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 two of the primary 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.

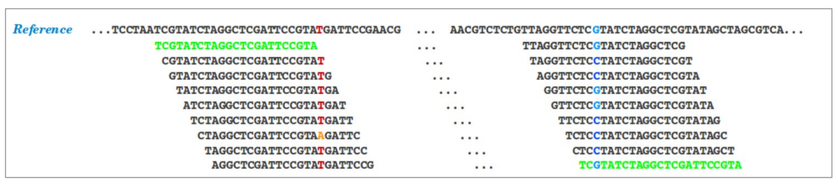
Results: We have developed an open-source pipeline called the Genome Mappability Analyzer (GMA) to compute the GMS of each position of a reference genome. The GMA builds on established input formats, and leverages the leading algorithms BWA and SAMtools for intermediate processing, so it can be applied 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. Notably, virtually every SNP mutation reported by the 1000 genomes pilot project was from the regions of the genome with high GMS score.

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 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. 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 minimizing the edit distance of each read. 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.



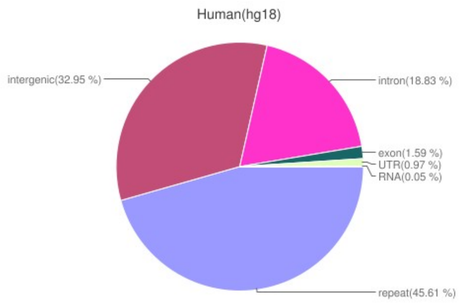
Base Quality Score

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

Read Mapping Quality Score

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

The primary complication of short read mapping is that a read may map equally well or nearly equally well to multiple positions because of repetitive sequences in the genome. Notably, nearly 50% of the human genome consists of repetitive elements of various forms including certain repeats that occur thousands of times throughout the entire genome (International Human Genome Sequencing Consortium, 2001).

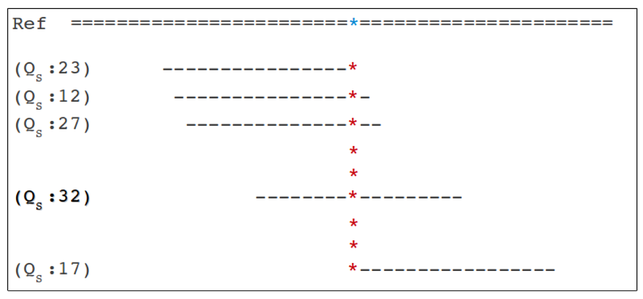


view	bases (bp)	%
repeat	1,405,061,157	45.61
intergenic	1,015,000,634	32.95
intron	580,090,705	18.83
exon	48,978,698	1.59
UTR	29,859,581	0.97
RNA	1,428,705	0.05
Whole Genome	**3,080,419,480**	**100.0**

Need a global view for a sequencing experiment!!!

Methods

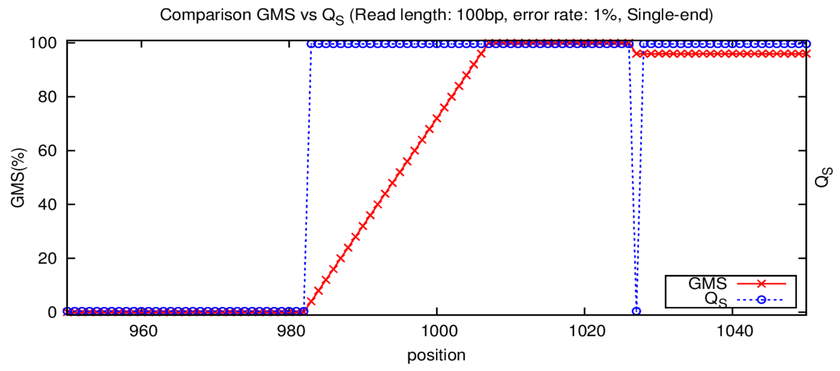
1. Genome Mappability Score (GMS)



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.

$$GMS(u) = \frac{100}{C} \sum_{\forall z \ni u} p_s(u|x,z) = \frac{100}{C} \sum_{\forall z \ni u} (1 - 10^{-\frac{Q_s(u|x,z)}{10}})$$

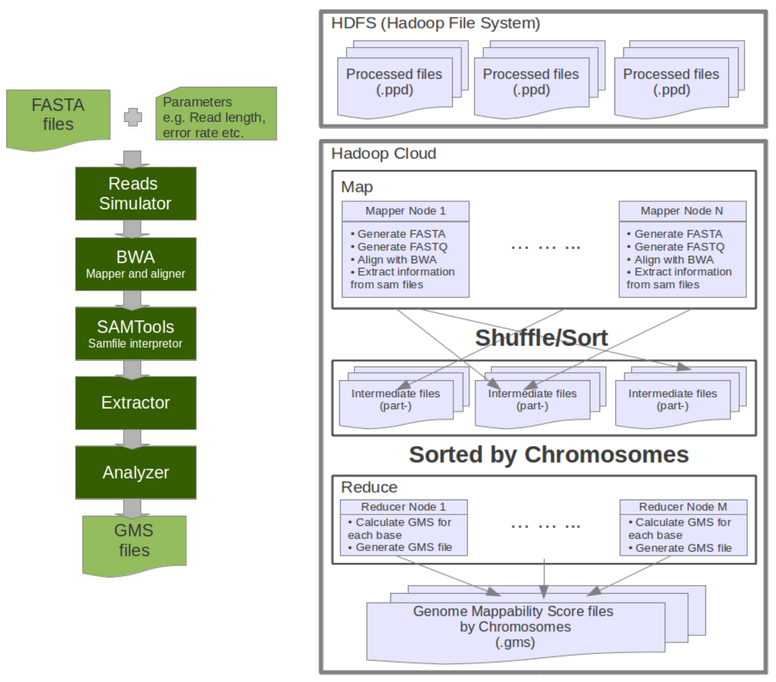
- GMS takes all possible reads spanning every position in the genome
- For the specific position $*$ sequenced using l-bp reads, there will be l possible reads spanning, each with a potentially different mapping probability $p_s(u|x, z)$.
- GMS is the average of the mapping probability of these spanning 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.



- Unlike the mapping quality score, which is assigned to individual reads, the GMS is to be computed at every position with all possible cases of reads.
- 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 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.



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

Results

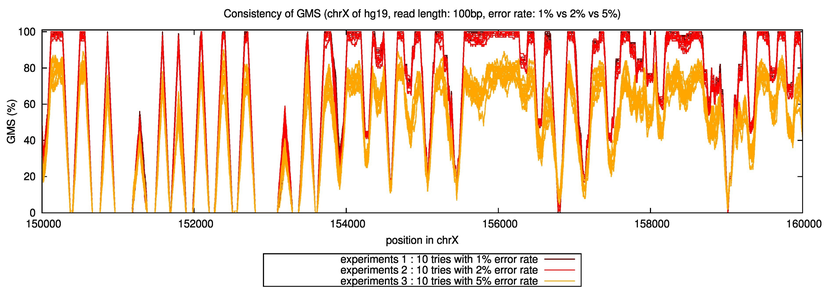
1. GMS Profiles

We computed the GMS profiles with common resequencing parameters such as 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	paired/single	whole (%)	transcribed (%)
yeast (sc2)	12 Mbp	paired	94.85	95.04
		single	94.25	94.62
fly (dm3)	130 Mbp	paired	90.52	96.14
		single	89.70	95.94
mouse (mm9)	2.7 Gbp	paired	89.39	96.03
		single	87.47	94.75
human (hg19)	3.0 Gbp	paired	89.02	97.40
		single	87.79	96.38

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 read characteristics.



3. Sequencing Technologies to GMS

The GMS can play an important role at evaluating the information gains using other sequencing technologies, and measure the fraction of the genome that is accessible using one technology over another.

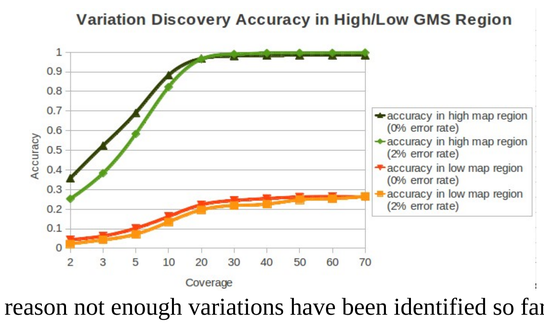
Sequencing Technology	Read length(bp)	Error (%) (substitution insertion deletion)	Low GMS region (%)	High GMS region (%)
SOLiD-like	75	0.10 N/A N/A	11.14	88.86
Illumina-like	100	0.10 N/A N/A	10.51	89.49
Ion Torrent-like	200	0.04 0.01 0.95	9.35	90.65
Roche/454-like	800	0.18 0.54 0.36	8.91	91.09
PacBio-like	2000	1.40 11.47 3.43	100.00	0.00
PacBio EC-like	2000	0.33 0.33 0.33	8.61	91.39

4. Variation Discoveries and Dark Matter

- Chromosome X (173M) of hg19, the 8th largest chromosome linked to inherented genetics.
- Variation detection accuracy is 4 times as high (98.96%) in high GMS regions compared to low GMS regions (20.45%)
- The detection failure errors are dominated by false negatives
- Among all 5,022 false negatives, 3.255 (70%) are located in low GMS region
- Considering only 12% of human genome is low GMS region, it is surprising that the concentration of false negatives almost entirely within low GMS regions.

	0% error rate		2% error rate	
	Low GMS Region	High GMS Region	Low GMS Region	High GMS Region
Total Simulated Mutations	4,498	144,855	4,406	146,477
Correct SNVs	1,096	141,969	901	144,960
False Positive	0	48	0	78
False Negative	3,402	2,886	3,505	1,517
Accuracy(%)	24.37	98.01	20.45	98.96

- False negatives are typically caused when the variation is not sufficiently sampled with high coverage
- To measure this effect, we repeated the experiment at 2, 3, 5, 10, 20, 30, 40, 50, 60, 70-fold coverages
- The accuracy rate improves with increasing coverage up to around 20-fold coverage.
- Accuracy rate hit close to 100%, almost perfect level, in high GMS region
- However, accuracy is not improved in low GMS region, even though high level of coverages are used.
- Therefore 27% is a upper limit in low GMS region that detection mechanism can reach in current cutting-edge bio-technology.



- That is the reason not enough variations have been identified so far (dbSNPs).

GMS Distribution Ratio in Human Genome (hg19)

	dbSNPs				1000 Genome Project
	Whole (%)	Transcription (%)	SNPs (%)	Clinical SNPs (%)	SNPs (%)
Low GMS	11.31	3.37	2.55	1.94	0.01
High GMS	88.69	96.63	97.45	98.06	99.99