High-quality draft assemblies of mammalian genomes from massively parallel sequence data

Sante Gnerre[a], Iain MacCallum[a], Dariusz Przybylski[a], Filipe J. Ribeiro[a], Joshua N. Burton[a], Bruce J. Walker[a], Ted Sharpe[a], Giles Hall[a], Terrance P. Shea[a], Sean Sykes[a], Aaron M. Berlin[a], Daniel Aird[a], Maura Costello[a], Riza Daza[a], Louise Williams[a], Robert Nicol[a], Andreas Gnirke[a], Chad Nusbaum[a], Eric S. Lander[a,b,c,1], and David B. Jaffe[a,1]

[a]Broad Institute of MIT and Harvard, Cambridge, MA 02142; [b]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139; and [c]Department of Systems Biology, Harvard Medical School, Boston, MA 02115

Contributed by Eric S. Lander, November 23, 2010 (sent for review October 8, 2010)

Massively parallel DNA sequencing technologies are revolutionizing genomics by making it possible to generate billions of relatively short (~100-base) sequence reads at very low cost. Whereas such data can be readily used for a wide range of biomedical applications, it has proven difficult to use them to generate high-quality de novo genome assemblies of large, repeat-rich vertebrate genomes. To date, the genome assemblies generated from such data have fallen far short of those obtained with the older (but much more expensive) capillary-based sequencing approach. Here, we report the development of an algorithm for genome assembly, ALLPATHS-LG, and its application to massively parallel DNA sequence data from the human and mouse genomes, generated on the Illumina platform. The resulting draft genome assemblies have good accuracy, short-range contiguity, long-range connectivity, and coverage of the genome. In particular, the base accuracy is high (≥99.95%) and the scaffold sizes (N50 size = 11.5 Mb for human and 7.2 Mb for mouse) approach those obtained with capillary-based sequencing. The combination of improved sequencing technology and improved computational methods should now make it possible to increase dramatically the de novo sequencing of large genomes. The ALLPATHS-LG program is available at http://www.broadinstitute.org/science/programs/genome-biology/crd.

The high-quality assembly of a genome sequence is a critical foundation for understanding the biology of an organism, the genetic variation within a species, or the pathology of a tumor. High-quality assembly is particularly challenging for large, repeat-rich genomes such as those of mammals. Among mammals, "finished" genome sequences have been completed for the human and the mouse (1, 2). However, for most large genomes, efforts have focused on using shotgun-sequencing data to produce high-quality draft genome assemblies—with long-range contiguity in the range of 20–100 kb and long-range connectivity in the range of 10 Mb (e.g., refs. 3–5). Using traditional capillary-based sequencing, such assemblies have been produced for multiple mammals at a cost of tens of million dollars each.

Recently, there has been a revolution in DNA sequencing technology. New massively parallel technologies can produce DNA sequence information at a per-base cost that is ~100,000-fold lower than a decade ago (6, 7). In principle, this should make it possible to dramatically decrease the cost of generating high-quality draft genome assemblies. In practice, however, this has been difficult because the new technology produces sequencing "reads" of only ~100 bases in length (compared with >700 bases for capillary-based technology). These shorter reads are also less accurate. For both of these reasons, these data are more difficult to assemble into long contiguous and connected sequence. Excellent de novo assemblies using massively parallel sequence data have been reported for microbes with genomes up to 40 Mb (refs. 8–10 and many others). There have been some important pioneering efforts (11, 12) for large genomes, but they fall far short of the high-quality draft sequences that can be obtained with the earlier technology. Moreover, fundamental issues have been raised about the quality of de novo assemblies that can be constructed from such data (13).

Here, we describe an algorithm and software package ALLPATHS-LG for de novo assembly of large (and small) genomes. We demonstrate the power of the approach by applying it to massively parallel sequence data generated from both the human and the mouse genomes. The results approach the quality of assemblies obtainable with capillary-based sequencing in terms of completeness, contiguity, connectivity, and accuracy. The uncovered regions of the genome consist largely of repetitive sequences, with segmental duplications remaining a particularly important challenge. The results indicate that it should be possible to generate high-quality draft assemblies of large genomes at ~1,000-fold lower cost than a decade ago.

Results

Model for Input Data. De novo genome assembly depends both on the computational methods used and on the nature and quantity of sequence data used as input. For capillary-based sequencing, genome scientists ultimately converged around a fairly standard model, specifying the desired coverage from libraries of various insert sizes. For massively parallel sequencing data, we specify such a model in Table 1.

We adopted this model for several reasons. First, it requires constructing only a few libraries, reducing the laboratory burden and the amount of DNA required. Second, the fragment library has inserts that are short enough that the sequencing reads from each end overlap by ~20% and can be merged to create a single longer "read". (The current read length is ~100 bases; as read lengths increase, insert sizes should be ~1.8 times the read length.) Third, we obtain long-range connectivity by using "jumping libraries" (in which the middle of the insert is removed, ref. 14) because current technology cannot sequence fragments > ~1 kb.

Our model sets a target of 100-fold sequence coverage (to compensate for shorter reads and possibly nonuniform coverage), whereas the model for capillary sequencing required only 8- to 10-fold coverage. Despite using higher coverage, the proposed model is dramatically cheaper because the per-base cost of

Author contributions: S.G., I.M., D.P., F.J.R., J.N.B., B.J.W., T.S., G.H., D.A., M.C., R.D., L.W., R.N., A.G., and D.B.J. performed research; T.P.S., S.S., A.M.B., C.N., and E.S.L. analyzed data; S.G. and I.M. led the genome assembly team; S.G., I.M., D.P., F.J.R., J.N.B., B.J.W., T.S., G.H., and D.B.J. developed and implemented algorithms; T.P.S. and S.S. generated SOAPdenovo and ABySS assemblies; D.A., M.C., R.D., L.W., R.N., and A.G. performed laboratory research and development; and C.N., E.S.L., and D.B.J. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The sequence data reported in this paper have been deposited in the NCBI Short Read Archive (study names Human_NA12878_Genome_on_Illumina and Mouse_B6_Genome_on_Illumina 2) and in the DDBJ/EMBL/GenBank database (accession nos. AEKP00000000, AEKQ00000000, and AEKR00000000).

[1]To whom correspondence may be addressed. E-mail: lander@broadinstitute.org or jaffe@broadinstitute.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1017351108/-/DCSupplemental.

GENETICS

Table 1. Provisional sequencing model for de novo assembly

Libraries, insert types*	Fragment size, bp	Read length, bases	Sequence coverage, ×	Required
Fragment	180[†]	≥100	45	Yes
Short jump	3,000	≥100 preferable	45	Yes
Long jump	6,000	≥100 preferable	5	No[‡]
Fosmid jump	40,000	≥26	1	No[‡]

*Inserts are sequenced from both ends, to provide the specified coverage.
[†]More generally, the inserts for the fragment libraries should be equal to ∼1.8 times the sequencing read length. In this way, the reads from the two ends overlap by ∼20% and can be merged to create a single longer read. The current sequencing read length is ∼100 bases.
[‡]Long and Fosmid jumps are a recommended option to create greater continuity.

massively parallel sequencing is ∼10,000-fold lower than the current cost of capillary sequencing. [Coverage can be measured in different ways. For Illumina sequencing, we define coverage in terms of purity-filtered bases (ref. 6 and Table 2).]

We developed several laboratory techniques for making the libraries (see *SI Materials and Methods* for details): (*i*) For fragments, we adapted existing protocols with the goal of improving the representation of high GC-content DNA; (*ii*) for short jumps (∼3 kb), we used the Illumina protocol (6); (*iii*) for long jumps (∼6 kb), we used a protocol that we had previously developed, on the basis of a protocol for the SOLiD sequencing platform that involves circularization and EcoP15I digestion (7, 9); and (*iv*) for Fosmid jumps (∼40 kb), we developed two methodologies, "ShARC" and "Fosill" (described in *SI Materials and Methods*).

Sequencing Data. Using the model above, we generated sequence data from human and mouse genomes (Table 2), using the Illumina GAII and HiSeq sequencers (*SI Materials and Methods*). For the human, we sequenced the cell line GM12878 because it has been extensively sequenced and analyzed as part of the 1000 Genomes Pilot Project (15). (The cell line GM12878 is from the Coriell Institute. DNA from this cell line is denoted NA12878.)

For the mouse, we used C57BL/6J female DNA because it was the strain used for the draft and finished sequences of the mouse (2, 3). The data have been deposited in the NCBI Short Read Archive under study names Human_NA12878_Genome_on_Illumina and Mouse_B6_Genome_on_Illumina.

ALLPATHS-LG Assembly Method. We next needed to develop algorithms and a software package able to perform de novo assembly of large mammalian genomes. For this purpose, we made extensive improvements to our previous program ALLPATHS (9, 16), which can routinely assemble small genomes. The improved program is called ALLPATHS-LG and is freely available at http://www.broadinstitute.org/science/programs/genome-biology/crd. We outline some of the key innovations (for more details, see *SI Materials and Methods*):

i) Handling repetitive sequences. Repetitive sequence is the fundamental genomic feature that stymies assembly. We adapted ALLPATHS-LG to be more resilient to repeats, as follows. In its initial assembly representation (called a unipath graph), ALLPATHS collapses repeats of length ≥K, where K is chosen to be short enough that overlaps of length K between reads are abundant (16). In ALLPATHS-LG, we are able to use a larger K (in this work 96) by performing an initial step dubbed "read doubling," in which the two end sequences from a fragment are pasted together provided that the overlap between them is confirmed by another read pair or if that read pair fills in a gap (Fig. S1*A*). A given pair can have more than one such completion, as could happen, for example, if a single-nucleotide polymorphism (SNP) were to fall between the two ends of a pair (Fig. S1*B*).

ii) Error correction (cf. ref. 17). We describe the ALLPATHS-LG approach to error correction. For every 24-mer, the algorithm examines the stack of all reads containing the 24-mer. Individual reads may be edited if they differ from

Table 2. Experimental data for human and mouse assemblies

Species	Library type	No. of libraries	DNA used, μg	Mean size, bp	Read length	Sequence coverage, ×					Physical coverage, ×
						All	PF	Aligned	Unique	Valid	
Human	Fragment	1	3	155	101	51.9	41.8	38.4	37.9	36.5	27.8
	Short jump	2	20	2,536	101	45.9	40.7	33.7	31.7	19.7	249.4
	Fosmid jump	2	20	35,295	76*	5.3	4.0	3.0	0.4	0.3	49.5
	Total	5	43			103.1	86.5	75.1	70.0	56.5	326.7
Mouse	Fragment	1	3	168	101	58.6	53.1	49.6	46.6	45.3	37.6
	Short jump	3	20	2,209	101	48.0	40.7	35.1	32.0	19.9	219.1
	Long jump	5	50	7,532	26	13.5	9.3	9.2	5.5	2.9	408.3
	Fosmid jump	1	30	38,453	76	1.4	1.1	1.1	0.1	0.1	23.1
	Total	10	103			121.5	104.2	95.0	84.2	68.2	688.1

The data used as assembly input are shown. Tables S1 and S2 provide more detail. Library type: See Table 1. DNA used: Amount of DNA used as input to library construction. For each genome and each library type, a single aliquot was used. DNA source for human: Coriell Biorepository, NA12878. DNA source for mouse: Jackson Laboratory C57/BL6J (stock 000664). Size: Mean of observed fragment size distribution. Read length: Number of bases sequenced. The exception is the long jump libraries prepared with the EcoP15I digestion, which yield 26 bases of genomic information; these inserts were sequenced to 36 bases and then trimmed to 26 bases. Sequence coverage: All reads were used in the assembly, but we describe their properties here via a series of nested categories. All: Total number of bases in reads, divided by genome size, assumed to be the reference size of 3.10 Gb for human and 2.73 Gb for mouse. PF: Coverage by purity-filtered (PF) reads. Aligned: Coverage by aligned PF reads. Unique: Coverage by aligned PF reads, exclusive of duplicates, which were identified by concurrence of start and stop points of pairs on the reference. Valid: Coverage by unique pairs for which the fragment length was within 5 SDs of the mean. Physical coverage: Total coverage by valid pairs and the bases between them.
*Reads from one library had length 76, and those from the other had length 101.

the overwhelming consensus of the stack. If a given base on a read receives conflicting votes (arising from membership of the read in multiple stacks), it is not changed. This latter step is designed to reduce the incidence of incorrect error correction.

iii) Use of jumping data. Data from jumping libraries present the challenge that the junction point of the jump will often lie within one of the two reads. In addition, nonjumped fragments can result in reads occurring in the reverse orientation with respect to the genome (Fig. S2). Depending on the read length, these events can occur in ≥50% of the read pairs (Table 2). We designed ALLPATHS-LG so that it could work with such data: It trims bases beyond putative junction points and treats each read pair as belonging to one of two possible distributions.

iv) Efficient memory usage. Because data sets for mammalian genome assembly are large ($\sim 3 \times 10^9$ reads) and any read could a priori overlap any other read, many computations require large amounts of data to be kept concurrently in memory. For these reasons, algorithms that work for smaller genomes, including Velvet (8) and earlier versions of ALLPATHS, cannot handle these large data sets. We engineered ALLPATHS-LG to economize the data structures and make efficient use of shared memory parallelization. ALLPATHS-LG can assemble mammalian genomes on commercial servers (Dell R815, 48 processors, 512 GB RAM, $39,000), in a few weeks (3 wk for mouse, 3.5 wk for human).

v) Low coverage regions. If reads were truly randomly distributed across the genome with the ∼100× average coverage used here, bases with <10× coverage would be extremely rare, occurring far less than once per mammalian genome. In practice, however, recalcitrant sequence contexts (including those with low and high GC content) do cause low coverage (9, 18), sometimes even to zero. (For our human sequence data, 0.07% of genome bases have zero coverage, and the N50 covered stretch of bases with nonzero coverage is 467 kb, whereas 0.3% of genome bases have <10× coverage, and the N50 size of stretches covered to ≥10× coverage is 75 kb.) In regions of very low coverage, overlaps between reads may be short. In these cases, it is desirable to use a smaller value of K within the region. Accordingly, ALLPATHS-LG can use K as small as 15, but only where reads have been bounded to lie within a short gap between two other sequences. (Implementation of this step is discussed in *SI Materials and Methods* and Fig. S3.)

Uncertainty in Assemblies. The goal of assembly is to reconstruct the genome as accurately as possible. For some loci, however, the data may be compatible with more than one plausible answer. Examples include the location of a SNP or of a long homopolymer run whose length cannot be perfectly resolved with the available data. Rather than making an arbitrary choice (and thus sometimes introducing errors), we designed ALLPATHS to retain and reflect the alternatives. The algorithm creates an assembly graph whose edges are sequences and whose branches represent alternate choices.

Because such a graph could be extremely complicated for a large genome and thus unusable by typical applications, ALLPATHS-LG introduces a preliminary output format in which alternatives are expressed "linearly," e.g.,

. . .ATC{A,T}GGTTTTTTT{,T, TT}ACAC . . . .

This example exhibits both a SNP and a homopolymer of uncertain length. The same information could be readily encoded as a markup of the genome, for example via a "vcf" file [Variant call format (vcf), http://vcftools.sourceforge.net/specs.html]. In principle, other applications could exploit this information. In practice, this remains an important challenge for the field.

We note that the current version of ALLPATHS-LG captures only single-base and simple sequence indel uncertainties. It would thus be valuable to devise better ways to capture alternatives, many of which are still lost in the current ALLPATHS-LG process, giving rise to errors. Moreover, it would be desirable to assign probabilities to each alternative, reflecting their likelihood of being present in the sample.

Human and Mouse Assemblies. The resulting genome assemblies provide good coverage of the human and mouse genomes (Table 3). (In all our assembly analyses, we defined contig boundaries at runs of one or more Ns and discarded contigs <1 kb. For the SOAP mouse assembly, we observed that contig size could be tripled by first deleting isolated Ns, and therefore we did this.) We compared our ALLPATHS-LG assemblies to previously published assemblies obtained with capillary-based sequencing (3, 19), as well as with assemblies obtained with massively parallel sequencing data using the recently published program SOAPdenovo (referred to here as SOAP, ref. 11). For human, we specifically used the human assembly "YH" published by the authors of SOAP (12). [Two human assemblies are presented in ref. 12, from the Asian individual YH and an anonymous African. We chose the YH assembly for comparison because it has greater contiguity. We note that two de novo YH assemblies were available (one deposited at the National Center for Biotechnology Information, GenBank accession no. ADDF010000000, and one at http://yh.genomics.org.cn). We chose the latter because it has greater contiguity (*SI Materials and Methods*).] For mouse, we ran the program on our own data. [We also ran the program ABySS (21) on our mouse data set, with assistance from its authors (see *SI Materials and Methods* for details); however, the resulting *scaffolds* had an N50 length of only 4 kb. We therefore excluded it from further analyses.]

Human Genome. The ALLPATHS-LG assembly has an N50 contig length of 24 kb and scaffold length of 11.5 Mb. The contiguity is >4-fold longer and connectivity is >25-fold longer than obtained using the SOAP algorithm (5.5-kb contigs, 0.4-Mb scaffolds).

Importantly, these metrics approach the results from capillary-based sequencing (109-kb contigs, 17.6-Mb scaffolds).

For the ALLPATHS-LG assembly, the assembled sequence contains 91.1% of the reference genome and 95.1% of the exonic bases, and its scaffolds span all but 2.4% of the genome. In contrast, the SOAP assembly covers only 74.3% of the genome and 81.2% of exonic bases, and its scaffolds fail to capture 7% of the genome. The coverage statistics for the ALLPATHS-LG assembly are similar to those obtained for capillary-based sequencing (96.2%, 96.2%, and 2.5%). The difference in genomic coverage is largely attributable to repeat sequences (Table S3). Specifically, 68.3% of the sequence missing from the ALLPATHS-LG assembly and present in the capillary-sequencing-based assembly consists of LINE, SINE, and LTR retrotransposon elements. These elements compose 45.1% of the genome.

The ALLPATHS-LG assembly also showed good short-range and long-range accuracy, on the basis of comparison with finished reference genomes. To assess short-range accuracy, we broke the assembly into chunks of ∼1 kb and aligned these chunks to a haploid NA12878 reference (noting that heterozygous sites would count as errors half the time unless represented in the assembly). We classified each chunk as being correctly assembled if sequence differed by ≤1% and having a local assembly error if sequence differed by >1%. We assessed the base accuracy by measuring the number of errors in correctly assembled chunks.

The ALLPATHS-LG assembly has a base accuracy of 99.95% (Q33). We could not assess the base accuracy of the SOAP or

Table 3. Human and mouse assemblies

Assemblies:	Human				Mouse		
Assembly no.:	1		2	3	4	5	6
Sequence data:	Illumina		Illumina	ABI3730	Illumina	Illumina	ABI3730
Program:	ALLPATHS-LG		SOAP	Celera	ALLPATHS-LG	SOAP	ARACHNE
Completeness							
Covered, %	91.1		74.3	96.2	88.7	86.2	94.2
Captured, %	6.6		18.6	1.3	8.6	8.0	3.8
Uncaptured, %	2.3		7.0	2.5	2.7	5.7	2.0
Segmental duplication coverage, %	41.1		12.1	62.2	42.3	27.9	65.7
Exon bases covered, %	95.1		81.2	96.2	96.7	92.4	97.3
Continuity							
Contig N50, kb	24		5.5	109	16	16	25
Scaffold N50, kb	11,543		399	17,646	7,156	340	16,871
Contig accuracy							
Ambiguous bases, %	0.08		0	0	0.04	0	0
1-kb chunks vs. reference	NA12878	GRC	GRC	GRC	B6	B6	B6
(I) perfect	77.1				88.6	76.8	78.0
(II) ≤0.1% error rate	8.7				2.5	2.9	7.0
(III) ≤1%	10.2				5.7	6.1	11.7
(IV) ≤10%	3.1	3.6	5.5	3.6	2.8	11.8	2.4
(V) >10%	0.4	0.4	0.7	0.5	0.2	2.4	0.3
Base quality, from I–III	Q33				Q36	Q35	Q33
Misassembly % of 1-kb chunks, from IV–V	3.5	4.0	6.2	4.1	3.0	14.2	2.7
Scaffold accuracy							
Validity at 100 kb, %	99.1		99.5	99.7	99.0	98.8	99.1

An evaluation of human and mouse assemblies is shown. Contigs of size <1 kb were excluded from the analysis. Reference sequences are described in *SI Materials and Methods*. Assembly no.: Assemblies 1, 4, and 5 are from the data of this paper and are deposited in DDBJ/EMBL/GenBank under accession nos. AEKP00000000, AEKQ00000000, and AEKR00000000, respectively. The versions described in this paper are the first versions, AEKP01000000, AEKQ01000000, and AEKR01000000. For each ambiguity $\{x_1, \ldots, x_n\}$, we inserted x_1 into the fasta sequence and referred to $x_2, \ldots, x_n$ in a note at the locus. Assemblies 2, 3, and 6 are from refs. 3, 12, and 19). Completeness: Contigs were aligned to the reference sequence, with each contig assigned to at most one location. The covered fraction of a genome consists of the fraction of total bases in the reference (exclusive of gaps) that lie under a contig. The captured fraction consists of those bases that lie within a gap in a scaffold. All other bases are uncaptured. Exon coverage was computed from RefSeq gene annotations (http://genome.ucsc.edu/cgi-bin/hgTables). Segmental duplication coverage was computed from http://humanparalogy.gs.washington.edu/build36/oo.weild10kb.join.all.cull.xwparse and http://mouseparalogy.gs.washington.edu/She2008_download/WGAC.tab.gz. Continuity: We report the N50 sizes of contigs and scaffolds, excluding gaps in the latter case. Contig accuracy: We first report the fraction of bases labeled as ambiguous (*SI Materials and Methods*). We then divide the contigs into 1-kb chunks (as in ref. 9, which, however, used a chunk size of 10 kb). Each chunk is then aligned to the reference sequence using the Smith–Waterman algorithm, seeded on perfect 100-mer matches, to find the optimal placement, and the number of errors (mismatch plus indel bases) is computed. (Contigs having no 100-mer match were treated as novel sequence and ignored for purposes of this analysis. There was <1% of novel sequence in all cases.) The contig is then assigned to one of five mutually exclusive classes on the basis of its error rate. The percentages of chunks landing in each class are listed. Note that for assembly 1, contig accuracy was calculated with respect to two reference sequences. Base quality: Inferred Phred quality (20) of bases in chunk classes I–III. Misassembly %: Total fraction of bases in chunk classes IV–V. Scaffold accuracy: Validity at 100 kb (9): We report the probability that two 100-base sequences in the assembly, separated by 100 kb, and also present in the reference, have the same orientation and are separated by 100 kb ± 10%.

capillary-based assemblies, because we lack a finished reference sequence for the individuals.

The ALLPATHS-LG assembly contained a local assembly error in 3.5% of the 1-kb chunks (mean spacing between local assembly errors of ~29 kb). The capillary-based assembly has essentially the same local accuracy at 4.1% (mean spacing of ~24 kb), whereas the SOAP assembly has lower accuracy at 6.2% (mean spacing of ~16 kb). [When compared with a common reference sequence (Genome Reference Consortium, GRC), the respective local assembly error rates for the ALLPATHS-LG and capillary-based assemblies are 4.0 and 4.1%, respectively.]

To assess long-range accuracy, we randomly selected short sequences separated by 100 kb and determined whether their distance and orientation were essentially the same in the reference. The ALLPATHS-LG assembly had long-range accuracy of 99.1%. This accuracy is slightly lower than the 99.7% for the capillary-sequencing-based assembly, although this comparison is not completely fair because the Sanger assembly was edited using the NCBI reference sequence to correct misjoins. The long-range accuracy of the SOAP assembly was very good (99.5%).

Mouse Genome. The results were broadly similar for the mouse genome. The ALLPATHS-LG assembly has an N50 contig length of 16 kb and scaffold length of 7.2 Mb. The contig size is similar but the connectivity is >20-fold larger than that obtained from the SOAP algorithm (16-kb contigs, 0.3-Mb scaffolds). Our results again approach the published results from capillary-based sequencing (25-kb contigs, 16.9-Mb scaffolds).

The ALLPATHS-LG assembly contains 88.7% of the genome and 96.7% of exonic bases, and its scaffolds span all but 2.8% of the genome. These percentages are not as good as those obtained for capillary-based sequencing (94.2%, 97.3%, and 2.0%), with the difference again largely attributable to repeats. The coverage, however, is considerably better than that obtained with SOAP.

The ALLPATHS-LG assembly again shows good short-range and long-range accuracy. Base accuracy is 99.97% (Q36), slightly better than that of both the SOAP and the capillary-based as-

semblies. There is an assembly error in 3.0% of the 1-kb chunks, which is slightly worse than the capillary-based assembly (2.7%) and much better than the SOAP assembly (14.2%). The long-range accuracy with ALLPATHS-LG (99.0%) is similar to that for capillary-based sequencing (99.1%) and is somewhat better than that obtained with SOAP (98.8%).

Segmental Duplications. Segmental duplications present a special challenge for de novo genome assembly, even for the capillary-based assemblies. For both mouse and human, the capillary-based assemblies cover only ~60% of the segmental duplications in the reference sequence. A recent article (13) expressed concern about limitations of massively parallel sequence data because the SOAP YH human assembly covers only 12% of segmental duplication.

The ALLPATHS-LG assemblies of both human and mouse cover ~40% of the segmental duplications, showing that it is possible to use massively parallel sequence data to come much closer to the coverage obtained with capillary-based sequencing. Nonetheless, the assembly still falls short of the coverage of segmental duplications obtained with capillary-based sequencing, and even that coverage falls short of what is truly desired. Clearly, additional work is needed to represent these biologically important regions (e.g., refs 22 and 23.).

Understanding Gaps. We next sought to understand the nature of the gaps in the ALLPATHS-LG assemblies. Roughly three-quarters of the gaps are captured (that is, lie within a scaffold) whereas the remaining gaps are not spanned.

As noted above, the majority of the sequence within the gaps consists of repetitive elements (Table S3). For mouse, LINE elements are major contributors to gaps: The assembly covers only 66.7% of the genomic bases in LINE repeats, and 59.9% of gap bases are in LINE elements. For human, LINE and SINE elements together play a large role: They account for 61.9% of gap bases.

Although the majority of the bases in gaps are explained by long repeats, the majority of captured gaps are small (median 380 bases for human and 200 bases for mouse) and thus do not consist of long repeats. To better understand captured gaps, we aligned the reads from the fragment libraries to the reference sequences. For each gap, we defined the "gap neighborhood" to be the gap itself together with the 100 bases on each side of the gap. We calculated the minimum coverage m at any base within the gap neighborhood. The median value of m was 17 for human and 20 for mouse (vs. mean fragment read coverage of 38 for human and 50 for mouse). For each gap neighborhood, we also calculated the proportion of similarly sized regions in the genome having lower coverage. If the gap neighborhoods had typical coverage, the median proportion would be 50%. In fact, it was 24% for human and 15% for mouse. In short, the gaps have substantially lower than average coverage—although they have some sequence coverage (but not enough for ALLPATHS-LG to assemble across them).

Discussion

High-quality draft genome assemblies of vertebrate genomes have provided an essential foundation for comparative analysis of the human genome, as well as for studies of individual organisms. However, it cost tens of millions of dollars each to generate such genome assemblies with capillary-based sequencing, limiting the number that could be obtained. In this work, we demonstrate that it is possible to generate high-quality assemblies at a cost that is ~1,000-fold lower by using data from massively parallel sequencing (Illumina) and the algorithm presented here, ALLPATHS-LG.

Using ALLPATHS-LG, we created assemblies of the human and mouse genomes, on the basis of ~100-fold shotgun coverage in four library types. The assemblies approach the quality obtained with capillary-based sequencing: (*i*) They have good long-range connectivity, with scaffold lengths of 11.5 and 7.2 Mb, respectively. These values are within a factor of 2 of the sizes for capillary-based assemblies. (*ii*) They have good accuracy, with nucleotide accuracy of at least 99.95% and assembly accuracy comparable to capillary-based assemblies. (*iii*) They have good coverage, with ~90% coverage of the genome and 95–97% coverage of exonic sequences. The genomic coverage is lower than for capillary-based assemblies, but coverage of exonic sequence is comparable. The missing sequence consists largely of repeat elements, but also includes some recent duplications (13).

The quality of the ALLPATHS-LG assemblies is considerably better than that obtained with other recent approaches for assembling massively parallel sequencing data. For example, the human scaffolds are >25 times longer than those in recently published assemblies using data from Illumina and 454 sequencing (12, 24).

We wanted to be certain that the superior performance of the ALLPATHS-LG algorithm relative to other approaches was not due simply to differences in the quality of the underlying raw data (such as the series of innovations in library construction that we used). To test this, we assembled the same dataset for the mouse genome with SOAP; we had extensive input from its authors, who generously carried out experiments to find optimal parameters for use with their algorithm. The ALLPATHS-LG assemblies had much greater long-range connectivity and significantly higher short-range accuracy then the current version of SOAP.

A recent critique (13), citing the limitations of the SOAP human assemblies (12), has raised the concern that it may impossible to achieve high-quality de novo genome assemblies with massively parallel sequencing data. The current study indicates that considerably better assemblies can be achieved, through improvements in both algorithms and data. First, we note that the SOAP assembly for mouse is better than for human—which may be due to increases in read length and improvements to the SOAP algorithm in the interim. Second, the ALLPATHS-LG assemblies are much better than the SOAP assemblies and approach the quality of capillary-based assemblies in many respects. Nonetheless, additional improvements will be needed to produce assemblies that reach and exceed the quality of assemblies based on long reads from capillary sequencing.

Because the costs of massively parallel sequencing data are so low, we used ~100-fold coverage rather than the ~10-fold coverage typically used for capillary-based sequencing. The higher coverage compensated for the shorter length and lower per-base accuracy of the sequencing reads, as well as for coverage biases. However, the higher coverage is also potentially enabling. When we inspect defects in the ALLPATHS-LG assemblies, we find that in most cases there are enough data at the loci to manually "resolve" them. In other words, ALLPATHS-LG is not yet exploiting the full power of the data. This is a sharp contrast to the lower coverage used with capillary-based sequencing, where algorithms were frequently confronted with regions having very low amounts of data. Thus, we anticipate that an improved version of the algorithm should yield even better results.

With ALLPATHS-LG, we introduced a preliminary syntax for expressing alternatives in an assembly—for example, TTTT{T, TT}, denoting five or six Ts. This approach is superior to simply picking one choice at random, inserting Ns, or assigning a low-quality score to the bases. By explicitly capturing the alternatives, the representation makes it possible, in principle, for downstream analyses to exploit the additional information.

Computational hardware requirements are an important issue in large-genome assembly. For mammalian-sized genomes, ALLPATHS-LG requires ~3 wk on commodity shared-memory hardware. SOAP is much faster: It takes ~3 d on the same hard-

ware. Thus, for now, ALLPATHS-LG is slower but produces higher-quality assemblies. We anticipate that with algorithmic improvements it can be speeded up, although there may be a trade-off between speed and accuracy.

As sequencing costs drop, investigators will want to sequence more genomes. To facilitate this work, it is important to have essentially automated laboratory and computational processes for producing high-quality genome assemblies. In this work, we defined a practical sequencing model, on the basis of a relatively small number of libraries. Similarly, using data from this model, we designed ALLPATHS-LG to run "out of the box" using default arguments, rather than requiring "tuning" for each genome; we have done so in this work.

For widespread application, it will be valuable to optimize the sequencing model. In particular, it will be valuable to explore the effects of coverage, insert sizes, and read length; the optimal values will likely depend on rapidly changing details of the sequencing technology, including cost. As an initial step, we explored the effect of using different coverage levels. Taking the reads that align to mouse chromosome 1, we performed assemblies in which the coverage in fragment reads and/or jumping reads was reduced by either 25% or 50% (Tables S4 and S5). We found that that decreasing coverage by half for both read types reduced contig N50 from 26 to 18 kb and reduced coverage from 93.5 to 92.1%, thus suggesting that higher coverage confers an advantage in assembly quality, but that lower coverage might be appropriate in some cases. We note, however, that for de novo sequencing projects, the true genome size is rarely known accurately, and the yield of valid constructs varies considerably from library to library (Table 2). Thus, aiming for the bare minimum of coverage may not be a good strategy at this time.

With continuing improvements in de novo assembly of massively parallel sequencing data, we are optimistic that it will be possible to greatly expand the application of genome sequence analysis, including to the recently proposed goal of sequencing 10,000 vertebrates (25) and to such medical applications as reconstruction of rearranged genomes in human tumors.

Materials and Methods

A detailed description of the algorithms and laboratory methods used in this work is described in *SI Materials and Methods*. This description includes outlines of the molecular biology protocols that we used in sequencing the human and mouse genomes. The description also includes lane-by-lane information about the data that were generated using these protocols.

ACKNOWLEDGMENTS. We thank the Broad Institute Sequencing Platform for generating the data for this work. We thank Ryan Hegarty, Purnima Kompella, Robert Lintner, Na Li, Adam Navidi, Scott Nisbett, Karen Ponchner, Taryn Powers, Nicole Stange-Thomann, Scott Steelman, Diana Tabbaa, Phung Trang, and Elizabeth Upsall for molecular biology development in support of this project. We thank Michael Ross for computational work on bias that contributed to this work. We thank Jane Wilkinson and Carsten Russ for scientific project management. We thank RuiBang Luo for extensive advice on the use of the SOAPdenovo assembler; Shaun Jackman for help with ABySS; Alexej Abyzov, Mark Gerstein, and Mark DePristo for sharing NA12878 reference data; and Leslie Gaffney for help with figures. This work was carried out with the aid of federal funds provided by the National Institutes of Health, Department of Health and Human Services, via Grants U54HG003067 and R01HG003474 through the National Human Genome Research Institute and Contract HHSN2722009000018C through the National Institute of Allergy and Infectious Diseases.

1. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431:931–945.
2. Church DM, et al.; Mouse Genome Sequencing Consortium (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol* 7:e1000112.
3. Waterston RH, et al.; Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562.
4. Lindblad-Toh K, et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438:803–819.
5. Warren WC, et al. (2008) Genome analysis of the platypus reveals unique signatures of evolution. *Nature* 453:175–183.
6. Bentley DR, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59.
7. McKernan K, Blanchard A, Kotler L, Costa G (2006) Reagents, methods, and libraries for bead-based sequencing. Patent WO/2006/084132.
8. Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18:821–829.
9. Maccallum I, et al. (2009) ALLPATHS 2: Small genomes assembled accurately and with high continuity from short paired reads. *Genome Biol* 10:R103.
10. Nowrousian M, et al. (2010) De novo assembly of a 40 Mb eukaryotic genome from short sequence reads: Sordaria macrospora, a model organism for fungal morphogenesis. *PLoS Genet* 6:e1000891.
11. Li R, et al. (2010) The sequence and de novo assembly of the giant panda genome. *Nature* 463:311–317.
12. Li R, et al. (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* 20:265–272.
13. Alkan C, Sajjadian S, Eichler EE (2010) Limitations of next-generation genome sequence assembly. *Nat Methods*, 10.1038/nmeth.1527.
14. Collins FS, Weissman SM (1984) Directional cloning of DNA fragments at a large distance from an initial probe: A circularization method. *Proc Natl Acad Sci USA* 81:6812–6816.
15. The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.
16. Butler J, et al. (2008) ALLPATHS: De novo assembly of whole-genome shotgun microreads. *Genome Res* 18:810–820.
17. Pevzner PA, Tang H, Waterman MS (2001) An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci USA* 98:9748–9753.
18. Kozarewa I, et al. (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* 6:291–295.
19. Levy S, et al. (2007) The diploid genome sequence of an individual human. *PLoS Biol* 5:e254.
20. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8:186–194.
21. Simpson JT, et al. (2009) ABySS: A parallel assembler for short read sequence data. *Genome Res* 19:1117–1123.
22. She X, Cheng Z, Zöllner S, Church DM, Eichler EE (2008) Mouse segmental duplication and copy number variation. *Nat Genet* 40:909–914.
23. Marques-Bonet T, et al. (2009) A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* 457:877–881.
24. Schuster SC, et al. (2010) Complete Khoisan and Bantu genomes from southern Africa. *Nature* 463:943–947.
25. Hayden EC (2009) 10,000 genomes to come. *Nature* 462:21.