nature biotechnology

Rapid genome sequencing with short universal tiling probes

Arno Pihlak[1,2], Göran Baurén[2,3], Ellef Hersoug[1,2], Peter Lönnerberg[1,2], Ats Metsis[1,2] & Sten Linnarsson[1,2]

The increasing availability of high-quality reference genomic sequences has created a demand for ways to survey the sequence differences present in individual genomes. Here we describe a DNA sequencing method based on hybridization of a universal panel of tiling probes. Millions of shotgun fragments are amplified *in situ* and subjected to sequential hybridization with short fluorescent probes. Long fragments of 200 bp facilitate unique placement even in large genomes. The sequencing chemistry is simple, enzyme-free and consumes only dilute solutions of the probes, resulting in reduced sequencing cost and substantially increased speed. A prototype instrument based on commonly available equipment was used to resequence the Bacteriophage λ and *Escherichia coli* genomes to better than 99.93% accuracy with a raw throughput of 320 Mbp/day, albeit with a significant number of small gaps attributed to losses in sample preparation.

DNA sequencing technology has improved at an exponential rate since 1977, when the first practical DNA sequencing method was described[1], and public databases currently hold more than 100 Gbp of sequence. Technology improvements such as four-color dye terminators[2], capillary electrophoresis[3] and robotic sample preparation have enabled sequencing factories with annual throughput of several gigabases. This tremendous increase in sequencing capacity has resulted in a wealth of new genetic information: whole genome sequences of more than 600 prokaryotes and 100 eukaryotes, including vertebrates such as the human, chimpanzee, zebrafish, mouse, rat and dog; metagenomic survey sequencing[4]; over 5 million mapped human single-nucleotide polymorphisms (SNPs)[5], and initial attempts to locate genes for common disease[6,7].

With the availability of so many high-quality reference genomes, several groups are developing resequencing methods that promise drastically reduced costs and increased throughput[8–10] (reviewed in refs. 11,12). In general such methods combine a massively parallel DNA display technology (bead cloning[13], emulsion PCR[14], in-gel PCR[15], solid-phase PCR[16], single molecules[17]) with a compatible sequencing chemistry such as pyrosequencing[18,19], sequencing by ligation[9,13] or cyclic reversible termination[20,21].

Shotgun sequencing by hybridization

Here we present a DNA sequencing method termed 'shotgun sequencing by hybridization' (shotgun SBH). The method is conceptually similar to tiling arrays[22] and to regular SBH[23–26], in that sequence is reconstructed from a complete tiling of the target sequence with short probes. However, resequencing is achieved hierarchically using a small universal set of probes compatible with any genome and proceeds in four steps: (i) *in situ* rolling-circle amplification (RCA) of millions of randomly dispersed circular single-stranded DNA fragments; (ii) sequential controlled hybridization of 582 pentamer probes, generating a so-called hybridization spectrum for each target; (iii) alignment of hybridization spectra to the reference genome; (iv) reconstruction of the target sequence using the combined hybridization patterns of all aligned fragments.

A massively parallel DNA display platform based on *in situ* RCA[27] was developed. Genomic DNA was fragmented enzymatically and converted to single-stranded, circular molecules having a 200-bp insert and a 50-bp universal linker. The integrity of the fragment preparation was verified by subcloning and Sanger sequencing (data not shown). The circular templates were annealed to surface-bound primers on a microscope glass slide via the universal linker and amplified by RCA to form covalently attached, tandem-repeated products that spontaneously formed sub-micrometer structures.

The approach has several desirable characteristics for DNA sequencing. First, it is simple to perform (**Fig. 1a**). Second, the amplified templates generate easily detectable signal when visualized with fluorescent universal reporter probes (**Fig. 1b**) or with short sequence-specific probes (**Fig. 1c**). Third, the templates remain stable over hundreds of wash cycles (**Supplementary Fig. 1** online), yet are readily accessible to hybridization due to their loose, single-stranded nature. Finally, the array density can be controlled to give 0.5–10 million resolvable features per cm^2.

Next, we developed a universal panel of tiling probes. Because we would hybridize the probes sequentially, we limited not only their number but also their length. Pentamers were essentially the best possible choice: shorter probes would be difficult to hybridize properly, whereas longer probes would require a much larger probe set,

[1]Laboratory for Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Scheeles väg 1, SE-171 77 Stockholm, Sweden. [2]Genizon Svenska AB, Nobels väg 12A, SE-171 77 Stockholm, Sweden. [3]Present address: GE Healthcare, Björkgatan 30, SE-751 84 Uppsala, Sweden. Correspondence should be addressed to S.L. (sten.linnarsson@ki.se).

Received 9 October 2007; accepted 29 April 2008; published online 25 May 2008; doi:10.1038/nbt1405

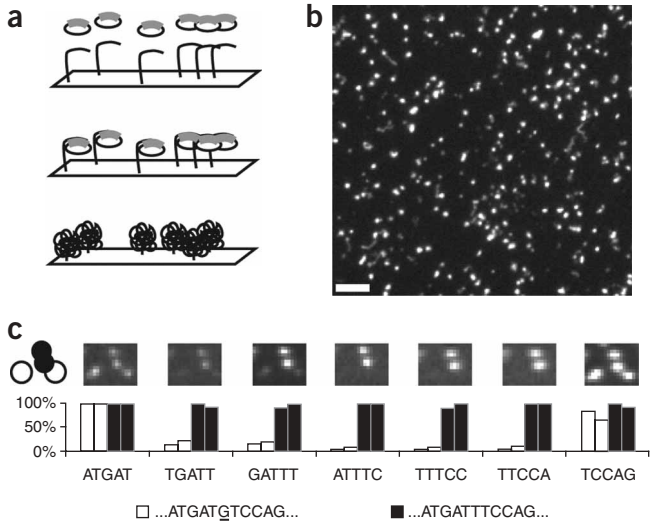

Figure 1 A massively parallel DNA display platform based on *in situ* RCA. (**a**) Single-stranded closed circular DNA templates were prepared from genomic DNA, with each circle carrying a universal linker sequence (thick gray line) and an insert fragment (thin black line). Circles were annealed to covalently bound primer on the surface of a microarray slide and subsequently amplified *in situ* using phi-29 polymerase. (**b**) Epifluorescence microscopy image showing RCA products on the surface of a slide, visualized by hybridization of a Cy3-labeled universal probe targeting the linker sequence. Scale bar, 10 μm. (**c**) Detection of a substitution in synthetic fragments by five overlapping probes. An RCA array was produced with a 1:1 mixture of two synthetic sequences (indicated below the graph) differing in a single substitution (underlined). The location and inferred identity of four features is indicated by white (target containing a substitution) and black (control target) circles in the upper left corner. Bar charts show feature intensities (normalized to the first probe, ATGAT and to the maximum intensity in each image). Note how the single substitution was detected by five overlapping probes.

resulting in very long instrument run times (e.g., 2,048 hexamers would take almost 3 weeks to hybridize on our current instrument). Furthermore, because the melting points of pentamer probes were initially too low to be practical, we introduced locked nucleic acid (LNA)[28] monomers, added a degenerate nucleotide at both ends of the pentamer, and added tetramethylammonium chloride (TMAC) to the hybridization buffers. As a result, each probe was a mixture of 16 heptamer oligonucleotides that functioned as one pentamer (**Fig. 2a** and **Supplementary Table 1** online). Because shotgun fragments were to be obtained from both strands of the genome, half of all 1,024 possible pentamers would suffice to tile the reference genome at every position on either strand. We used this minimal set of 512 probes plus an additional 70 probes, which had been designed for pilot experiments to fully tile a synthetic fragment (**Fig. 1c**), for a total of 582 probes.

Overall, the probes within this set showed reasonable melting points and excellent match/mismatch discrimination, as determined by melting curve analysis with perfect match and single-mismatch DNA targets. The average melting point T_m was 49.0 °C (**Fig. 2c**) and the average single-nucleotide match/mismatch discrimination ΔT_m was 30.4 °C (**Fig. 2d**). Less than twenty probes showed $T_m < 20$ °C or $\Delta T_m < 10$ °C. Although we assayed the full probe set for match/mismatch discrimination only at the central nucleotide position, we observed no difference in performance at all five nondegenerate positions when we used the probes for sequencing (data not shown).

In the remaining sections, the results of three independent genome sequencing runs are reported (summary statistics shown in **Table 1**).

Sequencing Bacteriophage λ

We sequenced the 48,502-bp Bacteriophage λ genome to demonstrate the feasibility of shotgun SBH, and obtain a first estimate of its performance (experiment A, **Table 1**). An RCA array, prepared from a single DNA sample fragmented to 200 bp, was subjected to serial hybridization with the universal probe and the 582 specific probes. Images were acquired after each hybridization cycle. Next, after removing weak features with a manually set threshold in the first image (that is, the image of the universal probe), we identified a total of 14,237 features (that is, tandem-repeated templates), representing 2.8 Mbp of raw sequence and 60-fold nominal genome coverage. For each feature we obtained a hybridization spectrum by collecting normalized intensity values for all the specific probes.

We used a customized algorithm to align the hybridization spectra to a hypothetical composite reference genome comprised of the *Saccharomyces cerevisiae* chromosome V sequence and the λ genome spliced in at position 7,000. The yeast chromosome sequence served to control for and quantify alignment errors. Ninety-five percent of all spectra aligned to the λ genome (**Fig. 3**), with significantly higher average alignment scores compared to the scores obtained for the false

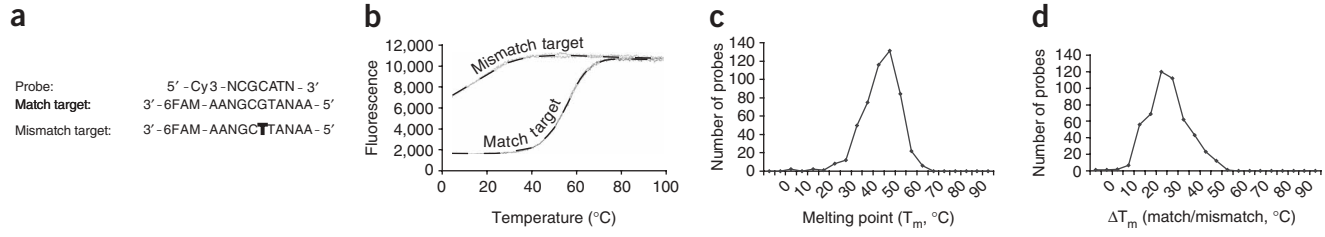

Figure 2 Probe design and characterization. (**a**) All probes were heptamers having two flanking degenerate positions and a 5′ Cy3 label. Each probe was tested against two 6FAM-labeled targets: one perfect match and one carrying a mismatch at the central position. Mismatch nucleotides were selected randomly. (**b**) The melting point T_m was determined by melting curve analysis, where hybridization was indicated by the appearance of fluorescence resonance energy transfer between the 6FAM and Cy3 labels when they were brought in close proximity, detected as a quenching of the 6FAM signal at low temperatures. The figure shows typical match and mismatch melting curves. Dashed lines are overlaid on the raw data for clarity. (**c**) Histogram showing the distribution of T_m values for all 582 probes. The average T_m was 49.0 °C. (**d**) Histogram showing the distribution of match/mismatch ΔT_m. The probes showed good mismatch rejection, with average $\Delta T_m = 30.4$ °C (probably an underestimate of the true average because mismatch melting points below zero could not be measured).

Table 1 Summary of sequencing results

	Experiment		
	A	B	C
Species	Bacteriophage λ	*E. coli* K12	*E. coli* K12
Genome size	48,502 bp	4.6 Mbp	4.6 Mbp
Number of reads	14,237	3.3 million	618,654
Fold coverage	60×	143×	27×
Repeat fraction	0	3%	3%
Fraction called	96%	83%	80%
Overall accuracy	99.96%	99.93%	99.94%
False-positive rate	20	0.07%	0.06%
False-negative rate	1 of 48	2.7%	3.5%
Median Q_{phred}	N/A	47	46

hits that fell onto the yeast chromosome sequence. Assuming the λ genome and the 12 times larger yeast chromosome have the same rates of false hits, 99.9% of the λ alignments were placed correctly.

In addition, if the misalignment rate and the mutation rate are both small, we can infer with high confidence the match/mismatch status for each probe and aligned fragment from the reference sequence. After obtaining match and mismatch intensity histograms for each probe (**Fig. 4a**), we converted these histograms into log-odds curves in which the logarithm of the odds in favor of a probe being a 'match' is plotted as a function of the observed intensity (**Fig. 4b**). This conversion enabled us to take any observed intensity and translate it into a probability. In particular, every position in the reference genome could be examined and for each possible call at that position the probability of the observed intensities of probes could be calculated.

Typically, a single base substitution would cause ten probes to change relative to the reference sequence. For example, five probes (AGCTG, GCTGG, CTGGA, TGGAA and GGAAT) would detect the central position in AGCT**G**GAAT, and these would be replaced by five others (AGCTC, GCTCG, CTCGA, TCGAA and CGAAT) if that central position were replaced with a **C** (compare also with the synthetic targets hybridized in **Fig. 1c**). For each probe, the odds in favor of its hybridization at each position in the reference genome could be calculated. Next, we derived a consensus sequence by calculating a Bayesian posterior probability for each possible call at each position along the genome that is based on the log-odds of each overlapping probe (**Fig. 4c**). A quality score q was calculated as the log-odds difference between the best and second-best calls, and a threshold $q_{min} = 0.7$ was applied; 2,109 positions (4% of the genome) with $q < q_{min}$ were reported as 'N'.

We introduced mock substitutions[9] in the reference genome to estimate the accuracy of the called sequence (see Methods). The ability of the basecaller to revert these substitutions was assessed. The overall basecalling accuracy (that is, total number of accurate calls divided by total number of called bases) was 99.96% and 47 of 48 mock substitutions were called correctly. The remaining error was a single false-negative call (failing to call a true substitution); no miscalls (that is, calling an incorrect substitution) were observed; 20 false-positive errors (that is, calling a substitution at a wild-type position) were made. Some of the errors were placed in GC-runs. For example, an A was erroneously called at position five of GCGGCGGCGGGG. This may indicate in some cases that strong local secondary structures in the target molecule prevent probe hybridization.

Sequencing *E. coli*

We then proceeded to resequence the 4.6-Mbp genome of *E. coli* (experiment B, **Table 1**). A total of 3.3 million image features, corresponding to 660 Mbp of raw sequence and 143-fold nominal coverage, were collected.

First, the distribution of coverage depth across the genome was examined. The most salient feature of this distribution was a pronounced wave across the genome, with a maximum near the origin of replication and a minimum near the terminus (**Fig. 5**). This wave probably reflects true intragenomic differences in DNA content in rapidly growing *E. coli* cultures. Inside any given *E. coli* cell, multiple nested replication forks co-exist, resulting in double- or quadruple-copy DNA content near the origin, but single-copy content near the terminus[29]. This distribution also demonstrates our ability to detect and quantify copy-number changes in this range (that is, 2–3 copies), even in the absence of a haploid control genome. Intriguingly, there was a clear difference in coverage between the leading and lagging strands for both replichores (the portion of the genome between origin and terminus). In replichore 1—going clockwise from the origin—the leading strand showed higher coverage than the lagging strand, which presumably is synthesized more slowly and may have numerous nicks and single-stranded regions. The same pattern was evident for replichore 2—going counter-clockwise from the origin—in which the leading strand corresponds to the 'reverse' strand.

Next, we performed basecalling. The resulting assembly covering 83% of the genome was 99.93% accurate, and 3,205 out of 3,295 mock substitutions were called correctly. The remainder of the substitutions were all false negatives; no miscalls were observed. The false-negative rate (that is, defined as total number of negative calls at mutated positions divided by total number of mock substitutions) was 2.7% and the false-positive rate (that is, total number of substitutions

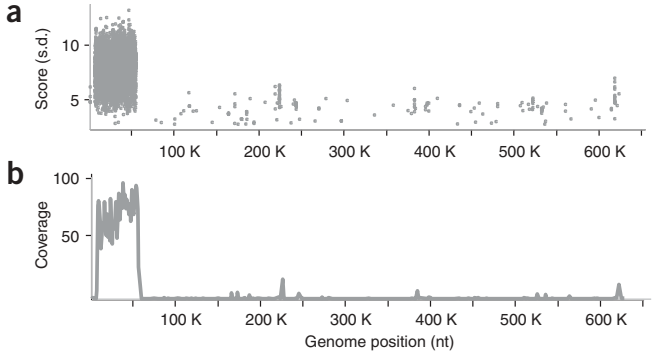

Figure 3 Fragments aligned to the reference genome in the Bacteriophage λ assembly. A composite reference genome was constructed by splicing the 48,502 nt λ genome (accession NC_001416.1) at position 7,000 in the sequence of yeast chromosome 5 (accession NC_001137.2). The total length of the composite genome was 625,371 nucleotides. (**a**) A plot of the score (in s.d. from the average score along the composite genome) for each alignment, showing that very few (5%) fragments align outside the λ genome, and with lower average scores. For clarity, only 10% of the alignments are shown. Only alignments with s.d. > 6 were used in subsequent analyses. (**b**) Histogram of the genome coverage (number of hits per 200 bp, equivalent to the effective fold-coverage because all fragments were 200 bp long) clearly showing the specificity of the genome alignment for the λ sequence. A few hotspots could be seen in the yeast genome (e.g., at position 230 K), which may represent sequences of low complexity that tend to attract poor quality fragments.

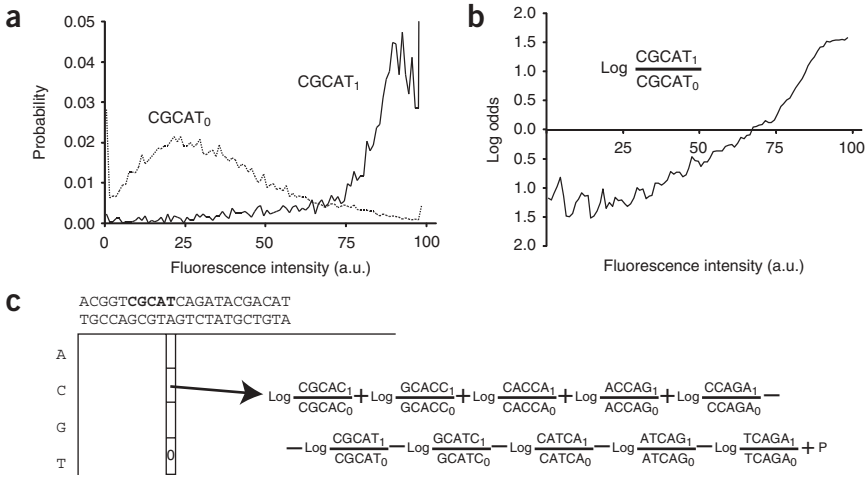

Figure 4 Probabilistic basecalling algorithm. (**a**) The intensity distribution for each probe was split in two components, match and mismatch, here shown for probe CGCAT and denoted $CGCAT_1$ and $CGCAT_0$, respectively. Given a genome alignment, and under the reasonable assumption that most sequences would be conserved, the aligned fragments could be separated into those that contained CGCAT and those that did not. The two histograms were converted to probability distributions by normalizing their areas to 1.0. As a result, the likelihood that a given intensity measurement represents a 'match' could be determined by simply looking up the intensity in the $CGCAT_1$ distribution. (**b**) For convenience, and to avoid round-off errors, all computations were performed on a log-odd scale, defined as the base-ten logarithm of the ratio of the probabilities given by the histograms in **a**. The log-odds for any given intensity measurement gives the logarithm of the odds in favor of a probe being a 'match' versus it being a 'mismatch', and it can be found in the log-odds curve. (**c**) Basecalling was performed by examining one position at a time in the reference genome and collecting log-odds terms for each probe overlapping that position as indicated. For each possible call, there are five positive terms and five negative terms plus the prior log-odds in favor of a substitution, $P = -1.2$. By convention, odds were computed against the reference, so that the log-odds for not calling a substitution was always zero.

called at wild-type positions divided by total number of non-mutated positions) was 0.07%. These results demonstrate the utility of shotgun SBH both for SNP discovery and for calling known SNPs with high accuracy.

Because the basecaller was designed to only call substitutions, we reasoned that insertions or deletions (indels) in the sequence might cause additional errors (that is, false substitutions). To examine this effect, we introduced single-nucleotide mock indels at 1/5,000 bp and repeated basecalling. The 5 bp context—corresponding to the length of one probe—around each indel was examined for false-positive calls and/or sequence flagged as unreliable ('N' calls). More than 90% of indels (92% of insertions, 94% of deletions) were correctly flagged as unreliable, or were called as wild-type sequence. The remainder of these indels were incorrectly called as substitutions, but did not significantly reduce the overall sequence accuracy. In the future, we expect to extend the basecaller to correctly consider short indels.

We intentionally generated excessive fold coverage to determine the limits of oversampling. Examination of the accuracy as a function of raw sequence depth (**Fig. 6a**) revealed saturation after about 30-fold nominal coverage. In an independent sample (experiment C, **Table 1**) with 27-fold nominal coverage, very similar levels of accuracy, albeit with slightly more gaps in the sequence, confirmed that the excess depth of coverage was unnecessary. Remaining errors at high coverage were presumably due to systematic sources, such as strong secondary structures or unfavorable combinations of poorly performing probes. This was confirmed by the fact that false-positive errors were highly concordant: errors in experiment B were more than 300 times more likely to coincide with errors in experiment C than expected by chance (observed 333, expected 1; $\chi^2 = 79{,}600$, $P < 10^{-10}$). 13% of all errors in experiment B coincided with an error in experiment C and an additional 55% coincided with a low quality score ($q < 0.7$). Conversely, only 3% of errors in experiment B coincided with a quality score higher than median in experiment C, whereas by definition 50% of all positions in C were of such high quality. Finally, there was a high degree of correlation in quality scores between the two experiments (**Fig. 6**). Thus in experiments involving multiple samples, poor-quality regions should be relatively confined, leaving large high-quality regions for comparative analysis.

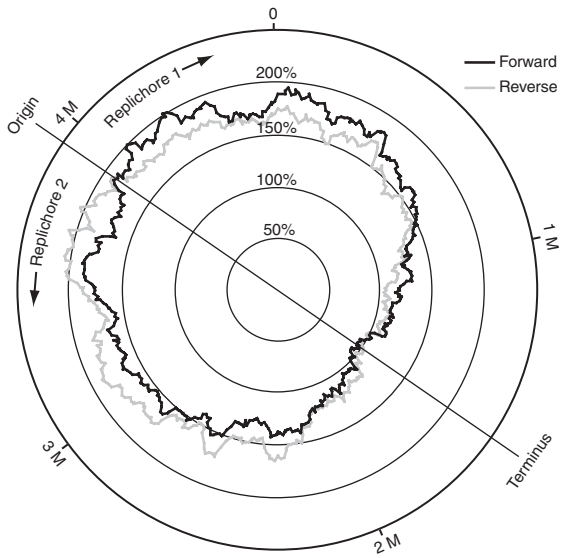

Figure 5 The depth of coverage along the *E. coli* chromosome was strongly skewed toward the origin of replication. The plot shows the ten-bin running average depth of coverage in 10-kb bins, normalized to the depth at the terminus as indicated by concentric circles. Coverage was lowest near the terminus, which was presumably always haploid, and increased toward the origin in both replichores, reaching an almost diploid level. Note that the data were obtained from a growing bacterial culture and thus represents the average ploidy along the chromosomes of millions of individual dividing cells. In both replichores, the leading strand showed slightly higher coverage all the way from origin to terminus, and the two replichores were separated almost perfectly by the transition points at origin and terminus. We speculate that the difference may reflect the fact that at any given point, the lagging strand contains RNA primers and nicks generated during the synthesis of Okazaki fragments. The outer ring shows nucleotide positions (M, million nucleotides), arrows indicate the direction of replication for the two replichores, the origin is indicated at position 3,923,882 (midpoint of *oriC*) and the terminus at 1,588,787 (midpoint of the *Dif* site).

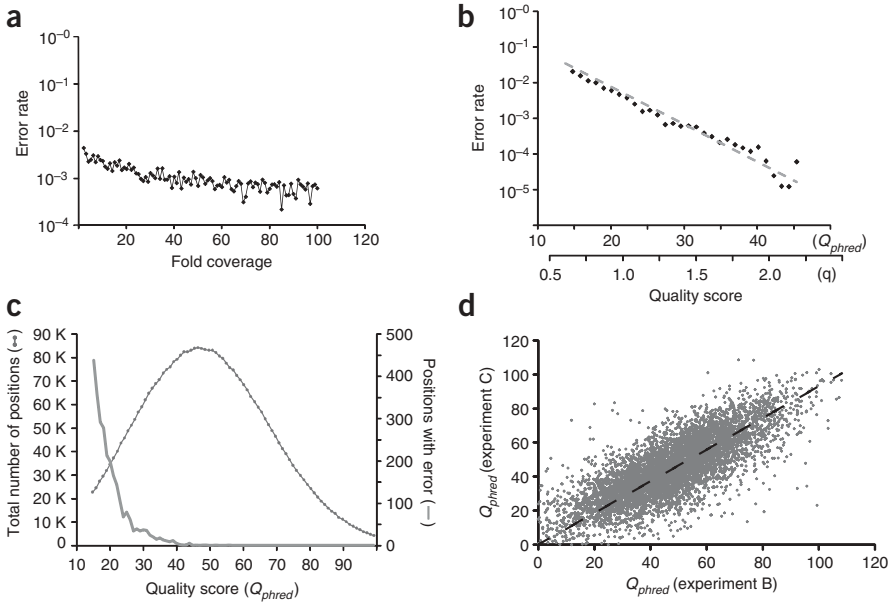

Figure 6 Assembly statistics for the *E. coli* genome. (**a**) The error rate as a function of fold coverage at individual positions. For example, the first data point at left shows the combined error rate for all positions covered by a single fragment. It can be seen that accuracy increases rapidly up to about 30-fold coverage and then saturates at an error rate of $\sim 10^{-3}$, suggesting the presence of systematic errors at that frequency. (**b**) Error rate as a function of quality score. The secondary horizontal axis shows the interim quality score *q*, taken as the difference in log-odds between the best and second-best call at each position. A linear fit ($R^2 = 0.95$; indicated by dashed gray line) was used to compute the constant of proportionality, which was then used to convert the interim score into a *phred*-equivalent standard quality measure Q_{phred} (shown on the primary horizontal axis). The scatter plot only extends to about Q45, whereas half the assembled bases were in Q47 or better. This is because the error rate could not be reliably estimated at qualities better than Q45, where too few errors occurred in each histogram bin. (**c**) Assembly-wide distribution of Q_{phred} scores (shown on left axis; filled circles) and the distribution of errors (on right axis; gray line). The median score was Q47, corresponding to an expected error rate of 1/50,000 bases called. (**d**) Reproducibility of quality scores between experiments. The figure shows all Q_{phred} scores obtained from an arbitrarily chosen 6-kb region in experiments B and C, revealing a high degree of concordance of quality scores between experiments (Pearson's correlation coefficient 0.77). Other regions showed similar results.

Compared with the λ sequence, a smaller proportion of the genome could be confidently called (83% versus 96%). Most of the missing *E. coli* sequence was due to AT-rich regions with zero coverage (called as N) or less than fivefold depth of coverage, giving poor sequence quality (that is, mostly called as N). The 2,287 gaps were generally short (median length 130 bp), consistent between samples (90.3% concordance between experiment B and C, **Table 1**) and could not be rescued by changing the fragmentation method (nebulization instead of DNaseI; data not shown) or by generating excess depth of coverage (compare experiments B and C). We performed quantitative PCR to determine if gaps were generated during sample preparation, or if they were a result of defects in the sequencing chemistry. We targeted 48 gaps and 48 nongaps in aliquots taken at the end of the sample preparation stage (see Methods). There was a very significant apparent loss of material in all gap regions, whereas only one nongap showed notable loss (**Supplementary Fig. 2** online). Median loss was 600-fold for gaps compared with 2.3-fold for nongaps; the difference was highly significant ($P < 0.001$, $D = 0.9783$ by the Kolmogorov-Smirnov's test, which was used because both the Kolmogorov-Smirnov and the Jarque-Bera tests rejected the hypothesis that the data had a normal distribution). We conclude that a defect in sample preparation caused the specific loss of AT-rich regions, and that the sequencing step was actually quite accurate in calling these regions as gaps.

To provide an easily interpreted quality measure, we constructed a *phred*-equivalent quality score termed Q_{phred} and used it to summarize the assembly quality[30] (**Fig. 6b,c**). The median *phred* score was $Q_{phred} = 47$, and hence half of all bases called had an expected error rate less than about 1/50,000—the average error rate for the whole assembly was significantly higher, because a small number of low-quality positions contribute a disproportionately large number of errors. In fact, across all the 2,095,116 bases having Q47 or better, we observed only four errors (all of them false-positive calls). Thus, Q_{phred} reliably identifies regions of high-quality sequence.

DISCUSSION

We developed a rapid and relatively inexpensive genome resequencing method. A single-molecule display platform based on *in situ* RCA combined with a hierarchical genome tiling approach revealed sequence differences in the context of a reference genome. The method should be scalable to larger genomes than the viral and bacterial ones sequenced herein. Whereas the current algorithms only deal with substitutions in a haploid sequence, new versions of the basecaller may be able to call short indels, large deletions and heterozygous sequence. Insertions longer than a few bases, however, would require a *de novo* assembly algorithm (perhaps along the lines of ref. 31).

Sequencing by hybridization must tackle two important challenges[32]. First, conditions must be found for short oligonucleotide probes to efficiently discriminate between match and mismatch hybridization. Most previous attempts immobilized the probes and provided target DNA in solution, thus requiring a single hybridization temperature and buffer for all probes. In addition, investigators commonly used stringent washes followed by drying. However, because short probes have very fast kinetics, it is difficult to control the washing step. In contrast, we hybridized probes sequentially to immobilized targets without drying, thereby enabling the use of optimal equilibrium reaction conditions for each probe. Under these sequential hybridization conditions, shorter probes actually outperform longer ones, because a single mismatch nucleotide has a proportionately larger impact on overall probe stability and melting point (see **Figs. 1c** and **2**).

Second, it has been shown for SBH that only sequences of length less than about 2^k can be reconstructed with high probability (due to the occurrence of repeated probe sequences), that is, only ~32 bp in the case of 5-mers[33]. However, theory suggests that this limitation can

be overcome by using a reference sequence as template for the reconstruction[34] or by using multiple partially overlapping clones[31]. We improved on these ideas by implementing a spectrum alignment algorithm, using an entire genome as reference for a set of overlapping shotgun fragments. This alleviates the problem both because the reference genome can be used to guide basecalling locally, and because adjacent partially overlapping fragments can often be used to separate repeated probes.

The long fragment lengths obtained by shotgun SBH should facilitate assembly of vertebrate genomes. A fixed 200-bp fragment length was used to simplify downstream assembly algorithms. Several considerations influenced the choice of fragment length. First, fragment length equals read length, because every position of each fragment was probed. Longer fragments should therefore increase throughput. However, because longer fragments are more likely to contain duplicate, uninformative probes, shorter fragments should be more accurate (**Supplementary Table 2** online). Additionally, depending on the genome size, there would be a minimal length required to uniquely place a fragment in the reference genome. For example, human genome resequencing requires at least 60-bp long fragments[35], and even longer ones when there are substitutions and/or inaccuracies in the raw sequence reads. Finally, longer fragments yield lower signal after RCA (amplicons contain fewer copies, and thus bind fewer probe molecules during hybridization), affecting signal quality. As a consequence of these trade-offs, 200 bp was chosen as a reasonable fixed fragment length. The most important factor preventing the use of longer fragments was the signal intensity; thus with improved detection methods it may be possible to at least double the fragment length.

Nevertheless, scaling-up assembly to gigabase-sized, highly repetitive genomes poses a number of additional challenges. Microsatellites and similar simple repeats probably cannot be resolved, whereas highly conserved interspersed repeats and segmental duplications prevent the unique placement of sequenced fragments. All in all, about half of the human genome is nonrepetitive[36], out of which we estimate that more than 99% can be sequenced (**Supplementary Table 2**).

The ultimate goal of human genome resequencing requires that throughput and cost be carefully optimized. In developing shotgun SBH, we sought to maximize throughput and minimize cost by using simple reagents and instrumentation and a very high degree of multiplexing. Current maximal throughput of the prototype instrument was achieved when using the full imaging surface of 405 images. Cycle times were 11.5 min (divided approximately equally between imaging and the fluidic cycle), typically yielding 1.6 Gbp of raw sequence in less than 5 d, for an overall raw sequencing speed of 3,800 bp/s. The sequencing chemistry consumed only simple oligonucleotide probes and buffer, and as a consequence, costs were dominated by equipment and plasticware. The crude reagent cost was $0.32/megabase (**Supplementary Table 3** online), which would translate to $960 or $28,800 per human genome at single-fold or thirty-fold coverage, respectively. Including the amortized cost of equipment, the overall cost was $0.5/megabase. By comparison, Shendure[9] reported a speed of 140 bp/s and a cost of $110/megabase in an assembly covering 70% of the *E. coli* genome, whereas Margulies[8] achieved a throughput of 1,700 bp/s at a reported[37] cost of $200/megabase of raw sequence when sequencing *Mycoplasma genitalium*.

Raw sequence throughput can be increased. For example, the fluidics cycle time can probably be decreased substantially. Hybridization kinetics were observed to be fast (on the order of a few seconds), so the fluidics cycle speed was dominated by the speed of liquid handling and temperature change. As suggested by others[9], the fluidics cycle time could be effectively eliminated by using two flow cells and alternating these between imaging and fluidics. Furthermore, in the present work, relatively sparse arrays were used to avoid excessive numbers of unresolved overlapping image features. However, the maximum number of nonoverlapping features would be obtained at much higher densities. Thus if overlapping features could be efficiently detected and ignored, the raw sequence yield per slide could be at least tripled. The combined effect of these improvements, and the possible doubling of the read-length suggested above, could increase the throughput as much as tenfold and reduce cost further.

The initial probe set reported here can also be improved. For example, it was synthesized with approximately equimolar ratios at degenerate positions ('N'), so that the amounts of each of the 16 individual oligonucleotides comprising each probe were approximately equal. However, oligonucleotides within each probe that have N = G or C bind tighter than oligonucleotides that have N = A or T. This difference in binding resulted in a narrowing of the temperature range for optimal hybridization, because the mismatch-T_m of GC-rich oligonucleotides was close to the match-T_m of AT-rich probes (**Supplementary Fig. 3a** online). The range was widened in test cases by balancing the relative concentrations of GC versus AT at degenerate positions (that is, by increasing the amounts of adenosine and thymine relative to guanine and cytosine during oligonucleotide synthesis at degenerate positions; **Supplementary Fig. 3b**). Further, 97 of the 582 probes were capable of forming self-dimers, resulting in weak signals for these probes. Self-dimerization could be eliminated by shortening the probes to hexamers (**Supplementary Fig. 3c**), selectively disrupting the self-dimer (which loses two interactions) relative to target hybridization (which loses only one). Together, these improvements may be expected to substantially increase overall sequencing accuracy when implemented in a fully revised probe set.

Careful attention was paid to maximize sequence accuracy, as well as to provide a reliable (and *phred*-compatible) quality measure for every nucleotide position. This should simplify the interpretation of the called sequence and its use in downstream analyses.

To fully realize the potential of shotgun SBH, sample preparation procedures must be improved. A large number of gaps were observed in AT-rich regions, which were shown by quantitative PCR to be caused by ~600-fold losses during sample preparation. In essence, these regions behaved as "uncloneable" DNA in Sanger sequencing and were therefore not available for sequencing. Because the losses occurred after fragmentation, but before circularization and RCA, and because we did not observe significant losses in the gel-purification step (data not shown), we suspect that the PCR step is the most likely culprit. The problem might be resolved by eliminating PCR amplification, or replacing it with linear RCA. Alternatively, emulsion PCR could be used to avoid many of the artifacts of bulk PCR amplification[38], as each template molecule is amplified in a separate microreactor. These suggested modifications could potentially close the large majority of gaps.

METHODS

Probes. Oligonucleotides were from Sigma Proligo. Probes were of the general formula 5′-Cy3-NXXXXXN-3′ (X are specified bases, N are degenerate positions), with LNA nucleotides at positions 1, 2, 4, 6 and 7; DNA nucleotides at positions 3 and 5. For example, one probe was 5′-Cy3-NCGCATN-3′. Each probe was quality controlled by mass spectrometry and capillary electrophoresis (not shown), and functionally validated as follows. For each probe, perfect match (e.g., 5′-AANATGCGNAA-6FAM-3′) and mismatch (e.g., 5′-AANATGGGNAA-6FAM-3′) targets were synthesized using DNA monomers.

The melting temperature T_m and the match/mismatch discrimination ΔT_m were calculated from melting curves obtained for the probe against the two targets separately. Hybridization (in 2.5M TMAC, 50 mM Tris-HCl pH 8.00, 0.05% Tween-20) was measured by fluorescence resonance energy transfer between the 6FAM and Cy3 dyes in a real-time PCR instrument (7900HT, Applied Biosystems). An initial set of 70 probes were designed to completely tile the synthetic fragment 5′-GGCTGGCTGGTGAACTTCCGATAGTGC GGGTGTTGAATGATTTCCAGTTGCTACCGATTT-3′. Subsequently, a complete set of 512 probes was added, for a total of 582 probes. Probe sequences and melting points are reported in **Supplementary Table 1**.

Sample preparation. 4 μg genomic DNA (Bacteriophage λ from New England Biolabs (NEB); *E. coli* K12 strain MG1655 from LGC Promochem) was fragmented enzymatically in 50 mM Tris-HCl pH 7.5, 50 μg/ml BSA, 10 mM $MnCl_2$ and 0.04 U DNaseI (NEB) in a total volume of 120 μl. Two reactions were incubated at 25 °C for 10 and 15 min, then stopped with 4.2 μl 0.5 M EDTA and purified on silica spin columns (PCR Cleanup, Qiagen). Fragmented samples were blunt-ended by Klenow enzyme treatment (55 μl eluted DNA, 30 μM dNTP, 0.03 U/μl NEB Klenow enzyme in 70 μl NEB2 buffer), then purified on silica spin columns and recovered in 55 μl elution buffer. 10 μl of each purified reaction product was separated on a 2% E-gel (Invitrogen) for 25 min to visualize the size distribution. Based on the gel result, either one sample was chosen or the two samples were pooled and fragments in the range 150–300 bp were excised and purified. Without the removal of fragments shorter than 150 bp before adaptor ligation, we observed a high frequency of chimeric inserts (data not shown). 5 pmol fragmented DNA, 187 pmol each of left (5′-GCAGAATCCGAGGCCGCCT-3′ and 5′-GACAAGGCGGCCTCGGATTCTGC-3′) and right (5′-AGTGGCGTGTCTTG GATGC-3′ and 5′-CGATAACGCATCCAAGACACGCCACT-3′) double-stranded adaptors, 5 μl Quick Ligase and 50 μl Quick Ligation buffer (NEB) were incubated at 25 °C for 15 min in a total volume of 105 μl, then purified on silica spin columns. To produce blunt-ended fragments, 20 μl 5× Phusion buffer HF (Finnzyme), 2 μl 10 mM dNTP were added to the sample, which was heated to 72 °C before 2 units Phusion polymerase (Finnzyme) was added and incubation continued for 5 min. The sample was cleaned up on a silica spin column and eluted in 30 μl H_2O. Samples were separated on an 8% nondenaturing PAGE gel run at 250V over night. The SYBR Gold (Molecular Probes)-stained gel was scanned on a Typhoon 9200 (Amersham Biosciences). A gel piece including the 250 ± 10 bp range was excised using a scalpel, collected in 50 μl of 10 mM Tris pH 8 and incubated for 3 h at 37 °C. To maximize yield and minimize PCR errors, eight amplification reactions were set up from each eluted sample. 0.2 mM dNTP, 400 nM biotinylated primer (5′-biotin-GACAAGGCGGCCTCGGATTCTG-3′), 400 nM phosphorylated primer (5′-phosphate-CGATAACGCATCCAAGACACGC-3′), 1 μl eluted template and 1 μl Phusion *Taq* polymerase (Finnzyme) in a total volume of 100 μl in Phusion buffer HF was thermocycled (98 °C 10 s, 72 °C 20 s) for 25 cycles. The reactions were sequentially purified over a single silica spin column. To remove primer dimer artifacts, the concentrated eluate was purified from a 2% agarose gel (Qiagen Gel Extraction Kit). From this point on, all procedures were carried out in polyallomer tubes (Beckman) to minimize loss of material due to adsorption. The phosphorylated DNA strand was isolated as follows. 100 μl paramagnetic streptavidin-coated beads (M280, Dynal) were washed twice in 200 μl B&W buffer (Dynal), then left in 100 μl B&W. 100 μl purified PCR product (having one biotinylated and one phosphorylated strand) was added and left for 20 min at 20 °C. After two washes in 200 μl B&W and two in 200 μl 10 mM Tris pH 8.0, the phosphorylated strand was eluted in 100 μl 0.1 M NaOH for 3 min. The supernatant was transferred to a fresh tube, 25 μl 1.0 M Tris pH 7.5 was added and the sample was cleaned up on a silica spin column. The single-stranded linear DNA was annealed at 0.03 μM to 0.06 μM biotinylated linker (5′-biotin–TGCGTTATCGGAC AAGGCGG-3′) in 30 μl ligation buffer (Fermentas) by incubation for 2 min at 65 °C followed by cooling to 25 °C over 15 min. 70 μl ice-cold T4 DNA ligase in ligation buffer (both from Fermentas) was added and the mix was incubated at 25 °C for 1 h. Circular product was purified on 25 μl Dynabeads (M280, Dynal). The beads were first washed twice in 100 μl B&W buffer, then 100 μl B&W buffer and 100 μl ligation product was added, let stand for 20 min, then washed twice in 100 μl B&W. Circular DNA was eluted in three fractions (30 μl H_2O, 30 μl 40 mM NaOH, 30 μl H_2O), the fractions were pooled and 5 μl of 1M Tris-HCl pH 8.0 was added. After purification, contaminating linear fragments were virtually undetectable by PAGE, and were of little concern since they would not amplify by RCA. The final circular DNA library was stored at –20 °C.

Array synthesis. Activated microarray slides (Genorama SAL-1 Ultra, Asper Biotechnology) were coated with aminated primer (5′-NH-AAA AAAAAAAGCGTGTCTTGGATGCGTTATCG–3′) at 1.0 μM in 100 mM carbonate buffer pH > 9.0, 15% DMSO, 0.001% Triton X-100 by incubation for 50 min at 30 °C, then blocked in 1% NH_4OH twice for 2 min. Before hybridization, the slide was incubated in SSB (2× SSC, 0.1% SDS) 2 min at 65 °C, 3 min at 50 °C, 5 min at 30 °C, then rinsed in TWB (2× SSC, 0.1% Tween-20) followed by MGB (1.5 mM Tris pH 8.0, 10 mM $MgCl_2$). The circular template DNA library was then annealed, typically at 1:200 dilution, in SSB 2 min at 65 °C, 3 min at 50 °C, 10 min at 30 °C, followed by wash in SSB 5 min at 30 °C, then rinsed in TWB followed by two rinses in MGB. Amplification buffer (1 mM dNTP, 0.1× BSA, 0.1 u/μl Phi29 polymerase in Phi29 DNA Polymerase Reaction Buffer, both from NEB) was added to the slide, which was incubated at 30 °C for 3 h. The slide was then rinsed in MGB and washed in SSB 2 min at 65 °C, 3 min at 50 °C, 2 min at 30 °C, then rinsed in TWB followed by two rinses in MGB. The slide was finally dried at 30 °C for 2 min and ready for mounting on the instrument.

Instrument. An integrated and automated instrument was built as follows. A Nikon TE2000PFS motorized inverted microscope was fitted with a Scan IM motorized stage (Märzhäuser), a Cy3 filter cube (Semrock), a 120W metal halide illumination system (X-Cite 120 PC, EXFO), an electro-mechanical shutter (Uniblitz VS35 with VCM-D1 controller, Vincent Associates) and a monochrome 4 megapixel cooled CCD camera (Spot Xplorer, Diagnostic Instruments). All images reported here were acquired with a 20× magnification Nikon PlanFluor ELWD objective through 1 mm glass slides. A custom flat rectangular flow cell capable of holding two slides was machined in aluminum, black anodized and coated with 4 μm Parylene (Plasma Parylene Coating Service). The flow cell was permanently fixed on a Peltier module (Melcor) in place of the hot plate. A plastic adaptor ring was used to mount the flow cell assembly onto the microscope stage. When a standard 25 × 75 mm glass slide was held onto the flow cell by vacuum suction (Vacuum Pump System, C&L Instruments) between two o-rings, an interior 10 × 50 × 0.15 mm chamber was formed with inlet and outlet at either end, inducing laminar flow across the glass surface. The flow cell was connected by tubing to a Tecan MSP9250 autosampler, from which reagents could be aspirated through the flow cell. All parts of the instrument were controlled by a custom software application.

Overview of a sequencing run. Each run was performed with a full set of 582 probes in 96-well plates. Between probe plates, two universal probes (the all-DNA 5′-Cy3-GCCGCCTTGTC-3′ and the mixed LNA/DNA 5′-Cy3-NCGAG GN-3′) targeting the adaptor sequence and two buffer-only negative controls were hybridized. Universal probes are henceforth denoted 'UNIP'. The whole run was fully automated except that buffers had to be replenished daily.

Hybridization. Each hybridization cycle was performed as follows. 400 nM probe (25 nM for UNIP) in TMAC buffer (3M TMAC, 50 mM Tris-HCl pH 8.0, 0.4% β-mercapto-ethanol and 0.05% Tween-20) was aspirated from a 96-well microtiter plate into the flow cell held at 45 °C. The temperature was briefly raised to 65 °C, then adjusted to the desired hybridization temperature (T_m, –33 °C), and excess probe was removed by two washes in TMAC buffer. After image acquisition, the temperature was raised to 45 °C in preparation for the next cycle.

Image acquisition. Before the first imaging cycle, an autofocus routine was performed as follows. A stack of images bracketing the expected focal plane was acquired and the best focus was determined by maximizing the focus criterion $\sum_{i,j}(p_{i+2,j}-p_{i,j})(p_{i+3,j}-p_{i+1,j})$ where $p_{i,j}$ denotes the pixel value at i,j. This ensured that the CCD sensor was perfectly in focus at the start of the experiment. It was then kept in focus indefinitely by the Nikon opto-mechanical Perfect Focus System. Images were acquired in a grid with 1.25 mm spacing at 1 s exposure.

Feature extraction. All local maxima (in a 7 × 7 neighborhood with clipped corners) were detected in the first UNIP image and a threshold was applied to remove weak features. The threshold was set once per experiment and was verified by visual examination. Only the features identified in the first UNIP image were then extracted for analysis from subsequent images. Subsequent images were registered onto the first UNIP image by scanning through a range of translations systematically, maximizing the sum of products of pixel values for all detected features.

Feature quantification and normalization. To allow for a small local image offset, the local maximum pixel value in a 3 × 3 neighborhood of each feature in each image was taken as its raw value for the corresponding probe. A background value was calculated for each feature and image by taking the second lowest pixel value in the corners of a 15 × 15 square. To monitor the reduction in signal with time (number of hybridizations), each set of 96 probes was flanked by UNIP and blank hybridizations. The intensity value of each feature in each image was normalized by first subtracting the background value, then dividing by the interpolated signal of the two flanking UNIP hybridizations.

Spectrum alignment. Each extracted feature corresponded to a DNA fragment from the original sample library. The vector of normalized intensity values of each feature across the full set of probes, that is, the 'hybridization spectrum' of the fragment, was used to find its position in the reference genome as follows. A window of width equal to the expected fragment length was scanned across the reference sequence. For each window position, the presence or absence of each probe sequence in the window was recorded. An alignment score was calculated as follows

$$\frac{\sum_{+}\hat{I}}{\sqrt{n_{unique}}} - \frac{\sum_{-}\hat{I}}{\sqrt{n_{unique}}}$$

(where $\hat{I}$ denotes the normalized intensity of a probe minus the median intensity of the probe across all spectra, n_{unique} is the total number of distinct probes in the window and where the first sum is over the probes present in the window, whereas the second sum is over the probes absent from the window). Thus a probe with a relatively large normalized value would contribute a high score to positions where it was present, and vice versa. The position with the maximum score was reported. Scores were expressed in terms of s.d. from the mean score across the genome. The score is admittedly heuristic, but was shown to yield accurate alignments on simulated data with similar sources of errors to those we think are present in real data (not shown).

Calculating hybridization probabilities. Basecalling was designed to operate on a probabilistic representation of hybridization. For each aligned fragment and probe, the probability of observed normalized intensity values conditional on the presence or absence of the probe was needed. For each probe, these probability densities are functions of the observed intensity, given by all fragments where the probe did or did not in fact occur. To obtain these distributions, it was assumed that the experimental genome was almost identical to the reference genome, that is, that the divergence was low. The number of occurrences of a probe in each fragment were taken from the corresponding window in the reference sequence as given by spectral alignment. Fragments predicted to have more than one occurrence of the probe, or where the number of occurrences could not be determined with confidence were not taken into account. After normalization to unit area these histograms were used to directly calculate the required probabilities. To illustrate this, histograms for probe CGCAT are shown in **Figure 4a**. As was typical of other probes, there was a significant overlap between the distributions, indicating that base calls could not be confidently based on single probes. For the *E. coli* experiment, the differential melting points for each probe—depending on the presence of GC ('strong') or AT ('weak') base pairs in the target at positions corresponding to the two degenerate nucleotides—were taken into account by generating separate histograms for the four cases (weak-weak, weak-strong, strong-weak and strong-strong) of flanking nucleotides. In subsequent computations it was more convenient to work with log-odds scores, a measure of the odds in favor of the presence of the probe over its absence. The log odds as a function of normalized intensity was taken as the base-10 logarithm of the ratio between positive and negative probabilities; this is again illustrated for probe CGCAT in **Figure 4b**. These curves were capped at their extremes to minimize errors due to the low number of cases in the tails of the histograms. Note that the zero-crossing of the log-odds curve corresponds to the crossing of the positive and negative histograms in **Figure 4a**.

Basecalling. To compute the final sequence, a Bayesian model was constructed. Given the reference sequence and a number of aligned fragments, the goal of sequence reconstruction was to find the most likely modification of the reference sequence as indicated by the probe hybridization probabilities. The current algorithm was designed to consider single-nucleotide changes only, but the extension to small indels should be straightforward. Basecalling proceeded nucleotide by nucleotide across the entire reference genome, ignoring repeats (defined as 200-bp windows whose theoretical hybridization spectrum was >50% identical to another 200 bp window in the genome). At each position, the four possible substitutions were considered (one of which would be identical to the reference sequence). For each substitution, typically five overlapping probes would change from present to absent and five from absent to present. There could be more than five in the cases where both strands were probed, as a result of having more than 512 probes. For each probe, we calculated the average of the measured probe intensity for all fragments containing that probe and overlapping the position by at least 20 bases (to guard against slight misalignments). This average intensity was used to calculate the posterior log-odds of a substitution, by taking the sum of log-odds of each probe given by the log-odds distribution for that probe (**Fig. 4b**), subtracting the log-odds for probes that would disappear as a result of the substitution. This is illustrated for one position and substitution in **Figure 4c**. Finally a prior probability term P was added to account for the prior expectation of a substitution. To assess accuracy, two kinds of mock substitutions were introduced: SNPs at a rate of 10^{-3} and private mutations at a rate of 10^{-4}. The only difference between them was that the two common SNP alleles were treated as a priori equally probable ($P = 0.0$) and more probable than the two rare alleles ($P = -1.2$), whereas private mutations were treated as completely unknown and thus received the standard bias ($P = -1.2$). The scheme was designed to mimic the ultimate target application, human genome resequencing. Next, an interim quality score q was calculated by taking the difference between the log-odds for the base called and the second most probable base call at each position. This measure should be roughly proportional to the logarithm of the error rate, that is, $q \propto log\ P_e$ as confirmed by the scatter plot in **Figure 6b**. The constant of proportionality was determined using a linear fit ($R^2 = 0.95$), and this was used to convert q to Q_{phred} (because $Q_{phred} = -10\ log\ P_e$). A call was made as 'N' if the raw quality q was less than a predefined q_{min}, and as 'R' if in a repeat; otherwise the base with maximum posterior odds was called. Coverage was reported as the fraction of non-N non-R bases; similarly, accuracy was reported as the fraction of accurate non-N non-R bases. The free parameters (q_{min} and P) were adjusted to balance false-positive and false-negative calls and the overall coverage. In all experiments reported here $q_{min} = 0.7$ (corresponding roughly to $Q_{phred} < 15$) and $P = -1.2$, that is, biased slightly against substitutions.

Quantitative PCR. To determine if gap regions were depleted during sample preparation, quantitative real-time PCR was performed as follows. 48 gaps and 48 nongaps were targeted, selected randomly from all gaps and nongaps larger than 1 kb (to allow some room for the design of good Q-PCR amplicons). Short (~100 bp) amplicons were designed to ensure amplification efficiency and to fit within the selected 200-bp fragment length. Targeted regions were distributed across the entire genome in proportion to the local depth of coverage (since gaps were more common in regions of low coverage and vice versa). Amplification was performed in a 7900 HT real-time PCR instrument (Applied Biosystems) fitted with a 384-well block. 20 µl reactions containing 0.25 ng/µl sample DNA, 200 µM dNTP (NEB), 3.5 mM $MgCl_2$, 8 pmol each primer, 0.5× SYBR Green (Invitrogen) and 2.25 U TaqExpress (Genetix) in 1× TaqExpress buffer were assembled and amplified in a two-step program (20 s at 94 °C, 20 s at 55 °C) followed by a melting-curve assay.

Each targeted region was assayed on purified genomic *E. coli* DNA and on an aliquot from a sample prepared for sequencing, taken just before the

circularization step. That stage was chosen because we had previously experienced difficulty with Q-PCR on circular DNA and because we had observed that most if not all depletion occurred at the PCR step (data not shown). Both samples were assayed in duplicate on a single 384-well RT-PCR plate, and the entire experiment was performed twice for a total of four data points per sample and amplicon. In addition, a dilution series (1×, 2×, 4× and 8×) of genomic DNA was assayed separately. A threshold was applied at 2,000 fluorescence units and the cycles-to-threshold value (CT-value) was calculated for each reaction. Primers were from Invitrogen; raw data, primer sequences and all calculations are provided as **Supplementary Data** online. Primer amplification efficiency was 90% on average (as directly measured using the genomic DNA dilutions). Of the 96 amplicons, 10 were discarded because they showed >25 CT-value on genomic DNA (average was about 17) or <50% efficiency. One amplicon was omitted due to aberrant amplification curve (early rise followed by a shallow and jittery slope). Interestingly, nine of the discarded amplicons were targeting gaps, indicating that gaps may have been more difficult to amplify by PCR, on average, than nongaps.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank M. Nilsson and U. Landegren for early advice on RCA; J. Sagemark for initial bioinformatics analysis of the feasibility of the concept; P. Bérubé for the *E. coli* DNA preparation; M. Belouchi, P. van Eerdewegh, J. Hooper, B. Houle, R. Paulussen for helpful discussions; and P. Ernfors for advice and discussions. This work was supported by Swedish Research Council grant 522-2006-6511.

AUTHOR CONTRIBUTIONS

A.P. developed the short LNA probes, participated in the development of sample preparation, the rolling-circle arrays and the hybridization cycle. G.B. participated in the development of sample preparation and the hybridization cycle. A.M. participated in the development of sample preparation and the rolling-circle arrays. P.L. participated in the development of image analysis and basecalling software and algorithms. E.H. developed the custom Peltier assembly, built the instrument and participated in the development of instrument control and image analysis software. S.L. conceived of the concept of shotgun SBH, participated in probe design, the development of sample preparation, rolling-circle arrays, hybridization cycle, and of the instrument control, image analysis and basecalling software; analyzed the experiments, directed the research and drafted the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturebiotechnology/.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Sanger, F., Nicklen, S. & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467 (1977).
2. Prober, J.M. *et al.* A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238**, 336–341 (1987).
3. Luckey, J.A. *et al.* High speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Res.* **18**, 4417–4421 (1990).
4. Venter, J.C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
5. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
6. Klein, R.J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
7. Maraganore, D.M. *et al.* High-resolution whole-genome association study of Parkinson disease. *Am. J. Hum. Genet.* **77**, 685–693 (2005).
8. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
9. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
10. Blazej, R.G., Kumaresan, P. & Mathies, R.A. Microfabricated bioprocessor for integrated nanoliter-scale Sanger DNA sequencing. *Proc. Natl. Acad. Sci. USA* **103**, 7240–7245 (2006).
11. Bennett, S.T., Barnes, C., Cox, A., Davies, L. & Brown, C. Toward the 1,000 dollars human genome. *Pharmacogenomics* **6**, 373–382 (2005).
12. Shendure, J., Mitra, R.D., Varma, C. & Church, G.M. Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.* **5**, 335–344 (2004).
13. Brenner, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630–634 (2000).
14. Ghadessy, F.J., Ong, J.L. & Holliger, P. Directed evolution of polymerase function by compartmentalized self-replication. *Proc. Natl. Acad. Sci. USA* **98**, 4552–4557 (2001).
15. Mitra, R.D. & Church, G.M. In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res.* **27**, e34 (1999).
16. Bing, D.H. *et al.* Bridge amplification: a solid phase PCR system for the amplification and detection of allelic differences in single copy genes. *Genetic Identity Conference Proceedings, Seventh International Symposium on Human Identification*, Scottsdale, AZ, September 18–20, 1996 (Promega Corp., Madison, WI, 1996).
17. Braslavsky, I., Hebert, B., Kartalov, E. & Quake, S.R. Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci. USA* **100**, 3960–3964 (2003).
18. Hyman, E.D. A new method of sequencing DNA. *Anal. Biochem.* **174**, 423–436 (1988).
19. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. & Nyren, P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* **242**, 84–89 (1996).
20. Metzker, M.L. *et al.* Termination of DNA synthesis by novel 3′-modified-deoxyribonucleoside 5′-triphosphates. *Nucleic Acids Res.* **22**, 4259–4267 (1994).
21. Canard, B. & Sarfati, R.S. DNA polymerase fluorescent substrates with reversible 3′-tags. *Gene* **148**, 1–6 (1994).
22. Hinds, D.A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).
23. Drmanac, R., Petrovic, N., Glisin, V. & Crkvenjakov, R. Sequencing of megabase plus DNA by hybridization: theory of the method. *Genomics* **4**, 114–128 (1989).
24. Drmanac, S. *et al.* Accurate sequencing by hybridization for DNA diagnostics and individual genomics. *Nat. Biotechnol.* **16**, 54–58 (1998).
25. Bains, W. & Smith, G.C. A novel method for nucleic acid sequence determination. *J. Theor. Biol.* **135**, 303–307 (1988).
26. Lysov, Y.P., Florent'ev, V.L., Khorlin, A.A., Khrapko, K.R. & Shik, V.V. Determination of the nucleotide sequence of DNA using hybridization with oligonucleotides. A new method. *Dokl. Akad. Nauk SSSR* **303**, 1508–1511 (1988).
27. Lizardi, P.M. *et al.* Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nat. Genet.* **19**, 225–232 (1998).
28. Koshkin, A.A. *et al.* LNA (Locked Nucleic Acids): synthesis of the adenine, cytosine, guanine, 5-methylcytosine, thymine and uracil bicyclonucleoside monomers, oligomerisation, and unprecedented nucleic acid recognition. *Tetrahedron* **54**, 3607–3630 (1998).
29. Donachie, W.D. The cell cycle of *Escherichia coli*. *Annu. Rev. Microbiol.* **47**, 199–230 (1993).
30. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. ii. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
31. Shamir, R. & Tsur, D. Large scale sequencing by hybridization. *J. Comput. Biol.* **9**, 413–428 (2002).
32. Drmanac, R. *et al.* Sequencing by hybridization (SBH): advantages, achievements, and opportunities. *Adv. Biochem. Eng. Biotechnol.* **77**, 75–101 (2002).
33. Arratia, R., Martin, D., Reinert, G. & Waterman, M.S. Poisson process approximation for sequence repeats, and sequencing by hybridization. *J. Comput. Biol.* **3**, 425–463 (1996).
34. Pe'er, I., Arbili, N. & Shamir, R. A computational method for resequencing long DNA targets by universal oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **99**, 15492–15496 (2002).
35. Whiteford, N. *et al.* An analysis of the feasibility of short read sequencing. *Nucleic Acids Res.* **33**, e171 (2005).
36. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
37. Church, G., Shendure, J. & Porreca, G. Sequencing thoroughbreds. *Nat. Biotechnol.* **24**, 139 (2006).
38. Williams, R. *et al.* Amplification of complex gene libraries by emulsion PCR. *Nat. Methods* **3**, 545–550 (2006).