Li *et al. BMC Bioinformatics* 2014, **15**:2
http://www.biomedcentral.com/1471-2105/15/2

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Diminishing return for increased Mappability with longer sequencing reads: implications of the *k*-mer distributions in the human genome

Wentian Li[1*], Jan Freudenberg[1] and Pedro Miramontes[2]

Abstract

Background: The amount of non-unique sequence (non-singletons) in a genome directly affects the difficulty of read alignment to a reference assembly for high throughput-sequencing data. Although a longer read is more likely to be uniquely mapped to the reference genome, a quantitative analysis of the influence of read lengths on mappability has been lacking. To address this question, we evaluate the *k*-mer distribution of the human reference genome. The *k*-mer frequency is determined for *k* ranging from 20 bp to 1000 bp.

Results: We observe that the proportion of non-singletons *k*-mers decreases slowly with increasing *k*, and can be fitted by piecewise power-law functions with different exponents at different ranges of *k*. A slower decay at greater values for *k* indicates more limited gains in mappability for read lengths between 200 bp and 1000 bp. The frequency distributions of *k*-mers exhibit long tails with a power-law-like trend, and rank frequency plots exhibit a concave Zipf's curve. The most frequent 1000-mers comprise 172 regions, which include four large stretches on chromosomes 1 and X, containing genes of biomedical relevance. Comparison with other databases indicates that the 172 regions can be broadly classified into two types: those containing LINE transposable elements and those containing segmental duplications.

Conclusion: Read mappability as measured by the proportion of singletons increases steadily up to the length scale around 200 bp. When read length increases above 200 bp, smaller gains in mappability are expected. Moreover, the proportion of non-singletons decreases with read lengths much slower than linear. Even a read length of 1000 bp would not allow the unique alignment of reads for many coding regions of human genes. A mix of techniques will be needed for efficiently producing high-quality data that cover the complete human genome.

Keywords: Next-generation sequencing, Read alignment, Repeat sequences, Genome redundancy, Long-tail distribution, *k*-mers

Background

Many applications of next-generation-sequencing (NGS) in human genetic and medical studies depend on the ability to uniquely align DNA reads to the human reference genome [1-6]. This, in turn, is related to the level of redundancy caused by repetitive sequences in the human genome, well known from the earlier human whole-genome shotgun sequencing [7,8], and the read length k. When the read length k is too short, it is theoretically impossible to have a reference sequence with size comparable to the human genome that does not contain any repeats of k bases. It has been shown using graph theory that the longest DNA sequences avoiding any repeats of k-mers can be constructed by packing all unique k-mers shifting one position at the time [9]. The number of different k-mer types is $4^k/2$ (k odd) or $(4^k + 2^k)/2$ (k even) if both a subsequence and its reverse complement are considered to belong to the same k-mer type. Solving $4^k/2 \approx 3 \times 10^9$ leads to the conclusion that read length k must be at least greater than 17 for all reads to be uniquely

*Correspondence: wtli2012@gmail.com
[1]The Robert S. Boas Center for Genomics and Human Genetic, The Feinstein Institute for Medical Research, North Shore LIJ Health System, 350 Community Drive, Manhasset, USA
Full list of author information is available at the end of the article

alignable to a hypothetical reference sequence that has the size of the human genome.

However, in reality the human genome did not evolve by a first principle to be consistently compact and incompressible. Redundant sequences in the human genome have resulted from duplication, insertion of transposable elements, and tandem repeats due to replication slippage, and more than half of the human genome can be traced to repetitive transposable elements. Although locally duplicated sequences can be deleterious [10] or disease-causing [11], a certain level of redundancy is a requirement for biological novelty and adaptation [12-14]. For higher eukaryotes, a slower removal of the deleterious repeats due to low mutation rates and smaller population sizes [15] lead to a higher level of genome-wide redundancy. This in turns may lead to more protein sequences with internal repeats and perhaps new fold or new functions such as the case for connection tissue, cytoskeletal, and muscle proteins [16].

Therefore, $k = 17$ is a very unrealistic estimation of the minimal read length required for a perfectly successful NGS reads alignment. Accordingly, NGS technologies utilize reads with various larger lengths: $k = 70$ for *Complete Genomics*, $35 \sim 85$ for *ABI SOLiD*, $75 \sim 150$ pair-end for *Illumina HiSeq*, 400 for *Ion Torrent PGM*, $450 \sim 600$ for *Roche 454 GS FLX Titanium XLR70*, etc. [17]. Currently, the technology is pushing towards read lengths of $k = 1000$ (e.g., *Roche 454 GS FLX Titanium XL+*) or even $k = 10000$ [18,19]. Needless to say, the longer the read length, the higher the chance that reads can be aligned to the reference genome. Ultimately, high quality genomes will be obtained by a mix of technologies. To find this optimal mixture, a quantitative understanding of the repeat structure of the human genome is required.

Our analysis of the repeat structure is different from some earlier investigations of read mappability [3,5]. In these studies, the actual reads from the current sequencing technology are used. There are two shortcomings in these approaches: (i) it is impossible to extrapolate the result to read lengths which is beyond the current technology; (ii) a certain proportion of reads are never mappable because the corresponding regions in the reference genome are not finished. Using the existing reference genome makes it possible to treat k-mers as hypothetic reads whose length k can be as long as possible, and unfinished regions can be excluded from the analysis.

In this paper we quantitatively address the question how alignment improves for greater read length. To this end, we artificially cut the human reference genome into overlapping k-windows (k-mers, k-tuples, or k-gram [20]), each considered to be possible a "read", and count the number of appearances (or "tokens", borrowing a terminology from linguistics [21]) of each k-mer type across the full reference sequence. Those k-mer types that appear in the genome only once ($f = 1$) are labeled singletons, and the remainder ($f > 1$) are non-singletons. Intuitively, the percentage of non-singleton reads is expected to decrease with increasing read length k. Obtaining the functional form of this decay enables us to predict the percentage of difficult-to-align reads for longer read lengths.

These seemingly simple calculations already encounter a "big data" problem on a regular-sized computer. In particular, storing counts in a hash table requires large amount of RAM. Suppose a k-mer needs K byte to store (e.g. $K = k/4$), a hash table to count all k-mers in the human genome would require $3K$ GByte RAM, which quickly becomes implausible when k is greater than 100. Using a solution that is similar to other applications where the hard disk [22-24] or computing time [25] is traded with RAM, we use a new public-domain program DSK which utilizes the less expensive hard disk and longer CPU time to compensate a lack of RAM [26]. Other efficient k-mer count procedures have been proposed in [27-29].

The mathematical relationship between the fraction of non-singleton k-mers and k predicts the fraction of putative reads that can be mapped uniquely. Another statistic of interest is the distribution of k-mer frequencies when k is fixed at a given value. This distribution has a head and a tail, a head for low frequency k-mers (including singletons), and a tail for high frequency k-mers. In the situation when these distributions exhibit long-tails [30] and power-law-like trends [31], thus fitting a straight line in log-log scale, the head end is best characterized by the frequency distribution [21], whereas the tail end is better characterized by the rank-frequency distribution commonly related to Zipf's law in quantitative linguistics [32]. Our analysis of these distributions provides information on the level of redundancy in the human genome at various scales.

The identification of regions in the human genome that cannot be uniquely mapped by reads (which can be called "non-uniqueome" following the term "uniqueome" used in [3]) is important in any NGS-based studies. These regions may contribute the most the false-positive and false-negative variant callings. These may also be hotspots for structural variations such as copy-number-variation [33,34]. We will specifically examine the location of some of these regions at the $k = 1000$ level.

Methods

Genome sequence data

The human reference genome GRCh37 (hg19) was downloaded from UCSC's Genome Browser (http://genome.ucsc.edu/). The intermittent strings of N's (marking unfinished basepairs that cannot be sequenced with the applied technology [35]) are used to partition the 22 autosomes and 2 sex chromosomes into 322 subsequences, and k-mers overlapping two chromosome partitions are not allowed.

For an additional analysis on repeat-filtered sequences, strings of lowercase letters in the reference genome (which mark repetitive sequences identified by the RepeatMasker program, http://www.repeatmasker.org/) are used to partition the genome into 3,456,905 subsequences with all transposable elements removed.

We further use the database *Dfam* version 1.2 (May 2013) (http://dfam.janelia.org/) [36] to annotate genomic regions by repeat sequences. *Dfam* contains the genomic locations of more than a thousand (1132) of transposable elements (TE) subfamily types. A hit is recorded whenever our genomic region overlaps with a TE. *Dfam* also provides information on tandem repeats by the program Tandem Repeat Finder [37].

Segmental duplication annotation of the human genome, which is either based on unusually high read coverage of whole-genome shotgun sequence segments from the Celera Genomics [38], or by a self-alignment by BLAST [39] on the RepeatMasker filtered genome ("fuguization") [40,41], is obtained from the Segmental Dups track ("Duplications of $>$ 1000 bases of non-RepeatMasker sequence") at Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=genomicSuperDups).

Counting k-mers

A k-mer type includes both the direct and the reverse complement substring; AAGC/GCTT is an example of such a 4-mer type. We use a state-of-art k-mer counting program DSK [26] (http://minia.genouest.org/dsk/), version 1.5031 (March 26, 2013). Most of the DSK calculations were carried out on a Linux computer with 48 GByte RAM and around 900 GByte disk space, except a calculation at $k = 1000$ which was run on another Linux computer with the same RAM but 30 TByte of disk space. The parameter setting of DSK was determined by a trial-and-error process. The output of the DSK program consists of a list of k-mers. The BLAT program from UCSC's Genome Browser is used to map frequent k-mers back to the reference genome.

Frequency distribution, rank frequency plot, and data fitting

Suppose a k-mer type appears in the genome f times (f is frequency, or copy number); frequency distribution (FD) is the number of k-mer types with frequency f. Individual k-mer types can be ranked by their f, highest f ranks number 1, second highest f ranks number 2, etc. The ranked f's of k-mer types as a function of rank r is the rank-frequency distribution (RFD).

The functions used here in fitting the RFD can all be expressed as linear regression, include Weibull function: $\log(f) \sim \log(\log((\max(r) + 1)/r))$ [42]; quadratic logarithmic: $\log(f) \sim \log(r) + (\log r)^2$ [43]; and reverse Beta: $\log(r) \sim \log(f) + \log(\max(f) + 1 - f)$. The latter function is derived from the Beta rank function [44-46] by reversing the f and r. All linear regressions are carried out by the *R* function *lm* (http://www.r-project.org/).

Results

Percentage of non-singleton reads vs. read length: piece-wise power-law function

In Figure 1 we show the percentage of non-singleton reads/tokens (p_{ns}) as a function of k-mer length k in log-log scale. The p_{ns} is 28.35% at $k = 20$, 8.16% at $k = 50$, 4.26% at $k = 80$, 3.40% at $k = 100$, 2.44% at $k = 150$, 1.33% at $k = 400$, 1.18% at $k = 500$, and 0.82% at $k = 1000$. If k is shorter than the "shortest unique substring" length, which is 11 in the human genome [47], singletons do not exist (i.e., $p_{ns} = 100\%$).

Visual inspection of the trend suggests the use of piecewise power-law function in fitting the data. We fit the points in $k = 20 - 80$ and $k = 200 - 1000$ ranges separately by linear regressions in the log-log scale: $\log_{10} p_{ns} = a + b \log_{10} k$ (or $\log p_{ns} \sim \log k$). The fitted $(\hat{a}, \hat{b})$ is (1.58366, -1.5478) and (-0.4371, -0.5495) for the two segments, equivalent to $p_{ns} = 38.34/k^{1.548}$ and $p_{ns} = 0.365/k^{0.55}$. The steep decay in the first segment shows a stronger increase of the amount of uniquely mappable sequences with read length, which implies that obtaining read lengths of at least around 100 is more cost-efficient with respect to reducing the amount of non-mappable reads. Of course, longer reads have extra benefits such as more robust alignments in the presence of polymorphisms or the ability to determine the length of longer repeat polymorphisms. The power-law function also indicates that the reduction of non-specific, difficult-to-align reads with longer read length is not linear.

If we assume our fitting function can be extrapolated to larger k's for which a direct analysis of k-mer frequencies is restricted by computational constraints, the proportion of non-singleton reads can be predicted. For example, this leads to the prediction of a 0.2% non-singleton rate at the 10kb read length.

It is known that repetitive sequences create considerable obstacle in NGS alignment [48]. Though TE's may exhibit subtle correlation with functional units in the genome [49], it is generally assumed that their biological role is indirect. Accordingly, we also looked at the non-singleton k-mer percentages in RepeatMasker filtered sequences (Figure 1). As expected, the percentage of uniquely mappable sequence is much higher than in the all-inclusive sequence for short k-mers (e.g. $k < 100$). Interestingly, the

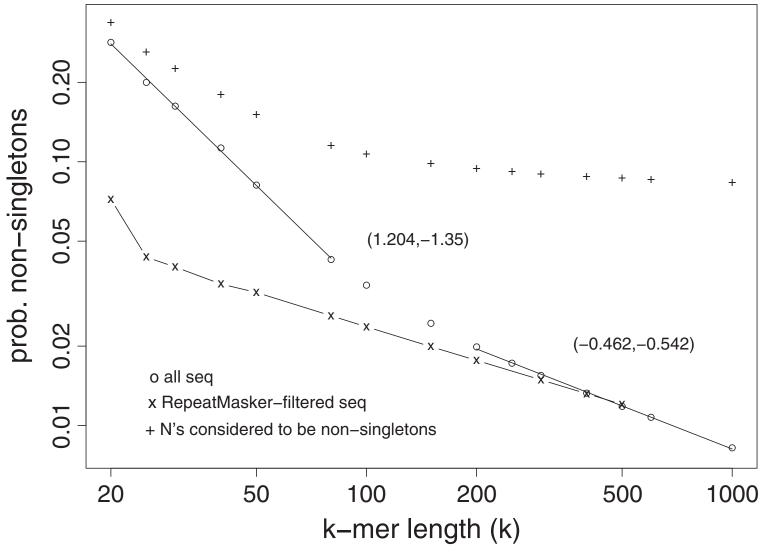

Figure 1 Proportion of non-singleton *k*-mers/tokens in the human genome (24 chromosomes) as a function of *k* (in log-log scale). Circles (o) show the results for all finished basepairs, whereas crosses (x) for the result from RepeatMasker-filtered sequences. Pluses (+) are results when unfinished sequences (234 Mbase) are included as non-singletons.

differences between the two disappear for longer k-mers (e.g. $k = 500$). A note of caution is that 89% of these RepeatMasker-filtered subsequences are shorter than 1kb, making the statistics less reliable at longer k's.

Maximum *k*-mer frequency decreases with *k* slowly

Another measure of the level of redundancy at length scale k is the maximum frequency ($\max(f)$) of k-mer types. For example, base A/T homopolymers of length 20 appear most often with 898,647 copies; at $k = 400$, AT repeats have more copy numbers ($f = 150$) than other 400-mers; the $\max(f)$ for $k = 1000$ is equal to 24 for a sequence which is not filtered by the RepeatMasker. The $\max(f)$ as a function of k is shown in Figure 2 in log-log scale.

For RepeatMasker-filtered sequences, $\max(f)$ quickly decays below 100 and then falls only slowly, indicating that RepeatMasker usually finds shorter repeats. At $k =$ 200–500, the k-mer with the $\max(f) \sim 50$ is a low-complexity sequence, with internal repeats of GGGGG GAACAGCGACAC/GTGTCCGCTGTTCCCCCC. Despite its high prevalence, this low-complexity sequence is not masked by RepeatMasker in the human reference genome.

Fitting the linear regression model $\log_{10} \max(f) = a + b \log_{10} k$ (or $\log \max(f) \sim \log k$) leads to $(a, b) =$ (8.99, -2.62). Extrapolating this regression to longer k's predicts that at $k = 2724$, $\max(f) = 1$. This prediction should be viewed with caution as $\max(f)$ is mainly determined by "outlier" events thus un-reproducible in principle, and the linear function in Figure 2 does not fit the data perfectly. Any extrapolation, exemplified by both Figure 1 and Figure 2, is based on the assumption that the fitted function in the observed range will continue as the same outside the range. There is no guarantee that this assumption is true in the present case.

Frequency distributions at fixed k values exhibit power-law-like trend

The frequency distribution (FD) describes the distribution of k-mer types according their copy numbers in the genome. When plotted in log-log scale, low-frequency k-mer types and the less redundant portion of the sequence are highlighted. Figure 3 shows five FDs at $k = 30, 50, 150,$ 500, and 1000 in log-log scale. The FDs at $k = 30$ and 50 span a wider frequency range, and the power-law trend is obvious.

A similar FD for $k = 40$ in human genome was shown in [50,51], and a slope of −2.3 in linear regression (in log-log scale) in the $f = 3$–500 range was reported. When we fit the $k = 50$ FD by linear regression in log-log scale, a very similar fitting slope value is obtained (−2.38, for $f =$ 3-200). However, it is clear from Figure 3 that the slopes are steeper for $k = 150$ (−2.7 for $f = 2$-100), $k = 500$ (−3.5 for $f = 2$-40), and $k = 1000$ (−5.3 for $f = 2$–19, or −5.9 from $f = 2$-9), indicating that the slope is not a universal parameter.

From the short read alignment perspective, the long tail at the high copy-numbers shows that many sequences cannot be uniquely mapped at smaller k values (e.g. $k = 30, 50$). However, the tail is much shortened at $k = 1000$. As expected, the tail for RepeatMasker-filtered sequences at various k values are much shorter (Figure 3, grey lines).

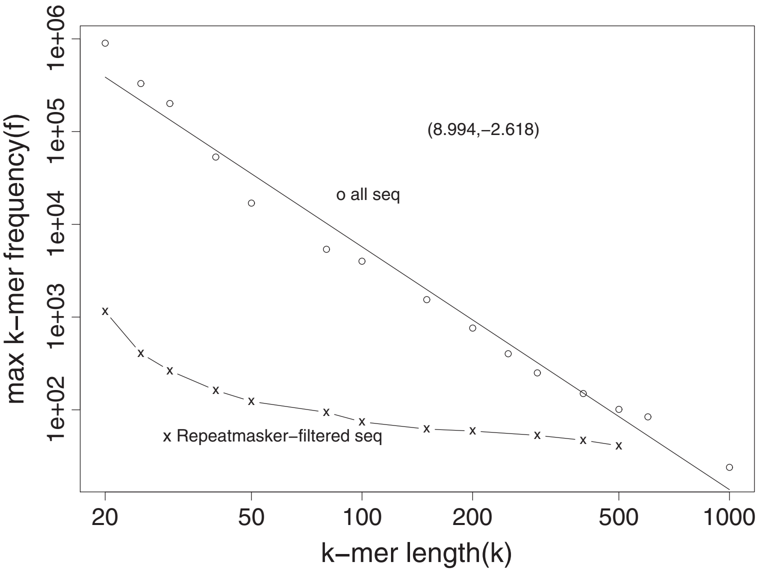

Figure 2 Maximum frequencies of *k*-mers as a function of *k* (in log-log scale). Circles (o) show the results for all finished bases, whereas crosses (x) for the result from RepeatMasker-filtered bases.

Rank-frequency distributions at fixed k values mostly follows a concave curve in log-log scale

Although a rank-frequency distribution (RFD) can be converted to cumulative FD [42], in log-log scale, it zooms into the high-frequency tail of the frequency distribution. Figure 4 shows five RFD at *k*'s from 30 to 1000. While the RFD at $k = 30$ may maintain a power-law or piecewise power-law trend, those at larger *k* values become more concave. This concave Zipf's curve is commonly observed in city size distributions [52,53].

For RFDs deviating from the Zipf's law, functions with two parameters may be used to account for the concave or convex shape of the curve in log-log scale [42]. We found that the quadratic logarithmic function, but not the Weibull function, fits the RFDs well (Figure 5). The Beta rank function usually exhibit "S" shapes [45], whereas the RFD in Figure 4 shows a "Z" shape. This motivated us to use a novel reverse Beta function to fit the data (Figure 5). The "Z" shaped log-log RFD means that if the power-law function is the default functional relationship between

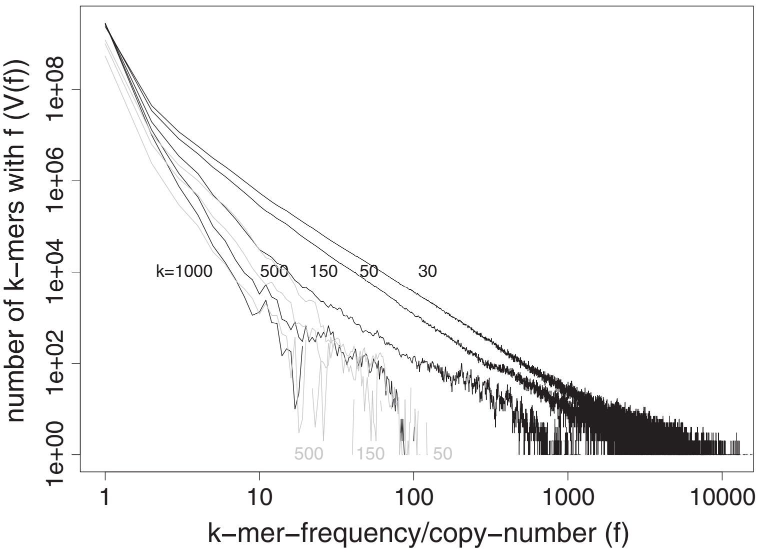

Figure 3 Frequency distributions of *k*-mers at *k* = 30, 50, 150, 500, and 1000 (in log-log scale). The distributions for *k*-mers in repeat-filtered sequences at $k = 50, 150, 500$ are shown in grey lines.

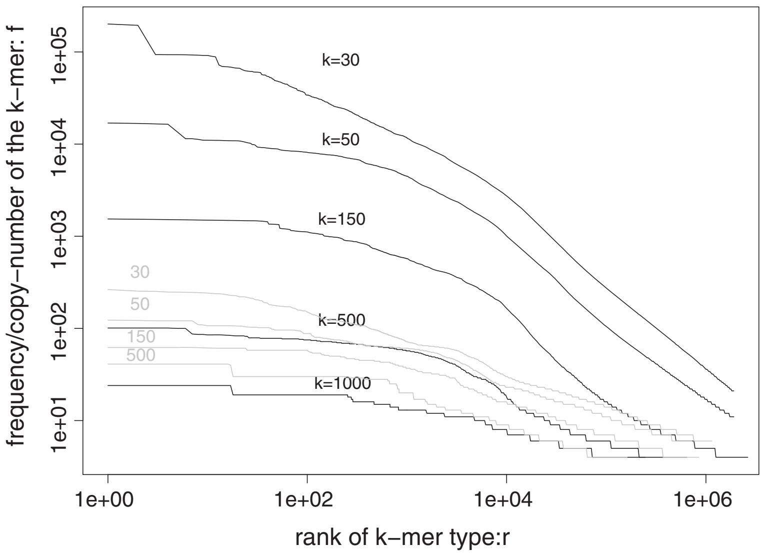

Figure 4 Rank-frequency distributions for *k*-mers at *k* = 30, 50, 150, 500, and 1000 (in log-log scale). The corresponding rank-frequency distributions for RepeatMasker-filtered sequences at $k = 30, 50, 150, 500$ are shown in grey lines.

frequency and rank, frequencies of the intermediately-ranked k-mers decrease faster than the two tails. The “S” shaped log-log RFD implies the opposite.

Mapping $f \geq 10$ 1000-mer to the reference genome

For $k = 1000$, there are 6107 k-mer types with frequency f larger or equal to 10. Due to the fact that these are overlapping k-mers, they are mapped to only 172 chromosomal regions, each of a few kb (the 172 locations, number of high-frequency 1000-mers, and the distance from the left-neighboring chromosome regions are included in Additional file 1: Table S1).

A total of 70 out of these 172 regions (or 40%) are clustered in four larger stretches on chromosomes 1 and X

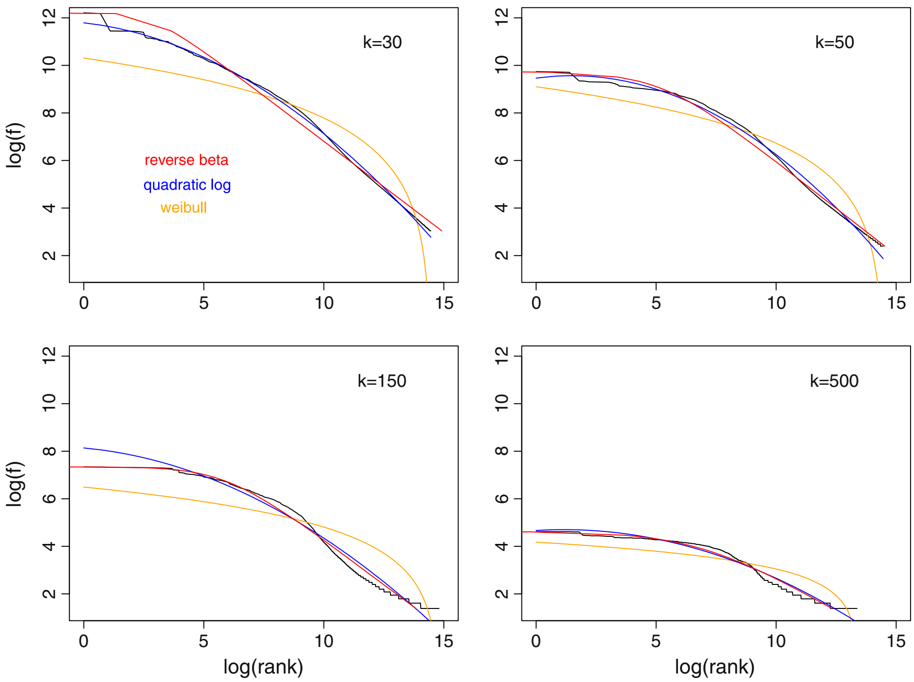

Figure 5 Fitting rank-frequency distribution of *k*-mers at k = 30, 50, 150, 500 using three functions. Red: quadratic logarithmic ($\log f \sim \log(r) + \log((r))^2$, f: frequency of a k-mer type, r: rank of a k-mer type, and the $\sim$ symbol represents linear regression); blue: reverse Beta rank function ($\log(r) \sim \log(f) + \log(max(f) + 1 - f)$); Orange: Weibull function ($\log(f) \sim \log(\log((max(r) + 1)/r))$).

and contain long tandem repeats (60, 70 kbase on chromosome 1q21.1, 1q21.2, and 41, 56 kbases on Xq23, Xq24). The two stretches on chromosome 1 contain copies of the neuroblastoma breakpoint family genes (*NBPF*) [54-56]. The Xq24 region contains cancer/testis antigen family genes (*CT47A*) [57,58], whereas the Xq23 region has no genes, but contains the macrosatellite *DXZ4* [59-61] which exhibits periodic appearance of other functional elements, such as H3K27Ac or H3K4me2 [62] histone modification marks.

Besides these long stretches, 39 out of 172 regions (or 23%) overlap with 34 genes: *ZNF3850, EPHA3, COL6A6, CD38, KCNIP4, FRAS1, ANTXR2, HSD17B11, FAM190A, DKK2, FBXL7, AK123816, FAM153A, FAM65B, LAMA2, MYCT1, NOD1, TPST1, PSD3, KCNB2, NR4A3, C9orf171, CACNA1B, DLG2, CCDC67, UACA, HOMER2, SMG1, CDH13, PRKCA, LILRA2, TTC28, MTMR8*, and *SLC25A43*. Obtaining high quality data on genetic variants in these genes is therefore likely to remain a challenge even with longer reads.

The distribution of transposable elements in the 172 regions is analyzed using the *Dfam* database. Interestingly, 1q21.1, 1q21.2, Xq23 regions discussed above do not overlap with any transposable elements. The Xq24 region contains a subfamily of Alu, AluSc8 (length $\sim$ 304, with mismatch-included copy number in the human genome $\sim$ 24000). Outside the four long stretches of genomic regions, however, almost all overlap with LINE-1 retrotransposons [63] (98/102, or 96%; 98/172, or 57%). Among these, the dominant LINE-1 subfamily is L1P1_orf2 (84/102, or 82%; 84/172 or 49%). The length of L1P1_orf2 is roughly 2174, and its mismatch-included copy number in the human genome is more than 16000.

Other LINE-1 subfamilies overlapping these regions include L1P1_5end, L1M2_5end, L1PA2_3end, and L1ME3G_3end. Three regions also overlap with a DNA transposon, Tigger3d. All transposable element information in these regions are listed in the Additional file 1: Table S1. Additional file 1: Table S1 also shows the tandem repeats result, such as TG-, AC-, or TTTA-repeat. Unlike transposable elements, these tandem repeats comprise a very small proportion of the region.

The Segmental Duplications Track in the Genome Browser provides repeat information that is different from the transposable elements. These repeats are usually large (> 1-15kb), and information is obtained either from the whole-genome shotgun sequencing reads, independent from the reference genome; or from the reference genome itself by self-alignment. We have listed overlapping information between our 172 regions and those in the Segmental Duplications Track in the Additional file 1: Table S1. Reassuringly, the four large regions on chromosomes 1 and X overlap with the previously identified segmentally duplicated regions, even though the methodology of the two approaches are very different.

By inspecting the Additional file 1: Table S1, it can be seen that the 172 regions either contain LINE transposable elements or overlap with the segmental duplication track. The large stretch on Xq24 overlaps with both segmental duplication track and transposable elements. However, the transposable element contained is the Alu element, which is a SINE instead of LINE. Possible connections between segmental duplication and Alu elements have been discussed before [64], and it is possible that the Alu element appeared in this region before the onset of duplication.

Discussion

Long *k*-mers in the reference genome as surrogate for sequencing reads

The *k*-mer distribution has many application in sequence analysis, such as measuring similarity between two genomes [65], correcting sequencing error [66], finding repeat structures [67], determining the feasibility of gene patents [68]. In many applications, only short *k*-mers are considered to be relevant, such as $k = 6$ [69], $k \leq 7$ [70], $k = 8$ [71], $k = 11$ [72]. This paper essentially uses long *k*-mers taken from the reference genome as surrogate for reads from future NGS technologies. Computationally speaking, counting long *k*-mers is more challenging and we are not aware of any prior publications on the long *k*-mer distributions in the human genome for *k* as long as 1000.

As compared to other papers on mappability of genome sequencing reads [3,5], our more theoretical approach has the advantage of being able to discuss long reads (e.g. $k = 1000$) where such data is not available from the current NGS technology. Our approach also separates the two causes of poor mappability: one due to the unfinished sequence in the reference genome and another due to the redundancy in the finished sequences. The unfinished bases are mainly located in the centromeres, short arms of acrocentric chromosomes and other heterochromatic regions, and rich in repetitive sequences. If we always treat this unfinished sequences (total 234 Mbases) to be non-singletons regardless of *k*, p_{ns} would flatten out around 0.1 (see Figure 1).

A baseline knowledge of redundancy of the human genome at length *k* level

Figures 1, 2 and 3 provides a baseline knowledge of the redundancy of the human genome at the *k*-mer level. Our results give a quantitative description of the effect of read length *k* on the mappability of reads from the finished region of the human genome.

Reference assembly is easier than *de novo* assembly, and our approach does not directly apply to *de novo* sequencing “assemblability”. However mappability and assemblability are closely related, as repetitive sequences cause problems in both situations [73]. The current *de novo* assemblies still do not perform consistently [74,75] and a quantitative assessment of the impact of repetitive sequences on reference assembly could be a useful piece of information for *de novo* assembly as well. Note that some discussion on k-mer-based assembly actually refers to k'-mer ($k' << k$) [76,77].

Highly redundant regions at $k = 1000$ level and copy-number-variation regions

The chromosome 1 and X regions which we have identified by showing at least 10 copy numbers of 1000-mers are discussed in the literature as regions with common copy-number-variations (CNV). CNVs in the 1q21.1 region, if not *NBPF*-specific, have been linked to congenital cardiac defects [78-80], autism [81,82], mental retardation [83], head size abnormalities [84], schizophrenia [85,86], and neuroblastoma [87]. With so many abnormalities mapped to this region, these are collectively called the chromosome 1q21.1 duplication syndrome in the Online Mendelian Inheritance in Man (OMIM 612475).

The Xq23 region, if not macrosatellite *DXZ4* specific, has been identified as likely CNV regions linked to developmental and behavioral problems [88]. Chromatin configuration at *DXZ4* region is reported to differ between male melanoma cells and normal skin cells [89]. The Xq24 and the *CT47A* gene are listed as a region of structural variants associated with intellectual disability [90] and mental retardation [91].

A well-known mechanism for CNV formation is non-allelic homologous recombinations (NAHR) between repetitive elements [92]. More copies of a repetitive sequence give more opportunities that NAHR could occur, resulting in a natural connection between repetitive sequences and CNV. The fact that simple counting of 1000-mer frequencies leads to CNV regions with medical implications indicates that understanding the k-mer distribution is an important part of genomic analyses.

Although the four highlighted large regions also appear in the Segmental Duplication track for > 1000 bp RepeatMasker-filtered sequences in the UCSC Genome Browser, the two methodologies are somewhat different. Here, we use the reference genome as starting point, length scale is upper-limited at 1000 bp, zero-mismatch, and high copy numbers (≥ 10). In SegDup track, the reference may or may not be used (in the latter case, raw reads are the starting point), length scale is lower-limited at one or few kbs, mismatches are allowed, and low copy number (e.g. 2) is allowed. From this may lead to the development of strategy where our approach can be used to check the consistency of the reference genome with raw read data.

Discussions of extensions to a next-generation-sequencing data

In a realistic setting of NGS, there are sequencing errors and single-nucleotide polymorphisms (SNP); alignment to the reference genome may allow mismatches; and there is a wide adoption of paired-end/mate-pair strategy [93-96]. It is a daunting challenge to provide a definitive answer under these situations [4] for long k-mer lengths such as $k = 1000$. Some concepts in this paper, e.g., the k-mer frequency distribution in Figure 3, cannot be used if mismatches are considered.

We can however speculate about some consequences when practical complications are introduced. Suppose a DNA fragment (of length k) is split into two ends (of length $k' < k/2$ each) which are to be sequenced, and an insert (of length $k - 2k'$). At $k' = k/2$, one is essentially sequencing the whole DNA fragment, and aligning two k'-mers next to each other is equivalent to aligning a $2k'$-mer. The result in Figure 1 implies that the proportion of non-mappable reads/tokens decreases with k' as $1/(2k')^b$. When $k \ll 2k'$, aligning two paired-end k'-mers is more likely to be unique than when the two k'-mers are next to each other, as the correlation between two k'-mers decrease with distance [97]. We may speculate that the proportion of non-uniquely-mapped reads as a function of k' and k is: $\sim f(k-2k')/(2k')^b$, where the unknown function $f(k - 2k')$ is 1 if $k = 2k'$, and decreases with $k - 2k'$.

There have been recent attempts to fill in the sequence of inserts between two ends in the pair-end strategy [98-101]. A typical example would consider a segment length k of 600-800 bp, and read length k' of 100 bp [101]. We then can consider the best scenario that the sequence of the whole segment of length k can be determined. This will merely shift the length scale from the two times the read length ($2k'$) to the segment length (k), and all our results still apply.

The effect of sequencing errors, single-nucleotide polymorphism, alignment allowing mismatches, can be discussed in the framework of k-mer space (with reverse complement). The observed k-mers in the human genome consist of a subspace of the k-mer space, and a link between two k-mers is established when the Hamming distance between the two is 1. Sequencing errors and polymorphisms either generate a new k-mer in this subspace, or move along a link to a previously existing k-mer. If new k-mers are generated, links between k-mers will be recalculated. One can argue that sequencing error and polymorphism would have less impact if the error/mutation does not lead to the creation of a new k-mer, or, even when a new k-mer is created, if the new k-mer does not have new links to other k-mers. In the case where sequencing errors and polymorphisms generate two or more mutations, links between k-mers with both 1- and

2-Hamming distances should be considered. The framework of discussion is similar, though more complicated.

Long-tails and the regime of diminishing return of longer reads

Our analysis shows that all distributions discussed in this paper are better viewed in log-log scale, proving the existence of power-law distributions or long-tails. This has been observed in the past for other genomic distributions, such as correlation function [97,102-104], power spectrum of base composition [105-108], frequency distribution of gene or protein family size [109-112], sizes of ultraconserved regions [113], and in models with duplications [114-117]. Ongoing duplications increase the copy number geometrically, which explains the presence of long-tails.

A consequence of the long-tail in Figure 1 is that with increasing read (or *k*-mer) lengths, the proportion of reads that cannot be mapped to a unique genomic region (within the finished sequences) decreases as a power-law function, as compared to a linear or exponential function. Numerically, if not economically, this defines a regime of diminishing return. It is important to emphasize that we have only directly observed an diminishing return in the range of 200-1000 bp. This diminishing return may be extended further beyond 1kb, until it reaches a point of accelerating return if the read length is longer than the size of any segmental duplication region (which can be 200kb for gene-containing duplications [118]). The use of paired-end strategy usually does not increase the length scale by orders of magnitude, thus it may still be confined to the diminishing return regime. To assess the economic return with NGS technology with longer reads, other factors should be considered, such as the choice of less redundant target regions such as the exome [119], read length and sequencing error tradeoff, and the overall cost of longer-read sequencing.

Conclusion

We have established that, up to 1000 bases, the mappability of reads decreases slower than linear with read length, when mappability is measured as the proportion of non-singletons in human reference genome. The slow decrease is similar to other observed long tail distributions in genomics. Anticipating that the highest-quality human genome sequences will be obtained by a combination of various technologies, the analysis of *k*-mer distribution at different scales is a prominent factor for determining how these technologies can be optimally combined. We also identified the most redundant 1000-mers in the human genome, which include the region responsible for the chromosome 1q21.1 duplication syndrome, as well as other regions which are rich in segmental duplication and macrosatellites.

Availability of support data

The data set supporting the results of this article is included within the article and its additional file.

Additional file

Additional file 1: The additional file includes the supplementary Table S1: 172 chromosome locations with high-frequency (*f* ≥ 10) 1000-mers.

Abbreviations

BLAT: BLAST like alignment tool; BLAST: Basic local alignment search tool; CPU: Central processing unit; CNV: Copy number vatiations; DSK: Disk streaming of kmers; DNA: Deoxyribo-nucleic acids; FD: Frequency distribution; GRCh37: Genome reference consortium human (build) 37; LINE: Long interspersed elements; NAHR: Non-allelic homologous recombinations; NGS: Next-generation sequencing; OMIM: Online Mendelian Inheritance in Man; RAM: Random-access memory; RFD: Rank-frequency distribution; SINE: Short interspersed elements; SNP: Single nucleotide polymorphism; TE: Transposable elements; UCSC: University of California at Santa Cruz.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

WL conceived of the study and contribute to the analysis of the data. JF carried out the mapping of redundant 1000-mers to the reference genome. PM carried out the fitting of rank-frequency distribution. WL, JF, PM contribute to draft of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Oliver Clay, Andrew Shih, Astero Provata, Yannis Almirantis for discussions, and the authors of DSK for timely responding to our inquiries and fixing bugs. WL acknowledges the support from the Robert S Boas Center for Genomics and Human Genetics, and JF was supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health under Award Number 1R03AR063340.

Author details

[1]The Robert S. Boas Center for Genomics and Human Genetic, The Feinstein Institute for Medical Research, North Shore LIJ Health System, 350 Community Drive, Manhasset, USA. [2]Departamento de Matemáticas, Facultad de Ciencias, Universidad Nacional Autónoma de México, Circuito Exterior, Ciudad Universitaria, 04510 DF México, México.

Received: 25 August 2013 Accepted: 17 December 2013
Published: 3 January 2014

References

1. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB: **PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls.** *NatBiotech* 2009, **27:**66–75.
2. Cahill MJ, Köser CU, Ross NE, Archer JAC: **Read length and repeat resolution: exploring prokaryote genomes using next-generation sequencing technologies.** *PLoS ONE* 2010, **5:**e11518.
3. Koehler R, Issac H, Cloonan N, Grimmond SM: **The uniqueome: a mappability resource for short-tag sequencing.** *Bioinformatics* 2011, **27:**272–274.
4. Derrien T, Estellé J, Marco Sola M, Knowles DG, Raineri E, Guigó R, Ribeca P: **Fast computation and applications of genome mappability.** *PLoS ONE* 2012, **7:**e30377.
5. Lee H, Schatz MC: **Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score.** *Bioinformatics* 2012, **28:**2097–2105.
6. Storvall H, Ramsköld D, Sandberg R: **Efficient and comprehensive representation of uniqueness for next-Generation sequencing by minimum unique length analyses.** *PLoS ONE* 2013, **8:**e53822.

7. Weber JL, Myers EW: **Human whole-genome shotgun sequencing.** *Genome Res* 1997, **7:**401–409.
8. Green ED: **Strategies for the systematic sequencing of complex genomes.** *Nat Rev Genet* 2001, **2:**573–583.
9. Fraenkel AS, Gillis J: **Appendix II. Proof that sequences of A, C, G, and T can be assembled to produce chains of ultimate length avoiding repetitions everywhere.** *Prog Nucl Acids Res Mol Biol* 1966, **5:**343–348.
10. Stoppa-Lyonnet D, Carter PE, Meo T, Tosi M: **Clusters of intragenic Alu repeats predispose the human C1 inhibitor locus to deleterious rearrangements.** *Proc Natl Acad Sci* 1990, **87:**1551–1555.
11. Conrad B, Antonarakis SE: **Gene duplication: a drive for phenotypic diversity and cause of human disease.** *Ann Rev Genomics Hum Genet* 2007, **8:**17–35.
12. Ohno S: *Evolution by Gene Duplication*. New York: Springer-Verlag; 1970.
13. Nowak MA, Boerlijst, Cooke J, Maynard Smith J: **Evolution of genetic redundancy.** *Nature* 1997, **388:**167–171.
14. Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, Meltesen L, Brenton M, Hink R, Burgers S, Hernandez-Boussard T, Karimpour-Fard A, Glueck D, McGavran L, Berry R, Pollack J, Sikela JM: **Lineage-specific gene duplication and loss in human and great ape evolution.** *PLoS Biol* 2004, **2:**E207.
15. Krakauer DC, Plotkin JB: **Redundancy, antiredundancy, and the robustness of genomes.** *Proc Natl Acad Sci* 2002, **99:**1405–1409.
16. Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D: **A cencus of protein repeats.** *J Mol Biol* 1998, **293:**151–160.
17. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M: **Comparison of next-generation sequencing systems.** *J Biomed Biotech* 2012, **2012:**251364.
18. Eisenstein M: **Companies 'going long' generate sequencing buzz at Marco island (news).** *Nat Biotech* 2013, **31:**265–266.
19. Heiner C, Wang S, Ashby M, Guo Y, Underwood J: **Greater than 10 kb read lengths routine when sequencing with Pacific Biosciences' XL release.** *J Biomol Tech* 2013, **24(suppl):**S43.
20. Brown PF, deSouza PV, Mercer RL, Pietra VJ, Lao JC: **Class-based n-gram models of natural languages.** *J Comp Linguist* 1992, **18:**467–479.
21. Baayen RH: *Word Frequency Distribution*. Dordrecht: Kluwer Academic Publishers; 2001.
22. Phoophakdee B: **TRELLIS: genome-size disk-based suffix tree indexing algorithm.** *Ph.D Thesis,* Rensselaer Polytechnic Institute, Troy, NY, 2007.
23. Phoophakdee B, Zaki MJ: **TRELLIS+: an effective approach for indexing genome-scale sequences using suffix trees.** *Pacif Sym Biocomp* 2008, **2008:**90–101.
24. Li Q, Yu C, Li Y, Lam TW, Y SM, Kristiansen K, Wang J: **SOAP2: an improved ultrafast tool for short read alignment.** *Bioinformatics* 2009, **25:**1966–1967.
25. Chu HT, Hsiao WWL, Tsao TT, Hsu DF, Chen CC, Lee SA, Kao CY: **SeqEntropy: genome-wide assessment of repeats for short read sequencing.** *PLoS ONE* 2013, **8:**e59484.
26. Rizk G, Lavenier D, Chikhi R: **DSK, k-mer counting with very low memory usage.** *Bioinformatics* 2013, **29:**652–653.
27. Kurtz S, Narechania A, Stein JC, Ware D: **A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes.** *BMC Genomics* 2008, **9:**517.
28. Marçais G, Kingsford C: **A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.** *Bioinformatics* 2011, **27:**764–770.
29. Melsted P, Pritchard JK: **Effecient counting of k-mers in DNA sequences using a bloom filter.** *BMC Bioinfo* 2011, **12:**333.
30. Anderson C: *The Long Tail: Why the Future of Business is Selling Less of More*. New York: Hyperion; 2006.
31. Clauset A, Shalizi CR, Newman MEJ: **Power-law distributions in empirical data.** *SIAM Rev* 2007, **51:**661–703.
32. Zipf GK: *Human Behavior and the Principle of Least Effort*: Addison-Wesley; 1949.
33. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE: **Segmental duplications and copy-number variation in the human genome.** *Am J Hum Genet* 2005, **77:**78–88.
34. Perry GH, Tchinda J, McGrath SD, Zhang J, Picker SR, Cáceres AM, Iafrate AJ, Tyler-Smith C, Scherer SW, Eichler EE, Stone AC, Lee C: **Hotspots for copy number variation in chimpanzees and humans.** *Proc Natl Acad Sci* 2006, **101:**8006–8011.
35. Genovese G, Handsaker RE, Li H, Altemose N, Lindgren AM, Chambert K, Pasaniuc B, Price AL, Reich D, Morton CC, Pollak MR, Wilson JG, McCarroll SA: **Using population admixture to help complete maps of the human genome.** *Nat Genet* 2013, **45:**406–414.
36. Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones TA, Jurka J, Smit AF, Finn RD: **Dfam: a database of repetitive DNA based on profile hidden Markov models.** *Nucleic Acids Res* 2013, **41:**D70–D82.
37. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27:**573–580.
38. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE: **Recent segmental duplications in the human genome.** *Science* 2002, **297:**1003–1007.
39. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403–410.
40. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE: **Segmental duplications: organization and impact within the current human genome project assembly.** *Genome Res* 2001, **11:**1005–1007.
41. Cheung J, Estivill X, Khaja R, MacDonald JR, Lau K, Tsui LC, Scherer SW: **Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence.** *Genome Biol* 2003, **4:**R25.
42. Li W, Miramontes P, Cocho G: **Fitting ranked linguistic data with two-parameter functions.** *Entropy* 2010, **12:**1743–1764.
43. Li W, Miramontes P: **Fitting ranked English and Spanish letter frequency distribution in US and Mexican presidential speeches.** *J Quant Linguist* 2011, **18:**337–358.
44. Mansilla R, Köppen E, Cocho G, Miramontes P: **On the behavior of journal impact factor rank-order distribution.** *J Infometrics* 2007, **1:**155–160.
45. Martínez-Mekler G, Alvarez Martínez R, Beltrán del Río, M, Mansilla R, Miramontes P, Cocho G: **Universality of rank-ordering distributions in the arts and sciences.** *PLoS ONE* 2009, **4:**e4791.
46. Miramontes P, Li W, Cocho G: **Some critical support for power laws and their variations.** arXiv preprint. arXiv:nlin.AO/1204.3124, 2012.
47. Haubold B, Pierstorff N, Möller F, Wiehe T: **Genome comparison without alignment using shortest unique substrings.** *BMC Bioinfo* 2005, **6:**123.
48. Treangen TJ, Salzberg SL: **Repetitive DNA and next-generation sequencing: computational challenges and solutions.** *Nat Rev Genet* 2012, **13:**36–46.
49. Li W, Sosa D, Jose MV: **Human repetitive sequence densities are mostly negatively correlated with R/Y-based nucleosome-positioning motifs and positively correlated with W/S-based motifs.** *Genomics* 2013, **101:**125–133.
50. Sindi SS: **Describing and Modeling Repetitive Sequences in DNA.** *Ph.D Thesis*, Univ. of Maryland; 2006.
51. Sindi SS, Hunt BR, Yorke JA: **Duplication count distributions in DNA sequences.** *Phys Rev E* 2008, **78:**061912.
52. Gabaix X, Ioannides YM: **The evolution of city size distributions.** In *Handbook of Regional and Urban Economics*. Edited by Henderson V, Thisse JF. North-Holland; 2004.
53. Eeckhout J: **Gibrat's law for (all) cities.** *Am Eco Rev* 2004, **94:**1429–1451.
54. Vandepoele K, Van Roy N, Staes K, Speleman F, van Roy F: **A novel gene family NBPF: intricate structure generated by gene duplications during primate evolution.** *Mol Biol Evol* 2005, **22:**2265–2274.
55. Paar V, Glunčić M, Rosandić M, Basar I, Vlahović I: **Intragene higher order repeats in neuroblastoma breakpoint family genes distinguish humans from chimpanzees.** *Mol Biol Evol* 2011, **28:**1877–1892.
56. Dumas LJ, O'Bleness MS, Davis JM, Dickens CM, Anderson N, Keeney JG, Jackson J, Sikela M, Raznahan A, Giedd J, Rapoport J, Nagamani SS, Erez A, Brunetti-Pierri N, Sugalski R, Lupski JR, Fingerlin T, Cheung SW, Sikela JM: **DUF1220-domain copy number implicated in human brain-size pathology and evolution.** *Am J Hum Genet* 2012, **91:**444–454.
57. Chen YT, Iseli C, Venditti CA, Old LJ, Simpson AJ, Jongeneel CV: **Identification of a new cancer/testis gene family, CT47, among expressed multicopy genes on the human X chromosome.** *Genes Chromosomes Cancer* 2006, **45:**392–400.

58. Dobrynin P, Matyunina E, Malov SV, Kozlov AP: **The novelty of human cancer/testis antigen encoding genes in evolution.** *Int J Genomics* 2013, **2013:**105108.
59. Giacalone J, Friedes J, Francke U: **A novel GC-rich human macrosatellite VNTR in Xq24 is differentially methylated on active and inactive X chromosomes.** *Nat Genet* 1992, **1:**137–143.
60. Tremblay DC, Moseley S, Chadwick BP: **Variation in array size, monomer composition and expression of the macrosatellite DXZ4.** *PLoS ONE* 2010, **6:**e18969.
61. Schaap M, Lemmers R, Maassen R, van der Vliet PJ, Hoogerheide LF, van Dijk HK, Baştürk N, de Knijff P, van der Maarel SM: **Genome-wide analysis of macrosatellite repeat copy number variation in worldwide populations: evidence for differences and commonalities in size distributions and size restrictions.** *BMC Genomics* 2013, **14:**143.
62. Horakova AH, Moseley SC, McLaughlin CR, Tremblay DC, Chadwick BP: **The macrosatellite DXZ4 mediates CTCF-dependent long-range intrachromosomal interactions on the human inactive X chromosome.** *Hum Mol Genet* 2012, **21:**4367–4377.
63. Smit AF, Tóth G, Riggs AD, Jurka J: **Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences.** *J Mol Biol* 1995, **246:**401–417.
64. Bailey JA, Liu G, Richler EE: **An Alu transposition model for the origin and expansion of human segmental duplications.** *Am J Hum Genet* 2003, **73:**823–834.
65. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinfo* 2004, **5:**113.
66. Liu Y, Schröder J, Schmidt B: **Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data.** *Bioinformatics* 2013, **29:**308–315.
67. Li X, Waterman MS: **Estimating the repeat structure and length of DNA sequences using l-tuples.** *Genome Res* 2003, **13:**1916–1922.
68. Rosenfeld J, Mason CE: **Pervasive sequence patents cover the entire human genome.** *Genome Med* 2013, **5:**27.
69. Chen YH, Nyeo SL, Yeh CY: **Model for the distributions of k-mers in DNA sequences.** *Phys Rev E* 2005, **72:**011908.
70. Nikolaou C, Almirantis Y: **'Word' preference in the genomic text and genome evolution: different modes of n-tuplet usage in coding and noncoding sequences.** *J Mol Evol* 2005, **61:**23–25.
71. Xie H, Hao B: *Visualization of K-tuple distribution in procaryote complete genomes and their randomized counterparts*. Los Alamitos: IEEE Computer Society Press; 2002.
72. Chor B, Horn D, Goldman N, Levy Y, Massingham T: **Genomic DNA k-mer spectra: models and modalities.** *Genome Biol* 2009, **10:**R108.
73. Paszkiewicz K, Studholme DJ: **de novo assembly of short sequence reads.** *Brief Bioinfo* 2010, **11:**457–472.
74. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, Chitsaz H, Chou, WC, Corbeil J, Del Fabbro C, Docking TR, Durbin R, Earl D, Emrich S, Fedotov P, Fonseca NA, Ganapathy G, Gibbs RA, Gnerre S, Godzaridis E, Goldstein S, Haimel M, Hall G, Haussler D, Hiatt JB, Ho IY, et al.: **Assemblathon 2: evaluting de novo methods of genome assembly in three vertebrate species.** arXiv preprint. arXiv:q-bio.GN/1301.5406, 2013.
75. Muñoz JF, Gallo JE, Misas E, McEwan JG, Clay OK: **The eukaryotic genome, its reads, and the unfinished assembly.** *FEBS Lett* 2013, **587:**2090–2093.
76. Zerbino D, Birney E: **Velvet: Algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18:**821–829.
77. Liu B, Yuan J, Yiu SM, Li Z, Xie Y, Chen Y, Shi Y, Zhang H, Li Y, Lam TW, Luo R: **COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly.** *Bioinformatics* 2010, **28:**2870–2874.
78. Christiansen J, Dyck JD, Elyas BG, Lilley M, Bamforth JS, Hicks M, Sprysak KA, Tomaszewski R, Haase SM, Vicen-Wyhony LM, Somerville MJ: **Chromosome 1q21.1 contiguous gene deletion is associated with congenital heart disease.** *Circ Res* 2004, **94:**1429–1435.
79. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, González JR, Gratacós M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, et al.: **Global variation in copy number in the human genome.** *Nature* 2006, **444:**444–454.
80. Greenway SC, Pereira AC, Lin JC, DePalma SR, Israel SJ, Mesquita SM, Ergul E, Conta JH, Korn JM, McCarroll SA, Gorham JM, Gabriel S, Altshuler DM, Quintanilla-Dieck Mde L, Artunduaga MA, Eavey RD, Plenge RM, Shadick NA, Weinblatt ME, De Jager PL, Hafler DA, Breitbart RE, Seidman JG, Seidman CE: **De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot.** *Nat Genet* 2009, **41:**931–935.
81. Autism Genome, Project Consortium, Szatmari P, Paterson AD, Zwaigenbaum L, Roberts W, Brian J, Liu XQ, Vincent JB, Skaug JL, Thompson AP, Senman L, Feuk L, Qian C, Bryson SE, Jones MB, Marshall CR, Scherer SW, Vieland VJ, Bartlett C, Mangin LV, Goedken R, Segre A, Pericak-Vance MA, Cuccaro ML, Gilbert JR, Wright HH, Abramson RK, Betancur C, Bourgeron T, Gillberg C, et al.: **Mapping autism risk loci using genetic linkage and chromosomal rearrangements.** *Nat Genet* 2007, **39:**319–328.
82. Girirajan S, Dennis MY, Baker C, Malig M, Coe BP, Campbell CD, Mark K, Vu TH, Alkan C, Cheng Z, Biesecker LG, Bernier R, Eichler EE: **Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder.** *Am J Hum Genet* 2013, **92:**221–237.
83. Mefford HC, Sharp AJ, Baker C, Itsara A, Jiang Z, Buysse K, Huang S, Maloney VK, Crolla JA, Baralle D, Collins A, Mercer C, Norga K, de Ravel T, Devriendt K, Bongers EM, de Leeuw N, Reardon W, Gimelli S, Bena F, Hennekam RC, Male A, Gaunt L, Clayton-Smith J, Simonic I, Park SM, Mehta SG, Nik-Zainal S, Woods CG, Firth HV, et al.: **Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes.** *New Eng J Med* 2008, **359:**1685–1699.
84. Brunetti-Pierri N, Berg JS, Scaglia F, Belmont J, Bacino CA, Sahoo T, Lalani SR, Graham B, Lee B, Shinawi M, Shen J, Kang SH, Pursley A, Lotze T, Kennedy G, Lansky-Shafer S, Weaver C, Roeder ER, Grebe TA, Arnold GL, Hutchison T, Reimschisel T, Amato S, Geragthy MT, Innis JW, Obersztyn E, Nowakowska B, Rosengren SS, Bader PI, Grange DK, et al.: **Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities.** *Nat Genet* 2008, **40:**1466–1471.
85. The International, Schizophrenia Consortium: **Rare chromosomal deletions and duplications increase risk of schizophrenia.** *Nature* 2008, **455:**237–241.
86. Ikeda M, Aleksic B, Kirov G, Kinoshita Y, Yamanouchi Y, Kitajima T, Kawashima K, Okochi T, Kishi T, Zaharieva I, Owen MJ, O'Donovan MC, Ozaki N, Iwata N: **Copy number variation in schizophrenia in the Japanese population.** *Biol Psych* 2010, **67:**283–286.
87. Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K, Cole K, Mossé YP, Wood A, Lynch JE, Pecor K, Diamond M, Winter C, Wang K, Kim C, Geiger EA, McGrady PW, Blakemore AI, London WB, Shaikh TH, Bradfield J, Grant SF, Li H, Devoto M, Rappaport ER, Hakonarson H, Maris JM: **Copy number variation at 1q21.1 associated with neuroblastoma.** *Nature* 2009, **459:**987–991.
88. Isrie M, Froyen G, Devriendt K, de Ravel T, Fryns JP, Vermeesch JR, Van Esch H: **Sporadic male patients with intellectual disability: contribution of X-chromosome copy number variants.** *Euro J Med Genet* 2012, **55:**577–585.
89. Moseley SC, Rizkallah R, Tremblay DC, Anderson BR, Hurt MM, Chadwick BP: **YY1 associates with the macrosatellite DXZ4 on the inactive X chromosome and binds with CTCF to a hypomethylated form in some male carcinomas.** *Nucleic Acids Res* 2012, **40:**1596–1608.
90. Whibley AC, Plagnol V, Tarpay PS, Abidi F, Fullston T, Choma MK, Boucher CA, Shepherd L, Willatt L, Parkin G, Smith R, Futreal PA, Shaw M, Boyle J, Licata A, Skinner C, Stevenson RE, Turner G, Field M, Hackett A, Schwartz CE, Gecz J, Stratton MR, Raymond FL: **Fine-scale survey of X chromosome copy number variants and indels underlying intellectual disability.** *Am J Hum Genet* 2010, **87:**173–188.
91. Honda S, Hayashi S, Imoto I, Toyama J, Okazawa H, Nakagawa E, Goto Y, Inazawa J: **Copy-number variations on the X chromosome in Japanese patients with mental retardation detected by array-based comparative genomic hybridization analysis.** *J Hum Genet* 2010, **55:**590–599.

92. Gu W, Zhang F, Lupski JR: **Mechanisms for human genomic rearrangement.** *PathoGenet* 2008, **1:**4.
93. Hong GF: **A method for sequencing single-stranded cloned DNA in both directions.** *Biosci Rep* 1981, **1:**243–252.
94. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M: **Paired-end mapping reveals extensive structural variation in the human genome.** *Science* 2007, **318:**420–426.
95. Williams LJ, Tabbaa DG, Li N, Berlin AM, Shea TP, Maccallum I, Lawrence MS, Drier Y, Getz G, Young SK, Jaffe DB, Nusbaum C, Gnirke A: **Paired-end sequencing of Fosmid libraries by Illumina.** *Genome Res* 2012, **22:**2241–2249.
96. Ramachandran P, Palidwor GA, Porter CJ, Perkins TJ: **MaSC: mappability-sensitive cross-correlation for estimating mean fragment length of single-end short-read sequencing data.** *Bioinformatics* 2013, **29:**444–450.
97. Li W: **The study of correlation structures of DNA sequences: a critical review.** *Comput Chem* 1997, **21:**257–271.
98. Rodrigue S, Materna AC, Timberlake SC, Blackburn MC, Malmstrom RR, Alm EJ, Chisholm SW: **Unlocking short read sequencing for metagenomics.** *PLoS ONE* 2010, **5:**e11840.
99. Magoč T, Salzberg SL: **FLASH: fast length adjustment of short reads to improve genome assemblies.** *Bioinformatics* 2011, **27:**2957–2963.
100. Liu B, Yuan J, Yiu SM, Li Z, Xie Y, Chen Y, Shi Y, Zhang H, Li Y, Lam TW, Luo R: **COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly.** *Bioinformatics* 2012, **28:**2870–2874.
101. Ruan J, Jiang L, Chong Z, Gong Q, Li H, Li C, Tao Y, Zheng C, Zhai W, Turissini D, Cannon CH, Lu X, Wu CI: **Pseudo-Sanger sequencing: massively parallel production of long and near error-free reads using NGS technology.** *BMC Genomics* 2013, **14:**711.
102. Li W, Kaneko K: **Long-range correlation and partial 1/f$^{\alpha}$ spectrum in a noncoding DNA sequence.** *Euro Phys Lett* 1992, **17:**655–660.
103. Bernaola-Galván P, Carpena P, Román-Roldán R, Oliver JL: **Study of statistical correlations in DNA sequences.** *Gene* 2002, **300:**105–115.
104. Arneodo A, Vaillant C, Audit B, Argoul F, d'Aubenton-Carafa Y, Thermes C: **Multi-scale coding of genomic information: from DNA sequence to genome structure and function.** *Phys Rep* 2011, **498:**45–188.
105. Voss RF: **Evolution of long-range fractal correlations and 1/f noise in DNA base sequences.** *Phys Rev Lett* 1992, **68:**3805–3808.
106. Fukushima A, Ikemura T, Kinouchi M, Oshima T, Kudo Y, Mori H, Kanaya S: **Periodicity in prokaryotic and eukaryotic genomes identified by power spectrum analysis.** *Gene* 2002, **300:**203–211.
107. Li W, Holste D: **Spectral analysis of guanine and cytosine fluctuations of mouse genomic DNA.** *Fluc Noise Lett* 2004, **4:**L453–L464.
108. Li W, Holste D: **Universal 1/f noise, crossovers of scaling exponents, and chromosome-specific patterns of guanine-cytosine content in DNA sequences of the human genome.** *Phys Rev E* 2005, **71:**041910.
109. Huynen M, van Nimwegen E: **The frequency distribution of gene family sizes in complete genomes.** *Mol Biol Evol* 1998, **15:**583–589.
110. Qian J, Luscombe NM, Gerstein M: **Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model.** *J Mol Biol* 2001, **313:**673–681.
111. Koonin EV: **Are there laws of genome evolution?** *PLoS Comp Biol* 2011, **7:**e1002173.
112. Herrada A, Euíluz VM, Hernández-García E, Duarte CM: **Scaling properties of protein family phylogenies.** *BMC Evol Biol* 2011, **11:**155.
113. Salerno W, Havlak P, Miller J: **Scale-invariant structure of strongly conserved sequence in genomic intersections and alignments.** *Proc Natl Acad Sci* 2006, **103:**13121–13125.
114. Li W: **Expansion-modification systems: a model for spatial 1/f spectra.** *Phys Rev A* 1991, **43:**5240–5260.
115. Yanai I, Camacho CJ, DeLisi C: **Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification.** *Phys Rev Lett* 2000, **85:**2641–2644.
116. Teichmann SA, Babu MM: **Gene regulatory network growth by duplication.** *Nat Genet* 2004, **36:**492–496.
117. Massip F, Arndt PF: **Neutral evolution of duplicated DNA: an evolutionary stick-breaking process causes scale-invariant behavior.** *Phys Rev Lett* 2013, **110:**148101.
118. Zhang L, Lu HH, Chung WY, Yang J, Li WH: **Patterns of segmental duplication in the human genome.** *Mol Biol Evol* 2005, **22:**135–141.
119. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J: **Targeted capture and massively parallel sequencing of 12 human exome.** *Nature* 2009, **461:**272–276.

doi:10.1186/1471-2105-15-2
Cite this article as: Li *et al.*: Diminishing return for increased Mappability with longer sequencing reads: implications of the *k*-mer distributions in the human genome. *BMC Bioinformatics* 2014 **15**:2.