Genomic Dark Matter:
The reliability of short read mapping illustrated by the Genome Mappability Score (GMS)

Hayan Lee[1,2] and Michael C. Schatz[1,2]

[1] Department of Computer Science, Stony Brook University, NY, USA
[2] Simons Center for Quantitative Biology, Cold Spring Harbor Lab, NY, USA

Abstract

Motivation: Genome resequencing and short read mapping have become some of the prime tools of genomics and are used for such applications as investigating the relationship between sequence variations and disease phenotypes, measuring gene transcription rates, profiling epigenetic activations, and numerous other important assays. The current state-of-the-art in short read mapping analysis uses the read quality values, edit distance, and mapping quality scores to evaluate the reliability of the read mapping used for computing the assay result. These attributes, however, are extremely sensitive to minute changes to read position or sequence quality, and are narrowly focused on individual reads. To address these
limitations, we propose the Gnomic Mappability Score (GMS) as a novel measure of the complexity of resequencing a genome with short reads. The GMS is a weighted probability that any read could be unambiguously mapped to each position in the genome and pinpoints the most problematic regions. As such, the GMS measures the fundamental composition of the genome itself, beyond the individual mapped reads in an experiment.
Results: We have developed an open-source pipeline called the Genome Mappability Analyzer (GMA) to compute the GMS of each position of a reference genome. The GMA builds on established input formats, and leverages the leading algorithms BWA and SAMTools for intermediate processing, so it can be applied to measure the GMS of any genome. The GMA can also be used to evaluate the tradeoffs of various experimental conditions including read length, library size, error rates, and coverage. Furthermore, we examined the accuracy of the widely used BWA/SAMTools single nucleotide polymorphism discovery pipeline under typical resequencing conditions, and found variation discovery errors are dominated by false negatives, especially in low GMS regions of the genome. These errors are fundamental to the mapping process and cannot be overcome at any coverage level. As such, the GMS should be considered in every resequencing project to pinpoint the dark matter of the genome in which no variations could possibly be discovered.
Availability: The GMA source code and GMS profiles for several model organisms are available open source at http://gmabio.sourceforge.net
Contact: hlee@cs.stonybrook.edu

Backgrounds for short read mapping and variation discovery

The most common approach to sequencing a genome today is called whole genome shotgun sequencing (the International Human Genome Sequencing Consortium, 2001), in which many copies of the genome are randomly sheared into short molecules which can then be individually sequenced .
As a random process, the number of molecules that originate from a given position of the genome will follow an approximately Poisson distribution (Lander and Waterman, 1988). It is therefore necessary to significantly oversample the genome to account for expected variations in coverage and to account for sequencing errors.
For genomes which have never been sequenced before, the only option is to assemble the reads de novo in which the reads are compared and merged with each other, metaphorically similar to assembling a jigsaw puzzle (Schatz et al., 2010a). For other genomes which have been assembled into a reference sequence, variations relative to the reference can be discovered by matching the short reads to the long genome, using algorithms called short read mappers. The most popular mapping algorithms, such as BWA, Bowtie, and SOAP, attempt to find the best alignment for each read that minimizes the number of differences between the read and the genome, optionally using the quality values to discount differences that are likely due to mere sequencing errors. These algorithms use sophisticated indices of the genome and various heuristics to make the computation efficient enough to map billions of reads in a tractable amount of time. Once the reads have been mapped, follow up algorithms can then analyze the alignments to see if there are any positions that the spanning reads significantly disagree with the reference, using the number of reads, the quality values of the bases, and other metric to distinguish sequencing errors from true variations.

Base Quality Score

$$qv = -10\log(P_e)$$

Read Mapping Quality Score

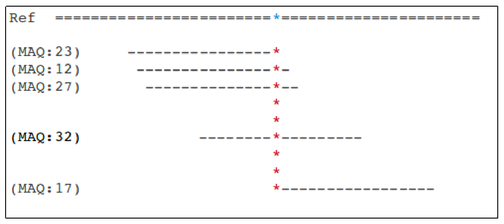
$$Q_s = -10\log[1 - P_s(u|x,z)]$$

$$p_s(u|x,z) = \frac{p_s(z|x,u)}{\sum_{v=1}^{L-l+1} p_s(z|x,u)}$$

Nothing Provides Global View!!!

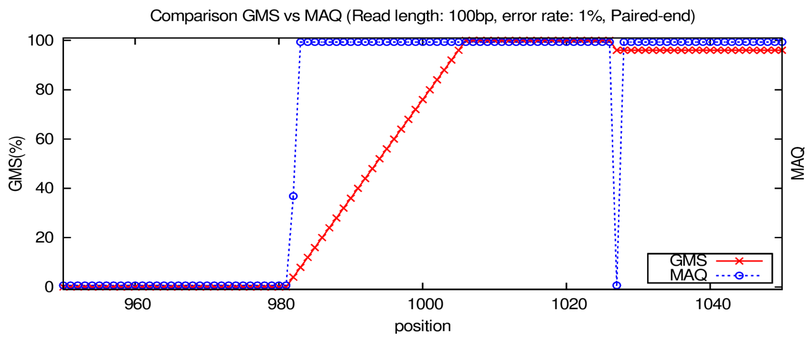
Methods

1. Genome Mappability Score (GMS)



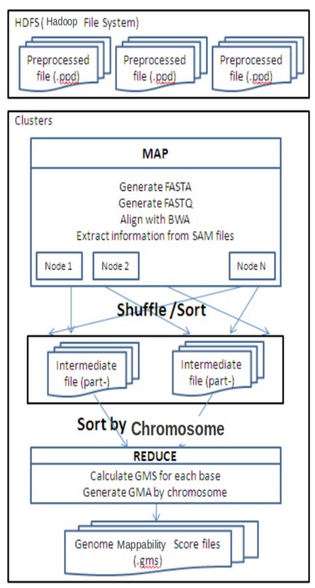
Here we introduce a new probabilistic metric called the Genome Mappability Score (GMS), that builds on the mapping quality scores to build a profile of certainty of mapping reads across the genome. The core of the GMS is to consider all possible reads spanning every position in the genome, as illustrated in figure. For the specific position $*$ sequenced using l-bp reads, there will be l possible reads spanning, each with a potentially different mapping probability $p_s(u|x, z)$. The GMS is the average of the mapping probability of these spanning reads, as defined in Equation 4 following the notation of notations of MAQ as described above. In this way, a GMS of 100% means the base can be precisely mapped by any spanning read, and if the GMS is zero, it cannot be reliably mapped by any read. Unlike the mapping quality score, which is assigned to individual reads, the GMS can be computed at every position, and is robust to biases in coverage or quality values that may artificially reduce the mapping quality score. The GMS is also naturally extended to consider other experimental conditions such as the expected error rate or the insert size for paired-end sequencing by simulating reads with these characteristics to be used for computing their corresponding mapping probability $p_s(u|x, z)$.

$$GMS(u) = \frac{100}{|z|} \sum_{all\, z \ni u} p_s(u|x,z) = \frac{100}{l} \sum_{all\, z \ni u} \left(1 - 10^{-\frac{Q_s(u|x,z)}{10}}\right)$$



2. Genome Mappability Analyzer (GMA)

The Genome Mappability Analyzer (GMA) is our pipeline and collection of tools for computing a profile of the GMS of a reference genome. GMA can be run in serial on a local machine and also in parallel on a cloud. For small genomes, local execution is recommended, while the cloud version is strongly recommended for large genomes.



Results

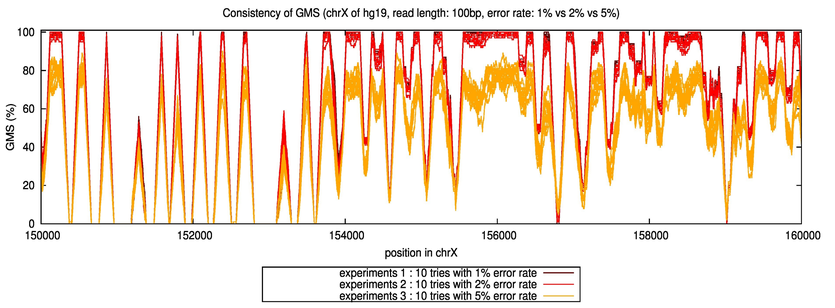
1. GMS Profiles

We computed the GMS profiles with common resequencing parameters of 100bp paired-end reads and an error rate of 2% of the human genome and three important model organisms: yeast (Saccharomyces cerevisiae), fruit fly (Drosophila melanogaster), and mouse (Mus musculus) and compared their GMS profiles to their reference annotations. The results of the analysis are displayed in the below table, and show that 86-95% of these genome sequences are highly mappable, meaning the GMS is at least 50%. Yeast at 1/600 the size of the human genome has the highest fraction of high GMS bases, because GMS depends on the repeat content, which is generally proportional to the genome size. Furthermore 91.1-95.1% of the transcribed sequences, and 91.2-95.1% of the exon sequences of these species are highly mappable. The remaining fraction low GMS values will be difficult or impossible to measure using today's sequencing technologies.

Species (build)	size	whole (%)	transcribed (%)	exon (%)
yeast (sc2)	12 Mbp	95.0	95.1	95.1
fly (dm3)	133 Mbp	88.9	91.7	92.8
mouse (mm9)	2.7 Gbp	86.5	91.1	91.2
human (hg19)	3.0 Gbp	86.1	94.2	94.4

2. Parameters to GMS

Given conditions such as read length, paired or single end and an error rate, the tendency of GMS does not change by mutations among individuals, which means it reflects species characteristics, not individual characteristics.



3. Variation Discoveries and Dark Matter

In the experiment, we use chromosome X (173M) of hg19, the 8th largest genome and important sex chromosome, linked to many inherented genetic diseases. The overall variation detection accuracy is very high, and is twice as high (99.83%) in high GMS regions compared to low GMS regions (42.25%). The detection failure errors are dominated by false negatives, which means the SNP calling program fails to find such variations. In particular, among all 3504 false negatives, 3255 (93%) are located in low GMS region, and only 249 (7%) are in high GMS region. Considering only 14% of human genome is low GMS region, it is huge difference. However, it is not surprising that errors are dominated by false negatives, as the SNP-calling algorithm will use the mapping quality score to filter out low confidence mapping. What is surprising is the extent of false negatives and the concentration of false negatives almost entirely within low GMS regions.

	Low GMS Region	High GMS Region
Total Simulated Mutations	5,636	145,094
Correct SNVs	2,381	144,845
False Positive	1	51
False Negative	3,255	249
Accuracy	0.4225	0.9983

False negatives are typically caused when the variation is not sufficiently sampled by enough reads, especially in light of the expected Poisson distribution in coverage. To measure this effect, we repeated the experiment at 10 coverage levels: 2, 3, 5, 10, 20, 30, 40, 50, 60, and 70-fold in order to observe how coverage contributes to variation detection errors even though 60 or 70-fold coverage is beyond what is commonly used. As expected the accuracy is extremely poor at 2 or 3-fold coverage, as many of the variations will not have any reads because of the Poisson distribution in coverage. The accuracy rate readily improves with increasing coverage as more of the variations with enough reads for the SNP-calling algorithm to find the variations. The improvement ends at around 20-fold coverage, though, because at this point almost every variation should have very deep coverage. At this point with 20-fold coverage and higher, the accuracy of the high mappability regions is very high, but the accuracy of low mappability regions remains very poor.