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

http://gma-bio.sf.net

hlee@cshl.edu, mschatz@cshl.edu

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://gma-bio.sf.net

Short read mapping and variation discovery

The most common approach to sequencing a genome today is called whole genome shotgun sequencing, in which many copies of the genome are randomly sheared into short molecules which can then be individually sequenced .

For 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. The most popular mapping algorithms, such as BWA (Li and Durbin, 2008), Bowtie (Langmead et al., 2009b), and SOAP (Li et al., 2009b) attempt to find the best alignment. Once the reads have been mapped, follow up algorithms can 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
The base quality score measures the error probability of each base from the primary image analysis.

$$qv = -10 \log_{10} p_e$$

Read Mapping Quality Score
The read mapping quality score measures the probability that the read has been mapped to the correct location in the genome. It builds on the base quality scores of the mismatched bases at the mapped position relative to the mismatched bases at other possible positions. The mapping quality score will be very low (or zero) if the read maps to multiple positions because of repeats.

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

$$Q_s = -10 \log_{10} [1 - p_s(u|x,z)]$$

Neither Provides a Global View!!!

References

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome research, 18(11), 1851–1858.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G.,Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009a). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079.
Li, R., Yu, C., Li, Y., Lam, T.-W. W., Yiu, S.-M. M., Kristiansen, K., and Wang, J. (2009b). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics (Oxford, England), 25(15), 1966–1967.
Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. (2009b). Ultrafast and memory efficient alignment of short DNA sequences to the human genome. Genome Biology, 10(3), R25+.
Schatz, M. C., Delcher, A. L., and Salzberg, S. L. (2010a). Assembly of large genomes using second-generation sequencing. Genome research, 20(9), 1165–1173

Methods

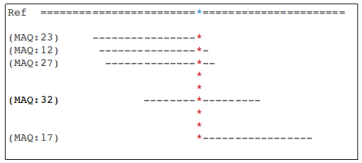
1. Genome Mappability Score (GMS)

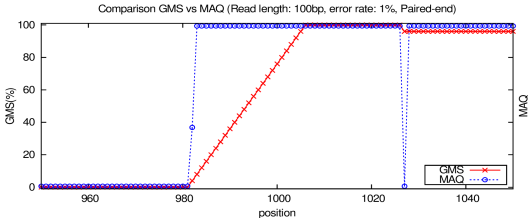
Here we introduce a new probabilistic metric called the Genome Mappability Score (GMS), that builds on the mapping quality scores to build a profile of certainty of mapping reads across the genome.

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

The GMS is computed by considering all possible reads spanning every position in the genome

- For the specific position * sequenced using *l*-bp reads, there will be *l* possible reads spanning, each with a potentially different mapping probability $p_s(u|x, z)$.
- GMS is the average of the mapping probability of these spanning reads
- GMS of 100% means the base can be precisely mapped by any spanning read
- If the GMS is zero, it cannot be reliably mapped by any read.



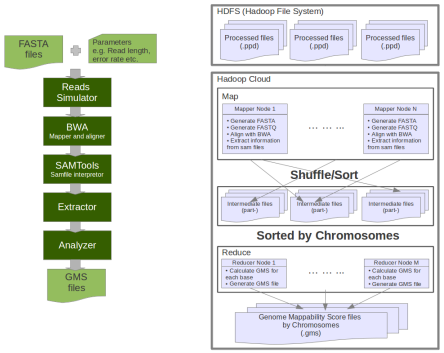


Advantages

- Unlike the mapping quality score, which is assigned to individual reads, the GMS is to be computed at every position.
- Unlike the mapping quality score, which is very sensitive to a minute change, GMS represent more stable characteristics of the genome and provide consistent and global view

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 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

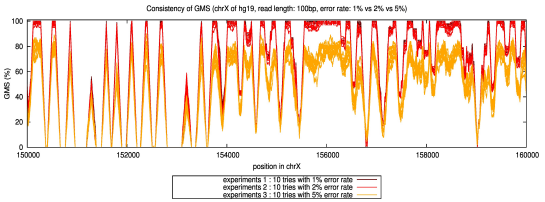
1. GMS Profiles

We computed the GMS profiles with common resequencing parameters: 100bp read length, paired-end and an error rate of 2%. The result shows that 86-95% of these genome sequences are highly mappable, meaning the GMS is at least 50%. The fraction of low GMS regions 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

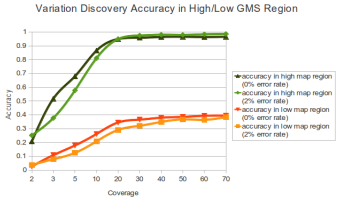
Analysis of simulated mutations into human chromosome X (173M)

- Simulate mutations and reads from a mutated sample in silico (wgsim), use BWA/SAMtools to identify them
- Variation detection accuracy is twice as high (99.83%) in high GMS regions compared to low GMS regions (42.25%), dominated by false negatives especially in low GMS region
- Among all 3504 false negatives, 3255 (93%) are located in low GMS region
- Considering only 14% of human genome is low GMS region, it is surprising that 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

Accuracy of low GMS sequences is independent of coverage

- False negatives are typically caused when the variation is not sampled with enough coverage
- The accuracy rate improves in both hi- and low- GMS regions up to ~20-fold coverage.
- Accuracy rate approaches 100% in high GMS region
- However, accuracy is not improved in low GMS region, even with very high coverage
- Therefore 42% is a upper limit in low GMS region that detection mechanism can reach in current cutting-edge bio-technology.



- GMS explains the skewed distribution in dbSNP and clinical SNPs

GMS Distribution Ratio in Human Genome (hg19)

	whole	transcription	coding	exon	SNPs	clinical SNPs
Low GMS	0.1131	0.0337	0.0311	0.0330	0.0255	0.0194
High GMS	0.8869	0.9663	0.9689	0.9670	0.9745	0.9806

rs445114 (GMS : 3.5972) PROSTATE CANCER
rs944289 (GMS : 3.7322) THYROID CARCINOMA
rs1016732 (GMS : 9.9999) AUTISM