Impact of using a personalized genome on histone ChIP-seq peak calls

Cristian Groza[1], Bing Ge[1], Tony Kwan[1,2], Nicole Soranzo[5,6,7,8], Tomi Pastinen[1,2,3], Guillaume Bourque[1,2,4]

Affiliations

1. Department of Human Genetics, McGill University, Montreal, QC, Canada.
2. McGill University and Genome Quebec Innovation Centre, Montreal, QC, Canada.
3. Center for Pediatric Genomic Medicine, Kansas City, MO, USA.
4. Canadian Centre for Computational Genomics, Montreal, Canada.
5. Department of Human Genetics, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1HH, UK.
6. Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Long Road, Cambridge, CB2 0PT, UK.
7. British Heart Foundation Centre of Excellence, Division of Cardiovascular Medicine, Addenbrooke's Hospital, Hills Road, Cambridge, CB2 0QQ, UK.
8. The National Institute for Health Research Blood and Transplant Unit (NIHR BTRU) in Donor Health and Genomics, University of Cambridge, Strangeways Research Laboratory, Wort's Causeway, Cambridge, CB1 8RN, UK.

Abstract

Next generation sequencing experiments typically involve the alignment of reads to a reference genome, a composite made from genomic sequences of multiple individuals. However, differences from the reference sequence are present in all individuals, because of the genetic diversity observed as single nucleotide substitutions and structural variants, such as insertions and deletions. We hypothesized that using a genomic reference that does not account for variations can lead to incorrectly mapped reads and biased downstream results such as ChIP-seq peak calls. To demonstrate this, we show that accounting for genetic variation using a modified reference genome (MPG) or a denovo personalized genome (DPG) can alter histone (H3K4me1, H3K27ac) ChIP-seq peak calls by either creating new personal peaks or by the loss of reference peaks. The best MPGs are found to alter roughly 1% of peak calls while this can increase to 5% for DPGs. At the same time, we show differences beyond the calls in terms of read counts in regions associated with the new, altered and unchanged peaks. Such differences between the reference and personalized alignments are abundant in DPGs but rarely seen in MPGs. We notice that indels, followed by SNVs have the highest chance to change peak calls in

MPGs. A counter-balancing factor is peak width, with wider calls being less likely to be altered. Read length also influences the impact of a personalized genome on histone peak tracks. Longer reads quickly reduce the impact in MPGs whereas DPGs maintain a strong impact despite fluctuations due to read length. Overall, the most significant altered peaks originate from DPGs, possibly due to the capture of SVs that are not present in MPGs. This suggests that the quality of personalized genomes may benefit greatly from the incorporation of SVs in addition to SNVs and indels.

Introduction

Standard ChIP-seq analysis pipelines rely on aligning reads to a standard reference genome, such as hg19, followed by peak calling. While this method is sufficient to detect the overwhelming majority of peaks, it does not account for genetic variation between individuals.

An individual's sequence and structural differences from the standard reference are bound to shift the mapping of certain reads. Provided that the mapping changes for a sufficient number of reads, an alignment to a personalized genome will present an altered peak (AP) track. This personalized peak track is expected to be mostly identical with the reference track. However, a closer look at the total set of peaks should reveal some that are unique to the personalized track and some that are unique to the reference track. Naturally, the impact of a personalized genome would be dependent on the amount of divergence from the reference genome.

To measure this impact, a personalized genome needs to be produced for every ChIP-seq dataset. A straightforward way of generating such a genome is to modify the reference genome with commonly available variant calls obtained from NGS technologies.

Studies involving the use of modified reference genomes are present in the literature. One such effort involves correcting the reference genome using phased SNV calls, realigning transcription factor and histone ChIP-seq data and then record allelic specific binding events [23]. However, this study does not compare the personalized peaks with the reference peaks and does not incorporate indels in the reference genome. Therefore, it is limited in understanding how the full spectrum of genetic variation may affect current ChIP-seq analyses. Pipelines such as AlleleSeq [22] do support indels and structural variations but remain restricted to detecting allellic specific events without comparing the results back to the reference. A study that did compare the modified genome with the reference was done in the context of RNA-seq [16], where it was shown that personalized mouse genomes can improve transcript abundance estimates.

This paper aims to evaluate the impact of personalized genomes on ChIP-seq analysis by first augmenting modified genomes with both indel and SNV calls and then comparing the results with the reference. This suffices to account for variation of small length, but not for larger scale structural variations. The inability to account for structural variation potentially removes much of the utility of a modified genome. For this reason, we later turn to denovo assembled personal genomes to capture a broader range of genetic differences. We were unable to find other papers that attempt to use denovo assemblies as personalized genomes in the analysis of ChIP-seq data.

Despite this advantage, denovo assembly is a much greater effort than genotyping and modifying the reference. Optimally, it requires at least 50X sequencing depth, which makes cost a significant factor [8]. Also, the computational time for

denovo assembly is much higher than aligning to a reference and calling variants [25], which is a drawback if not readily available. Some annotations needing to be adapted for the denovo assembly may require rerunning tools. For instance, a lifted RepeatMasker annotation should be sufficient for modified genomes but not for denovo assembled genomes.

Given the above trade-offs, an additional goal is to provide a comparison between denovo assembled and modified personalized genomes that can inform a choice between the two approaches, especially since the choice of assembly is known to be important in other types of experiments [24]. In addition, quantifying a personalized genome's impact should justify or reject their adoption in ChIP-seq pipelines or at least outline a scenario where the benefits outweigh the added complexity.

Methods

Data

Samples from the Blueprint project [5] were selected due to the availability of phased variation calls from low pass whole genome sequencing (8X) together with ChIP-seq datasets for the H3K4me1 and H3K27ac histone marks. In total, 151 H3K4me1 samples and 111 H3K27ac samples were used in this analysis.

We also selected NA12878 as a benchmark dataset due to the availability of phased variation calls from high coverage whole genome sequencing (200X) [10] in addition to several denovo assemblies. To match the Blueprint ChIP-seq samples for NA12878, FASTQs for H3K27ac, H3K4me1 marks and a control (input) were downloaded from the ENCODE project [9]. The accession numbers for these samples are *ENCFF000ASM*, *ENCFF000ASU* and *ENCFF002ECP* respectively. To generate additional supporting results for ChIP-seq, we used a low pass NA12878 WGS dataset from IGSR [6] (SRR622461).

Preparing personalized genomes

Two approaches were used to generate personalized genomes. First, *vcf2diploid* [22] was used to substitute the alternative sequence of the phased variation calls into the hg19 reference to create a modified personalized genome (MPG). The output are two FASTA files for each contig, forming the conventionally named maternal and paternal haplotypes. The contig FASTAs were concatenated according to their haplotype, resulting in one maternal and one paternal *FASTA*. It is to be noted that *vcf2diploid* does not process unordered contigs. Therefore, unordered contigs were removed from hg19 to ensure the same set of contigs between the standard and substituted versions. Also, *vcf2diploid* generates two chain files that allow the lifting of annotation tracks with coordinates in hg19 to the corresponding personalized haplotype using *liftOver*. This is necessary since the incorporated indels shift the coordinates of the maternal/paternal haplotype relative to hg19.

The second approach, applied only to the NA12878 dataset, consisted of using denovo assembled genomes from the Pendleton [18] and the 10X Genomics assemblies [11] to create denovo assembled personalized genomes (DPGs). The 10X Genomics assembly includes two pseudo-haps named Hap1 and Hap2 that will be used as a denovo assembled diploid genome. In the case of the denovo assemblies, the chain files had to be produced from a BLAT [12] alignment between the denovo assembly and hg19 with the UCSC tool set [13]. This allowed the lifting of annotation tracks from the denovo assembly to the hg19 reference. Note that hg19

contains alternative contigs that represent some loci multiple times. In denovo assemblies, we expect loci to be represented only once. Therefore, the above analysis was performed on a hg19 version that was stripped of alternative contigs.

To allow alignment to personalized MPG or DPG FASTAs using *bwa mem* [1], an index was created using *bwa index*. A *.fai* FASTA index was also created using *samtools faidx* [14], from which a *.chrom.sizes* file was extracted by keeping only the first two columns of the *.fai* file.

Aligning, peak calling and annotating

To match the lowest read length in all datasets and remove any effect of read length, all reads were trimmed to 36bp using *trimmomatic* [3]. The trimmed reads were aligned using *bwa mem* to hg19 and each personalized haplotype FASTAs. After marking duplicates with *picard* [19], peak calling was done on the corresponding BAM files using MACS2 [26] with *--nomodel* and the *--gsize* parameter set to 80% of the assembly length, which was computed from the contig lengths listed in the *.chrom.sizes* file.

For each alignment, a coverage annotation was produced with *bedtools bamtobed* [20], which required the previously made *.chrom.sizes* file. The output was a BED file listing all the aligned reads, which was reduced to only the first three fields that describe the mapped genomic interval of each read.

Lifting annotations

In the case of DPGs, coverage and peak annotations were lifted from the DPG to hg19. In the case of MPGs, it was easier to lift annotations from hg19 to the MPG. Therefore, this required the lifting of variant call annotations to the MPG in addition to the peak call and coverage annotations.

The variant call annotation was first converted from the VCF format to BED, separated by phase and type (SNP vs indel) and then lifted to the personalized haplotype. The outcome is a set of BED files listing the SNPs and indels separately for each respective haplotype.

liftOver was called with default arguments in BP samples, which require 95% sequence identity between lifted regions and target regions. This stringent *-minMatch* was not an issue since MPGs are almost identical to hg19 and virtually all peaks lift. In 10X and Pendleton samples, *-minMatch* was set 0.85 to reduce the number of unlifted peaks and reduce the number of false ref-only peaks. To evaluate lifting efficacy, the number of peaks that failed to lift was compiled for every sample. Once tracks are lifted to a common coordinate system, it becomes possible to overlap and compare the annotations from the personalized haplotype and the hg19 standard reference using *bedtools*.

Overlapping annotations

The lifted peak call annotations were overlapped using *bedtools intersect* and *bedtools subtract*. Peaks resulting from the intersect of the personalized and the hg19 peak tracks were categorized as **common**. Peaks resulting from subtracting the hg19 track from the personalized track were categorized as **personal-only**. Similarly, peaks resulting from subtracting the personalized track from the hg19 track were categorized as **ref-only**. The end result is a set of three BED files for each

personalized haplotype containing the common peaks, the personal-only peaks and ref-only peaks.

For the MPG approach, the number of variation calls in each peak was calculated. The haplotype specific indel and SNP tracks were intersected with the track of each category of peaks using *bedtools intersect -c* to list the number of variations overlapping each common, personal-only and ref-only peak.

Furthermore, the peak tracks were overlapped with the coverage tracks of the personalized and hg19 versions of the alignment using *bedtools intersect -c*. The output is the original peak track with an additional field listing the number of reads in each peak. As a result, the number of reads in regions corresponding to the peaks is known in the reference alignment and the personalized alignment.

Finding peaks with skewed coverage

To find peak called regions that have significant differences between their hg19 and personalized coverages, a statistical test is needed. This comparison is similar to differential expression in that read counts are compared between two conditions: the hg19 reference and the personalized assembly. For the purpose of differential expression, technical variation that occurs during the preparation of different libraries is underestimated by Poisson based tests (overdispersion) [2]. To account for this, various models have been developed. Unlike differential expression, our read counts are not compared between multiple sequencing experiments done under the two conditions. Instead, there is only one dataset that was aligned to two different assemblies, which implies that biological and technical variation is not present. Therefore, we do not attempt to correct for overdispersion, which is a concern when identifying differential expression of genes between two conditions. As such, a χ^2 test with a significance value α of 0.05 was performed to detect peaks with skewed coverage. We obtained an identical result with the edgeR package [21] by setting the dispersion parameter near 0. Peaks with null coverage in one of the alignment versions were artificially assigned one read to allow applying the test.

Peaks with insignificant differences were placed in the **no-skew** category. Peaks that had significant differences with a higher coverage in hg19 than in the personalized haplotype were categorized as **ref-skewed**. Similarly, peaks that had a higher coverage in the personalized haplotype than in hg19 were categorized as **personal-skewed**.

An overview of the above steps can be found in figures 1b and 1a.

Quantifying altered peak calls and probabilities

To quantify the fraction of AP calls, the number of ref-only and personal-only peaks was counted and then divided by the total number of peaks to obtain their frequency relative to the total number of peaks in their sample.

For each sample, the set of all peaks was divided into mutually exclusive categories according to the combination of overlapping variation calls (SNPs only, indels only, SNPs and indels, none). The same was repeated for ref-only and personal-only peaks.

For any given variation category, the counts of ref-only and personal-only peaks were divided by the sample wide peak count of the given category to obtain the probability that the peak call could be affected by that specific combination of variations. At the same time, the mean peak widths were recorded.

The above procedure was also repeated for untrimmed, full read runs to evaluate the influence of read length.

Comparing hg19 and Hap1 WGS alignments

If peak track differences occur between two assemblies, they should be corroborated by differences in the mapping of a sufficient number of reads between their raw alignments. That is, the proportion of reads with different mappings between hg19 and Hap1 should be sufficiently high to account for the number of ref-only and Hap1-only peaks. To show this, we used *Jvarkit cmpbamsandbuild* [15] to compare the hg19 and Hap1 alignments of the low pass NA12878 whole genome dataset. Reads with coordinate differences of less than 100 bp were considered to have identical mapping between the two builds. The IGSR WGS dataset was chosen instead of a ChIP-seq dataset because we expect a more uniform coverage of genomic regions. The same comparison was done between hg19 and the paternal NA12878 MPG to help contrast the results between MPGs and DPGs.

Finding the agreement between MPGs and DPGs

To get the agreement between the DPG and the MPG approaches, the personalized tracks needed to be lifted to a common coordinate system in hg19. This is necessary because the MPG APs were computed in MPG coordinates, while the DPG APs were computed in hg19 coordinates.

To do so, chain files were created through the previous BLAT method to lift the MPGs to hg19. Once the tracks of ref-only and personal-only peaks respective to the MPGs was lifted to hg19, *bedtools intersect* was used to find which peak called regions belonged to the same peak category in both approaches. The personal-only tracks of both approaches were intersected to produce an agreement track of peaks that were categorized as personal-only in both approaches. Similarly, the ref-only tracks of both approaches were intersected to produce an agreement track of peaks that were categorized as ref-only in both approaches.

Additionally, the tracks of common, ref-only and personal-only peaks in the DPG approach were intersected with the hg19 variation call track using *bedtools intersect -c*. This allowed to observe whether ref-only peaks and personal-only peaks remained enriched in hg19-relative variation calls compared to common peaks, despite the fact that they originate from peak calls in a denovo assembly and not hg19 itself.

Results

Liftover efficacy is acceptable in MPGs and DPGs

Since the type of personalized genome may affect lifting efficacy and the above procedures are highly dependent of the *liftOver* process, it is crucial to know how well the generated chain files lift between assemblies. After all, MPGs are much more similar to the reference than DPGs, so it would be unreasonable to expect an identical lifting efficacy. In addition, MPG chains were produced by *vcf2diploid* while the DPG chains were produced from an alignment with the reference.

As anticipated, lifting from the reference to an MPG yields fewer unlifted peaks than lifting from DPGs to the reference (Fig 1c). Nevertheless, both approaches remain in an acceptable window, with well below 1% of peaks failing to lift, with

the exception of the Pendleton assembly. Because the fraction of unlifted peaks is much smaller than the fraction of AP calls in both MPGs and DPGs, the results should remain valid.

MPGs create altered peaks of low confidence

To reiterate, AP calls take the form of personal-only and ref-only calls. Personal-only calls occur when an insignificant read pileup in a reference region acquires new reads in the personalized genome and becomes significant to the peak caller. Ref-only peaks are obtained by a reverse process where significant read pileups in a reference region loses reads in the personalized genome and are no longer called by MACS2.

The first object of interest is the number of AP calls found in this best case scenario with the MPG method. Since NA12878 is a genome with an unusually high coverage, its variation calls should be much more confident and comprehensive than the average sample encountered in datasets such as Blueprint. Knowing this result gives a good upper bound to the expected impact of MPGs on ChIP-seq peak calls. As Table 1 indicates, their numbers amount to 1% of the total number of peaks for each respective mark, with the vast majority of peaks being shared between the reference and the MPG. In addition, personal-only peaks are found at double the rate of ref-only peaks.

While the number of changed peak calls is not negligible, their quality is not representative of the typical peak. Personal-only peaks (Fig S2a) and ref-only peaks (Fig S2b) are confined to a region of narrower width and lower confidence than most common peaks. Larger and higher confidence peaks are rarely observed, if at all, which exposes the limitations of MPGs in accounting for large scale differences between a personalized genome and the reference.

Another informative metric is the number of reads overlapped by peaks in the MPG alignment compared to the reference alignment. This should reveal peaks that have a higher coverage in one of the alignments due to the presence of variations. Ideally, peaks that have AP calls should also have a skewed coverage. However, this is true for only a small number of cases. For most AP calls, their coverage distribution remains clustered within the distribution of the common and unaffected peaks calls (Fig 2a). At the same time, most affected peak calls fall into the no-skew category (Fig 2b) together with common calls, with a few peaks having a coverage skewed toward the reference or the MPG (Table S3). Looking at the WGS alignment differences between hg19 and the paternal NA12878 MPG, we observe only 3.7% of reads that change their mapping (Table S2b). This suggests that it may not be possible to generate large and confident AP calls by moving the mapping of such a small proportion of reads.

When the method is extended from a benchmark sample with rich and high quality calls to typical Blueprint samples with fewer and lower quality calls, the number of APs decreases as expected. In Blueprint, the average number of personal-only peaks is roughly half (%0.5 to 0.7%) compared to NA12878 (1.1%) (Table 1), with a few exceptions that surpass the benchmark (Fig S4a, S4b). Again, ref-only peaks occur less often than personal-only peaks.

A similar effect is also observed with skewed peaks. While not numerous in the benchmark to begin with (50 to 70), their number in the typical Blueprint sample barely reaches ten peaks (Fig S4c, S4d). As to the quality of altered Blueprint peaks, the same patterns and clustering of width, confidence and coverage observed in NA12878 are found again in Blueprint samples (Fig S6). The small differences

in coverage together with the weak confidence of APs, indicates that this approach can only alter the calls of regions that are very near the threshold of significance. Despite this, low quality peaks are not unique to this analysis. They appear in almost every experiment that involves peak calling and knowing what can alter their calling may still prove useful. A starting point would be to find why NA12878 produces more AP calls than Blueprint.

Due to the difference in WGS coverage, the set of variant calls in NA12878 is much richer than in Blueprint (Fig 3a) This can explain why Blueprint MPGs record fewer AP calls than the NA12878 MPGs. We confirmed this by creating a NA12878 MPG by downsampling the original set to 2.6M SNVs and 100K indels. As shown in Table 1, the downsampled MPG produced fewer AP calls relative to the full set for both H3K4me1 and H3K27ac marks. We should also keep in mind that the phasing of NA12878 variant calls is better than Blueprint variant calls, which could contribute better haplotypes and more AP calls.

Altered peaks in MPGs are driven by indels

Since the number of peaks altered by an MPG seems to be dependent on the number of variation calls available for its generation, measuring the impact of variations as a probability is more useful.

The likelihood that a peak call is altered depends on the combination of variants present in its region. We binned AP calls according to the overlapped combination of variants (see page 5) to bring out the main types of variants that are associated with AP calls in MPGs. Peak calls that do not contain variations have a near zero chance of being altered. Peaks overlapping at least one indel are the most likely to be altered followed by peaks overlapping at least one SNP. Interestingly, peaks containing at least one SNP and indel are the least likely to be altered (Fig 3b, 3c).

A factor that could explain these trends is the peak width associated to each peak category and histone mark. Since MPGs only contain small scale substitutions such as SNPs and indels, any differences in their alignment from the reference alignment are also expected to be of small scale. Therefore, large scale features in the coverage, including wider peaks, should be relatively unaffected by variations. It is notable that the mean width of peaks overlapping both indels and SNPs is the highest among the four combinations of variations, followed by peaks with at least one indel and peaks with at least one SNP. This should be expected, since peaks that span several variation calls are wider on average than peaks that only span one variation call. Similarly, H3K4me1 is known to be a wider mark than H3K27ac, which is reflected in the figure. Given these details, it should be clear why indels and SNPs do not interact in a synergistic fashion, in addition to why the probabilities are lower in H3K4me1 samples.

Concerning the type of histone marks, variation calls are less likely to alter H3K4me1 peaks than H3K27ac peaks, while preserving the same general trend. This holds true for personal-only and ref-only peaks. These reduced probabilities associated with ref-only peaks are consistent with their lower frequency relative to personal-only peaks.

DPGs can create confident altered peaks

If the moderate impact of MPGs on ChIP-seq calls is explained by the lack of larger scale variations in the sequence, then denovo assemblies, which do include such features, should have a larger effect. Indeed, the obtained altered peak calls

are roughly five times more numerous (Table 1) when using a DPG instead of an MPG. The increase also extends to skewed peaks, with their numbers reaching over three thousand peaks instead of several tens (Table S3).

Ref-only peaks are no longer characterized by narrow width and low confidence, but instead have a distribution identical to common peaks (Fig S3b). Although personal-only peaks do not reach an identical distribution as common peaks, there are considerable gains in terms of width and quality (Fig S3a). On top of low quality peaks already seen in MPGs, DPGs are able produce APs of higher confidence and width.

The reference coverage and personalized peak coverage are now well distinguished (Fig 4a), with many AP calls passing the significance threshold (Fig 4b). If these improvements are not from false positive peaks, DPGs may surmount the main limitation of MPGs when it comes to larger scale variations.

When summarizing the fate of individual WGS reads between the two assemblies, a consistent result appears (Table S2a). Roughly 10% of reads change their mapping in way that would lead to APs. This is a three fold increase from the equivalent analysis with MPGs. In the context of ChIP-seq, if 10% of reads were to similarly change their mapping between Hap1 and hg19, then they should account for the number of changed peaks.

In some extreme cases, APs register coverage in one version of the alignment but a null coverage in the other. These null coverage peaks could occur due to regions that are represented in one assembly but absent in the other. Such peaks are almost exclusively ref-only (null ref-only peaks), presumably because the reference is longer and more complete than a denovo assembly.

Segmental duplications (SDs) are known to be underrepresented in denovo assemblies [4] and could be the root of many AP call events. To elucidate this issue, we selected the most confident subset of Hap1-only and ref-only calls by removing any peak with a log(MACS2 score) < 4.0 This value was chosen because it excludes uncertain and uninteresting peak calls and most APs generated by MPGs. In the end, we reduce the initial set to only 828 Hap1-only and 2064 ref-only peaks. The confident subset of APs was then overlapped with a SD annotation. We also looked at the occurrence of SDs in null ref-only peaks because they may be caused by SDs that are not present in a denovo assembly.

Indeed, a majority (71.3%) of confident null ref-only peaks overlap SDs (Table S1), suggesting that these regions are not present in the 10X Genomics denovo assembly of NA12878 because no Hap1 reads lift to those intervals. This holds true even after including uncertain peak calls. As for ref-only peaks with positive Hap1 coverage, they register much fewer SDs (6.3%), although they become more enriched when excluding uncertain peaks (17.1%).

For confident Hap1-only peaks, more than half are associated with SDs (Fig 4c). This means that only 349 peaks out of the initial seven thousand Hap1-only peaks are sufficiently confident and located in regions free of SDs. Despite the more stringent criteria, these numbers are still comparable to the MPG results. After sorting the SD free Hap1-only peaks by confidence and read count, we notice that the top peak shows large differences between the hg19 alignment and the lifted Hap1 alignment (Fig S8b). In light of this result, the differences in MPGs are minor (Fig S8a).

Along the same lines, we overlapped the APs with the RepeatMasker annotation to detect enrichment of transposable elements in confident Hap1-only and ref-only peaks relative to the entire genome (RepeatMasker). The frequency of each repeat family for each region is the percentage of all overlapping repeats that belong to

said repeat family. The results show a 2 fold enrichment of confident Hap1-only peaks in Alu elements relative to their genome wide abundance (Fig 4d). Confident ref-only peaks show 1.5 fold enrichment in Alus. The same is not true for repeat families such as L1 and L2, which occur equally or less often relative to the genome.

There exists a small confident subset of 46 Hap1-only and 115 ref-only peaks that are free of both SDs and repeats. Despite the absence of known repeats or segmental duplications, these peaks can still have large differences in coverage between Hap1 and reference alignments (Fig S9).

MPG altered peaks replicate poorly in DPGs

If MPGs can account for small scale variations and DPGs can account for both small and large scale variations, then the AP calls obtained with the MPG approach should be a at least a partial subset of the AP calls obtained with the DPG approach.

When the AP call annotations of the two approaches are intersected, a small agreement is observed for personal-only peaks and ref-only peaks. Specifically, a tenth of personal-only peak calls in the MPGs are found again in the Hap1 and Hap2 DPGs (Fig S1b). Similarly, a tenth of ref-only peaks are recovered in this fashion. When intersecting Hap1-only, Hap2-only, paternal-only and maternal-only peak annotations, almost the same proportion of agreed personal-only peaks is retained, with around 50 peaks being dropped. Meanwhile, less than half of agreed ref-only peaks survive the same intersections. This low rate of replication could be explained by the uncertain nature of MPG APs.

The same overlaps can be performed between the two haploids of the same sample. Looking at Table 1, one wonders how many ref-only and personal-only calls occur in both haploids. To answer this question, personalized to hg19 chain files were generated for NA12878 and a small Blueprint sample to allow identifying shared personal-only and ref-only calls between haploids (Table S4). The ratio of replication of ref-only and personal-only calls between haploids are nearly identical between Blueprint and the NA12878 benchmark. This ratio is higher for personal-only calls compared to ref-only calls.

In the 10X DPG samples, the ratio of replication is higher for ref-only calls compared to personal-only calls, and overall much higher than in MPGs. This is expected, knowing that many ref-only peaks would be produced in hg19 intervals that are absent in both Hap1 and Hap2.

Another way of connecting the two approaches is to determine if AP calls in DPGs are enriched in hg19-relative variation calls in the same way as AP calls in MPGs. To demonstrate this, common and AP calls found in DPGs were lifted to hg19 and intersected with the hg19 SNP and indel annotations. The result is that DPG-only APs calls are enriched in variations compared to DPG common peak calls (Fig S1a). Specifically, Hap1-only peaks have a higher mean SNP and indel density compared to common peaks. As for ref-only peaks, they are only slightly enriched in variation calls.

Longer reads minimize the impact of MPGs

The initial analysis trimmed every sample to a read length of 36 bp to remove any influence of read length by making it uniform across all datasets. Although the short reads simplified the quantification and interpretation of the results, they are no longer representative of current datasets. To discern the relevance of personalized

genomes to longer reads produced by current technologies, the behavior of the MPGs and DPGs needs to be evaluated with untrimmed reads.

It is important to know that as the read length of a ChIP-seq dataset increases, the previous probabilities associated with MPGs become smaller. That is, if the reads are not trimmed, the impact of substituted sequences in MPGs diminishes. In fact, the number of APs and the associated probabilities are halved simply by using untrimmed reads (Fig 5a). This suggests that any use of MPGs could be made redundant by performing ChIP-seq experiments with longer reads.

Read length also affects the impact in DPGs, albeit not in the same uniform manner (Fig 5b). Full reads increase the number personal-only peaks but decrease the number of ref-only peaks in Hap1/Hap2 and Pendleton. The number of ref-only peaks rises quite abruptly in the trimmed versions of the samples. This appears to be due to the hg19 alignment registering up to 5000 more peaks than its denovo equivalents. The excess of calls relative to the denovo alignments then leads to a large number of ref-only calls. Despite these fluctuations, the impact remains strong in all cases, which cannot be said for MPGs.

Discussion

The above results show that the use of MPGs cannot recover any peaks of interest that are well above the noise in ChIP-seq signals. If a higher significance threshold is imposed during peak calling, most of the APs would be lost. Moreover, their uncertain nature makes validation by concordance with a biological replicate unlikely. Since increasing read length decreases the impact of MPGs, the small number of APs is expected to be even lower in a typical ChIP-seq dataset. Consequently, the lower effort involved in generating MPGs becomes irrelevant because it fails to bring tangible benefits to a ChIP-seq analysis.

In contrast, DPGs recover a broad range of peaks that are likely to survive more stringent filters and validation in biological replicates. On top of improved peak quality, they also produce several times more APs than the best available MPG. Unlike MPGs, their impact is not trivialized by increasing read length and should still yield comparable results in a modern dataset.

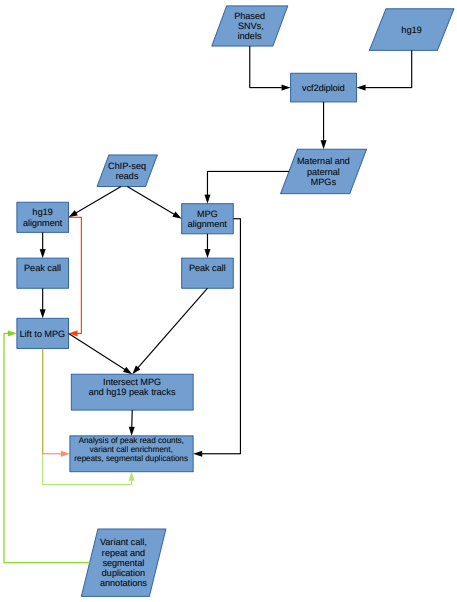
Since unlifted peaks are a byproduct of this analysis, their role needs to be addressed. Given that the direction of lifting was not identical, interpreting unlifted peaks must be done separately for MPGs and DPGs. In the case of MPGs, lifting was done from the reference to the MPG. Therefore, it is possible that the the reference counterparts for some MPG peaks are among the unlifted peaks. During track comparison, such MPG peaks would be wrongly classified as personal-only. The second possibility is that an unlifted peak is a ref-only peak that could not be placed in the MPG. Because the frequency of unlifted peaks (<0.01%) is much lower than the frequencies of personal-only (>0.5%) and ref-only peaks (>0.30%), the number of false personal-only or unplaced ref-only peaks is relatively small and can be neglected.

In the case of DPGs, lifting was done from the DPG to the reference for the practical reason of reusing hg19 annotations that would lift poorly to a denovo assembly. In terms of false altered calls or unplaced peaks, the opposite phenomenon occurs. If the counterpart of a reference peak does not lift from the DPG, then a false ref-only peak is obtained. Conversely, an unlifted peak could be a personal-only peak that could not be placed in the reference. A close look at high coverage APs with similar reference and Hap1 read numbers does reveal false ref-only peaks.

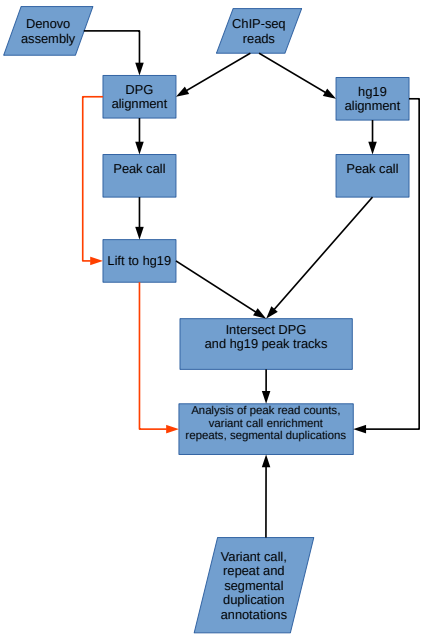
As predicted above, the equivalent peak call can be retrieved from the Hap1 unlifted peak list.

In conclusion, there is no credible benefit to replacing the reference with MPGs in ChIP-seq pipelines. On the other hand, DPGs could provide a slightly more accurate picture, especially when wishing to account for large structural differences from the reference assembly. For example, many tumor genomes feature widespread chromosomal changes that can be captured with denovo assembly. Even if denovo assembly is not possible, existing methods based on resequencing can reorganize the reference to reflect tumor genomes [17]. It is possible that such a genome could partially replace denovo assembly.

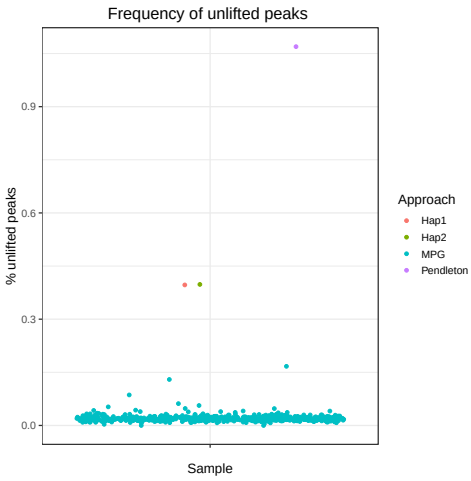
Figures



(a)

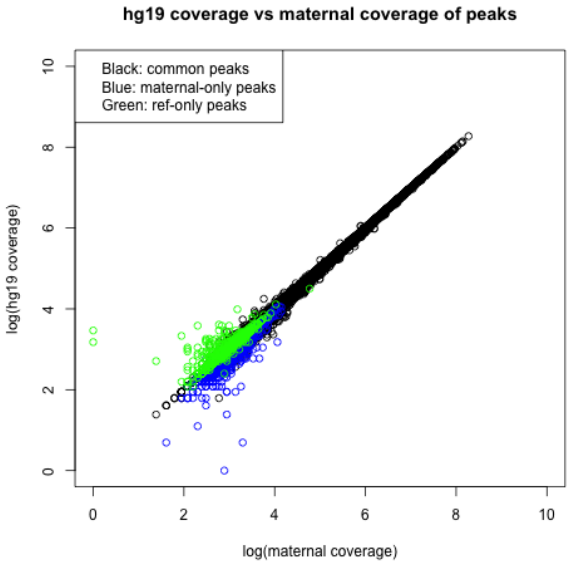


(b)

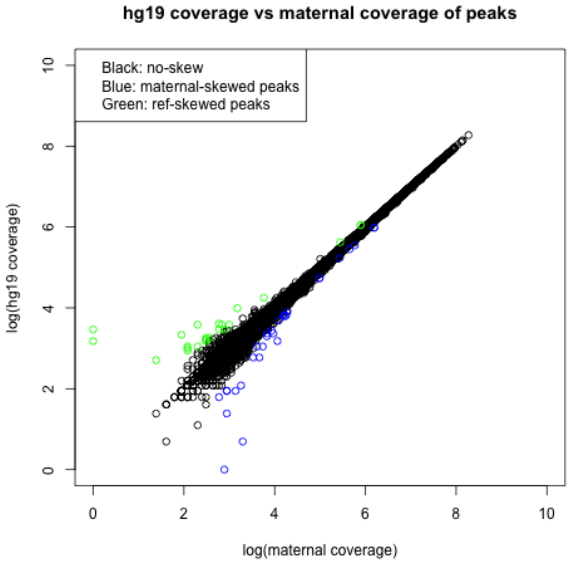


(c)

Figure 1: Flow chart of analysis for a) modified personal genomes and b) denovo personal genomes. c) Fraction of peaks that failed to lift from their native assembly to the target assembly across all samples.



(a)



(b)

Figure 2: Characterization of peak calls in hg19 versus the maternal MPG of NA12878. a) A comparison of the coverage of peak called regions in hg19 and the maternal genome. b) Identification of peak called regions that have a significant difference in coverage.

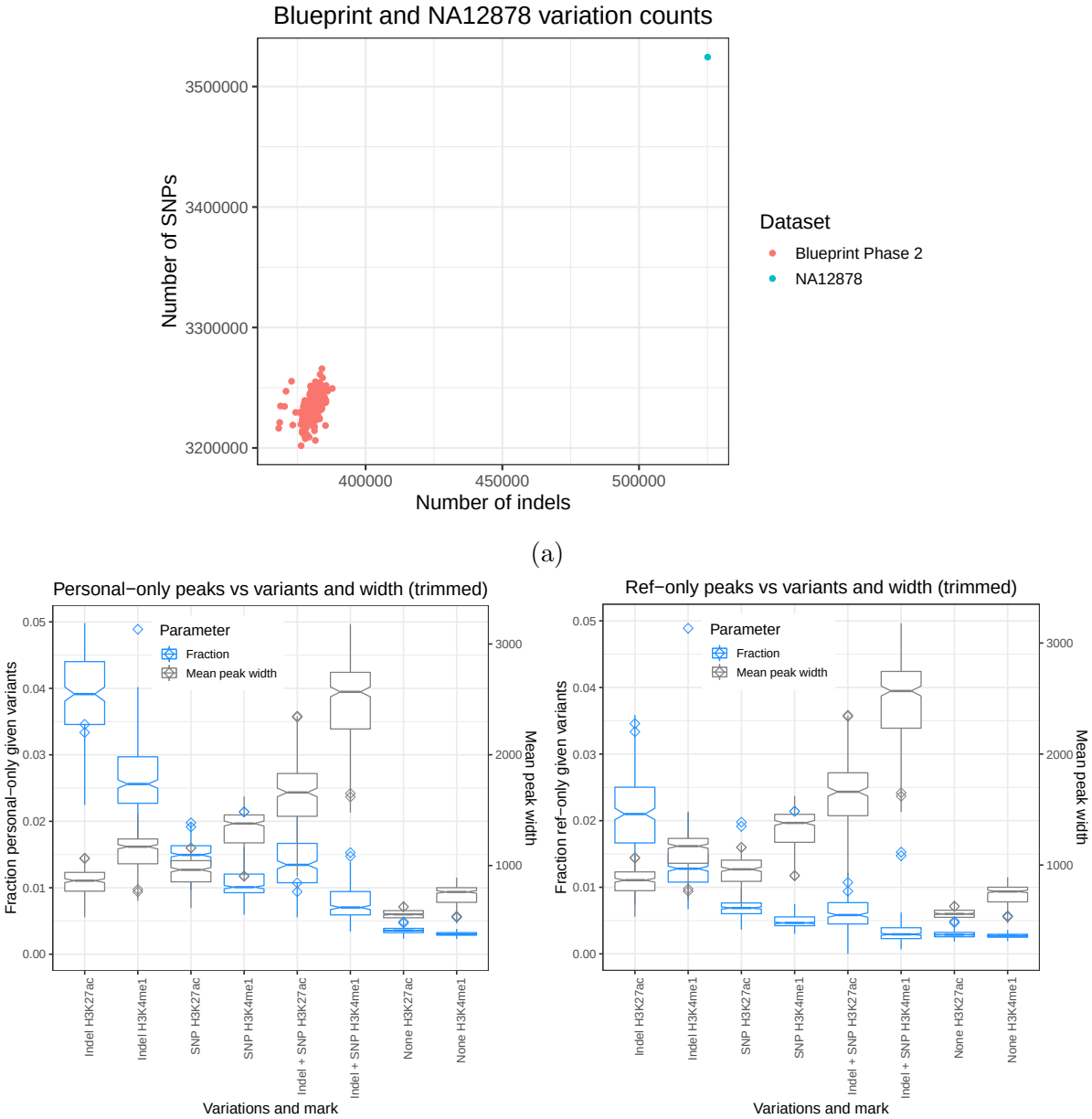


Figure 3: a) Difference of number in variation calls (SNPs and indels) between Blueprint samples and NA12878. b) An estimate of the probability that each combination of variation calls present in a region may cause a personal-only peak call compared to their average widths. c) Same estimates for ref-only events. The indicated points are the values for the NA12878 MPGs for fraction and width respectively.

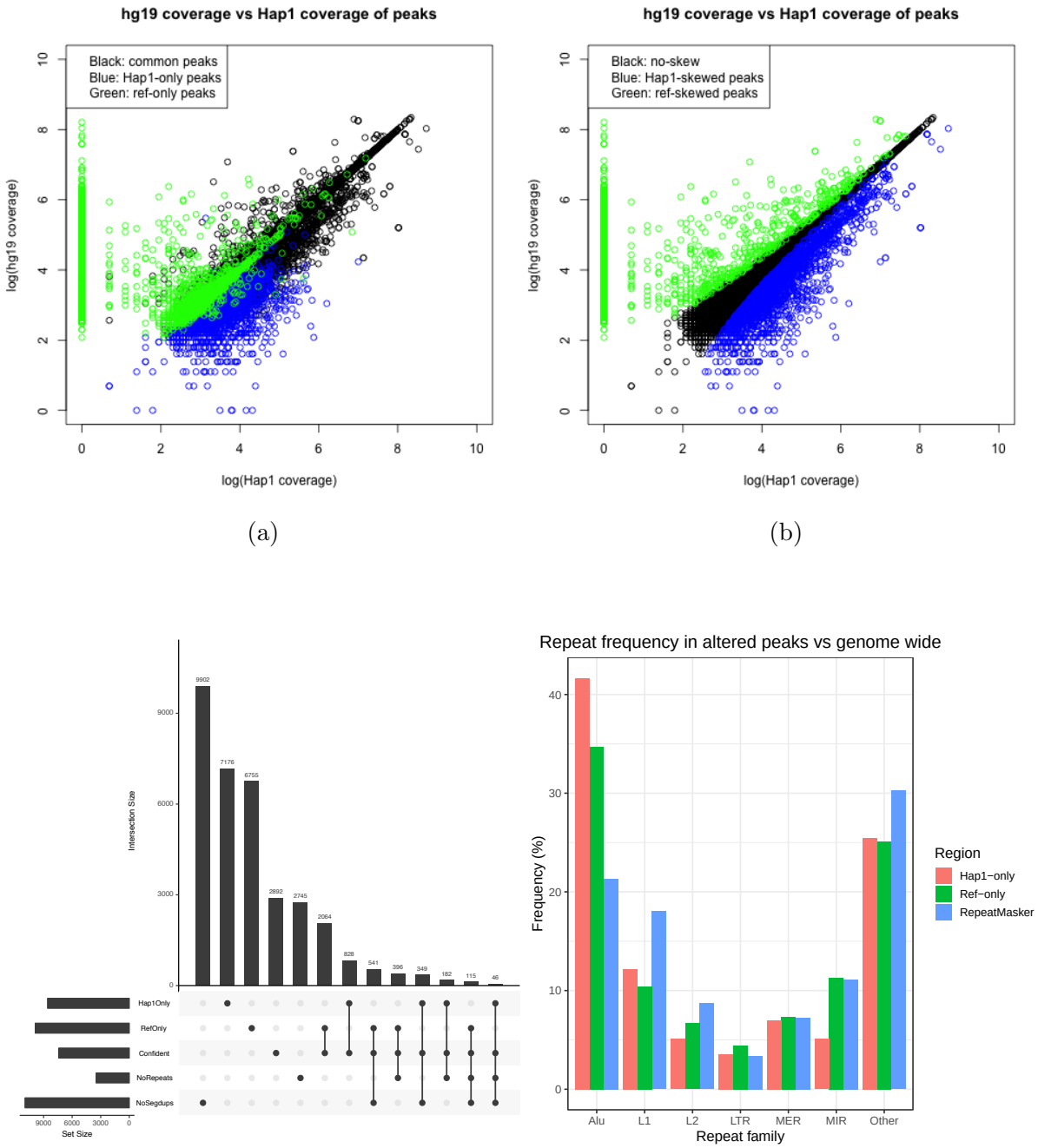
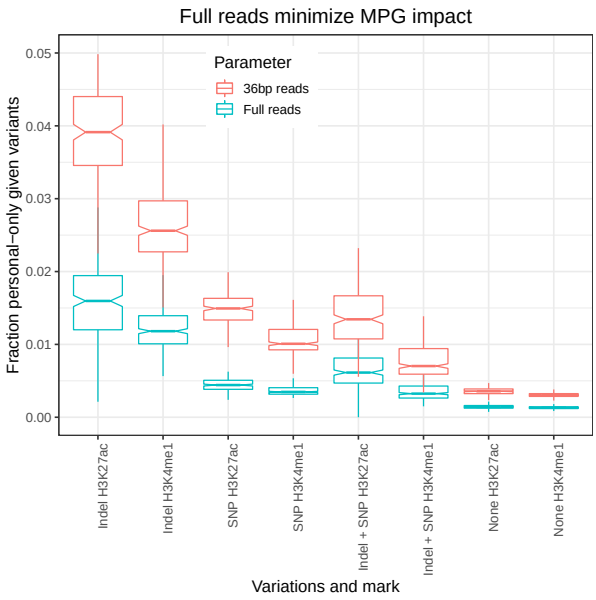
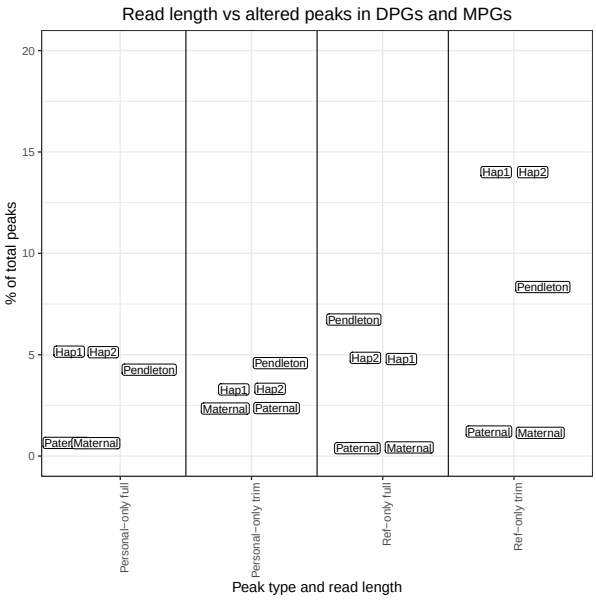


Figure 4: Characterization of peak calls in hg19 versus the denovo assembled Hap1 pseudohap of NA12878. a) A comparison of the coverage of peak called regions in hg19 and Hap1. b) Identification of peak called regions that have a significant difference in coverage. c) Summary of the overlap between altered peaks, confident peaks, repeats and segmental duplications [7]. d) The repeats that overlap altered peaks are enriched in Alu elements relative to their frequency in the RepeatMasker. The categories were chosen by grouping repeats by name prefix, summing their frequencies per group and taking the largest groups. Remaining groups were labeled as "other".



(a)



(b)

Figure 5: a) The probability that a variation affects a peak called on full reads is lower compared to trimmed reads. b) While read length does change the impact of DPGs, the altered peak frequencies remain large.

Tables

Dataset	Version	Mark	personal-only	ref-only	common
NA12878	Paternal MPG	H3K4me1	1636 (1.1%)	854 (0.6%)	146520
NA12878	Maternal MPG	H3K4me1	1622 (1.1%)	808 (0.6%)	146570
NA12878	Downsampled MPG	H3K4me1	1051 (0.7%)	550 (0.4%)	146688
Blueprint	Paternal MPG	H3K4me1	746 (0.6%)	456 (0.4%)	129496
Blueprint	Maternal MPG	H3K4me1	747 (0.6%)	455 (0.4%)	129497
Hap1	DPG	H3K4me1	7176 (4.8%)	6755 (4.6%)	141444
Hap2	DPG	H3K4me1	7130 (4.8%)	6774 (4.6%))	141442
Pendleton	DPG	H3K4me1	16245 (10.2%)	8912 (5.8%)	142347
NA12878	Paternal MPG	H3K27ac	660 (1.0%)	351 (0.5%)	68888
NA12878	Maternal MPG	H3K27ac	688 (1.0%)	335 (0.5%)	68909
NA12878	Downsampled MPG	H3K27ac	438 (0.6%)	218 (0.3%)	68953
Blueprint	Paternal MPG	H3K27ac	336 (0.7%)	196 (0.4%)	46830
Blueprint	Maternal MPG	H3K27ac	335 (0.7%)	195 (0.4%)	46831
Hap1	DPG	H3K27ac	2078 (3.2%)	9901 (13.5%)	63419
Hap2	DPG	H3K27ac	2091 (3.2%)	9899 (13.5%)	63441
Pendleton	DPG	H3K27ac	5208 (7.2%)	4980 (6.9%)	66811

Table 1: Number of altered peak calls in DPGs, MPGs for the H3K4me1 and H3K27ac marks. The numbers for Blueprint MPGs are the average of all samples. A NA12878 MPG that was created from downsampled variant calls is also included.

References

[1] Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. May 2013. URL: http://arxiv.org/abs/1303.3997.

[2] Keith A. Baggerly, Li Deng, Jeffrey S. Morris, and C. Marcelo Aldaz. "Differential expression in SAGE: accounting for normal between-library variation". In: Bioinformatics 19.12 (Aug. 2003), pp. 1477–1483. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btg173. URL: http://dx.doi.org/10.1093/bioinformatics/btg173.

[3] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data." In: Bioinformatics (Oxford, England) 30.15 (Aug. 2014), pp. 2114–2120. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btu170. URL: http://dx.doi.org/10.1093/bioinformatics/btu170.

[4] Mark J. P. Chaisson, Richard K. Wilson, and Evan E. Eichler. "Genetic variation and the de novo assembly of human genomes". In: Nature Reviews Genetics 16 (Oct. 2015). DOI: 10.1038/nrg3933. URL: http://dx.doi.org/10.1038/nrg3933.

[5] Lu Chen et al. "Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells". In: Cell 167.5 (Nov. 2016), 1398–1414.e24. ISSN: 00928674. DOI: 10.1016/j.cell.2016.10.026. URL: http://dx.doi.org/10.1016/j.cell.2016.10.026.

[6] Laura Clarke, Susan Fairley, Xiangqun Zheng-Bradley, Ian Streeter, Emily Perry, Ernesto Lowy, Anne-Marie Tass, and Paul Flicek. "The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data". In: Nucleic Acids Research (Sept. 2016), gkw829+. ISSN: 1362-4962. DOI: 10.1093/nar/gkw829. URL: http://dx.doi.org/10.1093/nar/gkw829.

[7] Jake R Conway, Alexander Lex, and Nils Gehlenborg. "UpSetR: An R Package For The Visualization Of Intersecting Sets And Their Properties". In: bioRxiv (Jan. 2017). DOI: 10.1101/120600. URL: http://biorxiv.org/content/early/2017/03/25/120600.abstract.

[8] Aarti Desai, Veer S. Marwah, Akshay Yadav, Vineet Jha, Kishor Dhaygude, Ujwala Bangar, Vivek Kulkarni, and Abhay Jere. "Identification of Optimum Sequencing Depth Especially for De Novo Genome Assembly of Small Genomes Using Next Generation Sequencing Data". In: PLoS ONE 8.4 (Apr. 2013), e60204+. DOI: 10.1371/journal.pone.0060204. URL: http://dx.doi.org/10.1371/journal.pone.0060204.

[9] Ian Dunham et al. "An integrated encyclopedia of DNA elements in the human genome." In: Nature 489.7414 (Sept. 2012), pp. 57–74. ISSN: 1476-4687. DOI: 10.1038/nature11247. URL: http://dx.doi.org/10.1038/nature11247.

[10] Michael A. Eberle et al. "A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree". In: Genome Research 27.1 (Jan. 2017), pp. 157–164. ISSN: 1549-5469. DOI: 10.1101/gr.210500.116. URL: http://dx.doi.org/10.1101/gr.210500.116.

[11] 10x Genomics. NA12878 10X Genomics Assembly. 10X Genomics, May 2016. URL: https://support.10xgenomics.com/de-novo-assembly/datasets.

[12] J. J. Kent. "BLAT - the BLAST-like alignment tool". In: Genome Research 12.4 (Apr. 2002), pp. 656–664. ISSN: 1549-5469. DOI: 10.1101/gr.229202. URL: http://dx.doi.org/10.1101/gr.229202.

[13] W. James Kent, Robert Baertsch, Angie Hinrichs, Webb Miller, and David Haussler. "Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes". In: Proceedings of the National Academy of Sciences 100.20 (Sept. 2003), pp. 11484–11489. ISSN: 1091-6490. DOI: 10.1073/pnas.1932072100. URL: http://dx.doi.org/10.1073/pnas.1932072100.

[14] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. "The Sequence Alignment/Map format and SAMtools." In: Bioinformatics (Oxford, England) 25.16 (Aug. 2009), pp. 2078–2079. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp352. URL: http://dx.doi.org/10.1093/bioinformatics/btp352.

[15] Pierre Lindenbaum. "JVarkit: java-based utilities for Bioinformatics". In: (May 2015). DOI: 10.6084/m9.figshare.1425030.v1. URL: https://figshare.com/articles/JVarkit_java_based_utilities_for_Bioinformatics/1425030.

[16] Steven C. Munger et al. "RNA-Seq Alignment to Individualized Genomes Improves Transcript Abundance Estimates in Multiparent Populations". In: Genetics 198.1 (Sept. 2014), p. 59. DOI: 10.1534/genetics.114.165886. URL: http://www.genetics.org/content/198/1/59.abstract.

[17] Layla Oesper, Anna Ritz, Sarah J. Aerni, Ryan Drebin, and Benjamin J. Raphael. "Reconstructing cancer genomes from paired-end sequencing data." In: BMC bioinformatics 13 Suppl 6.Suppl 6 (Jan. 2012). DOI: 10.1186/1471-2105-13-S6-S10. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3358655&tool=pmcentrez&rendertype=abstract.

[18] Matthew Pendleton et al. "Assembly and diploid architecture of an individual human genome via single-molecule technologies". In: Nat Meth 12.8 (Aug. 2015), pp. 780–786. DOI: 10.1038/nmeth.3454. URL: http://dx.doi.org/10.1038/nmeth.3454.

[19] Picard Tools. URL: http://broadinstitute.github.io/picard/.

[20] Aaron R. Quinlan and Ira M. Hall. "BEDTools: a flexible suite of utilities for comparing genomic features". In: Bioinformatics 26.6 (Mar. 2010), pp. 841–842. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btq033. URL: http://dx.doi.org/10.1093/bioinformatics/btq033.

[21] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." In: Bioinformatics (Oxford, England) 26.1 (Jan. 2010), pp. 139–140. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp616. URL: http://dx.doi.org/10.1093/bioinformatics/btp616.

[22] Joel Rozowsky et al. "AlleleSeq: analysis of allele-specific expression and binding in a network framework." In: Molecular systems biology 7.1 (Aug. 2011). ISSN: 1744-4292. DOI: 10.1038/msb.2011.54. URL: http://dx.doi.org/10.1038/msb.2011.54.

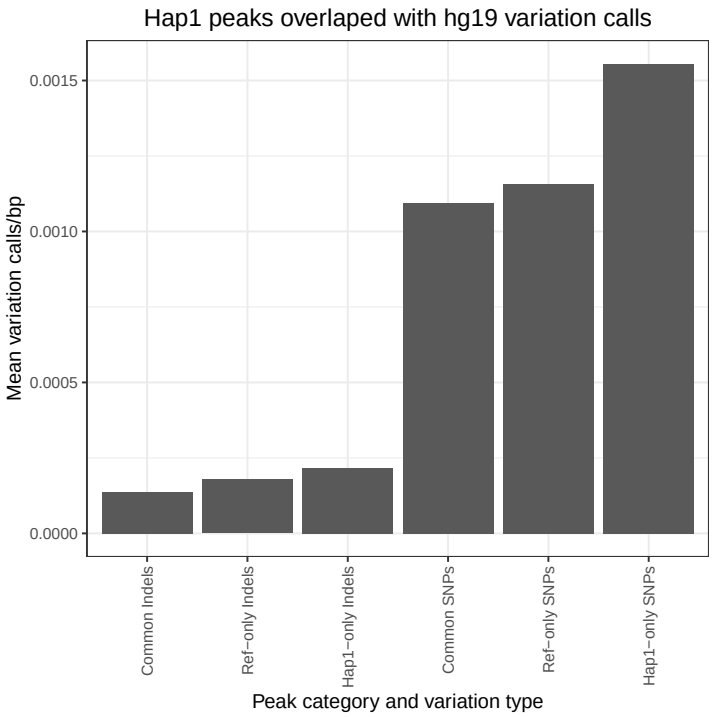
[23] Wenqiang Shi, Oriol Fornes, Anthony Mathelier, and Wyeth W. Wasserman. "Evaluating the impact of single nucleotide variants on transcription factor binding". In: Nucleic Acids Research 44.21 (Dec. 2016), pp. 10106–10116. ISSN: 0305-1048. DOI: 10.1093/nar/gkw691. URL: http://dx.doi.org/10.1093/nar/gkw691.

[24] Phillip Wulfridge, Ben Langmead, Andrew P. Feinberg, and Kasper D. Hansen. "Choice of reference genome can introduce massive bias in bisulfite sequencing data". In: 2016. DOI: https://doi.org/10.1101/076844. URL: http://dx.doi.org/https://doi.org/10.1101/076844.

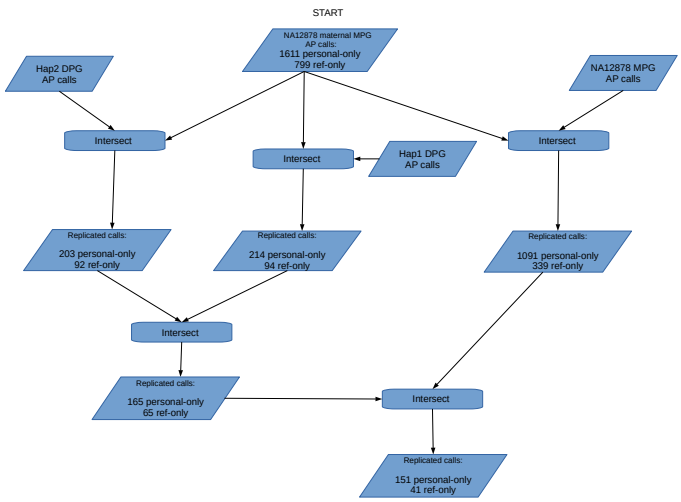
[25] Wenyu Zhang, Jiajia Chen, Yang Yang, Yifei Tang, Jing Shang, and Bairong Shen. "A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies". In: PLoS ONE 6.3 (Mar. 2011), e17915+. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0017915. URL: http://dx.doi.org/10.1371/journal.pone.0017915.

[26] Yong Zhang et al. "Model-based Analysis of ChIP-Seq (MACS)". In: Genome Biology 9.9 (Sept. 2008), R137+. ISSN: 1465-6906. DOI: 10.1186/gb-2008-9-9-r137. URL: http://dx.doi.org/10.1186/gb-2008-9-9-r137.

Supporting material

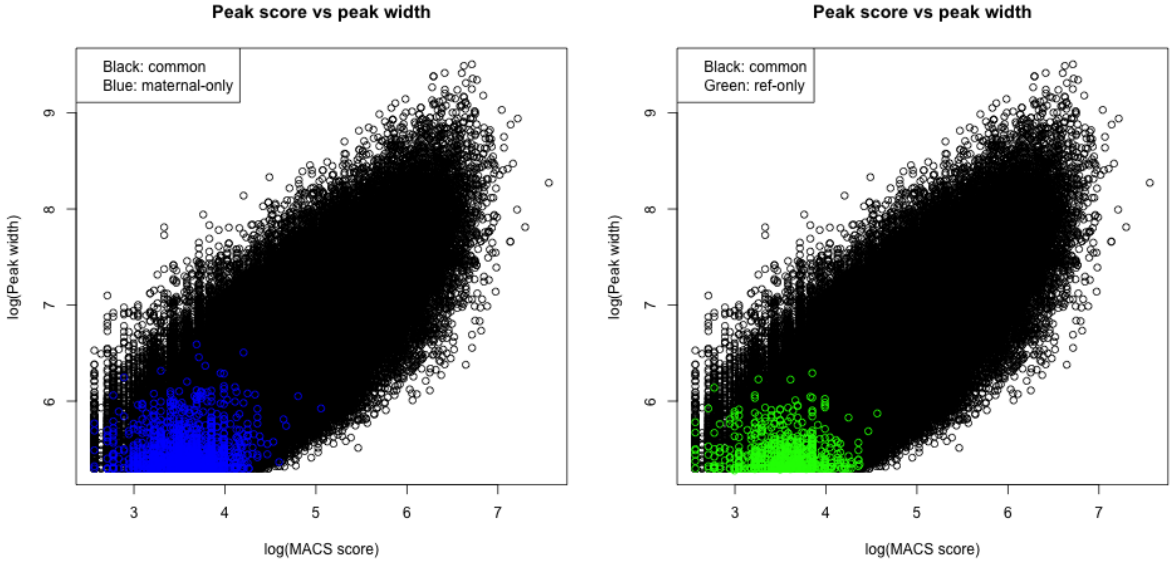


(a)



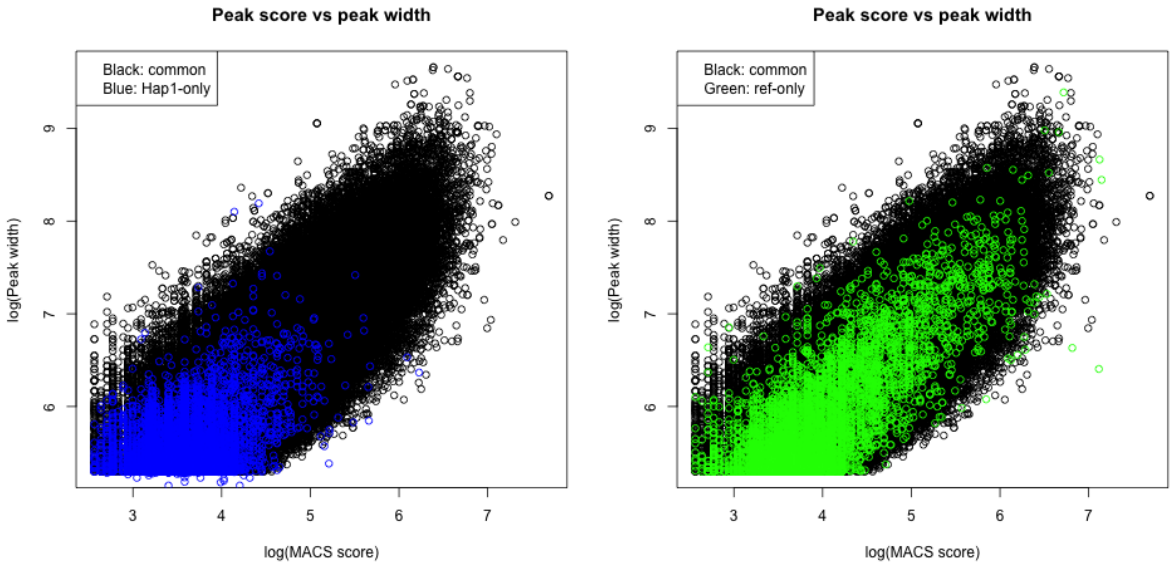
(b)

Figure S1: a) Hap1-only peaks are enriched in hg19-relative variant calls relative to common peaks. b) Intersections between the NA12878 maternal MPG, the paternal MPG and DPGs. The replicated peaks are a subset of the personal-only and ref-only peaks of the maternal MPG.



(a) (b)

Figure S2: Confidence and width of the NA12878 maternal MPG altered peak calls (H3K4me1). a) Peaks called only in the maternal haplotype against the common peak background. b) Peaks called only in hg19 against the common peak background.



(a) (b)

Figure S3: Confidence and width of the Hap1 DPG altered peak calls (H3K4me1). a) Peaks called only in Hap1 against the common peak background. b) Peaks called only in hg19 against the common peak background.

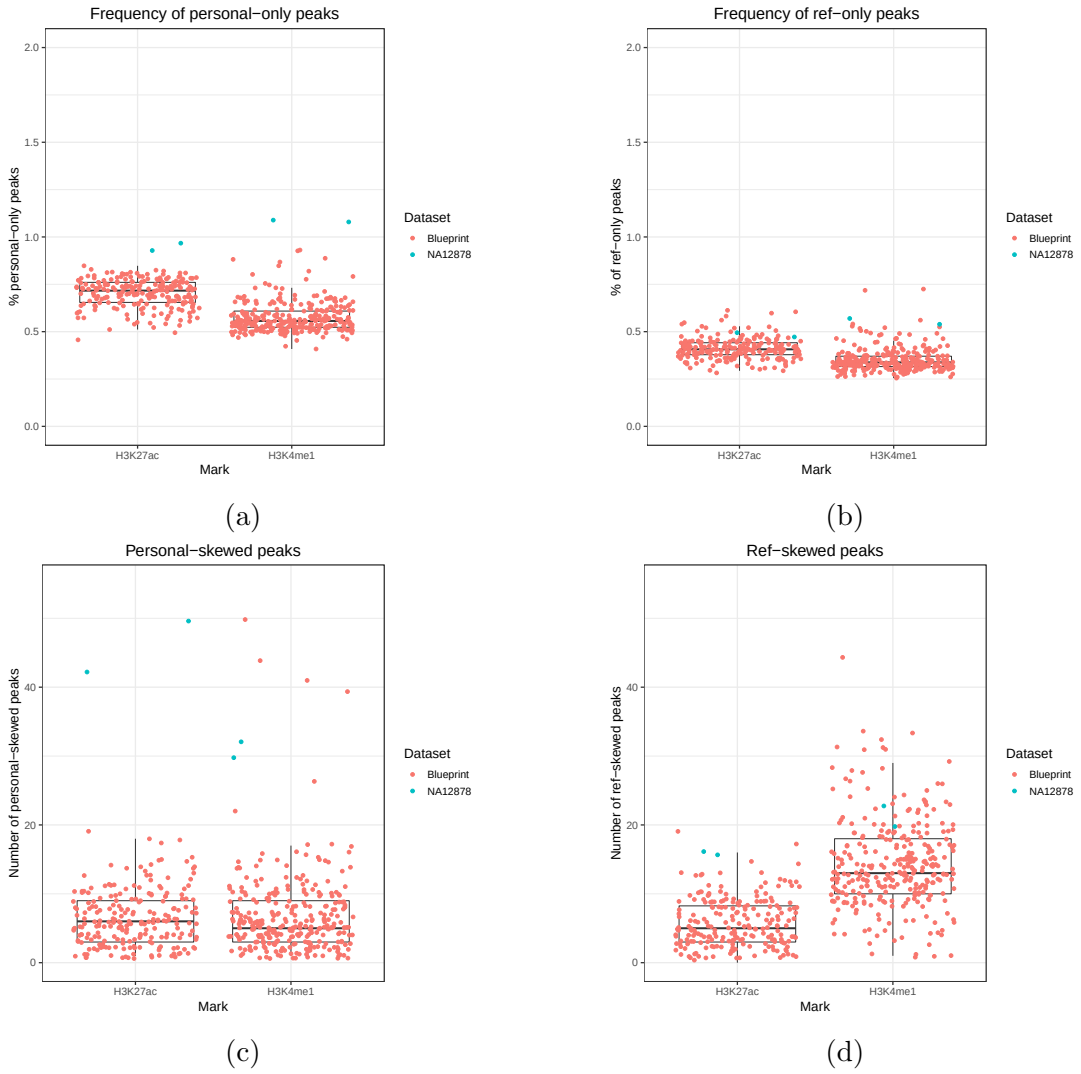


Figure S4: Comparison between NA12878 and Blueprint samples. a) Proportion of peaks that are called only in personalized haplotypes. b) Proportion of peaks that are called only in hg19. c) Number of peaks with higher coverage in the personalized haplotype than in hg19. d) Number of peaks with higher coverage in hg19 than the personalized haplotype.

Ref-only peaks	Number	Contain segmental duplications	Percent
Null coverage in Hap1	2547	1781	69.9
↪ Confident	1490	1062	71.3
Positive coverage in Hap1	4208	267	6.3
↪ Confident	574	98	17.1

Table S1: Ref-only peaks with null coverage in Hap1 are extremely enriched in segmental duplications.

Case (Hap1)	Number of reads	Percent of total alignments
Failed liftover	9769845	5.3
Lifted coordinates	165311184	89.4
↪ Equal in hg19, Hap1	158400964	85.7
↪ Unequal in hg19, Hap1	6910220	3.7
Unmapped in hg19, mapped in Hap1	49997	0.03
Mapped in hg19, unmapped in Hap1	749507	0.4
Unmapped in hg19, unmapped in Hap1	9038375	4.9
Total	184918908	100

(a)

Case (NA12878 MPG)	Number of reads	Percent of total alignments
Failed liftover	2030979	1.1
Lifted coordinates	175123570	89.4
↪ Equal in hg19, MPG	170876933	94.7
↪ Unequal in hg19, MPG	4246637	2.3
Unmapped in hg19, mapped in MPG	39169	0.02
Mapped in hg19, unmapped in MPG	19307	0.01
Unmapped in hg19, unmapped in MPG	9632679	5.2
Total	184918908	100

(b)

Table S2: a) Breakdown of *Jvarkit cmpbamsandbuild* comparison of hg19 and Hap1 WGS alignments. The proportion of failed liftover, mapped/unmapped differences and unequal coordinate reads is roughly 10%. b) In the same experiment with the NA12878 paternal MPG, the same proportion is only 3.7%.

Genome	Personal-skewed	Ref-skewed	Mark
NA maternal	30	23	H3K4me1
NA paternal	32	20	H3K4me1
Hap1	2216	3152	H3K4m1
Hap2	2221	3162	H3K4me1
Pendleton	1518	4320	H3K4me1
NA maternal	42	16	H3K27ac
NA paternal	50	16	H3K27ac

Table S3: Number of peaks with skewed coverage in NA12878 MPGs and DPGs for H3K4me1 and H3K27ac marks. The read length across all rows was kept at 36bp.

Mark	Read length	% replicated personal-only	% replicated ref-only
Between maternal and paternal NA12878			
H3K27ac	51 bp	61	49
H3K27ac	36 bp	67	39
H3K4me1	36 bp	67	41
Median between maternal and paternal in Blueprint			
H3K27ac	51 bp	64	36
H3K27ac	36 bp	63	39
H3K4me1	51 bp	65	37
H3K4me1	36 bp	65	38
Between 10X Hap1 and Hap2			
H3K27ac	51 bp	75	90
H3K27ac	36 bp	70	94
H3K4me1	36 bp	72	80

Table S4: The two haploids of the NA12878 MPG share a significant portion of personal-only and ref-only peaks. Similar proportions are observed in the MPGs of a 9 sample Blueprint subset. The same values are shown for the denovo NA12878 assemblies.

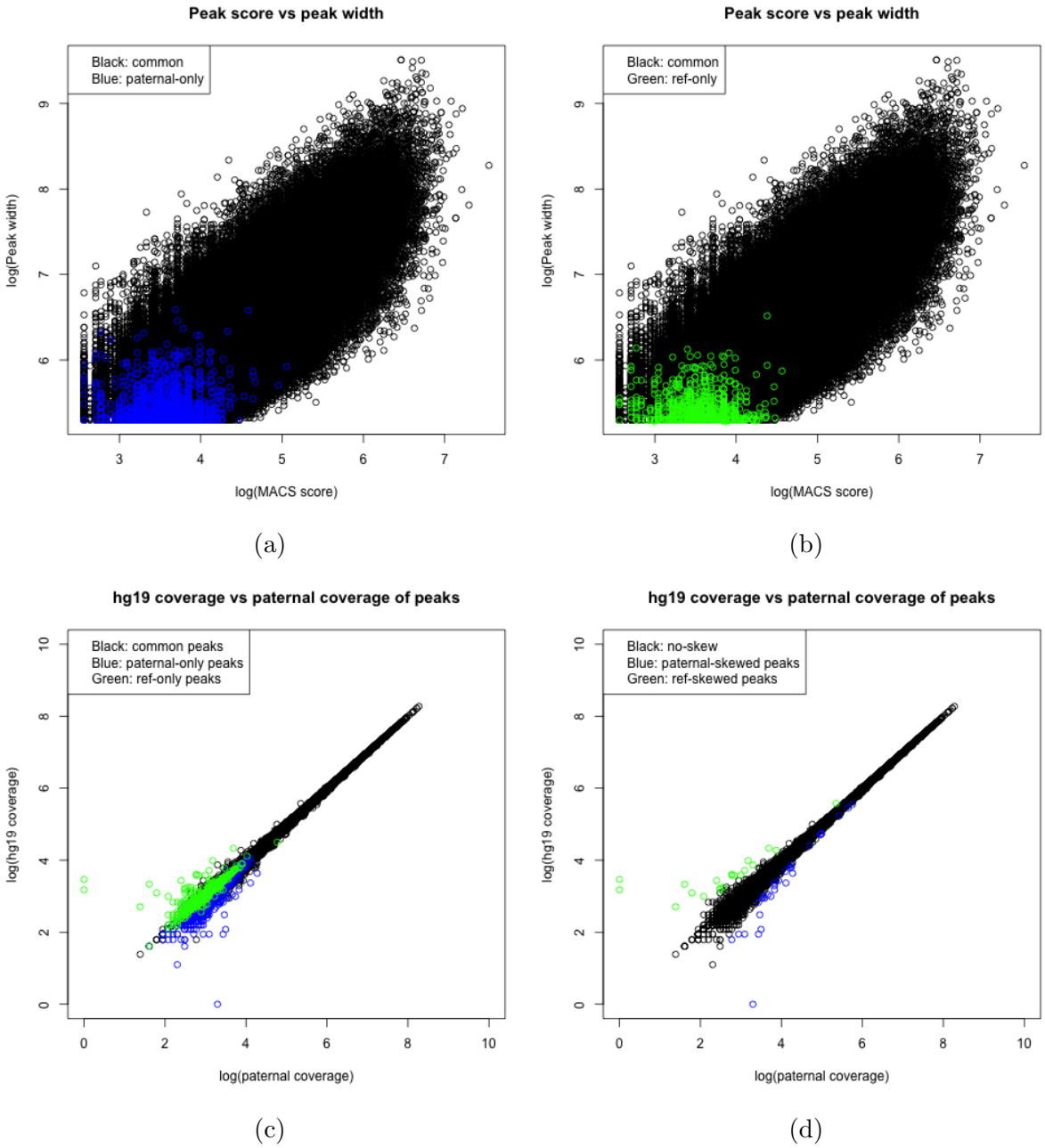


Figure S5: Characterization of NA12878 paternal MPG altered peak calls (H3K4me1).

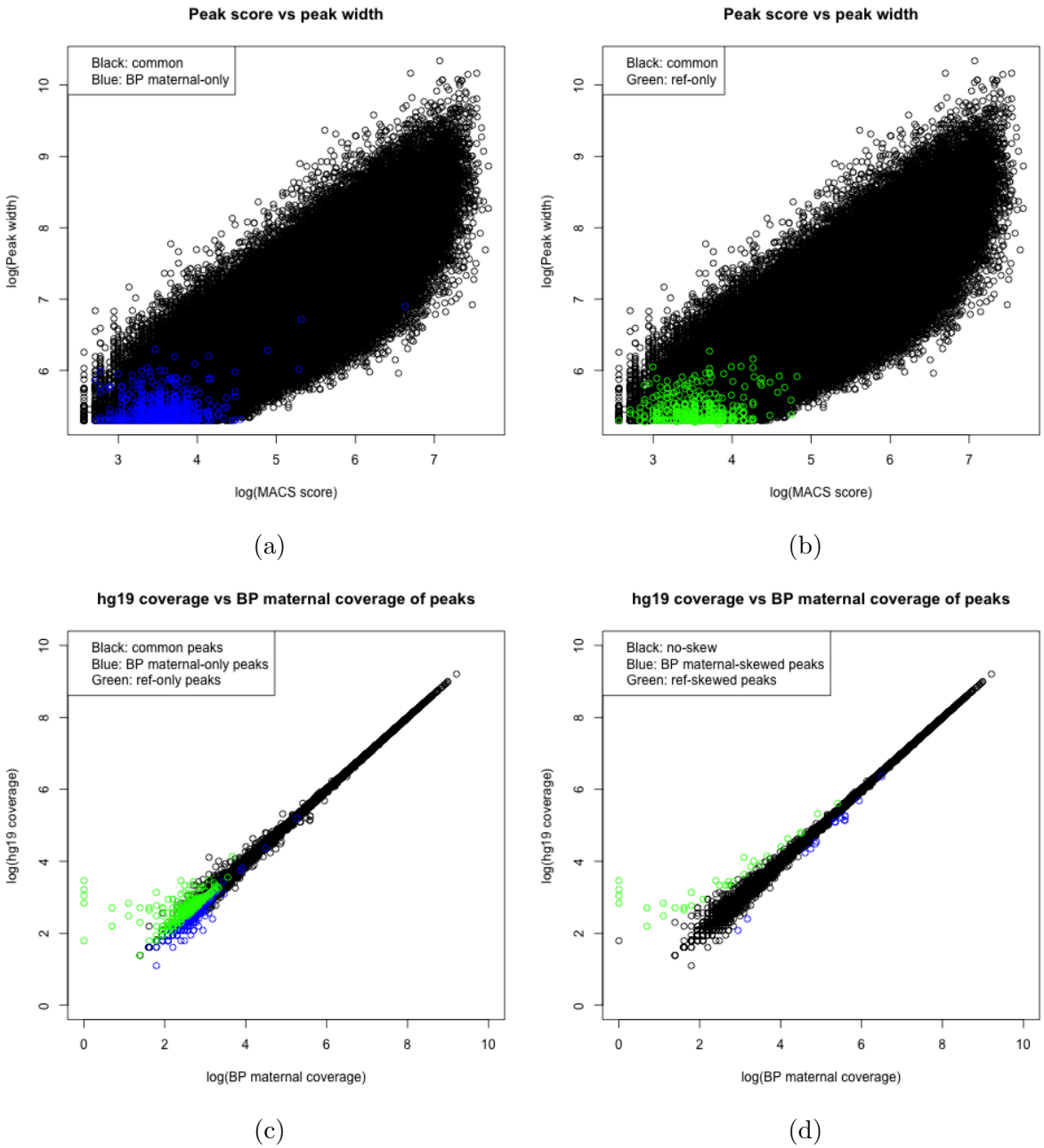
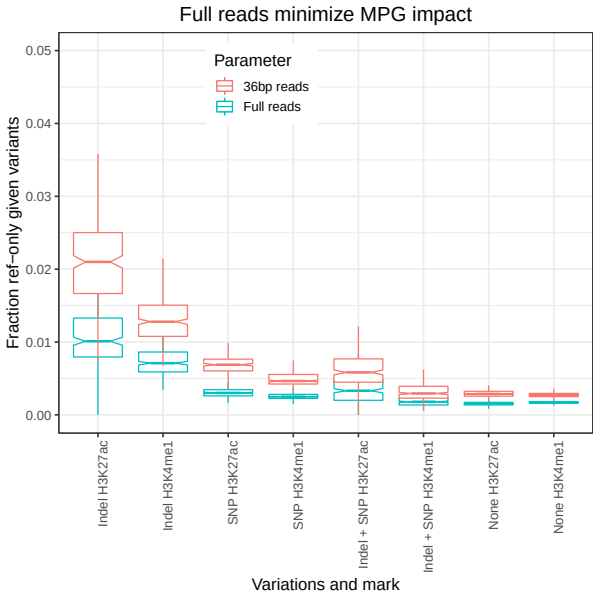


Figure S6: Characterization of AP calls in the maternal MPG of a typical Blueprint sample (H3K4me1).



(a)

Figure S7: Increasing read length reduces the association of variant calls with ref-only peaks.

(a)

(b)

Figure S8: a) Typical personal-only peak in a NA12878 MPG. The small variations have a visible effect on only a few reads. (b) Top SD free personal-only peak in the Hap1 DPG. Large scale changes in coverage become apparent with this approach.

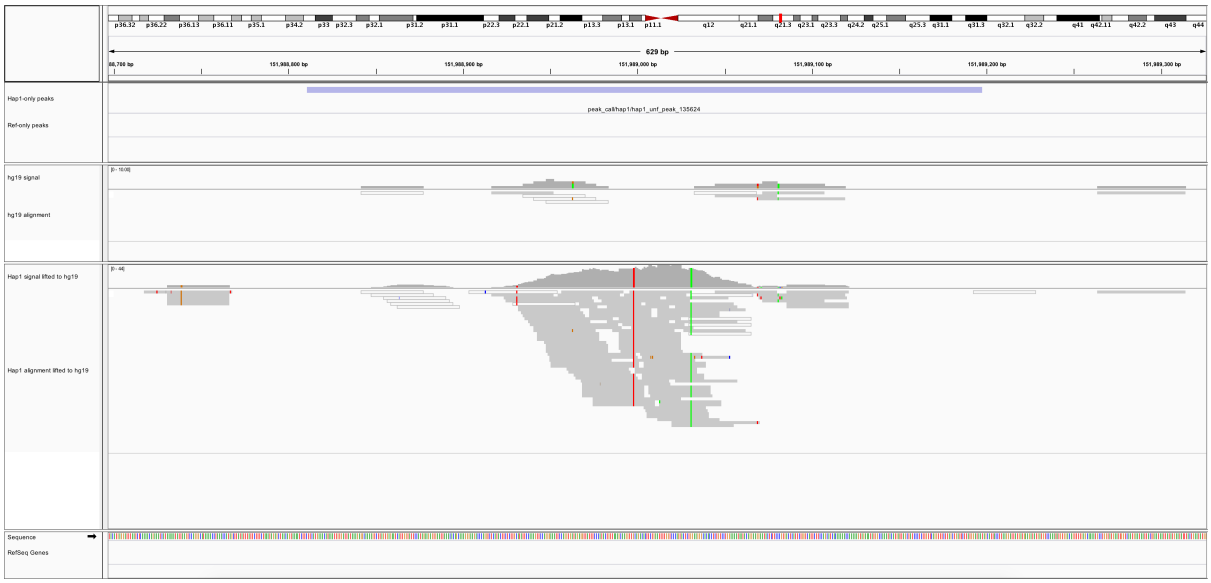


Figure S9: Hap1-only peak free of repeats and segmental duplications. Viewing this region in the UCSC genome browser shows overlaps with the hg19 self chain.