Supplementary material for Genomic Dark Matter

H. Lee\textsuperscript{a}\textsuperscript{*} and M. Schatz\textsuperscript{b}

\textsuperscript{a} Department of Computer Science, Stony Brook University, Stony Brook, NY
\textsuperscript{b} Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

\textsuperscript{*} Corresponding Author: hlee@cs.stonybrook.edu

Abstract

This is the supplementary materials for \textit{Genomic Dark Matter: The reliability of short read mapping illustrated by the Genome Mappability Score}.

Keywords: Short read mapping, mappability, genome mappability score

1 Background

1.1 Sequencing and Mapping

Sequencing DNA is a complex biochemical and computational process. Because of technical limitations of the chemical protocols, it is not possible to sequence entire intact chromosomes, and currently available second generation sequencing machines from Illumina can sequence at most 100 or 150 consecutive nucleotides [Bentley et al., 2008]. However, these instruments are massively parallel so that many billions of molecules can be simultaneously sequenced per $\sim2$ week run of an instrument generating a total of 500 Gbp for less than $10,000 in reagent costs. These instruments can also sequence "mate-pairs" or "paired-end" reads which are pairs of reads that are separated by an approximately known distance on the original molecule.

Sequencing on these machines uses a technique called sequencing-by-synthesis in which fluorescently-labeled nucleotides are imaged as they are incorporated into growing strands of DNA according to the template molecule [Bentley et al., 2008]. Sophisticated image analysis software then interprets the signal intensities into nucleotide sequences. The base calling algorithms also produces a sequence of quality values that are estimates of the probability of error for each nucleotide as first proposed by the Phred base calling algorithm [Ewing et al., 1998] [Ewing and Green, 1998]. The most common approach to sequencing a genome today is called whole genome shotgun sequencing [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. A critical parameter of a sequencing project is the desired average coverage of the genome, as controlled by the number of (partial) runs of the machines used to sequence a given sample. It is now

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

The overall accuracy of the process is typically 98\% or 99\% accurate per base, with errors concentrated towards the end of the reads (Figure 1). Quality values are commonly encoded according to the Phred quality scale, in which an error estimate of $p_e = 1\%$ (99\% accuracy) is written as $qv = 20$ according to Equation (1). Quality values at or above $qv = 20$ ($p_e \leq 1\%$), are typically classified as high quality bases.
Figure S 1: Quality value as a function of read position. The quality values at each read position were averaged from a sample of 100,000 100bp reads sequenced at the Broad Institute using an Illumina Genome Analyzer II (SRA study SRR001086). Approximately the first 50 bp have $qv \geq 20$, meaning the probability of error is less than 1%, while the latter half of the reads have considerably worse quality.

common to sequence enough molecules so that a total of 20-fold or 30-fold coverage of the genome are generated, e.g. 60Gbp - 90Gbp of raw sequencing for the 3Gbp human genome. At 20-fold coverage it is expected that most nucleotides from each copy of the genome will be sequenced by at least 5 individual reads, thus reducing the expected error rate to below .00001 when using bases with $Q \geq 20$ (Figure 2). This is especially important in the case of human genomes and other diploid organisms which have 2 copies of each chromosome and contain heterozygous variations. The molecules, and the derived sequences, will then be randomly selected between the two copies according to an approximately binomial distribution.

Figure S 2: Expected fraction of the genome at a given read depth given the targeted read coverage. When targeting 10x coverage, approximately 3% of the genome is expected to be at less than 5x coverage, while less than 1% of the genome is expected to be at less than 5x coverage when targeting 20x or 30x coverage. For diploid genomes, such as the human genome, heterozygous variations will have approximately half the targeted coverage (10x effected coverage when targeting 20x overall coverage).

The individual reads in a sequencing project are much shorter than the genome and originate from random positions, so it is necessary to use sophisticated software algorithms to map out an entire genome. 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., 2010]. 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 (Figure 1).

Base quality values and mapping quality scores are the most widely used measures of confidence in a genome sequencing project. As explained above, base quality values measure the probability that a given base was incorrectly sequenced based on the signal intensities observed during the sequencing process. Mapping quality scores build on base quality values to measure the probability of an incorrectly mapped read, and give less confidence to alignments with more bases that disagree with the reference genome, especially if those bases have high quality values. Furthermore, the mapping quality score is reduced if there are multiple possible mappings with the same minimum edit distance.

### 1.2 Supplementary Figures

![Comparison GMS vs Qs (Read length: 100bp, error rate: 1%, Single-end)](image)

**Figure S 3:** Mapping quality scores and GMS profile for 1% error reads near a repeat boundary. The x-axis shows the genome near the transition point between 950-1,050bp. The y-axis show the GMS score (red) and the mapping quality score (blue) normalized to the range 0-100. A one base difference in position drastically changes the mapping quality score, while the GMS changes much more gradually. Given an error around 1030th position, the mapping quality score plummeted while GMS is marginally reduced since the position still can be clarified by other reads. This shows how sensitive the mapping quality score is to small change. This figure follows the conditions of Fig. 2, except using 1% error for the simulated reads.
Figure S 4: GMA pipeline consists of 5 main components operating in a pipeline. Each component can be run separately, or in a batch with the GMA command "runall"

Figure S 5: The GMS distribution for all chromosomes in hg19. Most bases either have a very low (< 10) or very high (> 90) GMS value. As such, the threshold for deciding hi and lo GMS regions is generally robust as long as the value is near 50.
Figure S 6: Example region of chrX with 20 experiments, 10 for paired-end and 10 for single-end, to examine how paired/single-end affect GMS for chromosome X of hg19 at a given error rate 2% and read length 50bp. The experiments clearly prove that paired-end reads can be reliably be mapped where single-end reads cannot.

Figure S 7: Distribution in GMS value by sequencing technology. In general, the longer the characteristic read length of the platform, the greater the fraction of the genome with a higher GMS value, although higher rates of sequencing error can reduce the GMS value (especially uncorrected PacBio reads).
Figure S 8: Variation Accuracy Simulator (VAS) design. VAS is a pipeline consisting of variation and reads simulator (WGSIM), mapping and aligning tools (BWA), SAM format interpreter (SAMtools) and calling SNP program (BCFtools) and analyzer. It compares variations generated by WGSIM and variations found by VAS pipeline and outputs accuracy result. The pipeline is implemented for both local usage and Hadoop cloud pipeline like the GMA.
1.3 Supplementary Equations

The number of reads simulated depended on the targeted coverage as set by equation 2, where $c$ is target coverage, $L$ is a length of reference genome, $l$ is a length of a read and $d$ is 2 for paired-end and 1 for single end.

$$ n = \frac{c \times L}{l \times d} $$

(2)

1.4 Supplementary Tables

Table S1: Approximate user time for executing the GMA for each technology are measured. Since we used hadoop to exploit parallelism, the precise user time depends on the characteristics of the cloud such as number of cores, the performance of each core, memory size and network connection between the nodes. Therefore the purpose of this table is to give an approximate idea of how the major parameters can affect computing time, but a few seconds difference is not meaningful. To normalize the analysis, all human chromosomes are evaluated, and all technologies are evaluated as single end reads even though some technologies provide paired-end reads. While not the major focus of this study, we observed the 75bp SOLiD reads required more time than the 100bp Illumina reads, presumably because a greater fraction of the reads were repetitive. On the other hand, the other technologies were slower because of the higher error rates, and also because BWA uses a different algorithm for very long reads. The uncorrected PacBio reads were fast to align, because the algorithm aborts on reads with very high error rate when it cannot find an alignment seed.

| Technology  | coverage | length(bp) | substitution(%) | insertion(%) | deletion(%) | running time              |
|-------------|----------|------------|----------------|--------------|-------------|---------------------------|
| Illumina    | 100x     | 100        | 0.10           | n/a          | n/a         | 9 hrs 5 mins 15 secs     |
| Solid       | 100x     | 75         | 0.10           | n/a          | n/a         | 10 hrs 44 mins 19 secs   |
| Ion Torrent | 100x     | 200        | 0.04           | 0.01         | 0.95        | 83 hrs 33 mins 2 secs    |
| Roche/454  | 100x     | 800        | 0.18           | 0.54         | 0.36        | 52 hrs 21 mins 23 secs   |
| PacBio      | 100x     | 2000       | 1.40           | 11.47        | 3.43        | 28 hrs 54 mins 36 secs   |
| PacBio(EC)  | 100x     | 2000       | 0.33           | 0.33         | 0.33        | 39 hrs 0 mins 3 secs     |

Table S2: Distribution of GMS values in the *T.vaginalis* genome. Since *T. vaginalis* has high proportion of repeats, over half of the genome cannot be reliably mapped.

| GMS range   | whole(%) | coding(%) | gene(%)  |
|-------------|----------|-----------|----------|
| [0, 10]     | 33.28    | 30.31     | 30.37    |
| (10, 20]    | 5.49     | 5.55      | 5.52     |
| (20, 30]    | 4.56     | 4.54      | 4.52     |
| (30, 40]    | 4.12     | 3.96      | 3.94     |
| (40, 50]    | 4.59     | 3.88      | 3.86     |
| (50, 60]    | 2.75     | 2.28      | 2.26     |
| (60, 70]    | 2.55     | 1.98      | 1.96     |
| (70, 80]    | 2.47     | 1.78      | 1.77     |
| (80, 90]    | 2.57     | 1.66      | 1.66     |
| (90, 100]   | 37.63    | 44.06     | 44.16    |
| TOTAL       | 100.00   | 100.00    | 100.00   |
Table S 3: This table shows the distribution of the GMS values across the human genome (hg19), the transcriptome, SNPs from dbSNP, and known clinically relevant SNPs. The portion of GMS distribution should stay consistent between the different intervals, but the trends for the known SNPs and clinically relevant SNPs are very different, and are dominated by regions with very high GMS values.

| GMS range | whole | \( \text{ratio(\%)} \) | count | transcription | \( \text{ratio(\%)} \) | count | SNP | \( \text{ratio(\%)} \) | count | clinical SNP | \( \text{ratio(\%)} \) | count |
|-----------|-------|-----------------|-------|--------------|-----------------|-------|-----|--------------------|-------|---------------|--------------------|-------|
| [0, 10]   | 8.74  | 270,498,890     | 0.92  | 10,884,443   | 0.03            | 8,732 | 0.14 | 35 |
| (10, 20]  | 0.16  | 4,912,086       | 0.11  | 1,351,528    | 0.02            | 6,610 | 0.03 | 7 |
| (20, 30]  | 0.18  | 5,630,119       | 0.13  | 1,543,931    | 0.03            | 8,501 | 0.03 | 8 |
| (30, 40]  | 0.21  | 6,581,319       | 0.15  | 1,813,684    | 0.04            | 11,542 | 0.05 | 13 |
| (40, 50]  | 0.27  | 8,210,720       | 0.20  | 2,320,782    | 0.07            | 19,638 | 0.12 | 30 |
| (50, 60]  | 0.23  | 6,977,781       | 0.17  | 2,015,520    | 0.09            | 24,089 | 0.03 | 8 |
| (60, 70]  | 0.28  | 8,550,745       | 0.21  | 2,475,717    | 0.12            | 34,153 | 0.02 | 6 |
| (70, 80]  | 0.36  | 11,291,570      | 0.28  | 3,286,640    | 0.19            | 52,897 | 0.09 | 21 |
| (80, 90]  | 0.56  | 17,195,153      | 0.43  | 5,067,759    | 0.35            | 99,067 | 0.22 | 55 |
| (90, 100] | 89.02 | 2,755,829,028   | 97.40 | 1,152,608,483 | 99.06 | 27,993,106 | 99.26 | 24,434 |
| TOTAL     | 100.00 | 3,095,677,411   | 100.00 | 1,183,368,487 | 100.00 | 28,258,335 | 100.00 | 24,617 |

Table S 4: There are many important variations in low GMS region, even though those regions are hard to study. Some of the most important variations in low GMS region are listed, all having some important clinical role.

| ID(dbSNP) | GMS | chr(position) | Variation | Description                                                                     |
|-----------|-----|---------------|-----------|--------------------------------------------------------------------------------|
| rs33992775 | 0.0 | chr11(5270997) | G → C,T  | HEMOGLOBIN, GAMMA G; HBG2                                                        |
|           |     |               |           | HEMOGLOBIN, GAMMA A; HBG1                                                        |
| rs104893928 | 0.0 | chr5(70220935) | C → G  | SURVIVAL OF MOTOR NEURON 1; SMN1                                                |
| rs116840812 | 0.0 | chr3(82823346) | CCT → C | RIBOSOMAL PROTEIN S17; RPS17                                                    |
| rs104893611 | 0.0 | chr2(131355469) | G → A  | CRYPTIC PROTEIN; CFC1                                                           |
| rs104894914 | 0.0 | chrX(153457207) | T → C | OPSIN 1, MEDIUM-WAVE-SENSITIVE; OPN1MW                                          |
| rs34708054 | 1.5004 | chr16(000222972) | C → G,T | HEMOGLOBIN—ALPHA LOCUS 1; HBA1                                                   |
|           |     |               |           | HEMOGLOBIN—ALPHA LOCUS 2; HBA2                                                  |
| rs104894915 | 23.7267 | chrX(153453428) | C → A | OPSIN 1, MEDIUM-WAVE-SENSITIVE; OPN1MW                                          |
| rs116840811 | 28.9615 | chr15(082824835) | A → C | RIBOSOMAL PROTEIN S17; RPS17                                                    |
| rs1050501 | 33.4806 | chr1(161643798) | T → C | Fc FRAGMENT OF IgG, LOW AFFINITY IIb, RECEPTOR FOR; FCGR2B                     |
References
Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature, 456*(7218), 53–59.

Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research, 8*(3), 186–194.

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-Calling of Automated Sequencer Traces Using Phred. Accuracy Assessment. *Genome Research, 8*(3), 175–185.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860–921.

Lander, E. S. and Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics, 2*(3), 231–239.

Schatz, M. C., Delcher, A. L., and Salzberg, S. L. (2010). Assembly of large genomes using second-generation sequencing. *Genome research, 20*(9), 1165–1173.