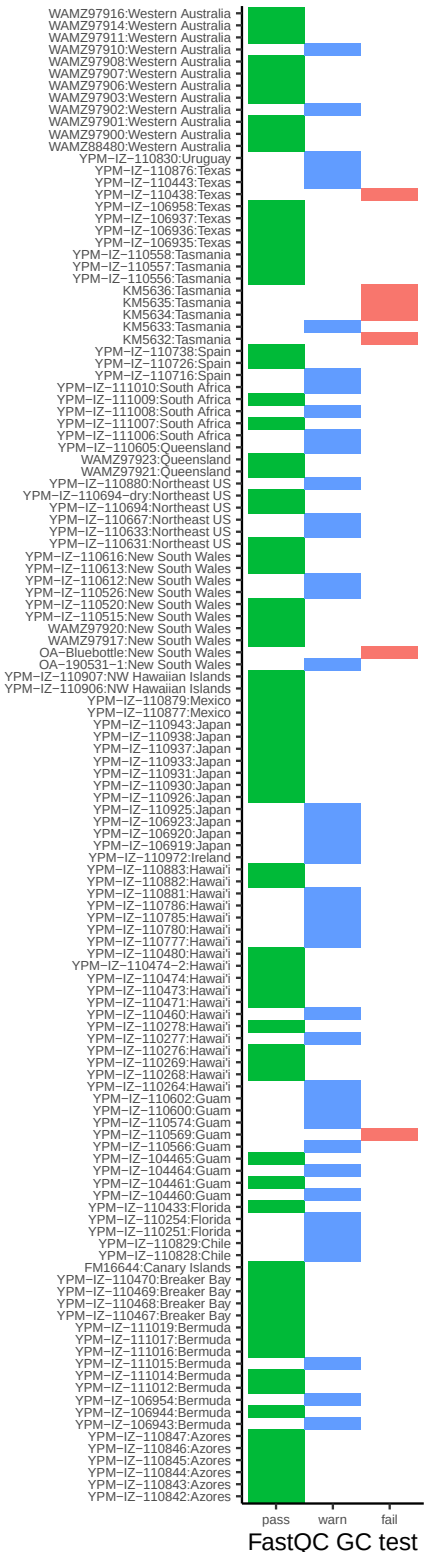
Comparing statistics on mapped reads vs. properly paired reads shows no significant outliers in terms of quality.


mapped
total reads
1B
300M
100M
100M
300M
1B
YPM−IZ−104465
YPM−IZ−110828
WAMZ88480
YPM−IZ−110269
YPM−IZ−110738
OA−190531−1
YPM−IZ−110469
YPM−IZ−106954
YPM−IZ−110877
OA−Bluebottle
YPM−IZ−110842
YPM−IZ−110254
YPM−IZ−104464
YPM−IZ−110251
KM5633
YPM−IZ−106943
YPM−IZ−110556
YPM−IZ−106920
YPM−IZ−106923
KM5632
YPM−IZ−110846
WAMZ97917
YPM−IZ−111017
KM5636
YPM−IZ−111016
YPM−IZ−110558
YPM−IZ−104460
YPM−IZ−106958
YPM−IZ−111014
YPM−IZ−110574
YPM−IZ−110876
YPM−IZ−110438
YPM−IZ−110474
YPM−IZ−110474−2
YPM−IZ−106944
YPM−IZ−110467
YPM−IZ−110468
YPM−IZ−110925
YPM−IZ−110443
YPM−IZ−104461
YPM−IZ−110930
KM5634
YPM−IZ−110926


properly paired
mapped
1B
300M
100M
100M
300M
1B
YPM−IZ−104465
YPM−IZ−110738
WAMZ88480
YPM−IZ−110828
YPM−IZ−110469
YPM−IZ−110269
YPM−IZ−106954
OA−190531−1
YPM−IZ−110254
YPM−IZ−110842
YPM−IZ−110877
YPM−IZ−110251
OA−Bluebottle
YPM−IZ−106943
YPM−IZ−104464
YPM−IZ−106923
KM5633
YPM−IZ−110846
YPM−IZ−106920
WAMZ97917
KM5632
YPM−IZ−110556
YPM−IZ−110880
YPM−IZ−110843
YPM−IZ−110716
YPM−IZ−111015
KM5635
KM5636
YPM−IZ−110925
YPM−IZ−110278
YPM−IZ−110438
YPM−IZ−110443
YPM−IZ−110474
YPM−IZ−104461
YPM−IZ−110467
YPM−IZ−110468
YPM−IZ−110930
YPM−IZ−110926
KM5634

`FastQC`: GC content

`FastQC` flagged no sequences as having low sequence quality, all were retained. But `FastQC` flagged several samples as having GC contents distinct from the expected distribution. Based on these results we will flag eight samples that fail the GC test and exclude those from our strict analyses (but retain them in general analyses presented in the manuscript).

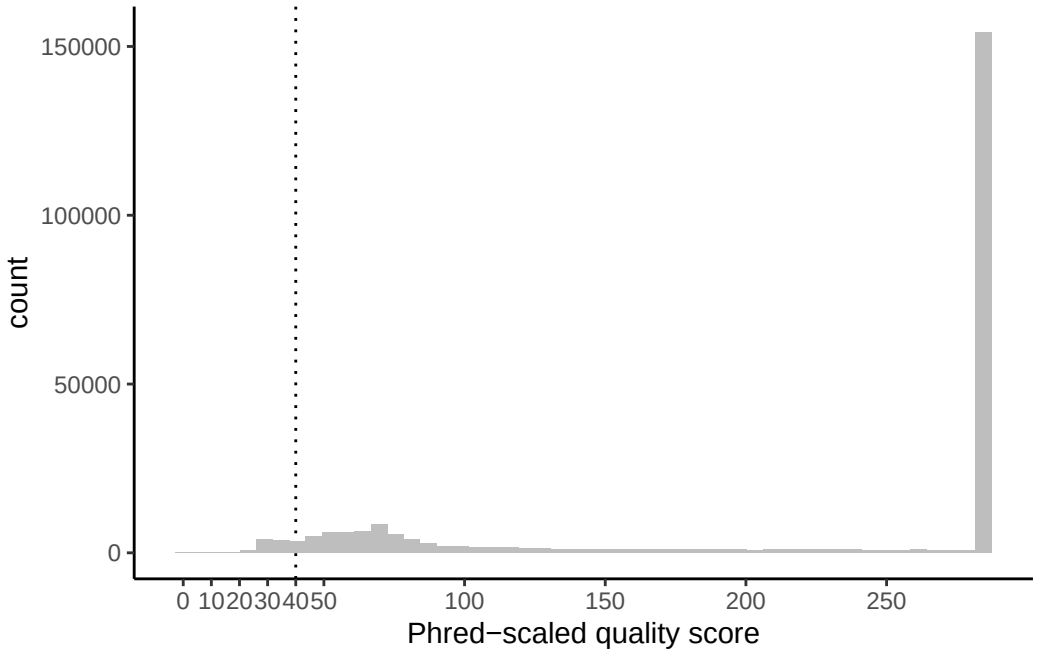

WAMZ97916:Western Australia
WAMZ97914:Western Australia
WAMZ97911:Western Australia
WAMZ97910:Western Australia
WAMZ97908:Western Australia
WAMZ97907:Western Australia
WAMZ97906:Western Australia
WAMZ97903:Western Australia
WAMZ97902:Western Australia
WAMZ97901:Western Australia
WAMZ97900:Western Australia
WAMZ88480:Western Australia
YPM−IZ−110830:Uruguay
YPM−IZ−110876:Texas
YPM−IZ−110443:Texas
YPM−IZ−110438:Texas
YPM−IZ−106958:Texas
YPM−IZ−106937:Texas
YPM−IZ−106936:Texas
YPM−IZ−106935:Texas
YPM−IZ−110558:Tasmania
YPM−IZ−110557:Tasmania
YPM−IZ−110556:Tasmania
KM5636:Tasmania
KM5635:Tasmania
KM5634:Tasmania
KM5633:Tasmania
KM5632:Tasmania
YPM−IZ−110738:Spain
YPM−IZ−110726:Spain
YPM−IZ−110716:Spain
YPM−IZ−111010:South Africa
YPM−IZ−111009:South Africa
YPM−IZ−111008:South Africa
YPM−IZ−111007:South Africa
YPM−IZ−111006:South Africa
YPM−IZ−110605:Queensland
WAMZ97923:Queensland
WAMZ97921:Queensland
YPM−IZ−110880:Northeast US
YPM−IZ−110694−dry:Northeast US
YPM−IZ−110694:Northeast US
YPM−IZ−110667:Northeast US
YPM−IZ−110633:Northeast US
YPM−IZ−110631:Northeast US
YPM−IZ−110616:New South Wales
YPM−IZ−110613:New South Wales
YPM−IZ−110612:New South Wales
YPM−IZ−110526:New South Wales
YPM−IZ−110520:New South Wales
YPM−IZ−110515:New South Wales
WAMZ97920:New South Wales
WAMZ97917:New South Wales
OA−Bluebottle:New South Wales
OA−190531−1:New South Wales
YPM−IZ−110907:NW Hawaiian Islands
YPM−IZ−110906:NW Hawaiian Islands
YPM−IZ−110879:Mexico
YPM−IZ−110877:Mexico
YPM−IZ−110943:Japan
YPM−IZ−110938:Japan
YPM−IZ−110937:Japan
YPM−IZ−110933:Japan
YPM−IZ−110931:Japan
YPM−IZ−110930:Japan
YPM−IZ−110926:Japan
YPM−IZ−110925:Japan
YPM−IZ−106923:Japan
YPM−IZ−106920:Japan
YPM−IZ−106919:Japan
YPM−IZ−110972:Ireland
YPM−IZ−110883:Hawai'i
YPM−IZ−110882:Hawai'i
YPM−IZ−110881:Hawai'i
YPM−IZ−110786:Hawai'i
YPM−IZ−110785:Hawai'i
YPM−IZ−110780:Hawai'i
YPM−IZ−110777:Hawai'i
YPM−IZ−110480:Hawai'i
YPM−IZ−110474−2:Hawai'i
YPM−IZ−110474:Hawai'i
YPM−IZ−110473:Hawai'i
YPM−IZ−110471:Hawai'i
YPM−IZ−110460:Hawai'i
YPM−IZ−110278:Hawai'i
YPM−IZ−110277:Hawai'i
YPM−IZ−110276:Hawai'i
YPM−IZ−110269:Hawai'i
YPM−IZ−110268:Hawai'i
YPM−IZ−110264:Hawai'i
YPM−IZ−110602:Guam
YPM−IZ−110600:Guam
YPM−IZ−110574:Guam
YPM−IZ−110569:Guam
YPM−IZ−110566:Guam
YPM−IZ−104465:Guam
YPM−IZ−104464:Guam
YPM−IZ−104461:Guam
YPM−IZ−104460:Guam
YPM−IZ−110433:Florida
YPM−IZ−110254:Florida
YPM−IZ−110251:Florida
YPM−IZ−110829:Chile
YPM−IZ−110828:Chile
FM16644:Canary Islands
YPM−IZ−110470:Breaker Bay
YPM−IZ−110469:Breaker Bay
YPM−IZ−110468:Breaker Bay
YPM−IZ−110467:Breaker Bay
YPM−IZ−111019:Bermuda
YPM−IZ−111017:Bermuda
YPM−IZ−111016:Bermuda
YPM−IZ−111015:Bermuda
YPM−IZ−111014:Bermuda
YPM−IZ−111012:Bermuda
YPM−IZ−106954:Bermuda
YPM−IZ−106944:Bermuda
YPM−IZ−106943:Bermuda
YPM−IZ−110847:Azores
YPM−IZ−110846:Azores
YPM−IZ−110845:Azores
YPM−IZ−110844:Azores
YPM−IZ−110843:Azores
YPM−IZ−110842:Azores
pass
warn
fail
FastQC GC test

Establishing appropriate filters

Site quality

Most sites have high phred quality scores (e.g. a Phred encoded score of 40 indicates a 1 in 10,000 chance of an erroneous call). Based on this, we set the following:

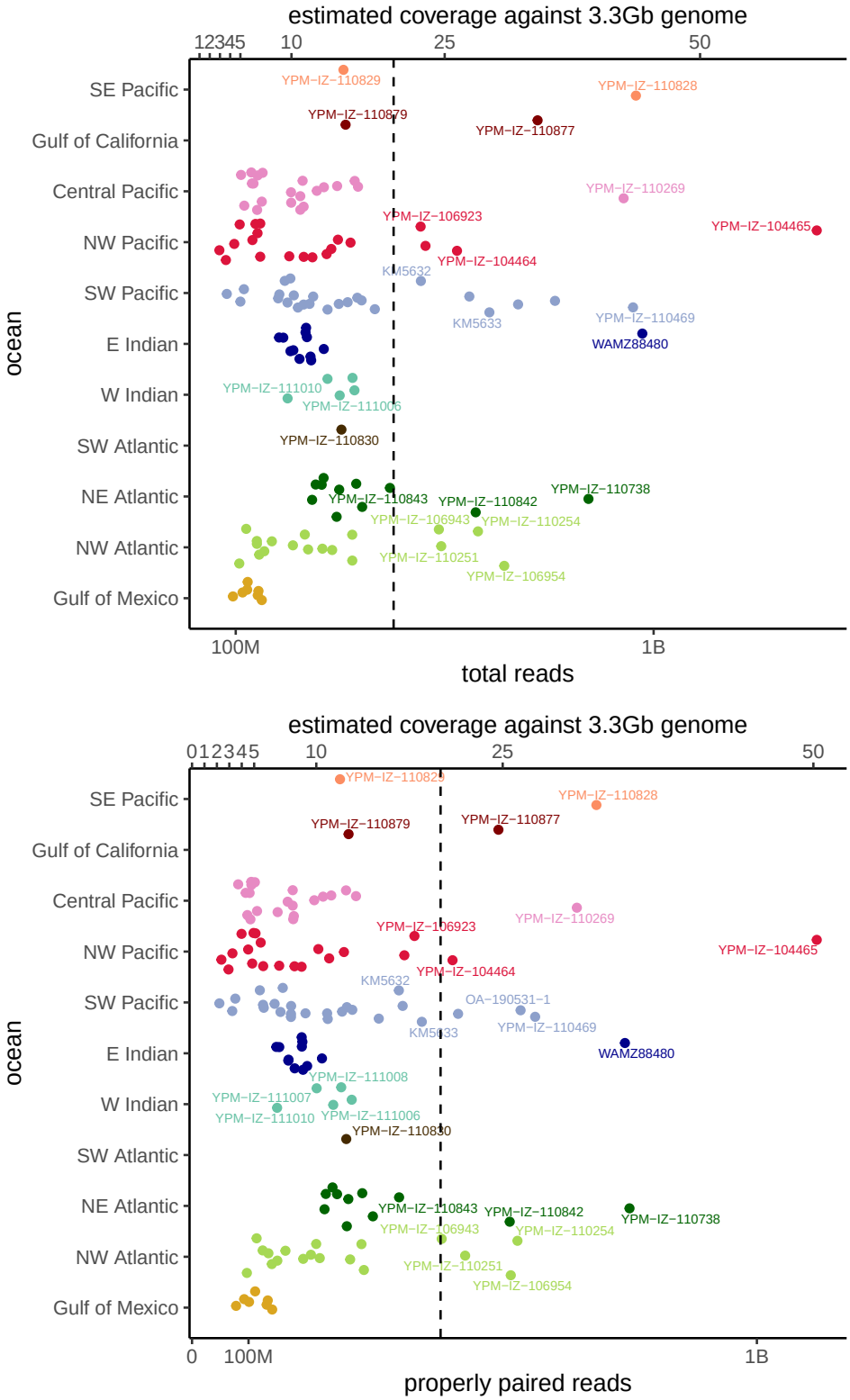
- minimum site quality score = 40



Coverage

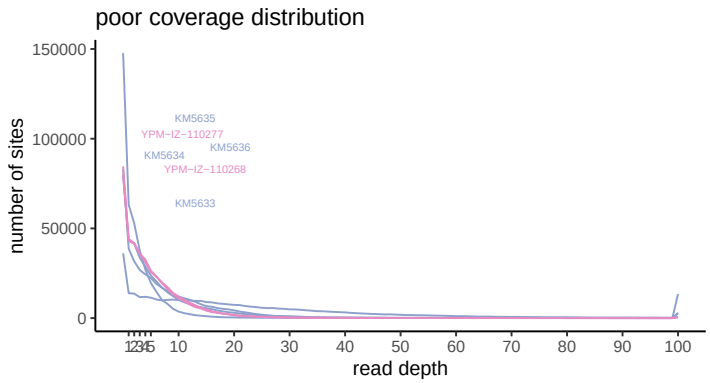
We can estimate coverage by dividing the number of basepairs sequenced for each sample by the length of the reference genome, 3.3Gb. Samples were sequenced to variable depths, with estimated coverage ranging from 5-60x.

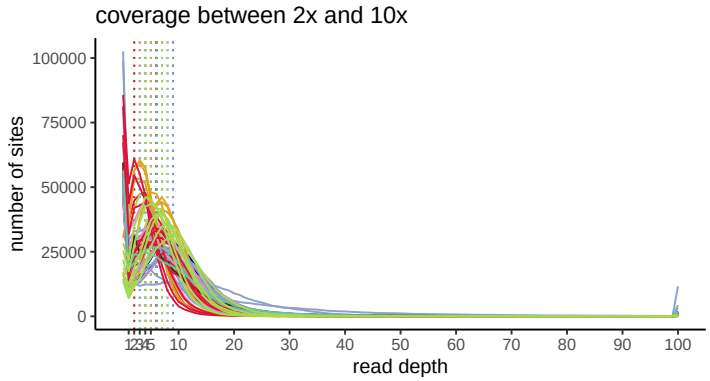
Samples with greater than 20x target coverage (dashed line) were considered high coverage, and were selected for genome size estimation and other high-coverage analyses.

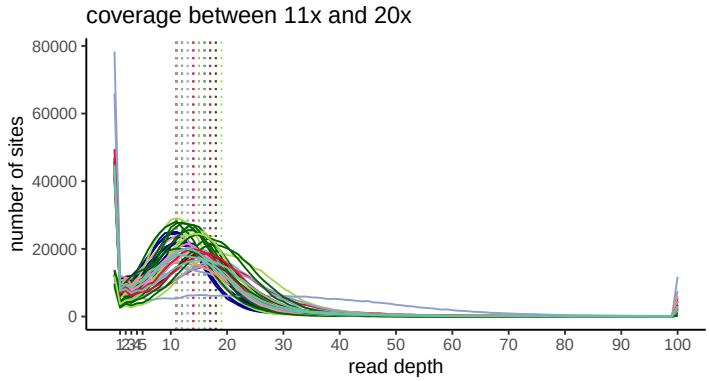

estimated coverage against 3.3Gb genome
12345
10
25
50
SE Pacific
Gulf of California
Central Pacific
NW Pacific
SW Pacific
E Indian
W Indian
SW Atlantic
NE Atlantic
NW Atlantic
Gulf of Mexico
ocean
YPM−IZ−110829
YPM−IZ−110828
YPM−IZ−110879
YPM−IZ−110877
YPM−IZ−110269
YPM−IZ−106923
YPM−IZ−104465
YPM−IZ−104464
KM5632
KM5633
YPM−IZ−110469
WAMZ88480
YPM−IZ−111010
YPM−IZ−111006
YPM−IZ−110830
YPM−IZ−110738
YPM−IZ−110843
YPM−IZ−110842
YPM−IZ−106943
YPM−IZ−110254
YPM−IZ−110251
YPM−IZ−106954
100M
1B
total reads
estimated coverage against 3.3Gb genome
012345
10
25
50
SE Pacific
Gulf of California
Central Pacific
NW Pacific
SW Pacific
E Indian
W Indian
SW Atlantic
NE Atlantic
NW Atlantic
Gulf of Mexico
ocean
YPM−IZ−110829
YPM−IZ−110828
YPM−IZ−110879
YPM−IZ−110877
YPM−IZ−110269
YPM−IZ−106923
YPM−IZ−104465
YPM−IZ−104464
KM5632
OA−190531−1
KM5633
YPM−IZ−110469
WAMZ88480
YPM−IZ−111008
YPM−IZ−111007
YPM−IZ−111010
YPM−IZ−111006
YPM−IZ−110830
YPM−IZ−110843
YPM−IZ−110842
YPM−IZ−110738
YPM−IZ−106943
YPM−IZ−110254
YPM−IZ−110251
YPM−IZ−106954
0
100M
1B
properly paired reads

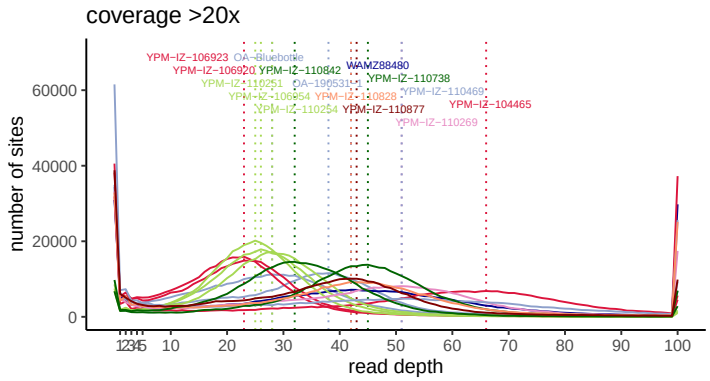
We use `angsd` to estimate realized depth across samples and sites using a subset of genomic regions. We categorize samples into those with poor coverage distributions, those with low coverage (between 2 and 10x), moderate coverage (11 and 20x) and high coverage (>20x). Poor coverage samples were the following:

- KM5633
- KM5634
- KM5635
- KM5636
- YPM-IZ-110277
- YPM-IZ-110268


poor coverage distribution
number of sites
read depth
KM5635
YPM−IZ−110277
KM5636
KM5634
YPM−IZ−110268
KM5633

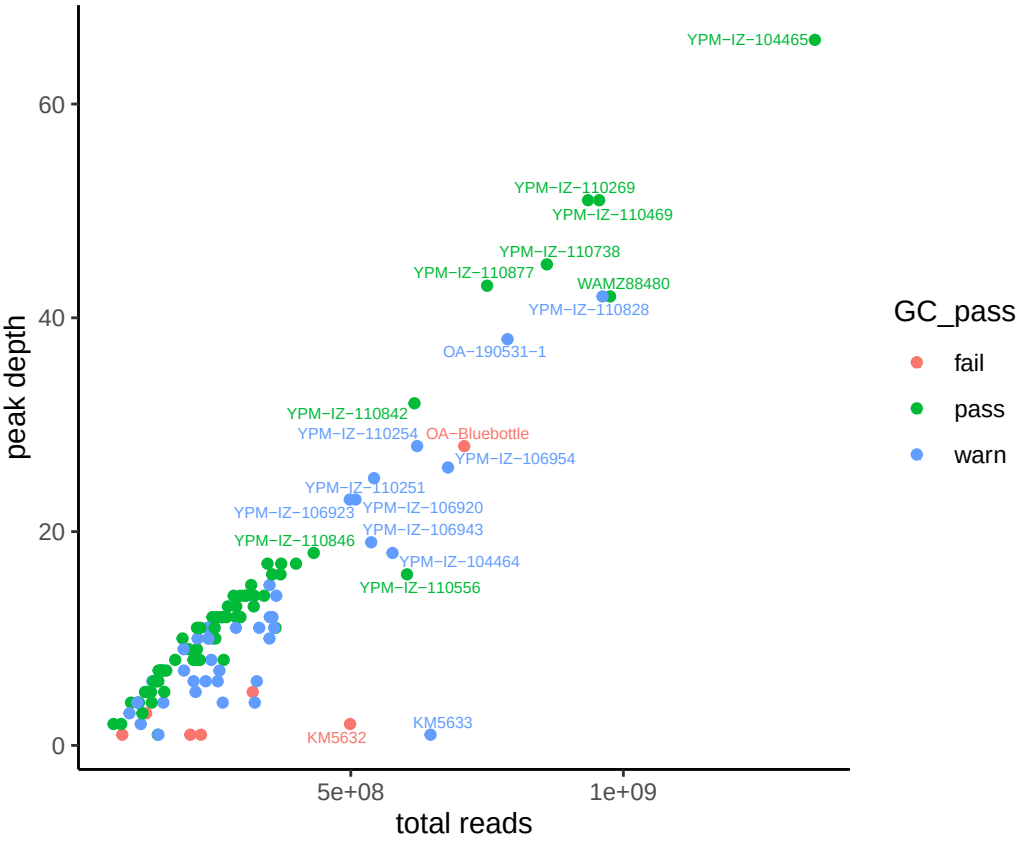

coverage between 2x and 10x
number of sites
read depth


coverage between 11x and 20x
number of sites
read depth


coverage >20x
number of sites
read depth
YPM−IZ−106923
OA−Bluebottle
YPM−IZ−106920
YPM−IZ−110842
WAMZ88480
YPM−IZ−110738
YPM−IZ−110251
OA−190531
YPM−IZ−110469
YPM−IZ−106954
YPM−IZ−110828
YPM−IZ−104465
YPM−IZ−110254
YPM−IZ−110877
YPM−IZ−110269

From the previous distributions we calculated peak depth, and then compared to total reads. We observed that GC distribution, as assessed using `FastQC` impact the relationship between input reads and realized depth.

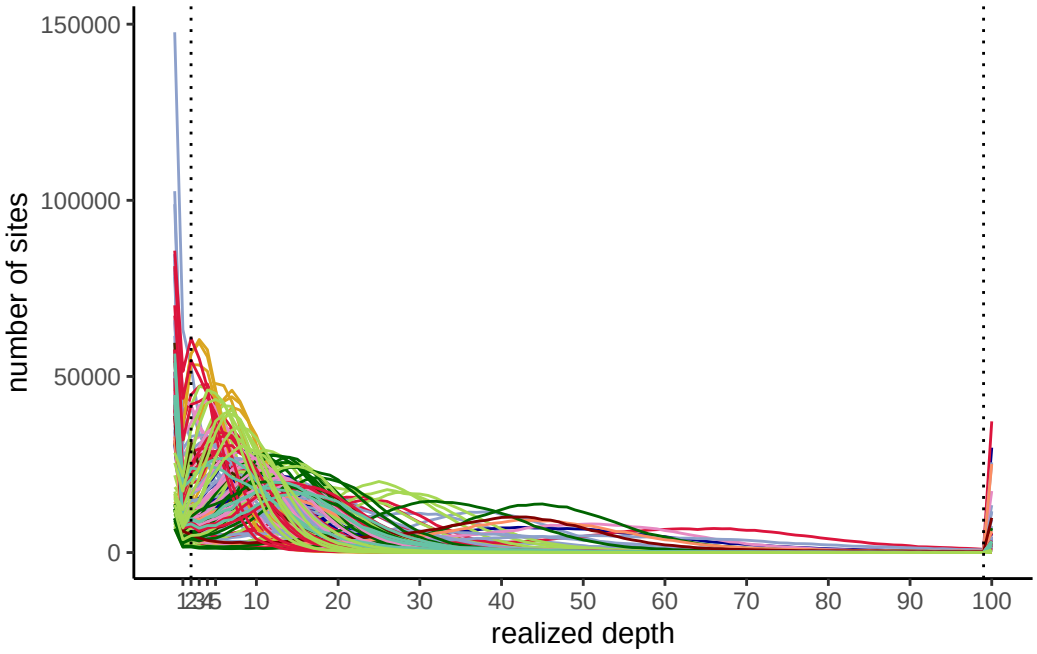
To address this, we run an additional robustness test using only samples that pass the GC test.



We used realized depth to establish the cutoff for minimum and maximum depth across sites.

For `ANGSD` analyses that are robust to low coverage, we did not perform any filtering based on depth, but for `BCF` based analsyes (e.g Fst calculations), we set the following filters across sites for each sample:

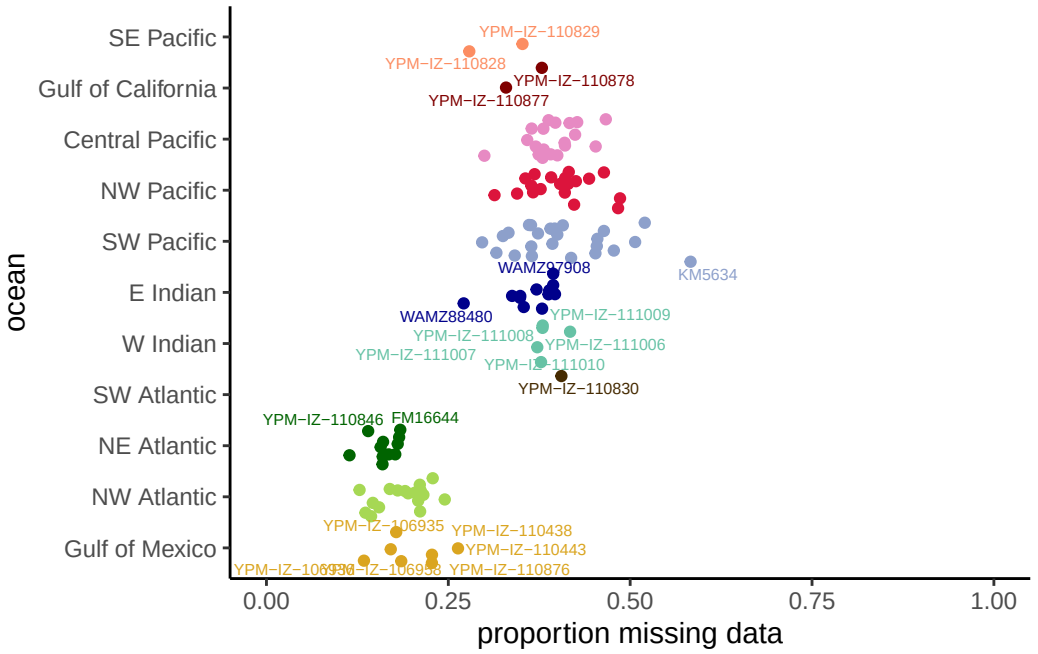
- minimum depth = 2x
- maximum depth = 99x



Missingness

We examined the proportion of missing sites across samples. This was calculated using `vcftools` on a random subsample of sites from across genome regions. From this we can see identify several samples with a high proportion of missing sites, in particular:

- KM5634



We can use the distribution of missing samples across sites to establish an appropriate filter for missingness. Here we set the following:

- tolerate missingness equal to or below 75%

This means we exclude sites for which >25% of samples are missing data.