Hybrid error correction de novo assembly of single-molecule sequencing reads

Article *in* Nature Biotechnology · July 2012

DOI: 10.1038/nbt.2280 · Source: PubMed

Jason Travis Howard
Howard Hughes Medical Institute

Ganeshkumar Ganapathy
Duke University

W. Richard Mccombie
Cold Spring Harbor Laboratory

Erich Jarvis
Duke University

nature
biotechnology

Hybrid error correction and *de novo* assembly of single-molecule sequencing reads

Sergey Koren[1,2], Michael C Schatz[3], Brian P Walenz[4], Jeffrey Martin[5], Jason T Howard[6], Ganeshkumar Ganapathy[6], Zhong Wang[5], David A Rasko[7], W Richard McCombie[3], Erich D Jarvis[6] & Adam M Phillippy[1]

Single-molecule sequencing instruments can generate multikilobase sequences with the potential to greatly improve genome and transcriptome assembly. However, the error rates of single-molecule reads are high, which has limited their use thus far to resequencing bacteria. To address this limitation, we introduce a correction algorithm and assembly strategy that uses short, high-fidelity sequences to correct the error in single-molecule sequences. We demonstrate the utility of this approach on reads generated by a PacBio RS instrument from phage, prokaryotic and eukaryotic whole genomes, including the previously unsequenced genome of the parrot *Melopsittacus undulatus*, as well as for RNA-Seq reads of the corn (*Zea mays*) transcriptome. Our long-read correction achieves >99.9% base-call accuracy, leading to substantially better assemblies than current sequencing strategies: in the best example, the median contig size was quintupled relative to high-coverage, second-generation assemblies. Greater gains are predicted if read lengths continue to increase, including the prospect of single-contig bacterial chromosome assembly.

Second-generation sequencing technologies, starting with 454 pyrosequencing[1] in 2004, Illumina sequencing-by-synthesis[2] in 2007 and others, have revolutionized DNA sequencing by reducing cost and increasing throughput exponentially over first-generation Sanger[3] sequencing. Despite the great gains provided by second-generation instruments, they have several drawbacks. First, they require amplification of source DNA before sequencing, leading to amplification artifacts[4] and biased coverage of the genome related to the chemical-physical properties of the DNA[5]. Second, current technologies produce relatively short reads, with median lengths of 100 bp for Illumina (max. 150 bp) and ~700 bp for 454 (max. 1,000 bp). Short reads make assembly and related analyses difficult, with theoretical modeling suggesting that decreasing read lengths from 1,000 bp to 100 bp can lead to a sixfold or more decrease in continuity[6].

Pacific Biosciences recently released their first commercial 'third-generation' sequencing instrument, the PacBio RS: a real-time, single-molecule sequencer. It aims to address the problems outlined above by requiring no amplification and reducing compositional bias[7,8], producing long sequences (e.g., median = 2,246, maximum = 23,000 bp using the latest PacBio chemistry)[9] and supporting a short turnaround time (24 h, sample to sequence)[8,10]. The long read lengths would be beneficial for *de novo* genome and transcriptome assembly as they have the potential to resolve complex repeats and span entire gene transcripts. However, the instrument generates reads that average only 82.1% (ref. 8)–84.6% (ref. 9) nucleotide accuracy, with uniformly distributed errors dominated by point insertions and deletions (**Supplementary Fig. 1**). This high error rate obscures the alignments between reads and complicates analysis because the pairwise differences between two reads is approximately twice their individual error rate, and is far beyond the 5–10% error rate[1,11,12] that most genome assemblers can tolerate; simply increasing the alignment sensitivity of traditional assemblers is computationally infeasible (**Supplementary Table 1** and **Supplementary Figs. 2** and **3**). Additionally, the PacBio technology utilizes hairpin adaptors for sequencing double-stranded DNA, which can result in chimeric reads if the sequencing reaction processes both strands of the DNA (first in the forward and then reverse direction). Although it is possible to generate accurate sequences on the PacBio RS by reading a circularized molecule multiple times (circular consensus or CCS), this approach reduces read length by a factor equal to the number of times the molecule is traversed, resulting in much shorter reads (e.g., median = 423 bp, max. = 1,915 bp). Thus, there is a great potential advantage to the long, single-pass reads if the error rate can be algorithmically managed.

To overcome the limitations of single-molecule sequencing data and unlock its full potential for *de novo* assembly, we developed an approach that utilizes short, high-accuracy sequences to correct the error inherent in long, single-molecule sequences (**Fig. 1**). Our PBcR (PacBio corrected Reads) algorithm, implemented as part of the Celera Assembler[11], trims and corrects individual long-read sequences by first mapping short-read sequences to them and computing a highly accurate hybrid consensus sequence: improving read accuracy from as low as 80% to over 99.9%. The corrected, 'hybrid' PBcR reads may then be *de novo* assembled alone, in combination with other data, or exported for other applications. As demonstrated below for several

[1]National Biodefense Analysis and Countermeasures Center, Frederick, Maryland, USA. [2]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA. [3]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. [4]The J. Craig Venter Institute, Rockville, Maryland, USA. [5]DOE Joint Genome Institute, Walnut Creek, California, USA. [6]Department of Neurobiology, Howard Hughes Medical Institute, Duke University Medical Center, Durham, North Carolina, USA. [7]Department of Microbiology & Immunology, Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland, USA. Correspondence should be addressed to S.K. (korens@nbacc.net) or A.P. (phillippya@nbacc.net).

Received 26 October 2011; accepted 18 May 2012; published online 1 July 2012; doi:10.1038/nbt.2280

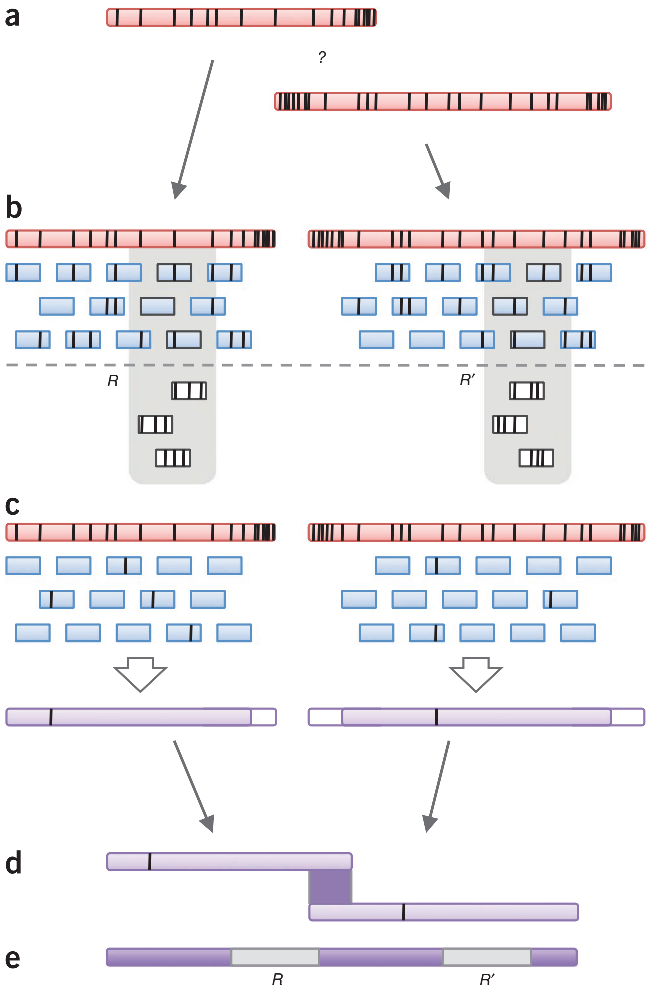

Figure 1 The PBcR single-molecule read correction and assembly method. (**a**) Errors, indicated by black vertical bars, in single-pass PacBio RS reads (pink rectangles) make it difficult to determine whether reads overlap. (**b**) Aligning high-fidelity short reads to error-prone long reads. Accurate alignments can be computed because the error between a short, high-accuracy sequence (~99% identical to the truth) and a PacBio RS sequence is half the error between two PacBio RS sequences. In this example, black bars in the short-reads indicate 'mapping errors' that are a combination of the sequencing error in both the long and short reads. In addition, a two-copy inexact repeat is present (outlined in gray) leading to pile-ups of reads at each copy. To avoid mapping reads to the wrong repeat copy, the algorithm selects a cutoff, *C*, and only the top *C* hits for each short read are used. The spurious mappings (in white) are discarded. (**c**) The remaining alignments are used to generate a new consensus sequence (purple), trimming and splitting long reads whenever there is a gap in the short-read tiling. Sequencing errors, indicated in black, may propagate to the PBcR read in rare cases where sequencing error co-occurs. (**d**) After correction, overlaps between long PBcR sequences can be easily detected. (**e**) The resulting assembly is able to span repeats that are unresolvable using only the short reads.

important genomes, including the previously unsequenced 1.2-Gbp genome of the parrot *Melopsittacus undulatus*, incorporation of PacBio data using this method leads to greatly improved assembly quality versus either first- or second-generation sequencing, indicating the promise of 'third-generation' sequencing and assembly.

RESULTS

De novo assembly of long reads

Genome assembly is the computational problem of reconstructing a genome from sequencing reads[13,14]. It and the closely related problem of *de novo* transcriptome assembly are critical tools of genomics, required to make order from a myriad of short fragments. The assembly problem is typically formulated as finding a traversal of a graph derived from sequencing reads using either the overlap-layout-consensus (OLC or string graph) paradigm, where the graph is constructed from overlapping sequencing reads, or the de Bruijn graph formulation, where the graph is constructed from substrings of a given length *k* derived from the reads. The complexity of the assembly graph is determined by both sequencing error and repeats, but repeats are the single biggest impediment to all assembly algorithms and sequencing technologies[15]. Under a de Bruijn graph formulation, repeats longer than *k* base pairs form branching nodes that must be resolved by threading reads through the graph or by applying other constraints, such as mate-pair relationships[16]. In contrast, only repeats longer than $l = r - 2 \times o$ (where *r* is the read length and *o* is the minimum acceptable overlap length) cause unresolved branches in a string graph. For short-read sequences, *k* and *l* are very similar, so the corresponding graphs are nearly equivalent. However, for long reads, *l* may be substantially longer than feasible values of *k*. Therefore, long sequences have great potential to simplify the OLC assembly problem. In the extreme case, if all repeats are spanned by reads of greater length, OLC assembly of a genome into its constituent chromosomes and/or plasmids would be trivial. In practice, longer reads increase the probability of spanning repeats and detecting overlaps[17], and thus produce better assemblies at lower sequencing coverage than short reads.

As a simple test, we evaluated the performance of multiple assemblers after correcting errors in lambda phage PacBio RS sequences with high-accuracy short-read sequencing technology (**Supplementary Table 1**); only the OLC assembler produced a single contig. To test the benefits of increasing read lengths, we simulated error-free data of varying length from the *Saccharomyces cerevisiae* S228c genome and compared the resulting N50 contig size (*N* such that 50% of the genome is contained in contigs ≥*N*, **Fig. 2**). OLC assembly becomes progressively more powerful for longer reads, displaying a nearly linear increase in contig size as read lengths grow. In contrast, the de Bruijn assemblies plateau and cannot effectively utilize the long reads without increasing *k* beyond practical values, owing to the inherent limitations of the graph construction and the complexity of the read-threading problem[16,18]. Therefore, we developed an algorithm to correct and assemble PacBio RS sequences using an OLC approach.

Correction accuracy and performance

We evaluated the PBcR correction and assembly algorithm on multiple short- and long-read data sets generated by Illumina, 454 and PacBio sequencing instruments, including three data sets with available reference sequences: Lambda NEB3011, *Escherichia coli* K12 and *S. cerevisiae* S228c (**Supplementary Table 2**). The correction accuracy and assembly continuity show diminishing returns after 50× of high-identity sequence; we recommend this coverage as a compromise between performance and accuracy (**Supplementary Figs. 4** and **5**). Using 50× of Illumina data to correct PacBio reads for each reference organism, the accuracy of the long reads improved from ~85% to >99.9%, and chimeric and improperly trimmed reads measured <2.5% and <1%, respectively (**Table 1** and **Supplementary Fig. 6**). The concurrence of the corrected reads with their references is testament to the automated trimming process, which is necessary for the removal of adaptor sequences that can otherwise be difficult to identify (Online Methods).

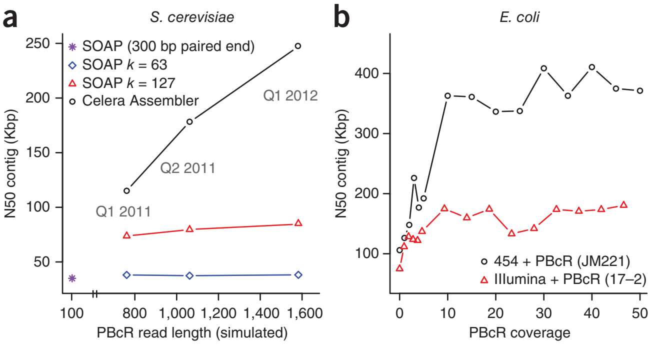

Figure 2 Long-reads yield assembly improvements, even at low coverage. (**a**) Effect of PacBio corrected read (PBcR) length on contig size is measured for the OLC assembler Celera Assembler[11] and the de Brujin assembler SOAPdenovo[49]. Contig size, after breaking contigs at structural assembly errors, is measured using the N50 metric. The baseline SOAPdenovo assembly (purple star) represents an assembly of 50× of real 76-bp Illumina paired-end (300 bp) reads from *S. cerevisiae* S228c. The effect of increasing PBcR read length was tested using 10× of simulated, error-free reads sampled from the *S. cerevisiae* genome. Read length was randomly sampled from actual length distributions of PBcR reads (from other genomes) to represent: the pre-release PacBio instrument (Q1, 2011), the first publicly available instrument (Q2, 2011), and the latest "C2" chemistry upgrade (Q1, 2012). (**b**) Effect of PBcR coverage is measured for *E. coli*, sequenced with a combination of PacBio and second-generation sequencing. The benefit of the PBcR sequences is visible even below 5×, which leads to a 50–100% increase in N50. Maximum contig N50 is reached by ~10×, where adding 10× of PBcR increases the N50 by as much as 3.5-fold (250%). The larger gain for the 454 hybrid assembly is due to the longer PBcR sequences available for *E. coli* JM221. The variation in N50 is due to random subsampling of sequencing data.

As a result, the corrected reads are slightly shorter than the originals, but length is not drastically affected (e.g., median 848 before correction versus 767 after correction for *E. coli* K12). During correction, reads may also be discarded because of unusually low quality or short length, and the percentage of reads that are successfully corrected is termed throughput. The observed throughput is generally around 60%, but varies widely depending on the quality of the individual runs. For example, throughput for the *S. cerevisiae* S228c reads appears unusually low, and that is likely because much of this sequencing was done using a pre-release PacBio RS instrument during testing at Cold Spring Harbor Laboratory. Nevertheless, in all cases the correction algorithm successfully identifies the usable data and outputs highly accurate long reads.

Hybrid *de novo* assembly

We evaluated the impact of PBcR reads on whole-genome assembly, either alone or in combination with the complementary reads. In addition to Celera Assembler, two other assemblers are reported to support PacBio reads: ALLPATHS-LG[19] and ALLORA[9]. However, neither program performs correction of or *de novo* assembly from uncorrected reads. Instead, ALLPATHS-LG uses the raw reads to assist in scaffolding and gap closure of short-read de Brujin assemblies. The downsides of this approach are that errors introduced in the short-read contigs may go uncorrected, and owing to computational limitations, this function is available only for genomes <10 Mbp with both an Illumina paired-end library <200 bp and a long-range Illumina jump library. Only the parrot genome presented here includes this required combination of Illumina and PacBio reads, but it is larger than the size limit and could not be evaluated. ALLORA, a long-read assembler based on AMOS[20–22], is computationally limited to small genomes and requires high-accuracy PacBio sequences, such as CCS, to operate. Inspired by our initial results, other researchers manually corrected low-accuracy PacBio sequences from the 2011 German *E. coli* outbreak using our consensus module and assembled it iteratively using ALLORA[9]. We have now evaluated our automated correction and assembly method on the same *E. coli* C227-11 genome, and have found it outperformed the previously published assembly (**Table 2**).

In all cases, from bacterial to eukaryotic, the incorporation of PBcR data produced substantially better assemblies than any other sequencing strategy tested—in the best cases, more than tripling the N50 contig size for equivalent depths of coverage (**Table 2**). These improvements also came without introducing additional assembly error, as measured against the three available reference genomes. The degree of improvement correlates with the median length of the corrected reads, with the newer, longer reads seeing bigger gains than the shorter reads of the older technology (**Table 2** and **Supplementary Fig. 7**). The observed gains are striking because they were entirely a result of resolving repeat structure rather than closing so-called sequencing gaps in the short-read coverage (Online Methods). This was due to the PBcR reads' unique ability to close difficult gaps left by second-generation technologies, such as interspersed, inverted and complex tandem repeats (e.g., VNTRs and STRs), that can be difficult to assemble even with paired ends (**Supplementary Fig. 8**).

Figure 3 summarizes the N50 results for various technologies and coverage for the *E. coli* genome. The three 'short-read' alternatives of 50× 454, 50× PacBio CCS and 100× Illumina paired-ends all produced similar assemblies. However, substituting half of the 454 coverage with corrected PacBio reads increased the N50 contig by threefold (e.g., 25× 454 + 25× PBcR); matching 50× short-read CCS coverage with 50× of PBcR reads resulted in a fivefold increase. Because PacBio sequencing can be completed in just hours, this pure PacBio example provides a promising method for rapid genotyping and sequencing in situations where time is critical, such as for an emerging disease outbreak.

An assembly of PacBio and CCS reads also outperformed an assembly of simulated Illumina short and long pairs by 44%, with an N50 of 527,198 versus 364,181 (**Supplementary Table 3**). In addition, the combination of PacBio reads and Illumina short-range paired data produces an assembly nearly identical to the idealized Illumina

Table 1 PacBio correction results

Organism	TP (%)	Idy (Reads) (%)	Idy (%) (Assembly)	Cov (%)	Chimer (%)	Trim (%)	Time (s)	Mem (GB)
Lambda NEB3011	74.03	99.90	100.00	100.00	1.82	0.10	121	0.12
E. coli K12	57.46	99.99	99.99	99.92	2.02	0.34	1,580	2.10
S. cerevisiae S228c	21.86	99.90	99.97	99.93	1.46	0.33	4,357	5.90

Corrected (PBcR) read accuracy as compared to reference sequence. Reads were mapped using Nucmer 3.23 (ref. 50). For all statistics, only reads >500 bp were included. %TP (throughput), the percentage of raw uncorrected bases that are in nonchimeric, correctly trimmed sequences after correction; %Idy (identity), average identity of good corrected reads to the reference; %Cov (coverage), average coverage of good corrected reads by a single match to the reference; %Chimer, the percentage of corrected bases within reads with a split mapping to the reference; %Trim, the percentage of corrected bases within reads with a single match to the reference over less than 99.5% of their length. The corrected sequences remain above 99% identity and 99% trim within the repetitive regions of the genome (**Supplementary Table 6**).

Table 2 PacBio assembly continuity

Organism	Technology	Reference bp	Assembly bp	No. of Contigs	Max. contig length	N50
Lambda NEB3011	Illumina 100× 200 bp	48,502	48,492	1	48,492 / 48,492	48,492 / 48,492 (100%)[a]
(median: 727 max: 3,280)	PacBio PBcR 25×		48,440	1	48,444 / 48,444	48,444 / 48,440 (100%)[a]
E. coli K12	Illumina 100× 500 bp	4,639,675	4,462,836	61	221,615 / 221,553	100,338 / 83,037 (82.76%)[a]
(median: 747 max: 3,068)	PacBio PBcR 18×		4,465,533	77	239,058 / 238,224	71,479 / 68,309 (95.57%)[a]
	PacBio PBcR 18× + Illumina 50× 500 bp		4,574,029	63	238,272 / 238,224	93,048 / 89,431 (96.11%)[a]
S. cerevisiae S228c	Illumina 100× 300 bp	12,157,105	11,034,156	192	266,528 / 227,714	73,871 / 49,254 (66.68%)[a]
(median: 674 max: 5,994)	PacBio PBcR 13×		11,110,420	224	224,478 / 217,704	62,898 / 54,633 (86.86%)[a]
	PacBio PBcR 13× + Illumina 50× 300 bp		11,286,832	177	262,846 / 260,794	82,543 / 59,792 (72.44%)[a]
E. coli C227-11	PacBio CCS 50×	5,504,407	4,917,717	76	249,515	92,580
(median: 1,217 max: 14,901)	PacBio PBcR 25× (corrected by 25× CCS)		5,207,946	80	357,234	98,774
	PacBio PBcR 25× + CCS 25×		5,269,158	39	647,362	227,302
	PacBio PBcR 50× (corrected by 50× CCS)		5,445,466	35	1,076,027	376,443
	PacBio PBcR 50× + CCS 25×		5,453,458	33	1,167,060	527,198
	Manually corrected ALLORA Assembly[9]		5,452,251	23	653,382	402,041
E. coli 17-2	Illumina 100× 300 bp	5,000,000	4,787,888	88	232,371	74,940
(median: 886 max: 10,069)	PacBio PBcR 50×		4,981,368	58	318,969	143,307
	PacBio PBcR 50× + Illumina 50× 300 bp		5,022,503	55	367,911	180,932
E. coli JM221	454 50×	5,000,000	4,714,344	66	308,063	106,034
(median: 1,216 max: 12,552)	PacBio PBcR 25×		5,005,429	30	631,386	314,500
	PacBio PBcR 25× + 454 25×		5,008,824	30	633,667	314,500
Melopsittacus undulatus	Illumina 194× (220/500/800 paired-end 2/5/10 Kb mate-pairs)	1.23 Gbp	1,023,532,850	24,181	1,050,202	47,383
	454 15.4X (FLX + FLX Plus + 3/8/20 Kbp paired-ends)		999,168,029	16,574	751,729	75,178
(median: 1,182 max: 14,596)	454 15.4X + PacBio PBcR 3.83× (corrected by 15.4× 454)		1,066,348,480	15,328	871,294	93,069
(median: 997 max: 13,079)	454 15.4X + PacBio PBcR 3.75X (corrected by 54× Illumina)		1,071,356,415	15,081	1,238,843	99,573

The median and maximum lengths of corrected PacBio sequences (PBcR) are given in parentheses. The corrected length is shorter than original PacBio RS sequences due to trimming and splitting chimeric sequences. **Supplementary Table 2** reports the original PacBio RS sequence lengths before correction. The three reference data sets (Lambda NEB3011, *E. coli* K12 and *S. cerevisiae* S228c) were generated using the prerelease PacBio RS, resulting in shorter read lengths. Pair separation (if applicable) is listed immediately after the coverage. Organism, the genome being assembled; technology, the read data used for assembly; reference bp, the assumed genome size used for the N50 calculation; assembly bp, the total number of base pairs in all contigs (only contigs ≥ 10,000 bp are included in all results); no. contigs, the number of contigs comprising the assembly; max. contig length, the maximum contig length. Assemblies for next-generation (Illumina/454) were generated by Celera Assembler[11,], SOAPdenovo[49] and ALLPATHS-LG[19] (where possible). Only the best assembly (based on continuity) in each case was reported.

[a]For genomes with an available reference, the max. and N50 contig was measured both before and after breaking contigs at structural assembly errors. The percentages in parenthesis indicate the ratio between corrected and original N50. A higher ratio indicates a more correct assembly. Full assembly quality statistics are listed in **Supplementary Table 7**, following the GAGE assembly evaluation methodology[12].

long-range libraries (**Supplementary Table 3**). As Illumina short-range libraries double the sequencing time, and long-range libraries are difficult to construct, these results suggest that long, single-molecule sequencing is a practical alternative to both. This comparison is based on the second-generation PacBio chemistry, with an uncorrected median read length of ~2 Kbp. As read lengths increase, our simulations predict that given adequate coverage of reads longer than around 5.5 Kbp (the size of the largest repeat), our algorithm will assemble the *E. coli* K12 chromosome into a single closed contig, without the need for paired reads (**Supplementary Note 1**).

Impact of long-read coverage on assembly

Long reads are capable of producing better assemblies, even at greatly reduced coverage. A comparison of the literature shows that a 10–20× Sanger assembly is better than a 100× Illumina assembly, albeit with prohibitively greater sequencing costs using the older technology[19,23]. We found that for *S. cerevisiae* S228c, an assembly using 13× of PBcR data (corrected by 50× Illumina) was comparable to an assembly of 100× of paired-end Illumina data (**Table 2**). This is true despite the fact that sequencing was done using a pre-release instrument. The corrected PacBio sequences also generated a more accurate assembly; whereas the 100× of Illumina produced a slightly longer raw N50, after splitting contigs at assembly errors, the N50 was larger for the PBcR assembly. Another striking example was *E. coli* JM221, for which the 25× PBcR assembly tripled the N50 of the 50× 454 assembly.

Given the evident ability of PBcR reads to improve assemblies, the added benefit of supplementing second-generation data was

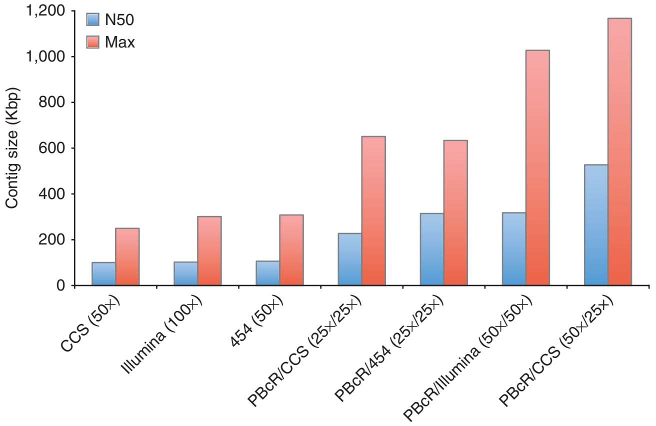

Figure 3 Contig sizes for various combinations of sequencing technologies. Assemblies are for *E. coli* C227-11 (assemblies including Illumina and PacBio CCS) and *E. coli* JM221 (assemblies including 454). Both genomes have similar repeat content, PacBio read length and coverage. Assemblies of only second-generation data are comparable and average N50 ≈ 100 Kbp. By comparison, adding 25× or 50× of PBcR to these data sets increases N50 as much as fivefold and results in a maximum contig size of greater than 1 Mbp (for the PBcR and CCS combination).

measured using *E. coli*. Between 1× to 50× of corrected PacBio data was added to the short-read data for an existing assembly (**Fig. 2**). The large and rapid gains after the addition of long-read sequencing were readily apparent. At just 10× coverage, nearly the maximum N50 was reached for the second-generation/PBcR assembly. The N50 measured a 2.5- and 3.5-fold improvement over the Illumina and 454 assemblies, respectively. These results demonstrate considerable improvements in continuity without the need for paired libraries and at relatively minor coverage. Thus, one might expect roughly double the N50 contig size with the addition of just 20× raw PacBio sequencing (assuming a throughput of >50% during correction).

Assembling the parrot genome

To demonstrate the applicability of the PBcR approach to vertebrate genomes, we used it to assemble the *Melopsittacus undulatus* genome. A total of 5.5× PacBio reads were corrected using 15.4× of 454 reads, producing 3.83× of sequences for a throughput of 69.62%. For comparison, the same PacBio RS sequences were corrected using 54× of Illumina, producing 3.75× of sequence. For the highest coverage data set, the correction took 6.8 d (20K CPU h) to complete. For reference, an ALLPATHS-LG Illumina assembly and a Celera Assembler 454 assembly each took over 1 week to complete, with the Celera Assembler using the same number of cores as PBcR. Thus, the correction represents an approximate doubling of the total assembly time.

Because the parrot genome had not been sequenced before and therefore did not have an available reference, correction accuracy was estimated by mapping PBcR reads to all parrot assemblies (except our own) submitted for the Assemblathon 2 (http://assemblathon.org/)[24]. For this diploid genome, each assembly is a mosaic of the two haplotypes, so only the best mapping for each PBcR read was considered. Using this method, we found that 99.9% of the 454-corrected PBcR reads had at least one mapping, and 97.0% mapped end-to-end with an average identity of 99.6%. Of the 3.0% of reads with fragmented mappings, 1.4% had breakpoints internal to a contig, which provides a rough estimate of chimerism. The remaining 1.6% mapped to contig boundaries and their accuracy could not be determined. In contrast, the Illumina-corrected reads showed a slightly increased rate of chimerism but maintained a similar identity (**Supplementary Note 2**). Considering likely haplotype switching in the reference assemblies, these slight increases in estimated error are not unexpected, but are likely amplified for the shorter Illumina reads, which are more difficult to uniquely map during correction. Nevertheless, in both cases the PBcR reads show good congruence with the independent assemblies, indicating that the correction algorithm succeeded for this difficult genome and can correct errors using both 454 or Illumina reads for complex vertebrate genomes, including human (**Supplementary Fig. 9**).

The PBcR reads were then co-assembled with 15.4× of 454 reads, which included 3-, 8- and 20-Kbp libraries to provide a diverse set of insert lengths, generating a PBcR-454 assembly and PBcR-454-Illumina assembly (where the Illumina data were used for correction only). For comparison, two additional assemblies were generated: one by running Celera Assembler with identical parameters but on the 454 data only, and a second by running ALLPATHS-LG on 194× of Illumina data, including 0.2-, 0.5-, 0.8-, 2-, 5- and 10-Kbp libraries. ALLPATHS-LG has been shown to be an effective short-read assembler for large genomes[12,24], and serves as an appropriate benchmark for assembling this genome using only Illumina data. A hybrid assembly of the 454 and Illumina data was not possible because Celera Assembler does not support high-coverage Illumina data and ALLPATHS-LG does not support 454. Interestingly, both the 454- and Illumina-corrected PBcR reads produced considerably better assemblies than the 454 data alone, demonstrating that the improvements were mostly attributable to the PacBio reads resolving repeats. To illustrate the effect of adding PBcR reads to an existing genome, the 454-corrected PBcR assembly is discussed below. Full statistics for both PBcR assemblies are included in **Supplementary Notes 3** and **4**.

The 454-PBcR assembly, with an N50 contig size of 93 Kbp, was more continuous than the second-generation assemblies in **Table 2** and more than twice that of previous avian genome assemblies sequenced using the gold-standard Sanger method. The zebra finch (*Taeniopygia guttata*) was sequenced to 6× coverage using Sanger sequencing, generating a maximum contig length of 425 Kbp and an N50 of 39 Kbp (ref. 25). The chicken *Gallus gallus* was also sequenced using Sanger to 6.6×, resulting in a maximum contig of 442 Kbp and an N50 of 36 Kbp (ref. 26). In contrast, for genomes assembled using only short-read sequencing, the N50 contig size rarely exceeds 30,000 bp (refs. 12,19,23). Much of the parrot genome continuity can be attributed to the long-read 454 data, including a mix of library sizes and the latest 454 FLX+ chemistry, but the addition of just 3.83× 454-PBcR sequences results in a 24% increase in N50, whereas the Illumina-corrected PBcR reads led to an increase of 32% (**Table 2**). The increased continuity of the Illumina-assisted assembly was likely due to the complementary benefit of multiple technologies, with the Illumina reads correcting PacBio reads that fill coverage gaps in the 454 data.

In addition to improved continuity, the overall quality of the contigs remained high after the addition of the PBcR reads. Long-range accuracy was supported by satisfaction of both assembled 454 pairs and mapped Illumina mate pairs (**Supplementary Note 3**), which serve as an effective indicator of assembly quality[15,27]. The percentage of bases not covered by satisfied 10-Kbp Illumina mates was virtually unchanged, and mate-pair coverage across the gaps closed by PBcR reads showed no observable deterioration (**Supplementary Fig. 10**). Additionally, of the 33,881 scaffold gaps in the 454 assembly, the 16,251 gaps closed by the 454-PBcR reads closely matched the corresponding gap size estimates from the 454 scaffolds (**Supplementary Fig. 11**).

Completeness and correctness of the 454-PBcR assembly was also supported by mapped zebra finch mRNA transcripts, which aligned

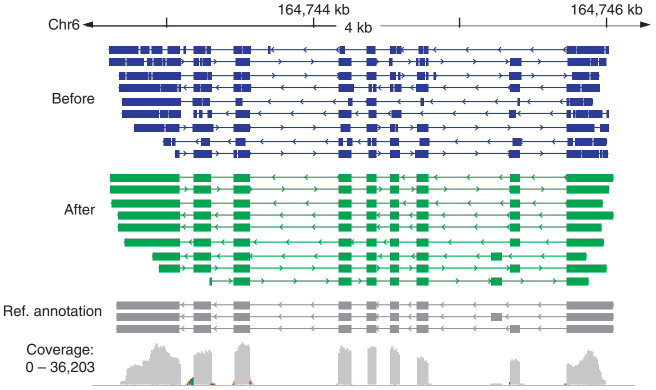

Figure 4 Error correction of RNA-Seq data provides more accurate mapping of transcripts. A genome browser view of cDNA alignments using uncorrected (dark blue) and Illumina-corrected (green) PacBio reads generated from *Zea mays* B73 cDNAs. The splice-aware aligner, BLAT[42], was used for aligning PacBio reads to the genome. Long gaps in the alignment correspond to introns in the PacBio reads but not the reference genome, and short gaps (only visible in the pre-corrected PacBio reads) are putative indel errors. The read coverage of the Illumina reads used for correction is also shown, along with the current reference gene annotation for this locus. The corrected PBcR sequences match the reference annotations end to end and include two isoforms. The colored bars in read coverage are an artifact of the aligner, indicating reads that have overhangs across exon junctions. Genome coordinates for chr6 are shown from the RefGen v2 genome assembly (http://maizesequence.org/).

to the PBcR assemblies with slightly higher coverage and fewer chimeras than the 454 assembly (chimeric mappings: 81 454-PBcR, 86 454, 85 Illumina; mapped coding bases: 23.95 Mbp 454-PBcR, 23.78 Mbp 454, 24.26 Mbp Illumina; **Supplementary Note 4**). Of the 15,275 finch mRNA sequences currently annotated in GenBank, ~95% are partially mappable to the parrot assemblies using the gmap spliced aligner[28]. Despite its smaller contigs, ALLPATHS-LG appears very effective at assembling and scaffolding exons, with its scaffolds containing an additional 1–2% of the transcript bases compared to the other assemblies as a result of the high Illumina coverage. All assemblies also showed similar identity to the mapped transcripts (89.17% 454-PBcR, 89.15% 454, 89.09% Illumina), but in terms of both exon coverage and identity, the PBcR assemblies were an improvement over the 454 assembly. For the 3,117 exons that were entirely contained in closed gaps, the average identity decreases slightly to 87.53%, and this 1.64% reduction from the average could be explained by limitations in the PBcR sequence accuracy or lowered sequencing depth across these difficult-to-sequence regions (**Supplementary Table 4**).

However, despite similarity between all assemblies at the exon level, the PBcR assemblies excelled at reconstructing the often-repetitive noncoding sequence: in the case of 454 correction, splitting 22% and 7% fewer transcripts across contigs than the Illumina and 454 assemblies, and covering a greater fraction of each transcript with a single contig (**Supplementary Fig. 12** and **Supplementary Table 5**). For example, 92% of the 454-PBcR–closed gaps occurred entirely outside of mapped finch exons, either within introns (18%) or between gene models (74%), and were enriched for extreme GC content (**Supplementary Table 4**). Such sequences are of particular importance for studying the parrot genome because "~40% of [zebra finch] transcripts in the unstimulated auditory forebrain are noncoding and derive from intronic or intergenic loci"[25].

Both the coding and noncoding sequences of genes with known relevance to vocal learning in birds are improved by the addition of PBcR reads (**Supplementary Note 5**). One striking example is the language and song–associated *FOXP2* gene[29–34], which is highly fractured in all but the PBcR assemblies (**Supplementary Fig. 13**). Additional examples include the neurotransmitter glutamate receptors GRIK3 (ref. 35), GRIN 2A and GRIN 2B (ref. 36), which contain intronic gaps closed only by the PBcR assemblies. The NAV3 (ref. 37) and PLEXIN A4 (ref. 38) axon guidance genes also showed improved reconstruction in the PBcR assemblies, with the full PLEXIN A4 transcript recovered by both PBcR assemblies and only 20.8% and 47.5% by the 454 and Illumina assemblies, respectively. Lastly, the published zebra finch and chicken assemblies both contain gaps ~700 bp upstream of *ERG1*, a major immediate-early gene that connects external stimuli to transcription in neurons[39,40]. The Illumina, 454 and 454-PBcR assemblies all contained a gap in this GC-rich (>70% GC) promoter region as well, but the 454-PBcR-Illumina assembly included the full sequence. In this case, the combination of Illumina, 454 and PacBio succeeded where all independent assemblies failed (including Sanger). We note that there were other examples where the 454 and Illumina assemblies outperformed the PBcR assemblies (**Supplementary Table 5**), and future work remains to best harness the complementary advantage of these multiple technologies.

Single-molecule RNA-Seq correction

Because the length of the single-molecule PacBio reads (ranging from a few hundred bases to several kilobases) from RNA-Seq experiments is within the size distribution of most transcripts, we expect many PacBio reads will represent full-length or near full-length transcripts. These long reads can therefore greatly reduce the need for transcript assembly, which requires complex algorithms for short reads[41], and confidently detect alternatively spliced isoforms. However, the predominance of indel errors makes analysis of the raw reads problematic. For example, in this study we generated 50,130 PacBio reads with a median size of 817 bp from a *Zea mays* B73 seedling mRNA sample, but only 11.6% (15,173) of the reads aligned to the reference genome by BLAT[42] at >90% sequence identity. In contrast, for the corrected PBcR sequences, the percentage of sequences that aligned at >90% identity increased dramatically to 99.1% (49,679 reads corrected in 3.6 d using 17.8× of Illumina data). Consistent with the results reported above for genome assembly, the corrected RNA-Seq sequences had very low error rates, with only 0.06% insertion and 0.02% deletion rates.

Many PacBio reads indeed represented close to full-length transcripts. However, the exon structure was not evident before the error correction by PBcR (**Fig. 4**). The post-correction sequences have virtually no errors and precisely identify splicing junctions. As a result, two of the isoforms at the displayed reference locus in the reference annotation were confirmed by PacBio RNA-Seq reads. To systematically test the ability of PacBio reads to validate annotated gene structure, we aligned the PacBio reads to the reference genome and looked for PacBio reads that matched the exon structure over the entire length of the annotated transcripts. Before correction, only 41 (0.1%) of the PacBio reads exactly matched the annotated exon structure. This number rose sharply after correction to 12, 065 (24.1%), suggesting that PBcR can greatly increase the usefulness of the PacBio RNA-Seq reads for transcript structure annotation or validation.

DISCUSSION

Current *de novo* assemblers are unable to effectively use the long-read sequencing data generated by present single-molecule sequencing technologies primarily because of the considerable error rate. Our approach exploits this technology by complementing it with shorter, high-identity sequences resulting in long, accurate transcripts and improved assemblies. Because the average contig size produced by our approach correlates with read length, assembly results are expected to improve as the read lengths of the technology improve. This strategy also benefits from the complementarity of multiple technologies, which proved powerful when combining Sanger sequencing with second-generation data when the latter first became available[43]. The result of our hybrid approach is higher quality assemblies with fewer errors and gaps, which will drive down the expensive cost of genome finishing and enable more accurate downstream analyses.

High-quality assemblies are critical for all aspects of genomics, especially genome annotation and comparative genomics. For example, many microbial genomic analyses depend on finished genomes[44], but producing finished sequence remains prohibitive with the cost of finishing proportional to the number of gaps in the original assembly. Eukaryotic genomics requires continuous assemblies to capture long, multi-exon genes and to determine genome organization and structural polymorphisms. In addition, recent work has suggested *de novo* assembly may be superior to read mapping approaches for discovering large structural variations, even when a reference genome is available[45]. This is especially significant for understanding the genetic variations of cancer genomes and other human diseases such as autism that frequently contain gene fusions, copy number variations and other large-scale structural variations[46,47]. It is clear that higher-quality assemblies, with long unbroken contigs, will have a positive impact on a wide range of disciplines.

Potential improvements to the PBcR algorithm include the addition of a gap-closure routine to fill sequencing gaps in the short-read data using the PacBio reads and incorporation of the single-molecule base calls during consensus calling. This is particularly important for GC-rich sequences that tend to be under-represented by second-generation sequencers, and for metagenomic and amplified samples that have severe coverage fluctuations. Nonuniform coverage will also require modifications to the repeat separation algorithm, as the current heuristic assumes uniform long-read coverage and error. This could include better utilization of paired-end information or variant clustering, which could also be applied to the problem of haplotype separation.

We have demonstrated that high error rates need not be a barrier to assembly. High-error, long reads can be efficiently assembled in combination with complementary short-reads to produce assemblies not possible with any prior technology, bringing us one step closer to the goal of "one chromosome, one contig." The rapid turnaround time possible with PacBio and other technologies, such as Ion Torrent[48], will make it possible to produce high-quality genome assemblies at a fraction of the time once required. Future studies are needed to explore the relative costs and trade-offs of the available technologies, but from our results we anticipate future sequencing projects will consist of a combination of both long- and short-read sequencing. Today, short-read insert libraries ≥9 Kbp are necessary for effective long-range scaffolding, for which the current PacBio reads provide limited assistance. However, if single-molecule technology continues to advance and reads begin to exceed the lengths of typical bacterial repeats (~6 Kbp) at reasonable cost and throughput, single-contig assemblies of some bacterial chromosomes will be possible without the need for expensive pair libraries. Additionally, we believe many long sought capabilities will be enabled, such as haplotype separation in eukaryotes, accurate transcriptome annotation and true comparative genomics that extends beyond an exon-centric view to include the whole genome.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. SRA: Lambda, SRS344250; *S. cerevisiae*, SRS344297; *Z. mays* SRA053579.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank Pacific Biosciences, Roche 454, Illumina, BGI and the Duke Genome Center for the generation and/or release of many of the data sets examined herein, and to the Assemblathon working group for the coordination and release of the parrot genome data. This publication was developed and funded in part under Agreement No. HSHQDC-07-C-00020 awarded by the US Department of Homeland Security for the management and operation of the National Biodefense Analysis and Countermeasures Center (NBACC), a Federally Funded Research and Development Center. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the US Department of Homeland Security. The Department of Homeland Security does not endorse any products or commercial services mentioned in this publication. The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231. This work was also funded in part by the US National Institutes of Health (NIH) R01-HG006677-12 (M.C.S.), NIH 2R01GM077117-04A1 (B.P.W.), the state of Maryland (D.A.R.), National Science Foundation IOS-1032105 to W.R.M., and Howard Hughes Medical Institute and NIH Directors Pioneer Award to E.D.J.

AUTHOR CONTRIBUTIONS

S.K. and A.M.P. conceived and designed the algorithm. S.K. implemented the algorithm and carried out the *de novo* assembly experiments. S.K., M.C.S. and A.M.P. drafted the manuscript, ran experiments and contributed analysis. B.P.W. modified the Celera Assembler to support long sequencing reads and developed the BOGART unitigger. J.M. and Z.W. sequenced *Z. mays* cDNA and performed analysis. J.H., G.G. and E.D.J. sequenced *M. undulatus* and performed analysis of vocal learning genes. D.A.R. provided and sequenced *E. coli* strains. W.R.M. sequenced *S. cerevisiae* S228c. All authors read and approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Published online at http://www.nature.com/doifinder/10.1038/nbt.2280.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
2. Bentley, D. Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.* **16**, 545–552 (2006).
3. Sanger, F., Nicklen, S. & Coulson, A. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467 (1977).
4. Niu, B., Fu, L., Sun, S. & Li, W. Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics* **11**, 187 (2010).
5. Dohm, J., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105 (2008).
6. Kingsford, C., Schatz, M. & Pop, M. Assembly complexity of prokaryotic genomes using short reads. *BMC Bioinformatics* **11**, 21 (2010).
7. Schadt, E.E., Turner, S. & Kasarskis, A. A window into third-generation sequencing. *Hum. Mol. Genet.* **19**, R227–R240 (2010).
8. Chin, C.-S. The origin of the Haitian cholera outbreak strain. *N. Engl. J. Med.* **364**, 33–42 (2011).
9. Rasko, D.A. *et al.* Origins of the *E. coli* strain causing an outbreak of hemolytic–uremic syndrome in Germany. *N. Engl. J. Med.* **365**, 709–717 (2011).
10. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
11. Miller, J.R. *et al.* Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**, 2818–2824 (2008).
12. Salzberg, S.L. *et al.* GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Res.* 557–567 (2012).
13. Pop, M. Genome assembly reborn: recent computational challenges. *Brief. Bioinform.* **10**, 354 (2009).

14. Miller, J., Koren, S. & Sutton, G. Assembly algorithms for next-generation sequencing data. *Genomics* **95**, 315–327 (2010).
15. Phillippy, A., Schatz, M. & Pop, M. Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol.* **9**, R55 (2008).
16. Pevzner, P.A., Tang, H. & Waterman, M.S. An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. USA* **98**, 9748–9753 (2001).
17. Schatz, M.C., Witkowski, J. & McCombie, W.R. Current challenges in *de novo* plant genome sequencing and assembly. *Genome Biol.* **13**, 243 (2012).
18. Nagarajan, N. & Pop, M. Parametric complexity of sequence assembly: theory and applications to next generation sequencing. *J. Comput. Biol.* **16**, 897–908 (2009).
19. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. USA* 1513–1518 (2011).
20. Pop, M., Phillippy, A., Delcher, A.L. & Salzberg, S.L. Comparative genome assembly. *Brief. Bioinform.* **5**, 237–248 (2004).
21. Schatz, M.C. *et al.* Hawkeye and AMOS: visualizing and assessing the quality of genome assemblies. *Brief. Bioinform.* published online, doi: 10.1093/bib/bbr074 (23 December 2011).
22. Sommer, D., Delcher, A., Salzberg, S. & Pop, M. Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics* **8**, 64 (2007).
23. Schatz, M.C., Delcher, A.L. & Salzberg, S.L. Assembly of large genomes using second-generation sequencing. *Genome Res.* **20**, 1165–1173 (2010).
24. Earl, D.A. *et al.* Assemblathon 1: a competitive assessment of *de novo* short read assembly methods. *Genome Res.* 2224–2241 (2011).
25. Warren, W.C. *et al.* The genome of a songbird. *Nature* **464**, 757–762 (2010).
26. Hillier, L. *et al.* Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004).
27. Vezzi, F., Narzisi, G. & Mishra, B. Feature-by-feature—evaluating *de novo* sequence assembly. *PLoS ONE* **7**, e31002 (2012).
28. Wu, T.D. & Watanabe, C.K. Gmap: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
29. Enard, W. *et al.* Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* **418**, 869–872 (2002).
30. Enard, W. FOXP2 and the role of cortico-basal ganglia circuits in speech and language evolution. *Curr. Opin. Neurobiol.* **21**, 415–424 (2011).
31. Lai, C.S., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F. & Monaco, A.P. A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* **413**, 519–523 (2001).
32. Haesler, S. *et al.* FoxP2 expression in avian vocal learners and non-learners. *J. Neurosci.* **24**, 3164–3175 (2004).
33. Haesler, S. *et al.* Incomplete and inaccurate vocal imitation after knockdown of FoxP2 in songbird basal ganglia nucleus Area X. *PLoS Biol.* **5**, e321 (2007).
34. Carroll, S.B. Evolution at two levels: on genes and form. *PLoS Biol.* **3**, e245 (2005).
35. Brose, K. *et al.* Slit proteins bind Robo receptors and have an evolutionarily conserved role in repulsive axon guidance. *Cell* **96**, 795–806 (1999).
36. Wada, K., Sakaguchi, H., Jarvis, E.D. & Hagiwara, M. Differential expression of glutamate receptors in avian neural pathways for learned vocalization. *J. Comp. Neurol.* **476**, 44–64 (2004).
37. Maes, T., Barcelo, A. & Buesa, C. Neuron navigator: a human gene family with homology to unc-53, a cell guidance gene from *Caenorhabditis elegans*. *Genomics* **80**, 21–30 (2002).
38. Matsunaga, E. & Okanoya, K. Vocal control area-related expression of neuropilin-1, plexin-A4, and the lig-and semaphorin-3A has implications for the evolution of the avian vocal system. *Dev. Growth Differ.* **51**, 45–54 (2009).
39. Morgan, J.I. & Curran, T. Stimulus-transcription coupling in neurons: role of cellular immediate-early genes. *Trends Neurosci.* **12**, 459–462 (1989).
40. Jarvis, E.D. & Nottebohm, F. Motor-driven gene expression. *Proc. Natl. Acad. Sci. USA* **94**, 4097–4102 (1997).
41. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
42. Kent, W.J. Blat–the blast-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
43. Goldberg, S. *et al.* A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc. Natl. Acad. Sci. USA* **103**, 11240–11245 (2006).
44. Fraser, C.M., Eisen, J.A., Nelson, K.E., Paulsen, I.T. & Salzberg, S.L. The value of complete microbial genome sequencing (you get what you pay for). *J. Bacteriol.* **184**, 6403–6405 (2002).
45. Li, Y. *et al.* Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome *de novo* assembly. *Nat. Biotechnol.* **29**, 723–730 (2011).
46. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
47. Sebat, J. *et al.* Strong association of *de novo* copy number mutations with autism. *Science* **316**, 445–449 (2007).
48. Rothberg, J.M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352 (2011).
49. Li, R. *et al.* *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
50. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).

ONLINE METHODS

Our strategy consists of two phases: a long-read correction phase and an assembly phase. Both are implemented as part of the Celera Assembler[11], but the output of the correction phase can be used as input to any other analysis or assembler capable of utilizing long FastA sequences. The outline of the correction algorithm is as follows: (i) high-identity short-read sequences are simultaneously mapped to all long-read sequences, (ii) repeats are resolved by placing each short-read sequence in its highest identity repeat copy, (iii) chimera and trimming problems are detected and corrected within the long-read sequences, and (iv) a consensus sequence is computed for each long-read sequence based on a multiple alignment of the short-read sequences. This approach was inspired by the intuition that whereas overlaps between single-pass PacBio reads average 31.6% pairwise differences (~16.8% + 16.8%, **Supplementary Fig. 3a**), overlaps between long-read sequences and high-identity sequences would be lower and easier to detect. As most second-generation sequence overlaps are found below 3% error (**Supplementary Fig. 3a**), the average overlap between PacBio reads and high-identity short-read sequences should be at most 17.5% (~16.8% + 1%) (**Supplementary Fig. 3b**).

The algorithm begins by computing all-versus-all overlaps between the low-accuracy, single-pass, (PacBio) long-read sequences and high-identity short-read sequences (Illumina, 454, PacBio CCS). The overlaps are computed only between fragments that have shared seed sequences of a pre-defined length (14 bp by default), and only short-read sequences aligned across their entire length to a long-read sequence are considered; support for partial overlaps to the ends of long reads is left for future versions. For efficiency, overlaps between reads of the same technology (e.g., short to short) are not computed during this phase.

Next, overlaps are converted into a tiling of short-read sequences along each long-read sequence. Each short-read sequence is permitted to map to more than one long-read sequence, as the long-read sequences are expected to cover the genome at more than 1× coverage. However, within a single long-read sequence, a short-read sequence is placed only in its highest identity location with ties broken randomly. In the case of repeats distributed across multiple long-read sequences, short-read sequences from all repeat copies will map to each copy of the repeat. To avoid tiling each repeat copy with the same set of reads, short-read sequences are separated into their appropriate copies by ranking their mappings by identity and permitting each short-read sequence to map only to its top C hits, where C is roughly defined to be the expected long-read sequencing depth. This effectively separates repeat copies when sequencing coverage and error is uniform. The value of C, a repeat threshold, is defined as follows:

Given a histogram

$$H = \sum_{n=1}^{n\le\max(n_i)} (n_i)\,\forall i$$

and a threshold $0 \le T \le 1$

$$\text{slope}\,(K) = \frac{H_K}{H_{K-1}}\,\forall K \ge 2$$

$$\text{total}\,(K) = \sum_{k=1}^{k\le K}\left(\frac{H_k}{\sum_{n=1}^{n\le\max(n_i)} H_n}\right)$$

$$C = \min(K)\ \text{s.t. total}\,(K) \ge T\ \text{and}$$

$$\text{slope}\,(K) > \text{slope}\,(K-1)\ \text{and}$$

$$H_K < H_{K-1}$$

Where n_i is the number of long-read sequences a short-read sequence i maps to $\forall i$. Theoretically, the histogram H has a peak equal to the long-read depth of coverage. It can be expected that a unique short-read sequence will map to, on average, this many long-read sequences. Thus, a short-read sequence from a two-copy repeat will map to roughly double this number. The chosen repeat threshold is the point in the curve past this peak that includes at least T % of the high-identity reads (**Supplementary Fig. 14**). In this way, each repeat copy will only be tiled by its best representative reads for correction. This approach can sometimes place reads in the wrong repeat copy. For instance, in cases where the error rate of two PacBio RS sequences from two separate repeat instances is significantly different, such that one is higher, Illumina sequences may preferentially map to the lower-error PacBio read. This would increase the mapped coverage of the low-error read by including some reads from the alternate copy, while decreasing the coverage of the high-error read. However, this problem should be alleviated as overall PacBio coverage is increased because the read accuracy distribution in the different repeat copies will converge after a few-fold redundancy. As evidence, systematic misplacement of reads in repeats, leading to inaccurate correction, coverage fluctuations or decreased throughput, has not been observed in any of our experiments (e.g., **Table 1**, **Supplementary Table 6** and **Supplementary Fig. 15**).

Finally, from the multiple-alignment of the tiled short-read sequences, the correction algorithm generates a new consensus sequence for each long-read sequence using the AMOS consensus module[20]. In the consensus, if there is a gap in the layout between adjacent overlapping short reads, this is considered an irreconcilable discrepancy between the short and long-read sequences, especially as the reads are generated from the same biological sample and it is assumed there is sufficient coverage in the short sequences to tile each long-read sequence. Therefore, any gap in coverage is indicative of either improper trimming of the long-read sequence or chimera formation, and the long-read sequence is broken at this point. If instead there is merely insufficient coverage leading to a true sequencing gap for the short-read sequences, this will result in an unnecessary split. However, the correction algorithm errs on the side of caution. Future work remains to resolve any unnecessary gaps caused by the conservative trimming, such as by recognizing and filling these gaps during scaffolding, using a consensus of multiple long-read sequences (**Supplementary Fig. 16**). The pipeline has been designed to run parallel either using a shared-memory machine or a distributed grid supporting Sun Grid Engine (SGE) (**Supplementary Note 6**).

The corrected, now high-identity, long-read sequences are provided in FastA format and can be assembled alone or co-assembled with other read types using standard OLC assembly techniques. To support *de novo* assembly using Celera Assembler, we have increased the input size limitation to allow a maximum read length of 65,536 bp.

Availability. The latest version of the Celera Assembler (including PBcR) along with usage instructions is available at http://wgs-assembler.sourceforge.net/. The software and data used for this manuscript are available at http://www.cbcb.umd.edu/software/PBcR/. The source code used in this study is also available as **Supplementary Datasets 1** and **2**.