Genome Assembly, Rearrangement, and Repeats

Haixu Tang[†]

School of Informatics, Center for Genomics and Bioinformatics, Indiana University, Bloomington, Indiana 47408

Received November 2, 2006

Contents

1. Introduction

Genomes evolve at different scales. At a small scale, a single nucleotide may be substituted, deleted, or inserted at a specific position in the genome. At the chromosomal scale, segments of genetic material may be acquired, removed, duplicated, and/or rearranged due to various mechanisms.

Eukaryotic genomes are usually much larger than prokaryotic genomes, and often carry many *repeats*, that is, DNA sequences appearing multiple times as similar copies in the genome. A typical example is the human genome, in which repeats constitute more than half of the whole genome. Genome rearrangements, which alter the chromosomal architecture during evolution, can be observed when comparing the order of genetic markers (e.g., genes) in two genomes sharing a common ancestor. Each genome rearrangement event disrupts homologous segments in two genomes, and creates *breakpoints* between them. Evidently, repeats are frequently observed around the breakpoint regions and, thus, are hypothesized to be one of the driving forces of genome rearrangement.

The richness of repeats in eukaryotes poses a great challenge for fragment assembly when sequencing these genomes. Although a few strategies were proposed to address this issue, several kinds of misassemblies may still exist even in the published, but not yet completely finished, genome sequences, especially for the ones sequenced using the whole genome shotgun (WGS) approach. Some repeats may be missed and left as gaps. Some repeats may be collapsed, resulting in a smaller number of copies and inaccurate sequence for each copy. Finally, assemblers may be confused by

Dr. Haixu Tang is an assistant professor in the School of Informatics, and an affiliated faculty of Center for Genomics and Bioinformatics (CGB) at Indiana University, Bloomington, since 2004. Professor Tang received his Ph.D. in Molecular Biology from the Shanghai Institute of Biochemistry in 1998; between 1999 and 2001, he was a post-Doc associate in the Department of Mathematics at the University of Southern California; between 2001 and 2004, he was an assiatant project scientist in the Department of Computer Science and Engineering, University of California, San Diego.

[†] Corresponding author: Haixu Tang, School of Informatics, Indiana University, 901 E. 10th Street, Bloomington, IN 47408, Tel, 812-856-1859; fax, 812-856-1995; e-mail, hatang@indiana.edu.

10.1021/cr0683008 CCC: $65.00 © 2007 American Chemical Society
Published on Web 07/18/2007

Table 1. Published Eukaryotic Genomes as of October, 2006[a]

group	species	common name	size (Mbp)	approximate gene no.
Fungus	*Aspergillus fumigatus*[4]	Filamentous fungus	29.4	10000
Fungus	*Candida glabrata*[5]	Hemiascomycete yeast	12.4	6000
Fungus	*Cryptococcus neoformans*[6]	Basidiomycetous yeast	20	6500
Fungus	*Candida albicans*[7]	fungal pathogen	12	7500
Fungus	*Cryptosporidium hominis*[8]	Intracellular parasite	9.2	4000
Fungus	*Cryptosporidium parvum*[9]	Intracellular parasite	9.1	3800
Fungus	*Debaryomyces hansenii*[5]	Debaryomyces yeast	12	6000
Fungus	*Encephalitozoon cuniculi*[10]	Intracellular parasite	2.9	2000
Fungus	*Ashbya gossypii*[11]	Filamentous ascomycete	9.2	4700
Fungus	*Kluyveromyces waltii*[12]	-	10.9	5230
Fungus	*Kluyveromyces lactis*[5]	-	10.7	6000
Fungus	*Magnaporthe grisea*[13]	Rice blast fungus	38.8	11000
Fungus	*Neurospora crassa*[14]	Filamentous fungi	40	10000
Fungus	*Phanerochaete chrysosporium*[15]	Lignocellulose degrading fungus	30	11700
Fungus	*Saccharomyces cerevisiae*[16]	Budding yeast	12	6000
Fungus	*Schizosaccharomyces pombe*[17]	Fission yeast	13.8	4800
Fungus	*Yarrowia lipolytica*[5]	-	20.5	6000
Protist	*Cyanidioschyzon merolae*[18]	Red alga	16	5331
Protist	*Entamoeba histolytica*[19]	Intracellular parasite	23.8	10000
Protist	*Dictyostelium discoideum*[20]	Social amoeba	33.8	12500
Protist	*Leishmania major*[21]	Kinetoplastid parasite	5.4	8300
Protist	*Plasmodium falciparum*[22]	Malaria parasite	23	5300
Protist	*Thalassiosira pseudonana*[23]	Diatom	34	11000
Protist	*Theileria parva*[24]	Intracellular parasite	8.3	4035
Protist	*Trypanosoma brucei*[25]	African trypanosome	26	9068
Protist	*Trypanosoma cruzi*[26]	African trypanosome	26	9068
Roundworm	*Caenorhabditis elegans*[27]	Nematode worm	97	19000
Roundworm	*Caenorhabditis briggsae*[28]	Nematode worm	104	19500
Insect	*Drosophila melanogaster*[29]	Fruit fly	137	14000
Insect	*Bombyx mori*[30]	Silkworm	429	18510
Insect	*Anopheles gambiae*[31]	Mosquito	278	14000
Urochordate	*Ciona intestinalis*[32]	Sea squirt	160	16000
Fish	*Takifugu rubripes*[33]	Fugu fish	365	26700
Fish	*Tetraodon nigroviridis*[34]	Puffer fish	342	28000
Bird	*Gallus gallus*[35]	Chicken	1200	20000
Mammal	*Canis familiariz*[36]	Dog	2400	19300
Mammal	*Homo sapiens*[37]	Human	2900	30000
Mammal	*Mus musculus*[38]	Mouse	2500	30000
Mammal	*Pan troglodytes*[39]	Chimpanzee	3100	30000
Mammal	*Rattus norvegicus*[40]	Rat	2750	28000
Plant	*Arabidopsis thaliana*[41]	Mustard weed	120	25000
Plant	*Oryza sativa*[42]	Rice	466	37500
Plant	*Populus trichocarpa*[43]	Black Cottonwood	550	45000

[a] Sequenced euchromatin is used to estimate the genome size in these genomes. Gene numbers are estimated based on the current annotation.

repeats and misjoin nonadjacent genomic fragments together. Therefore, the identification of repeats and genome rearrangements should be conducted with caution.

The recent availability of whole genome sequences from many eukaryotes has set up a playground for the comprehensive *in silico* analysis of both types of genomic variations across different species.[2] According to the statistics from National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi), there have been more than 300 ongoing or complete eukaryotic genome sequencing projects, including 21 complete genomes, 112 assembled genomes, and 181 genomes in progress. Table 1 lists the published eukaryotic genome sequences up to date. These organisms (17 fungal, 9 protists, 3 plants, and 14 animals, including 2 nematodes, 3 insects, 1 urochordate, 2 fishes, 1 bird, and 5 mammals) represent a broad range of evolutionary diversity, even though the strategy for selecting representative species that optimally balance the cost and their medical, agricultural, and evolutionary importance is not obvious.[3] To respond to the challenge of analyzing these massive genome sequence data, a new discipline named *comparative genomics* emerges, in which computational approaches have been adopted to study the function and the evolutionary processes acting on genomes. Many computational methods have been proposed and applied to the whole genome data in an attempt to characterize evolutionary difference between closely, as well as distantly, related genomes. In this review, we will describe the genomic sequence analysis methods for identifying the genomic variations at the chromosomal scale, such as transposition, segmental duplication, and genome rearrangements. We will also discuss the challenges of eukaryotic genome assembly caused by frequent genome rearrangements.

2. Classification of Repeats in Eukaryotic Genomes

As a seemingly obvious implication of the role of DNA as the genetic material, the genome size (i.e., the total genetic content) should be positively correlated to the complexity of the organism. Many counterexamples, however, were found in a broad-scale survey of animal genome size.[44] For example, the human genome is 200 times larger than the genome of yeast *Saccharomyces cerevisae*, but 200 times

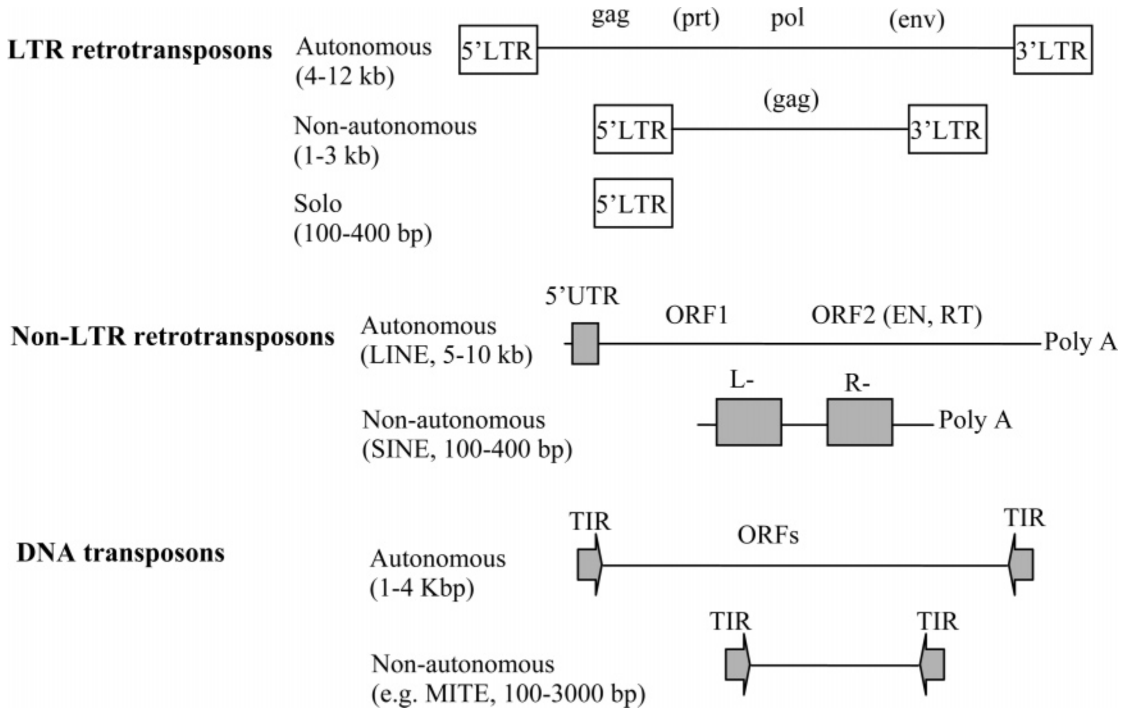

Figure 1. Classification of TE-derived repeats in eukaryotic genomes based on their transposition/duplication mechanisms.

smaller than the genome of a small creature, *Amoeba dubia*.[45] This confusion, known as the C-value paradox,[46] is now fully resolved by the observation that the majority of each eukaryotic genomes is composed of non-coding sequences, including a large quantity of repetitive sequences. For example, only less than 5% of human genome is coding sequence, whereas repeats constitute more than 50% of the genome![37] Some eukaryotic organisms possess special genome defense mechanisms that can prevent segmental duplications. In the fungi genomes, the mechanism called repeat-induced point mutation (RIP) can efficiently detect, mutate, and thus eliminate the repeated segments in the genome.[47] As a result, the repeat content in a eukaryotic genome also depends on the efficiency of the defensive mechanisms. For example, the analysis of the *Neurospora crassa* genome sequence shows an unusually low abundance of highly similar segments, which may be a result of its efficient genome defensive mechanism as RIP.[14]

2.1. Tandem Repeats versus Interspersed Repeats

According to their origin, repeats generally fall into five classes:[37] (1) microsatellites, that is, tandem repeats with short repeating units (2−5 bases in length), such as $(A)_n$, $(TA)_n$, or $(CGC)_n$; (2) minisatellites, that is, tandem repeats with long repeating units (10−100 base in length); (3) repeats derived from transposable elements; (4) low copy repeats derived from segmental duplications; and (5) processed pseudogenes, that is, retroposed copies of transcribed coding genes. Unlike the first two classes of tandem repeats, the copies of the remaining three classes of repeats can be present at different locations across a whole eukaryotic genome and, thus, are often called *interspersed repeats*. Tandem repeats are, in general, underrepresented in even complete genome sequence because of their high polymorphisms. In the rest of this paper, we will focus on two major classes of repeats in eukaryotic genomes, those derived from transposable elements and those derived from segmental duplications.

2.2. Transposable Elements

Transposable elements (TEs), also called *transposons* or *mobile genetic elements*, are known to be the cause of most repeats in the human genome.[48] As a conservative estimation, 45% of the human genome has been recognized to belong to this class. These TE-derived repeats fall into three categories (Figure 1): (1) long terminal repeat (LTR) retrotransposons, also called retrovirus-like elements; (2) non-LTR retrotransposons, including long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs); and (3) DNA transposons. The first two types of repeats are sometimes also called Class I (RNA-mediated) elements, where the third type is called Class II (DNA-mediated) elements.[49] Each class of repeats consists of both *autonomous* and *non-autonomous* elements. Autonomous elements are intact elements so that they encode the full set of proteins (enzymes) that are sufficient to move themselves. On the other hand, non-autonomous elements contain only partial sequence of an intact element, thus, rely on another active intact element of the same type to move them.

Similar to endogenous retroviruses (ENVs), LTR retrotransposons consist of a few overlapping open reading frames (ORFs) encoding proteins including group-specific antigen (gag), protease (prt), and polymerase (pol). The expression of these genes is ensured by promoter activities within the two long terminal repeats (LTRs, 100−400 bases in length) flanking the intact element. The pol protein contains domains such as retrotranscriptase (RT) and integrase (IN), and can retrotranspose the element into the genome in a “copy-and-paste” fashion. Some LTR retroelements contain an envelope-like gene, which implies that they may originate from the common ancestor as ENVs. LINEs (non-LTR retrotransposons) usually contain two non-overlapping ORFs, encoding protein domains with RT and endonuclease (EN) activities. A short 5′ UTR in front of the first ORF shows promoter activity and ensures the expression of these genes. SINEs are short (100−400 bases in length) and do not encode proteins; thus, they can only be moved by other active LINE

Table 2. Repeat Content in Eukaryotic Genomes[a]

		TR-derived repeats (fraction of genome %)				
species	sequencing method	LINE	SINE	DNA transposon	LTR retrotransposon	segmental duplications (fraction %)
C. elegans	hierarchical	0.3	0.1	5.3	<0.1	?
C. briggsae	WGS	22.4[b]				
D. discoideum[69]	hierarchical	4.6		1.4	4.4	?
Fruit fly[70]	Hybrid	0.5	0.3	1.5	2.6	?
Silkworm	WGS	6.7	1.4	1.7	11.1	?
Mosquito	WGS	0.2	3.5	1.1	11.2	?
Fugu fish	WGS	1.2	<0.1	0.5	0.8	?
Puffer fish	WGS	<0.1	0	<0.1	<0.1	?
Chicken	WGS	6.5	<0.1	0.8	1.3	∼2.8
Dog	WGS	5.4	6.7	0.13	0.17	∼0.6
Human	hierarchical	20.4	13.1	2.8	8.3	3.2
Mouse	WGS	19.2	8.2	0.9	9.9	1.2
Rat	Hybrid	23.1	7.1	0.8	9.0	2.9
Chimpanzee	WGS	23.1	7.0	0.8	9.0	2.5
Mustard weed	hierarchical	0.5	0.5	5.1	4.8	58[d]
Rice[c]	WGS	1.19	0.09	2.8	9.3	?
Rice[c]	hierarchical	1.12	0.06	13.0	18.1	60[d]

[a] Estimates are based on the original reports of the respective genomes (see Table 1 for references) unless noted otherwise. Analysis of segmental duplications is unavailable for lower eukaryotic species. [b] This number is estimated from a *de novo* repeat finding analysis by RECON (see section 3 for details) and, thus, accounts for all kinds of interspersed repeats, including segmental duplications. [c] Two different rice strains were sequenced simultaneously: one used the whole genome shotgun (WGS) strategy,[42] and the other used a map-based hierarchical sequencing approach.[71] The data from both reports were presented here for the comparison. [d] Plant genomes often consist of plenty of medium-age segmental duplications. Therefore, these duplicated segments are defined by pairwise alignments longer than 1000 bp with at least 50% identity. The results for the vertebrate genomes will not change if applying this criteria.

elements. The most common SINE element in the human genome, the Alu element, contains two GC-rich monomers (L- and R-), and ends with a poly-A tail. DNA transposons are flanked by two short terminal inverted repeats (TIRs, 2−10 bases in length) and contain a long ORF encoding protein domains with DNA binding and transposase (TR) activities. DNA transposons can be transposed in different modes, even though their duplications all occur during DNA replication.[50] Most eukaryotic DNA transposons follow the classical "cut-and-paste" mechanism, in which the element is cleaved and transferred to a new location by the TR. Some eukaryotic elements are transposed by a rolling circle (RC) mechanism, similar to the ones first discovered in prokaryotes.[51] Many non-autonomous DNA transposons are derived from the intact elements by internal deletions. For instance, Miniature Inverted Repeat Transposable elements (MITEs) are a collection of short (100−500 bp in length) non-autonomous DNA transposons without encoding a TR gene.[52]

TEs are generally assumed to be originated from various bacterial elements. In a recently proposed probable evolutionary scenario, DNA transposons and non-LTR retrotranposons may be derived from bacterial transposons and retroelements (e.g., group II introns),[53,54] respectively, whereas the LTR retrotransposons may be derived from the fusion of these two types of elements.[55]

2.3. Segmental Duplications

Although a majority of interspersed repeats in eukaryotic genomes are derived from transposable elements, many other repeats are derived from a different mechanism. Segmental duplications, sometimes referred to as low copy repeats (LCRs), are typically long (>1 kb), similar copies of a fragment of genomic sequence that are present in at least two locations in the genome. Segmental duplications originate from the duplicative transpositions of a piece of chromosomal DNA. Interestingly, these duplicated segments have no known distinguishable sequence features from the rest of the genome.[56,57] Therefore, to date, the molecular mechanism of segmental duplication is not fully understood, although it is often speculated that it may be related to mistakes in DNA replication. Many of the duplicated segments in mammalian genome are highly similar (with >90% sequence identity) and, thus, are often referred to as recent duplications. It is estimated that about 5% of the human genome is composed of recent segmental duplications that occurred in the last 35 millions years,[37,58] as compared to only 2.9% and 1.2% of rat and mouse (Table 2).[59,60] The distribution of these duplicated segments in mammalian genome is highly nonuniform, with some regions (e.g., Y chromosome, telomeres, centromeres, etc.) containing more duplications than others.[57,59−61]

Segmental duplications represent an important feature of mammalian evolution. On the one hand, highly homologous sequences may cause misalignment during meiosis and, thus, facilitate large-scale chromosomal rearrangements, such as pericentric inversions and segmental deletions. On the other hand, numerous paralogous genes in mammals are created through segmental duplications, especially those recombined mosaic genes resulted from joining promoter regions and exons of different ancestor genes. Many of them have developed new biological functions.[62,63] Segmental duplication regions have also shown high genic variations in human population, among which more and more have been linked to substantial phenotypic variations, in particular, common diseases.[64]

Because of the nonuniformity of the targets of segmental duplications, a duplicated segment may be inserted into another previously duplicated segment. An attractive hypothesis regarding the evolutionary history of these segmental duplications is that they emerge through a two-step process, involving the initial duplicative transposition of specific segments to focal regions within the genome followed by the duplication of multiple units in concert among these collector regions.[57,61,65] In this scenario, the segment at the original location, called the *ancestral duplication unit* or

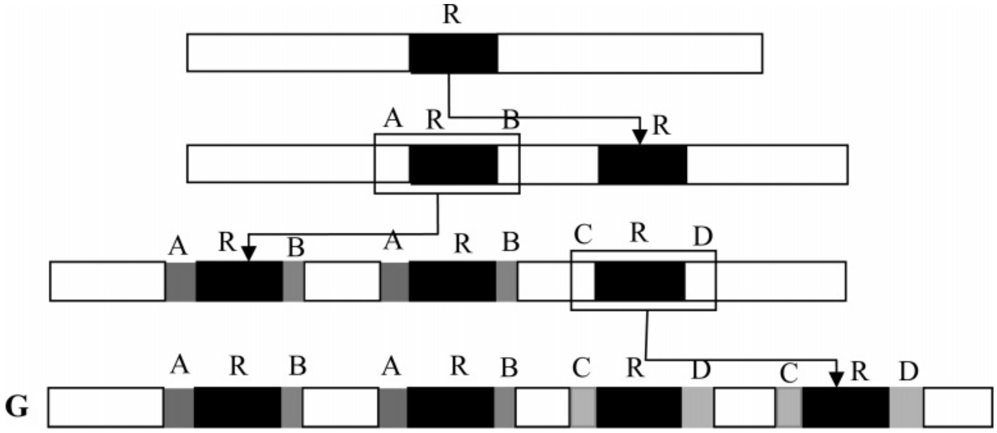

Figure 2. An imaginary example of complex segmental duplications. Three consecutive duplications of one ancestral duplication unit R. In the second duplication, R is duplicated together with its flanking segments A and B. In the third duplication, R is duplicated together with its flanking segments C and D. In the resulting genome G, in addition to the four copies of repeat R, there are also two copies of repeats A, B, C, and D that form the complicated mosaic structure.

duplicon,[57] jumps to various chromosomal locations and brings the flanking chromosomal materials over, resulting in repetitive copies in the genome. The identification of all ancestral duplication units was formulated as an open problem by Eichler and colleagues in the 1990s. Since then, ancestral duplication units for some human loci have been systematically identified by combining different sources of evidence, including the comparative study of the DNA sequences in the duplicated regions in the same genome (e.g., human), and their counterparts in another distant species (e.g., mouse), as well as the experimental detection of the presence or absence of the duplicated regions in a range of close species.[61] Figure 2 uses an imaginary example to illustrate the potential complexity of the evolutionary scenario of segmental duplications. It shows a simple duplication history of a single duplication unit *R*. During each duplication, the unit's flanking regions (A, B, C, D) may also be brought together with R to be duplicatively transposed to a different location in the genome. As a result, not only R, but also A, B, C, and D appear to be repeats in current genomics sequence G. And it is not easy to delimit the boundaries between these duplication units.

2.4. Repeat Content in Eukaryotic Genomes

The high similarity between recently duplicated segments poses great challenges to genome assembly (see section 4 for details). Consequently, segmental duplications may be artificially underrepresented in shotgun assembled genomes. It has been shown that long (>15 kb) and nearly identical (>97%) repeats cannot be adequately resolved by whole genome shotgun (WGS) assembly, and the resulting genome size may be significantly reduced.[66] When analyzing the repeat content of eukaryotic genomes, it should be taken into consideration how this genome was sequenced and assembled, especially when comparing the content of recent segmental duplications in multiple genomes.

Table 2 summarizes the repeat content through a rapid homologue-based computational screening in the assembled genome sequences (see section 3 for details of the methods). We note that these estimations present a low bound of repeat content in these genomes. The accuracy of these estimations depends on various factors, for example, the coverage of complete genomic sequence, the approach used for genome sequencing (hierarchical or WGS, see section 4), and the structural details that have been known for the repeat families in a specific genome. Novel repeat families, for example, transposons, are continuously reported in various genomes.[67] Although TEs exist universally in all eukaryotic genomes, their densities are very different from one genome to another. Generally, large genomes (e.g., mammalian genomes) contain more TE-derived repeats than smaller ones, indicating a significant contribution of TEs' activities to the expansion of these genomes. Furthermore, the distribution of different classes of TE-derived repeats may be very different across genomes. For examples, non-LTR retrotransposons derive a majority of repeats (LINEs and SINEs) in higher animals, whereas the repeats in plants are mainly derived from LTR retrotransposons and DNA transposons; SINEs are present in high density in all mammalian genomes but are very rare in the chicken genome; the dog genome has the least portion of TE-derived repeats among all sequenced mammalian genomes, but these elements consist of a large number of polymorphic sites (e.g., 8% among 87 000 young SINE elements).[36] Recent segmental duplications are sometimes difficult to place within whole genome shotgun assemblies, and special attention should be paid when drawing conclusions from these analyses.[68] For instance, in the chicken genome, a self-genome comparison reveals that 11% (∼123 Mb) of the genome sequence in pairwise alignment is longer than 1 kb with higher than 90% sequence identity. In an independent test, however, only 26% (32.3 out of 123 Mb) of them were confirmed.[35]

3. Computational Methods To Identify Repeats in Genomic Sequences

3.1. Tandem Repeat Finders

A number of programs can be used to find tandem repeats. Some of them, like REPuter,[72] can be used to identify both tandem and interspersed repeats, and will be discussed in the following sections. The other programs are designed to identify tandem repeats only. Tandem Repeats Finder (TRF),[73] STRING,[74] and Mreps[75] adopt an *ab initio* approach and report all types of tandem repeats within the input genome, although their definitions of a tandem repeat are slightly different. These programs are rapid enough for whole genome sequence analysis and allow a user-defined "fuzziness" (mismatches or gaps) among the tandem copies of repeating units. TROLL,[76] designed mainly for identifying microsatellites, adopts a similarity-based strategy, which limits the search to an appropriate dictionary of repeating units.

3.2. Repbase Update and RepeatMasker

Similar to the programs for tandem repeat finding, the programs for identifying interspersed repeats adopt two different approaches: similarity searching and *ab initio* repeat finding. The similarity searching approach is commonly used in eukaryotic genome annotation, which requires a library of repeat sequences, and a fast sequence comparison program. Repbase Update (RU, http://www.girinst.org/repbase/update/index.html)[77] is a well-maintained eukaryotic repeat (mainly TE-derived repeat) sequence database, developed since 1990. RU represents large repeat families/subfamilies by their consensus sequences and represents small repeat families by sequence examples. It contains many TE consensus sequences that are not deposited anywhere else. RU is used in many genome sequencing projects for masking repeats prior to fragment assembly (see section 4.1.1 for details), and for repeat annotation in complete genomes. RepeatMasker (http://www.repeatmasker.org/) is a typically used software tool for screening interspersed repeats (as well as low-complexity regions) in genomic sequences from a library such as RU. Other programs with similar functions include CENSOR[78] and MaskerAid.[79]

3.3. Other Similarity-Based Repeat Finders

Although RepeatMasker is popular in repeat annotation, its sensitivity relies on good representative sequences in the repeat library, since its screening is based on the comparison of DNA sequences. Precise repeat annotation requires more sensitive repeat finders, using the sequence comparison of protein domains that are encoded in TE-derived repeats. It has been shown that tblastx,[80] a program in BLAST family that compares two DNA sequences by comparing their six-frame translations, can identify many more TE families than RepeatMasker in the *Drosophila melanogaster* and *Anopheles gambiae* genomes.[81] Additional information, for example, the difference in base composition between TEs and the rest of the genome, can be incorporated in the repeat searching to further increase the sensitivity. TE-HMM used Hidden Markov Models (HMM) to represent TE families, accounting for the word frequency and the heterogeneity between coding and non-coding parts within TE sequences.[82] Recently, a computational pipeline that combines several similarity-based repeat annotation programs was developed.[70] This combined evidence approach has been shown to be able to identify more repeat elements than any single program, which implies that these methods may be complementary in repeat finding.

Considering the high similarity between two flanking long terminal repeats of LTR retrotransposons, programs are specifically designed to identify this class of repeats. LTR_STRUC is a program that searches for structural features of LTR elements to automatically identify these elements in genome sequences.[83] Its applications in insect,[84] animal,[85] and plant[86] genome annotation demonstrate its advantages over the other established similarity-based methods when there is low sequence homology between the known and query elements. Along the same lines, a recent algorithmic improvement has facilitated fast and accurate LTR element screening in whole mammalian genomes.[87]

3.4. *De Novo* Repeat Classification by Whole Genome Sequence Analysis

Similarity-based repeat annotation programs can identify repeats only if they are similar to one that has been previously identified and deposited in the library. Therefore, many novel repeat families may escape current similarity-based genome annotations and are waiting to be discovered. *De novo* repeat classification aims at identifying repeats from genomic sequences without relying on any previous knowledge; thus, it is capable of identifying new repeat elements. The conventional *de novo* repeat finding methods (e.g., REPuter[72] and RepeatFinder) perform a self-alignment of a target DNA sequence and represent the repeats in a pairwise fashion. It is not straightforward to classify these pairs of repetitive DNA segments into repeat families, because some copies of repeats may exist in a fragmented form, and some others are located so close that they can be identified together as a single segment. This issue was first addressed by Volfovsky et al.,[88] and Bao and Eddy[89] independently, resulting in two pioneer programs, RepeatFinder and RECON, which applied heuristic rules to classify the identified pairs of similar DNA sequences into repeat families, and then define their boundaries. These methods can be modified to a rigorous formulation using graph theory.[90] Nevertheless, these programs have been successfully used to identify novel repeat families, such as TE in plants.[67]

A major difficulty when applying RECON-like algorithms in analyzing large mammalian genomes is that an extremely large number of segment pairs may be discovered in the genome self-alignment step, due to very large repeat families in these genomes. For example, there are about 10^6 Alu repeats in the human genome;[91] all-against-all pairwise alignment will result in 10^{12} pairs of segments to be reported and analyzed. This issue can be addressed in two different ways. One method, implemented in the program PILAR,[92] utilizes the characteristic patterns of pairwise alignments for certain classes of repeats to filter redundant segment pairs, and then cluster them into repeat families. The other method, implemented in the program RepeatScout,[93] completely avoids the pairwise alignment step; instead, it adopts a much more efficient method, which progressively extends each seed (i.e., high-frequency substrings with the same length l) to a longer consensus sequence in a greedy fashion, following the fit alignment score between the consensus sequence and the repeat occurrences in the genome. RepeatScout can reconstruct the consensus sequence for each repeat family with approximately the same accuracy as RECON, but with a magnitude faster speed.

An emerging method for *de novo* repeat identification that is different from those described above makes use of the comparison of closely related genomes. For instance, in the comparative analysis of human and chimpanzee genomes, TE insertions specific to one lineage can be identified by examining the gap regions in their pairwise alignment.[94] This method has recently been systematically tested on four closely related *Drosophila* genomes.[95] Since it studies the TE-derived repeats in multiple genomes simultaneously, the history of TE insertions and their relationship with the evolution of the host genome may be revealed at the same time.

3.5. Identification of Segmental Duplications

Segmental duplication commonly covers a significant fraction of higher vertebrate genomes (Table 2). Since duplicated segments are usually in low copies in the genome and do not exhibit sequence characteristic patterns, they cannot be identified through similarity search. The most frequently used strategy for detecting segmental duplication is based on a self-alignment of a whole genome sequence.[58] To avoid the large number of alignments from the high-copy TE-derived repeats, their identification and removal prior to whole genome alignment is required. A computational pipeline developed by Eichler and colleagues was used to analyze higher vertebrate genomes,[35,58−60] and many segmental duplications have been identified. Using these results, Eichler and colleagues discovered some important features of segmental duplications, such as the relatively large fraction of duplicated region and the nonuniform distribution of the segmental duplications. Many novel gene families were also observed to be overlapping with recent segmental duplications and, thus, may be created through this mechanism.

The accuracy of the self-alignment-based method to detect duplicated regions, however, depends on the quality of genome assembly. It has been shown that, for genomes assembled using the whole genome shotgun (WGS) approach, those large and highly identical duplications are often collapsed. Consequently, the resulting genome size is significantly reduced, and many duplicated genes embedded in these duplications are missed.[66] To address this issue, a different strategy was proposed, which first maps the WGS reads onto the assembled genome and then detects duplicated regions and their copy numbers within the genome based on the read coverage of these regions.[96] This strategy can also be extended to identify segmental duplications in one genome (e.g., chimpanzee) using another well-assembled genome as the mapping template.[68]

The methods described above can detect duplicated regions, but do not classify them into repeat families. Because of the interacting duplication units caused by the complex duplication history (e.g., Figure 2), the existing repeat classification algorithms (section 3.4) are not applicable to classify duplicated segments. It is suggested that a mosaic structure of duplication units can represent well the complex duplication.[65] It involves, however, extensive manual integration of results from sequence comparison and experimental verification, to reveal the mosaic structure of segmental duplications, even for the short human chromosome 22.[97] A repeat graph approach has been recently proposed for automatically identifying duplication units and revealing their mosaic structures.[90] Interestingly, the repeat graph derived from the whole genome, in which duplication units are represented by edges and each duplicated region is represented a path, can be computed equivalently from sequence fragments of genome. This observation establishes a connection between the repeat classification and the repeat resolution in fragment assembly (see section 4 for details).

3.6. Domains in TE-Derived Repeats

Accumulating results indicate that TE-derived repeats may show a mosaic structure similar to that of segmental duplications. One cause may be the active recombination between various transposable elements.[98] A graph-theoretical approach similar to the repeat graph approach has been applied to sequences of several TE-derived repeat families, and has revealed novel shared *repeat domains*.[99] This implies that repeat elements may evolve in a way analogous to proteins, through not only point mutation, but also domain recombination.

4. Repeats and Genome Fragment Assembly

4.1. Repeat-Induced Genome Misassembly

Sanger's method,[100] used by the current DNA sequencing machines, can accurately acquire a DNA sequence of about 600−1000 bases. To sequence long DNA molecules, a shotgun strategy is often adopted.[101] The whole procedure starts from breaking the target DNA molecules into overlapping fragments, which are each processed by a DNA sequencing machine, and output to a short piece of DNA sequence (called a *read*). The reads need to be put together based on their overlaps to reconstruct the original sequence of the target DNA, by an automated computer program called an *assembler*. This strategy was successfully applied to a whole bacterial genome for the first time in 1995, leading to the 1.8 Mbp sequence of the whole genome of the bacterium *Haemophilus influenzae*.[102]

Fragment assemblers mainly follow the "overlap-layout-consensus" paradigm.[103−109] The overlaps between all possible pairs of reads are first detected in the overlap step, based on which of the reads are then aligned together in the layout step. Finally, the consensus sequence is constructed by taking the major nucleotide at each aligned position. In theory, it is not much more difficult to assemble long sequences than the short ones, as long as a sufficient number of reads are sequenced. The number of *contigs* (i.e., the group of overlapping reads) for a given read *coverage* (i.e., the ratio between the total length of reads and the length of genome) can be estimated using Lander-Waterman's probabilistic model.[110] A not-too-high read coverage (e.g., 9−12) seems sufficient to assemble entire vertebrate genomes.

However, repeats pose a major challenge to fragment assembly and genome sequencing.[111] Since different copies of repeats are very similar (sometimes identical) to each other, overlapping reads detected by the assembler may not be from the overlapping regions, but from two copies of repeats in the real DNA molecule. These *pseudo-overlaps* may cause two types of mistakes in the assembled sequence: base-calling errors and false rearrangements. When reads in multiple nearly identical repeats are mistakenly placed in the wrong repeat copy, nucleotides will be misassigned in the consensus sequence, giving *base-calling errors*.[112] And pseudo-overlaps between repeats may create large-scale rearrangements of DNA segments in the assembly (Figure 3).

The high repeat content in eukaryotic genome (Table 1) raises the question of whether the shotgun strategy can be applied to whole genome sequencing of eukaryotic genomes (e.g., the human genome), or if it can be efficiently conducted.[113,114] Conventional genome sequencing, including the worldwide Human Genome Project (HGP), adopted a *hierarchical* shotgun strategy.[37] It involves the construction and selection of overlapping large-insert clones called BAC (Bacterial Artificial Chromosome), each carrying an inserted human genome fragment (typically 100−200 kb), and the shotgun sequencing of every clone. Assembling the shotgun reads from individual clones is a relatively easy task because the interspersed repeat copies are often split into different clones and, thus, no longer produce pseudo-overlaps. The genome assembly is finally obtained by mapping the sequenced clones to obtain a tiling path.[115] Although it alleviates the challenge for genome assembly, the hierarchical strategy requires intensive efforts in constructing and selecting clones. In contrast, whole genome shotgun (WGS)

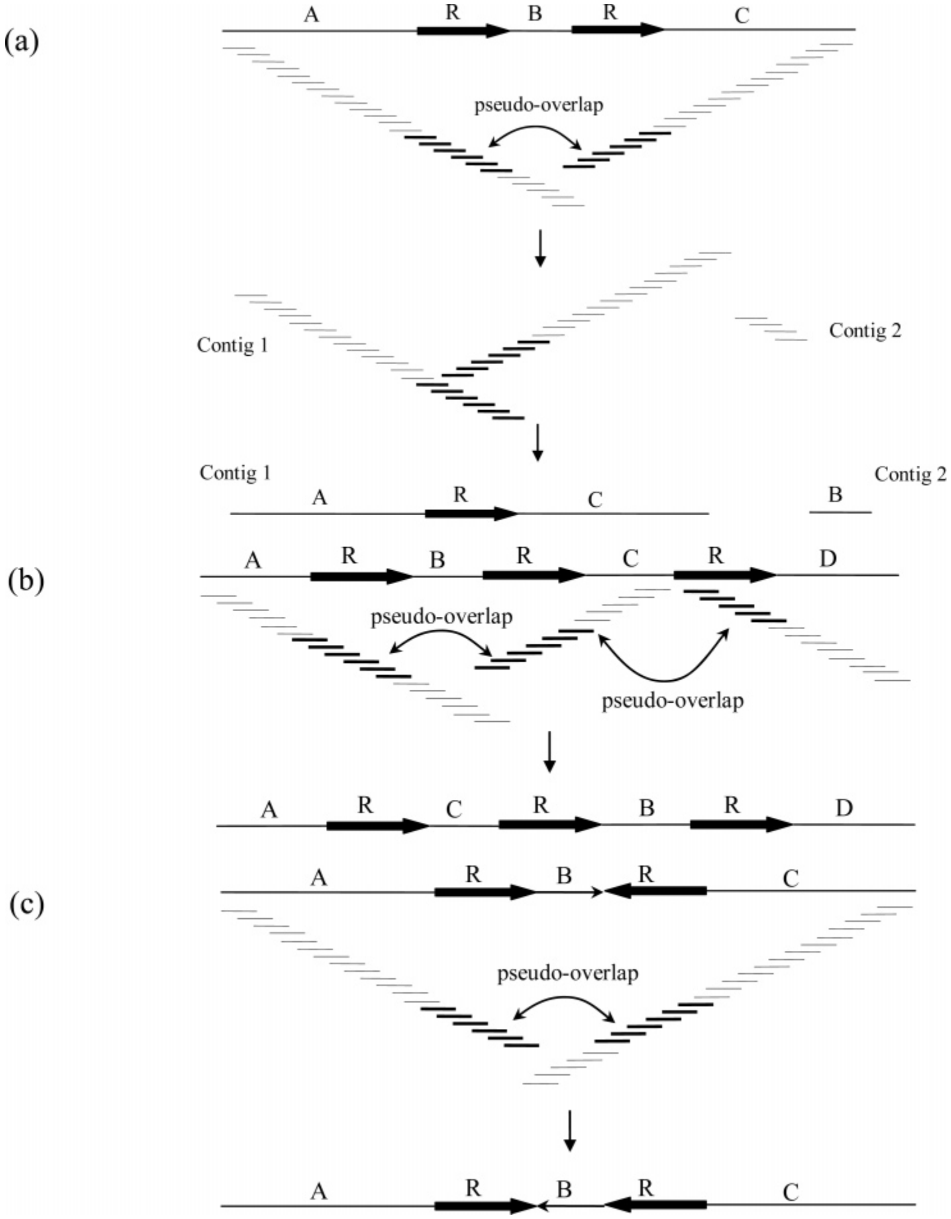

Figure 3. Large-scale misassemblies caused by the pseudo-overlaps between reads from the repeats. (a) Two closely located repeat copies (R) in the correct genomic sequence (top) may be collapsed into one copy in the misassembled genome sequence (bottom); (b) two DNA segments (B and C) flanked by three repeat copies in the correct genome sequence (top) may switch their order in this misassembled genome sequence; (c) a DNA segment flanked by two inverted repeat copies (B) in the genome sequence (top) may reverse its orientation in the misassembled genome sequence.

sequencing attempts to assemble the reads directly from the original large genome without the clone map. It speeds up the sequencing process but, at the same time, poses great challenges to the assembler, particularly to place the long repeats correctly. Since the rice genome has been sequenced using both the WGS and hierarchical approaches by two different groups, we compare these two assemblies to show the impact of sequencing approach on the repeat contents perceived in the genome assembly.[42,71] Even though the WGS rice genome sequencing reported larger set of contig sequences (466 vs 389 Mbp) and more protein coding genes (46 022 vs 37 544) than the hierarchical sequencing, the hierarchical sequencing method reported significantly more repeats, in particular TE-induced repeats (Table 2). We note that in WGS assembly, about 42% of shotgun reads were masked as highly repetitive sequences and disregarded in the assembly, in which many could be classified as partial TE-induced repeats.[42]

4.2. Repeat Resolution in Whole Genome Shotgun Sequencing

4.2.1. Repeat Masking

A simple yet effective approach to handle repeats in WGS assembly is not to assemble them. If the reads from the repeats can be detected and masked from the assembly process, the remaining reads from unique regions in the genome can be easily assembled into the contigs (sometimes called *unitigs*)[106] based on their all-real overlaps. Indeed, WGS sequenced genomes are assembled following this principle.[42,116,117] RepeatMasker, coupled with repeat libraries Repbase,[77] is often used in genome projects to detect and mask reads from known repeats.[42] For those genomes in which the repeat families have not been well-studied, statistical methods have to be used. Assuming the reads are sampled randomly from the genome with a fixed read

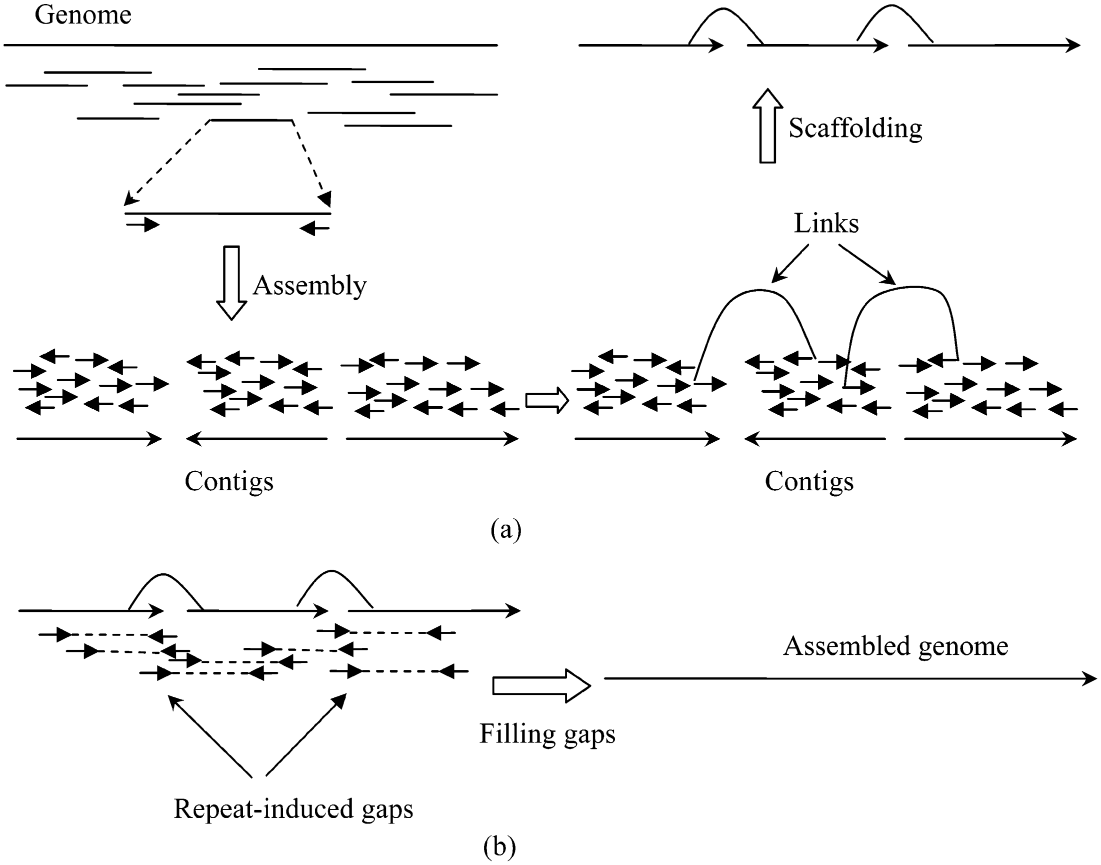

Figure 4. Improving WGS assembly by double-barreled sequencing. (a) Double-barreled sequencing is carried out by obtaining a pair of reads (*mate pairs*) simultaneously from both DNA strands of a medium-insert clone. These reads are first assembled into contigs using a shotgun assembler; the contigs are further linked together by the read pairs located at adjacent contigs in a scaffolding step. (b) The repeat-induced gaps in a scaffold may be filled by reassembling masked reads, which sometimes can be placed into the appropriate gaps based on their mates (linked by dashed lines) in the assembled contigs.

coverage (e.g., 8), those reads overlapping with a sufficiently large number of reads indicate that they may come from a repeat region. An even simpler method can identify repetitive *k*-tuples (i.e., words of length *k*) in reads based on their multiplicity (i.e., the number of reads containing it).[118,119]

The assembly of repeat-masked WGS reads generates a collection of contigs. Most of the gaps that separate these contigs are unassembled, highly identical repeat regions, rather than physical gaps with no read coverage. These contigs can be further grouped together into *scaffolds*, that is, a subset of contigs with determined order and orientation, based on double-barreled data (Figure 4). Most recently developed WGS assemblers contain such a scaffolding step.[119−124] Separate scaffolding programs are also implemented.[125]

4.2.2. The Repeat Graph and Eulerian Path Approach

Although repeat masking is useful in WGS assembly, not all repeats can be easily detected and masked. In particular, low copy repeats derived from segmental duplications are often undetectable because there is no clear sequence characteristic in these regions. As a result, collapsed repeat copies have been observed in WGS assembled genomes.[66]

A different approach to fragment assembly, the Eulerian path approach, attempts to represent repeats in a repeat graph rather than masking them. In graph theory, an Eulerian path is defined as a path that visits each edge in the graph once and only once. A graph in which an Eulerian exists is called an Eulerian graph. The greedy algorithm solution of the Eulerian path problem was first introduced by Leonhard Euler when solving the famous Seven Bridge in Konigsberg problem in 1736. The repeat graph is defined as an Eulerian graph representing each repeat in the genome that is longer than a predefined minimum length as an edge, and the genome as one of the Eulerian paths (Figure 5). Intuitively, if the genome sequence is given, the repeat graph can be constructed simply by first representing the genome as a linear line and then gluing together all similar repetitive regions. Importantly, it was shown that the repeat graph built from the fragments of genome (reads) is equivalent to the one built from the whole genome, if the minimum repeat length is predefined as the read length *l* (i.e., representing repeats longer than *l*).[90] So the fragment assembly problem can be transformed into the problem of constructing a repeat graph from a given set of reads, and afterward, assembling the genome sequence is equivalent to finding an Eulerian path in the repeat graph. The repeat graph can be built from reads using either the classical de Bruijn graph approach, if the reads are error-free or error-corrected,[126] or a generalized A-Bruijn graph approach, if the reads are error-prone.[90] In general, repeat resolution in fragment assembly can be viewed as a *de novo* repeat classification problem with the input of a set of fragments rather than the entire genome. The repeat graph provides a

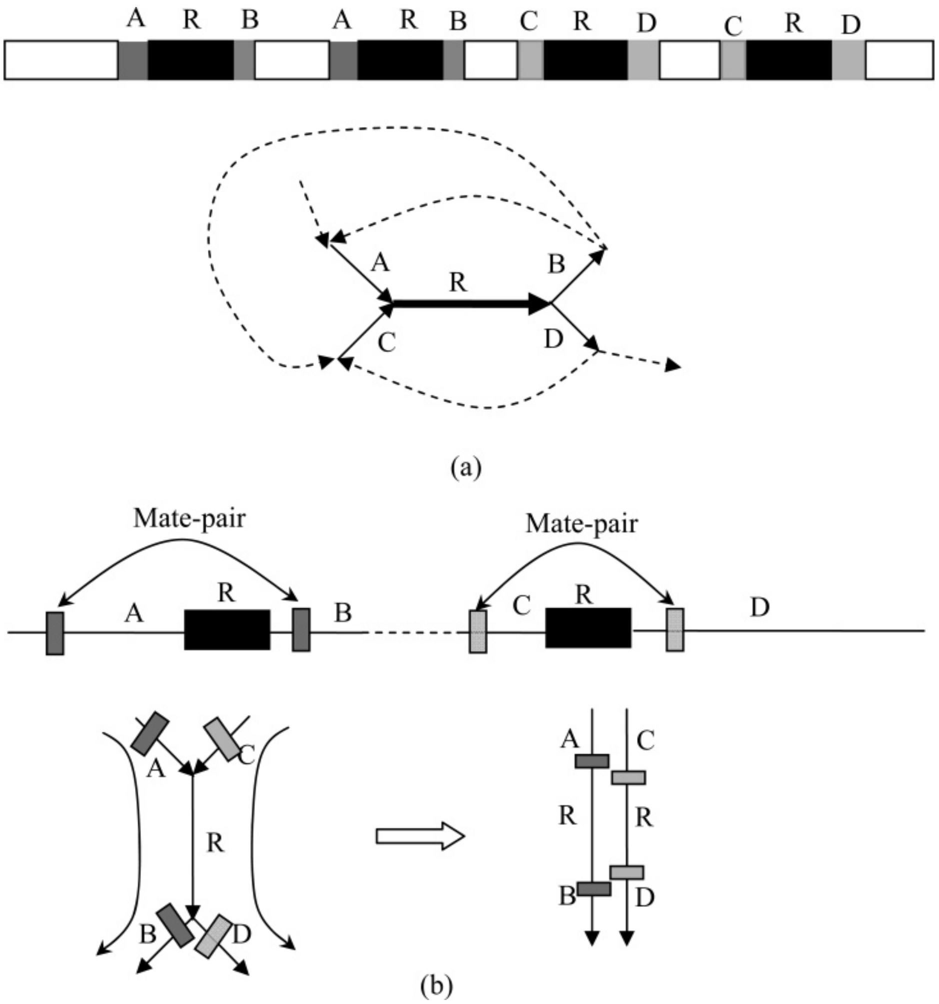

Figure 5. Repeat representation and resolution by repeat graph. (a) A complex repeat structure in a genome (top) created by a series of imaginary segmental duplications (see Figure 2) is represented by tangles in the repeat graph (bottom). Each repeat, A, B, C, D, and R, is represented as an edge with different copy number, e.g., R with 4 and A, B, C, and D all with 2. Unique regions are shown as dashed lines. (b) A tangle in the repeat that represents a repeat in the genome (top) is resolved by equivalent transformation using mate-pair derived paths (bottom).[1]

unified solution to both problems of genome assembly and repeat classification.

4.2.3. Repeat Resolution with Double-Barreled Data

The repeat graph reveals the best possible assembly one can achieve using shotgun sequencing data. It is often complex enough even for many bacterial genomes. An improved shotgun strategy, called *double-barreled sequencing* (or clone-end sequencing), greatly helps repeat resolution, and leads to longer contigs and fewer gaps. Double-barreled sequencing obtains a pair of reads (called *mate-pairs*) simultaneously from both DNA strands of a medium-insert clone, typically 3−5 or 30−45 kb in length (Figure 4a). Since the approximate distance between mate-pairs is known, many nearly identical repeats can be resolved using double-barreled data.

Within the “overlap-layout-consensus” framework of fragment assembly, the gaps induced by repeat masking can be filled in by placing masked reads into the corresponding gaps and reassembling them if their mates are placed into the assembled unique contigs (Figure 4b).[124] By contrast, within the repeat graph framework, double-barreled data can be used to transform (or simplify) the repeat graph by eliminating some repeat edges (Figure 5b).[1]

4.2.4. Identification and Separation of Collapsed Repeats in Assembled Genomes

The methods for repeat resolution described above are applied during the process of genome assembly. Nevertheless, there may be collapsed repeat copies in the assembled contigs. Identification and separation of these collapsed repeats have been formulated as a rigorous repeat separation problem.[127] Since then, various probabilistic and combinatorial methods have been proposed to solve this problem.[127,128] A recent study shows that the double-barreled data can significantly help the separation of nearly identical repeat copies.[112]

5. Genome Rearrangement

5.1. Classification of Genome Rearrangement

Genome rearrangements, referred to as the shuffling of large genomic segments between two species sharing a common ancestor, fall into two general classes: interchromosomal rearrangements, for example, inversions, and intrachromosomal rearrangements, for example, translocations, and fusion/fission of chromosomes.[129] The first analysis of large-scale genome rearrangement in eukaryotic genomes can be traced back to over 50 years ago, when Sturtevant

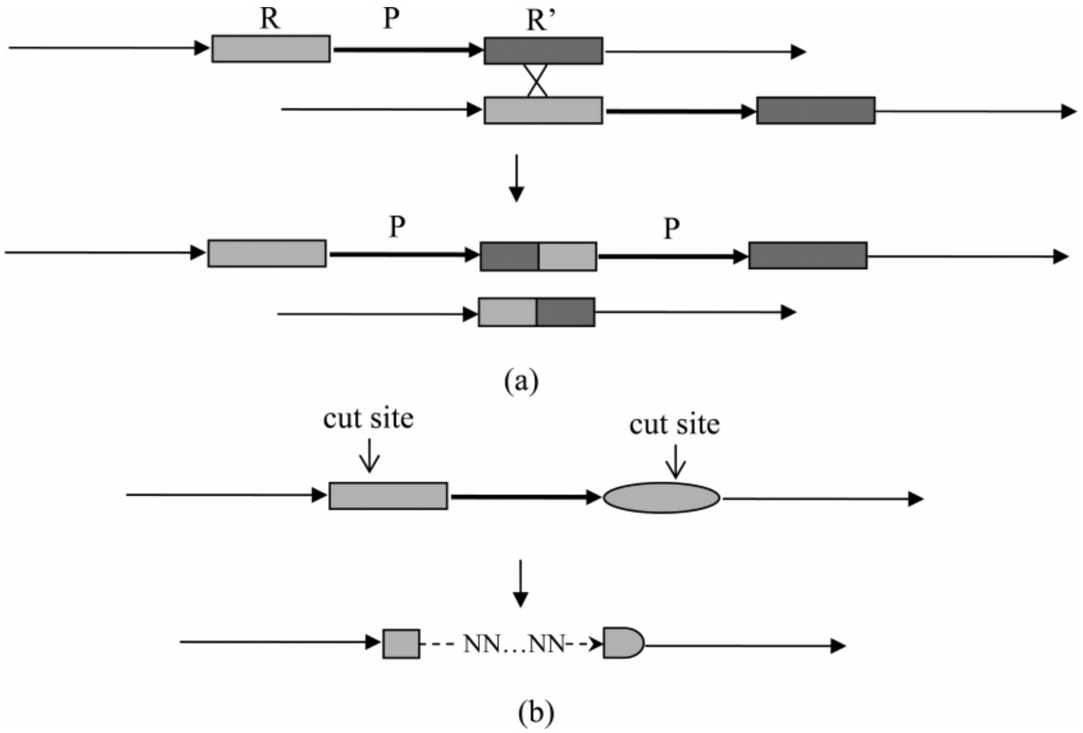

Figure 6. Segmental duplication/deletion/insertion rearrangement induced by Homologous Recombination (a) and Non-Homologous End Joining. (a) Two non-allelic repeats (R and R′) cause chromosomal misalignment and act as the substrates for recombination, resulting in a deletion with one recombinant product, and a duplication in the other recombinant product. (b) Two non-homologous segments (shown in rectangle and oval, respectively) act as the substrates for Non-Homologous End Joining (NHEJ), resulting in the deletion of intervening fragment, and the insertion of additional bases (NN...NN) at the junction.

and Dobzhansky studied the gene order in the various strains of *Drosophila*[130,131] and discovered many "blocks of genes rotated by 180° (inversions)".[132] Since then, geneticists have found many similar events in different eukaryotic genomes by comparing maps of genes or other types of genetic markers.[133] With many available eukaryotic genome sequences, the comparative analysis of two or multiple complete genomes may reveal homologous genomic regions in these genomes and their rearrangements at the nucleotide level (i.e., the highest possible resolution).

5.2. Homologous Blocks and Breakpoints

First introduced by Nadeau and Taylor,[134] a *homologous block* (or conserved block) between two genomes is referred to as segments with same gene order not disrupted by any genome rearrangements. Accordingly, the boundaries of maximum homologous blocks, where genome rearrangements occur, are called breakpoints. Although these notions consistently work well in comparative genetic map analysis,[133] they need to be revisited when applied to the comparison of genome sequences, at a much higher resolution than genetic maps. It has been shown that many homologous blocks identified based on genetic maps are not completely conserved at the sequence level, because they may contain multiple microrearrangements (i.e., rearrangements with a small span) within the blocks that are not detectable by low-resolution comparative genetic map analysis.[135] A new concept, *synteny blocks*, was proposed to generalize the original idea of homologous blocks for the comparative analysis of whole genome sequences.[136] Synteny blocks between two genomes are generally a chain of short regions with high similarity, intervened with dissimilar regions, which can be transformed to conserved blocks by only microarrangements within the blocks. In practice, the generation of synteny blocks is dependent on the algorithms used and the parameter settings (e.g., the maximum size to determine a microrearrangement).[38,136,137] Synteny blocks allow to distinguish large-scale rearrangements from many microarrangements, and focus on reconstructing the scenario of large-scale rearrangements. This strategy may avoid the influence of potential misassemblies in the WGS-assembled genomes, which causes some of the microrearrangements.[138]

5.3. Computational Methods for Genome Rearrangement Analysis

The computational problem of inferring genome distance of rearrangement between two genomes can be formulated as finding a series of genome rearrangements to transform the order of homologous or synteny blocks in one genome into the order of another one. The parsimony approach to the genomic distance problem was introduced by Palmer and colleagues.[139] It seeks a scenario to transform the block orders in a minimal number of rearrangements. This minimal number is called the genomic distance. In these studies, the most common genome rearrangement events, for example, translocations, fusion, and fission for multichromosomal genomes, and inversions (also called *reversals*) for both unichromosomal and multichromosomal genomes, are usually considered.

Nadeau and Taylor had already noticed the connection between the number of breakpoints and the genome distance, when they proposed the definition of breakpoints.[134] However, in early studies, breakpoints were counted independently without considering their potential relationships. For example, a single inversion can create two related breakpoints. This connection was first recognized by Kececioglu and Sankoff.[140] Its combinatorial property was later fully revealed by the introduction of the concept of *breakpoint* *graph* in a series of papers by Pevzner and colleagues,[141−143] which finally led to the polynomial solution to the genome distance problem for both the unichromosomal[144,145] and multichromosomal cases.[146,147]

5.4. Fragile versus Random Breakage Model

Under the assumption that breakpoints are uniformly distributed in the genomes (known as the "random breakage model"), Nadeau and Taylor estimated the number of conserved segments in human and mouse.[134] This estimate, along with the assumption of a random distribution, has been largely consistent with subsequent comparative studies of genetic maps[148] and genomic sequences.[38] It was, however, recently challenged by a combinatorial analysis by Pevzner and Tesler,[149] who tried to reconstruct the genome rearrangement scenario of 281 synteny blocks between the human and mouse genome. The most parsimonious scenario involves 245 rearrangement events, which leads to at least 190 (1.9 times) breakpoint reuses, and indicates 190 "short" (<1 Mb) synteny blocks located to one of the breakpoints from the original blocks. This observation contradicts the random distribution of breakpoints. Hence, an alternative model for chromosome evolution, called the fragile breakage model, is suggested, which postulates that there exist hotspots of rearrangement in certain fragile regions of the genome.[149] Although there is still debate between the random and fragile breakage models,[150,151] the fragile model appears to be consistent with observations from the genome rearrangement analysis of cancer and genetic diseases (see section 5.5).

5.5. Genome Rearrangement and Repeats

Genome rearrangements are of great medical interests. A growing group of genetic diseases, called genomic disorders, are coupled with abnormal function of a gene(s) located within a rearranged genomic segment.[152−154] At least two kinds of recombination mechanisms have been observed to be able to induce genome rearrangments, both of which are related to repeats (Figure 6). Homologous recombination (HR) may occur during DNA repair or other processes, resulting in non-allelic crossover, often between paralogous low copy repeats (LCR) induced by recent segmental duplications.[155] Non-homologous end joining (NHEJ), one of the mechanisms that repairs DNA double strand break (DSB), may also induce genome rearrangements around the junction. Although most NHEJs occur at the unique genomic regions, they were observed to be associated to TE-derived repeats, for example, Alu or LINE.[153] Even the origination of LCRs in the human genome appears to be associated with *Alu* elements.[156] Above all, the nonuniform distribution of repeats in eukaryotic genomes suggests the existence of hotspots for rearrangement breakpoints. The completion of the human genome and the full annotation of repeats can greatly help the identification of these regions that may be linked to genomic disorders. It should be kept in mind, however, that the repeats may induce not just the real rearrangements, but also virtual rearrangements (i.e., misassemblies) (Figure 3). These assembled breakpoint regions need to be carefully examined for assembly mistakes before drawing any conclusion.

6. Conclusion

The complete sequences of many eukaryotic genomes open the door for computational approaches to many biological problems. As compared with the intensive studies in coding regions of eukaryotic genomes, not much attention has been paid to the noncoding regions ("junk DNA"); in particular, repeats in noncoding regions have not been well-studied. Accumulating evidence shows, however, that many of them play essential roles in various cellular functions and are probably the driving forces of important evolutionary process, for example, genome rearrangements. Therefore, novel computational methods should be continuously developed toward identifying and classifying repeats, and ultimately understanding their potential functions.

7. Acknowledgment

I am indebted to Professors Michael S. Waterman and Pavel A. Pevzner for their generous supports and stimulating discussions. I also thank Professors Michael Lynch, Ben Raphael, Glenn Tesler, Uri Keich, and Sun Kim, Dr. Degui Zhi, and Mina Rho for their suggestions on different topics. This project is partially supported by the Indiana METACyt Initiative at Indiana University, funded in part through a major grant from Lilly Endowment, Inc.

8. References

(1) Pevzner, P. A.; Tang, H. *Bioinformatics* **2001**, *17*, S225.

(2) Eichler, E. E.; Sankoff, D. *Science* **2003**, *301*, 793.

(3) Pardi, F.; Goldman, N. *PLoS Genet.* **2005**, *1*, e71.

(4) Nierman, W. C.; Pain, A.; Anderson, M. J.; Wortman, J. R.; Kim, H. S.; Arroyo, J.; Berriman, M.; Abe, K.; Archer, D. B.; Bermejo, C.; Bennett, J.; Bowyer, P.; Chen, D.; Collins, M.; Coulsen, R.; Davies, R.; Dyer, P. S.; Farman, M.; Fedorova, N.; Fedorova, N.; Feldblyum, T. V.; Fischer, R.; Fosker, N.; Fraser, A.; Garcia, J. L.; Garcia, M. J.; Goble, A.; Goldman, G. H.; Gomi, K.; Griffith-Jones, S.; Gwilliam, R.; Haas, B.; Haas, H.; Harris, D.; Horiuchi, H.; Huang, J.; Humphray, S.; Jimenez, J.; Keller, N.; Khouri, H.; Kitamoto, K.; Kobayashi, T.; Konzack, S.; Kulkarni, R.; Kumagai, T.; Lafton, A.; Latge, J.-P.; Li, W.; Lord, A.; Lu, C.; Majoros, W. H.; May, G. S.; Miller, B. L.; Mohamoud, Y.; Molina, M.; Monod, M.; Mouyna, I.; Mulligan, S.; Murphy, L.; O'Neil, S.; Paulsen, I.; Penalva, M. A.; Pertea, M.; Price, C.; Pritchard, B. L.; Quail, M. A.; Rabbinowitsch, E.; Rawlins, N.; Rajandream, M.-A.; Reichard, U.; Renauld, H.; Robson, G. D.; de Cordoba, S. R.; Rodriguez-Pena, J. M.; Ronning, C. M.; Rutter, S.; Salzberg, S. L.; Sanchez, M.; Sanchez-Ferrero, J. C.; Saunders, D.; Seeger, K.; Squares, R.; Squares, S.; Takeuchi, M.; Tekaia, F.; Turner, G.; de Aldana, C. R. V.; Weidman, J.; White, O.; Woodward, J.; Yu, J.-H.; Fraser, C.; Galagan, J. E.; Asai, K.; Machida, M.; Hall, N.; Barrell, B.; Denning, D. W. *Nature* **2005**, *438*, 1151.

(5) Dujon, B.; Sherman, D.; Fischer, G.; Durrens, P.; Casaregola, S.; Lafontaine, I.; de Montigny, J.; Marck, C.; Neuveglise, C.; Talla, E.; Goffard, N.; Frangeul, L.; Aigle, M.; Anthouard, V.; Babour, A.; Barbe, V.; Barnay, S.; Blanchin, S.; Beckerich, J.-M.; Beyne, E.; Bleykasten, C.; Boisrame, A.; Boyer, J.; Cattolico, L.; Confanioleri, F.; de Daruvar, A.; Despons, L.; Fabre, E.; Fairhead, C.; Ferry-Dumazet, H.; Groppi, A.; Hantraye, F.; Hennequin, C.; Jauniaux, N.; Joyet, P.; Kachouri, R.; Kerrest, A.; Koszul, R.; Lemaire, M.; Lesur, I.; Ma, L.; Muller, H.; Nicaud, J.-M.; Nikolski, M.; Oztas, S.; Ozier-Kalogeropoulos, O.; Pellenz, S.; Potier, S.; Richard, G.-F.; Straub, M.-L.; Suleau, A.; Swennen, D.; Tekaia, F.; Wesolowski-Louvel, M.; Westhof, E.; Wirth, B.; Zeniou-Meyer, M.; Zivanovic, I.; Bolotin-Fukuhara, M.; Thierry, A.; Bouchier, C.; Caudron, B.; Scarpelli, C.; Gaillardin, C.; Weissenbach, J.; Wincker, P.; Souciet, J.-L. *Nature* **2004**, *430*, 35.

(6) Loftus, B. J.; Fung, E.; Roncaglia, P.; Rowley, D.; Amedeo, P.; Bruno, D.; Vamathevan, J.; Miranda, M.; Anderson, I. J.; Fraser, J. A.; Allen, J. E.; Bosdet, I. E.; Brent, M. R.; Chiu, R.; Doering, T. L.; Donlin, M. J.; D'Souza, C. A.; Fox, D. S.; Grinberg, V.; Fu, J.; Fukushima, M.; Haas, B. J.; Huang, J. C.; Janbon, G.; Jones, S. J. M.; Koo, H. L.; Krzywinski, M. I.; Kwon-Chung, J. K.; Lengeler, K. B.; Maiti, R.; Marra, M. A.; Marra, R. E.; Mathewson, C. A.; Mitchell, T. G.; Pertea, M.; Riggs, F. R.; Salzberg, S. L.; Schein, J. E.; Shvartsbeyn, A.; Shin, H.; Shumway, M.; Specht, C. A.; Suh, B. B.; Tenney, A.; Utterback, T. R.; Wickes, B. L.; Wortman, J. R.; Wye, N. H.; Kronstad, J. W.; Lodge, J. K.; Heitman, J.; Davis, R. W.; Fraser, C. M.; Hyman, R. W. *Science* **2005**, *307*, 1321.

(7) Jones, T.; Federspiel, N. A.; Chibana, H.; Dungan, J.; Kalman, S.; Magee, B. B.; Newport, G.; Thorstenson, Y. R.; Agabian, N.; Magee, P. T.; Davis, R. W.; Scherer, S. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 7329.

(8) Xu, P.; Widmer, G.; Wang, Y.; Ozaki, L. S.; Alves, J. M.; Serrano, M. G.; Puiu, D.; Manque, P.; Akiyoshi, D.; Mackey, A. J.; Pearson, W. R.; Dear, P. H.; Bankier, A. T.; Peterson, D. L.; Abrahamsen, M. S.; Kapur, V.; Tzipori, S.; Buck, G. A. *Nature* **2004**, *431*, 1107.

(9) Abrahamsen, M. S.; Templeton, T. J.; Enomoto, S.; Abrahante, J. E.; Zhu, G.; Lancto, C. A.; Deng, M.; Liu, C.; Widmer, G.; Tzipori, S.; Buck, G. A.; Xu, P.; Bankier, A. T.; Dear, P. H.; Konfortov, B. A.; Spriggs, H. F.; Iyer, L.; Anantharaman, V.; Aravind, L.; Kapur, V. *Science* **2004**, *304*, 441.

(10) Katinka, M. D.; Duprat, S.; Cornillot, E.; Metenier, G.; Thomarat, F.; Prensier, G.; Barbe, V.; Peyretaillade, E.; Brottier, P.; Wincker, P.; Delbac, F.; El Alaoui, H.; Peyret, P.; Saurin, W.; Gouy, M.; Weissenbach, J.; Vivares, C. P. *Nature* **2001**, *414*, 450.

(11) Dietrich, F. S.; Voegeli, S.; Brachat, S.; Lerch, A.; Gates, K.; Steiner, S.; Mohr, C.; Pohlmann, R.; Luedi, P.; Choi, S.; Wing, R. A.; Flavier, A.; Gaffney, T. D.; Philippsen, P. *Science* **2004**, *304*, 304.

(12) Kellis, M.; Birren, B. W.; Lander, E. S. *Nature* **2004**, *428*, 617.

(13) Dean, R. A.; Talbot, N. J.; Ebbole, D. J.; Farman, M. L.; Mitchell, T. K.; Orbach, M. J.; Thon, M.; Kulkarni, R.; Xu, J.-R.; Pan, H.; Read, N. D.; Lee, Y.-H.; Carbone, I.; Brown, D.; Oh, Y. Y.; Donofrio, N.; Jeong, J. S.; Soanes, D. M.; Djonovic, S.; Kolomiets, E.; Rehmeyer, C.; Li, W.; Harding, M.; Kim, S.; Lebrun, M.-H.; Bohnert, H.; Coughlan, S.; Butler, J.; Calvo, S.; Ma, L.-J.; Nicol, R.; Purcell, S.; Nusbaum, C.; Galagan, J. E.; Birren, B. W. *Nature* **2005**, *434*, 980.

(14) Galagan, J. E.; Calvo, S. E.; Borkovich, K. A.; Selker, E. U.; Read, N. D.; Jaffe, D.; FitzHugh, W.; Ma, L.-J.; Smirnov, S.; Purcell, S.; Rehman, B.; Elkins, T.; Engels, R.; Wang, S.; Nielsen, C. B.; Butler, J.; Endrizzi, M.; Qui, D.; Ianakiev, P.; Bell-Pedersen, D.; Nelson, M. A.; Werner-Washburne, M.; Selitrennikoff, C. P.; Kinsey, J. A.; Braun, E. L.; Zelter, A.; Schulte, U.; Kothe, G. O.; Jedd, G.; Mewes, W.; Staben, C.; Marcotte, E.; Greenberg, D.; Roy, A.; Foley, K.; Naylor, J.; Stange-Thomann, N.; Barrett, R.; Gnerre, S.; Kamal, M.; Kamvysselis, M.; Mauceli, E.; Bielke, C.; Rudd, S.; Frishman, D.; Krystofova, S.; Rasmussen, C.; Metzenberg, R. L.; Perkins, D. D.; Kroken, S.; Cogoni, C.; Macino, G.; Catcheside, D.; Li, W.; Pratt, R. J.; Osmani, S. A.; DeSouza, C. P. C.; Glass, L.; Orbach, M. J.; Berglund, J. A.; Voelker, R.; Yarden, O.; Plamann, M.; Seiler, S.; Dunlap, J.; Radford, A.; Aramayo, R.; Natvig, D. O.; Alex, L. A.; Mannhaupt, G.; Ebbole, D. J.; Freitag, M.; Paulsen, I.; Sachs, M. S.; Lander, E. S.; Nusbaum, C.; Birren, B. *Nature* **2003**, *422*, 859.

(15) Martinez, D.; Larrondo, L. F.; Putnam, N.; Gelpke, M. D.; Huang, K.; Chapman, J.; Helfenbein, K. G.; Ramaiya, P.; Detter, J. C.; Larimer, F.; Coutinho, P. M.; Henrissat, B.; Berka, R.; Cullen, D.; Rokhsar, D. *Nat. Biotechnol.* **2004**, *22*, 695.

(16) Goffeau, A.; Barrell, B. G.; Bussey, H.; Davis, R. W.; Dujon, B.; Feldmann, H.; F., G.; Hoheisel, J. D.; Jacq, C.; Johnston, M.; Louis, E. J.; Mewes, H. W.; Murakami, Y.; Philippsen, P.; Tettelin, H.; Oliver, S. G. *Sci. Mar.* **1996**, *274*, 563.

(17) Wood, V.; Gwilliam, R.; Rajandream, M. A.; Lyne, M.; Lyne, R.; Stewart, A.; Sgouros, J.; Peat, N.; Hayles, J.; Baker, S.; Basham, D.; Bowman, S.; Brooks, K.; Brown, D.; Brown, S.; Chillingworth, T.; Churcher, C.; Collins, M.; Connor, R.; Cronin, A.; Davis, P.; Feltwell, T.; Fraser, A.; Gentles, S.; Goble, A.; Hamlin, N.; Harris, D.; Hidalgo, J.; Hodgson, G.; Holroyd, S.; Hornsby, T.; Howarth, S.; Huckle, E. J.; Hunt, S.; Jagels, K.; James, K.; Jones, L.; Jones, M.; Leather, S.; McDonald, S.; McLean, J.; Mooney, P.; Moule, S.; Mungall, K.; Murphy, L.; Niblett, D.; Odell, C.; Oliver, K.; O'Neil, S.; Pearson, D.; Quail, M. A.; Rabbinowitsch, E.; Rutherford, K.; Rutter, S.; Saunders, D.; Seeger, K.; Sharp, S.; Skelton, J.; Simmonds, M.; Squares, R.; Squares, S.; Stevens, K.; Taylor, K.; Taylor, R. G.; Tivey, A.; Walsh, S.; Warren, T.; Whitehead, S.; Woodward, J.; Volckaert, G.; Aert, R.; Robben, J.; Grymonprez, B.; Weltjens, I.; Vanstreels, E.; Rieger, M.; Schafer, M.; Muller-Auer, S.; Gabel, C.; Fuchs, M.; Fritzc, C.; Holzer, E.; Moestl, D.; Hilbert, H.; Borzym, K.; Langer, I.; Beck, A.; Lehrach, H.; Reinhardt, R.; Pohl, T. M.; Eger, P.; Zimmermann, W.; Wedler, H.; Wambutt, R.; Purnelle, B.; Goffeau, A.; Cadieu, E.; Dreano, S.; Gloux, S.; Lelaure, V. *Nature* **2002**, *415*, 871.

(18) Matsuzaki, M.; Misumi, O.; Shin-i, T.; Maruyama, S.; Takahara, M.; Miyagishima, S.-y.; Mori, T.; Nishida, K.; Yagisawa, F.; Nishida, K.; Yoshida, Y.; Nishimura, Y.; Nakao, S.; Kobayashi, T.; Momoyama, Y.; Higashiyama, T.; Minoda, A.; Sano, M.; Nomoto, H.; Oishi, K.; Hayashi, H.; Ohta, F.; Nishizaka, S.; Haga, S.; Miura, S.; Morishita, T.; Kabeya, Y.; Terasawa, K.; Suzuki, Y.; Ishii, Y.; Asakawa, S.; Takano, H.; Ohta, N.; Kuroiwa, H.; Tanaka, K.; Shimizu, N.; Sugano, S.; Sato, N.; Nozaki, H.; Ogasawara, N.; Kohara, Y.; Kuroiwa, T. *Nature* **2004**, *428*, 653.

(19) Loftus, B.; Anderson, I.; Davies, R.; Alsmark, U. C. M.; Samuelson, J.; Amedeo, P.; Roncaglia, P.; Berriman, M.; Hirt, R. P.; Mann, B. J.; Nozaki, T.; Suh, B.; Pop, M.; Duchene, M.; Ackers, J.; Tannich, E.; Leippe, M.; Hofer, M.; Bruchhaus, I.; Willhoeft, U.; Bhattacharya, A.; Chillingworth, T.; Churcher, C.; Hance, Z.; Harris, B.; Harris, D.; Jagels, K.; Moule, S.; Mungall, K.; Ormond, D.; Squares, R.; Whitehead, S.; Quail, M. A.; Rabbinowitsch, E.; Norbertczak, H.; Price, C.; Wang, Z.; Guillen, N.; Gilchrist, C.; Stroup, S. E.; Bhattacharya, S.; Lohia, A.; Foster, P. G.; Sicheritz-Ponten, T.; Weber, C.; Singh, U.; Mukherjee, C.; El-Sayed, N. M.; Petri, W. A.; Clark, C. G.; Embley, T. M.; Barrell, B.; Fraser, C. M.; Hall, N. *Nature* **2005**, *433*, 865.

(20) Eichinger, L.; Pachebat, J. A.; Glockner, G.; Rajandream, M. A.; Sucgang, R.; Berriman, M.; Song, J.; Olsen, R.; Szafranski, K.; Xu, Q.; Tunggal, B.; Kummerfeld, S.; Madera, M.; Konfortov, B. A.; Rivero, F.; Bankier, A. T.; Lehmann, R.; Hamlin, N.; Davies, R.; Gaudet, P.; Fey, P.; Pilcher, K.; Chen, G.; Saunders, D.; Sodergren, E.; Davis, P.; Kerhornou, A.; Nie, X.; Hall, N.; Anjard, C.; Hemphill, L.; Bason, N.; Farbrother, P.; Desany, B.; Just, E.; Morio, T.; Rost, R.; Churcher, C.; Cooper, J.; Haydock, S.; van Driessche, N.; Cronin, A.; Goodhead, I.; Muzny, D.; Mourier, T.; Pain, A.; Lu, M.; Harper, D.; Lindsay, R.; Hauser, H.; James, K.; Quiles, M.; Madan Babu, M.; Saito, T.; Buchrieser, C.; Wardroper, A.; Felder, M.; Thangavelu, M.; Johnson, D.; Knights, A.; Loulseged, H.; Mungall, K.; Oliver, K.; Price, C.; Quail, M. A.; Urushihara, H.; Hernandez, J.; Rabbinowitsch, E.; Steffen, D.; Sanders, M.; Ma, J.; Kohara, Y.; Sharp, S.; Simmonds, M.; Spiegler, S.; Tivey, A.; Sugano, S.; White, B.; Walker, D.; Woodward, J.; Winckler, T.; Tanaka, Y.; Shaulsky, G.; Schleicher, M.; Weinstock, G.; Rosenthal, A.; Cox, E. C.; Chisholm, R. L.; Gibbs, R.; Loomis, W. F.; Platzer, M.; Kay, R. R.; Williams, J.; Dear, P. H.; Noegel, A. A.; Barrell, B.; Kuspa, A. *Nature* **2005**, *435*, 43.

(21) Ivens, A. C.; Peacock, C. S.; Worthey, E. A.; Murphy, L.; Aggarwal, G.; Berriman, M.; Sisk, E.; Rajandream, M.-A.; Adlem, E.; Aert, R.; Anupama, A.; Apostolou, Z.; Attipoe, P.; Bason, N.; Bauser, C.; Beck, A.; Beverley, S. M.; Bianchettin, G.; Borzym, K.; Bothe, G.; Bruschi, C. V.; Collins, M.; Cadag, E.; Ciarloni, L.; Clayton, C.; Coulson, R. M. R.; Cronin, A.; Cruz, A. K.; Davies, R. M.; De Gaudenzi, J.; Dobson, D. E.; Duesterhoeft, A.; Fazelina, G.; Fosker, N.; Frasch, A. C.; Fraser, A.; Fuchs, M.; Gabel, C.; Goble, A.; Goffeau, A.; Harris, D.; Hertz-Fowler, C.; Hilbert, H.; Horn, D.; Huang, Y.; Klages, S.; Knights, A.; Kube, M.; Larke, N.; Litvin, L.; Lord, A.; Louie, T.; Marra, M.; Masuy, D.; Matthews, K.; Michaeli, S.; Mottram, J. C.; Muller-Auer, S.; Munden, H.; Nelson, S.; Norbertczak, H.; Oliver, K.; O'Neil, S.; Pentony, M.; Pohl, T. M.; Price, C.; Purnelle, B.; Quail, M. A.; Rabbinowitsch, E.; Reinhardt, R.; Rieger, M.; Rinta, J.; Robben, J.; Robertson, L.; Ruiz, J. C.; Rutter, S.; Saunders, D.; Schafer, M.; Schein, J.; Schwartz, D. C.; Seeger, K.; Seyler, A.; Sharp, S.; Shin, H.; Sivam, D.; Squares, R.; Squares, S.; Tosato, V.; Vogt, C.; Volckaert, G.; Wambutt, R.; Warren, T.; Wedler, H.; Woodward, J.; Zhou, S.; Zimmermann, W.; Smith, D. F.; Blackwell, J. M.; Stuart, K. D.; Barrell, B. *Science* **2005**, *309*, 436.

(22) Gardner, M. J.; Hall, N.; Fung, E.; White, O.; Berriman, M.; Hyman, R. W.; Carlton, J. M.; Pain, A.; Nelson, K. E.; Bowman, S.; Paulsen, I. T.; James, K.; Eisen, J. A.; Rutherford, K.; Salzberg, S. L.; Craig, A.; Kyes, S.; Chan, M.-S.; Nene, V.; Shallom, S. J.; Suh, B.; Peterson, J.; Angiuoli, S.; Pertea, M.; Allen, J.; Selengut, J.; Haft, D.; Mather, M. W.; Vaidya, A. B.; Martin, D. M. A.; Fairlamb, A. H.; Fraunholz, M. J.; Roos, D. S.; Ralph, S. A.; McFadden, G. I.; Cummings, L. M.; Subramanian, G. M.; Mungall, C.; Venter, J. C.; Carucci, D. J.; Hoffman, S. L.; Newbold, C.; Davis, R. W.; Fraser, C. M.; Barrell, B. *Nature* **2002**, *419*, 498.

(23) Armbrust, E. V.; Berges, J. A.; Bowler, C.; Green, B. R.; Martinez, D.; Putnam, N. H.; Zhou, S.; Allen, A. E.; Apt, K. E.; Bechner, M.; Brzezinski, M. A.; Chaal, B. K.; Chiovitti, A.; Davis, A. K.; Demarest, M. S.; Detter, J. C.; Glavina, T.; Goodstein, D.; Hadi, M. Z.; Hellsten, U.; Hildebrand, M.; Jenkins, B. D.; Jurka, J.; Kapitonov, V. V.; Kroger, N.; Lau, W. W. Y.; Lane, T. W.; Larimer, F. W.; Lippmeier, J. C.; Lucas, S.; Medina, M.; Montsant, A.; Obornik, M.; Parker, M. S.; Palenik, B.; Pazour, G. J.; Richardson, P. M.; Rynearson, T. A.; Saito, M. A.; Schwartz, D. C.; Thamatrakoln, K.; Valentin, K.; Vardi, A.; Wilkerson, F. P.; Rokhsar, D. S. *Science* **2004**, *306*, 79.

(24) Gardner, M. J.; Bishop, R.; Shah, T.; de Villiers, E. P.; Carlton, J. M.; Hall, N.; Ren, Q.; Paulsen, I. T.; Pain, A.; Berriman, M.; Wilson, R. J. M.; Sato, S.; Ralph, S. A.; Mann, D. J.; Xiong, Z.; Shallom, S. J.; Weidman, J.; Jiang, L.; Lynn, J.; Weaver, B.; Shoaibi, A.; Domingo, A. R.; Wasawo, D.; Crabtree, J.; Wortman, J. R.; Haas, B.; Angiuoli, S. V.; Creasy, T. H.; Lu, C.; Suh, B.; Silva, J. C.; Utterback, T. R.; Feldblyum, T. V.; Pertea, M.; Allen, J.; Nierman, W. C.; Taracha, E. L. N.; Salzberg, S. L.; White, O. R.; Fitzhugh,

H. A.; Morzaria, S.; Venter, J. C.; Fraser, C. M.; Nene, V. *Science* **2005**, *309*, 134.

(25) Berriman, M.; Ghedin, E.; Hertz-Fowler, C.; Blandin, G.; Renauld, H.; Bartholomeu, D. C.; Lennard, N. J.; Caler, E.; Hamlin, N. E.; Haas, B.; Bohme, U.; Hannick, L.; Aslett, M. A.; Shallom, J.; Marcello, L.; Hou, L.; Wickstead, B.; Alsmark, U. C. M.; Arrowsmith, C.; Atkin, R. J.; Barron, A. J.; Bringaud, F.; Brooks, K.; Carrington, M.; Cherevach, I.; Chillingworth, T.-J.; Churcher, C.; Clark, L. N.; Corton, C. H.; Cronin, A.; Davies, R. M.; Doggett, J.; Djikeng, A.; Feldblyum, T.; Field, M. C.; Fraser, A.; Goodhead, I.; Hance, Z.; Harper, D.; Harris, B. R.; Hauser, H.; Hostetler, J.; Ivens, A.; Jagels, K.; Johnson, D.; Johnson, J.; Jones, K.; Kerhornou, A. X.; Koo, H.; Larke, N.; Landfear, S.; Larkin, C.; Leech, V.; Line, A.; Lord, A.; MacLeod, A.; Mooney, P. J.; Moule, S.; Martin, D. M. A.; Morgan, G. W.; Mungall, K.; Norbertczak, H.; Ormond, D.; Pai, G.; Peacock, C. S.; Peterson, J.; Quail, M. A.; Rabbinowitsch, E.; Rajandream, M.-A.; Reitter, C.; Salzberg, S. L.; Sanders, M.; Schobel, S.; Sharp, S.; Simmonds, M.; Simpson, A. J.; Tallon, L.; Turner, C. M. R.; Tait, A.; Tivey, A. R.; Van Aken, S.; Walker, D.; Wanless, D.; Wang, S.; White, B.; White, O.; Whitehead, S.; Woodward, J.; Wortman, J.; Adams, M. D.; Embley, T. M.; Gull, K.; Ullu, E.; Barry, J. D.; Fairlamb, A. H.; Opperdoes, F.; Barrell, B. G.; Donelson, J. E.; Hall, N.; Fraser, C. M. *Science* **2005**, *309*, 416.

(26) El-Sayed, N. M.; Myler, P. J.; Bartholomeu, D. C.; Nilsson, D.; Aggarwal, G.; Tran, A.-N.; Ghedin, E.; Worthey, E. A.; Delcher, A. L.; Blandin, G.; Westenberger, S. J.; Caler, E.; Cerqueira, G. C.; Branche, C.; Haas, B.; Anupama, A.; Arner, E.; Aslund, L.; Attipoe, P.; Bontempi, E.; Bringaud, F.; Burton, P.; Cadag, E.; Campbell, D. A.; Carrington, M.; Crabtree, J.; Darban, H.; da Silveira, J. F.; de Jong, P.; Edwards, K.; Englund, P. T.; Fazelina, G.; Feldblyum, T.; Ferella, M.; Frasch, A. C.; Gull, K.; Horn, D.; Hou, L.; Huang, Y.; Kindlund, E.; Klingbeil, M.; Kluge, S.; Koo, H.; Lacerda, D.; Levin, M. J.; Lorenzi, H.; Louie, T.; Machado, C. R.; McCulloch, R.; McKenna, A.; Mizuno, Y.; Mottram, J. C.; Nelson, S.; Ochaya, S.; Osoegawa, K.; Pai, G.; Parsons, M.; Pentony, M.; Pettersson, U.; Pop, M.; Ramirez, J. L.; Rinta, J.; Robertson, L.; Salzberg, S. L.; Sanchez, D. O.; Seyler, A.; Sharma, R.; Shetty, J.; Simpson, A. J.; Sisk, E.; Tammi, M. T.; Tarleton, R.; Teixeira, S.; Van Aken, S.; Vogt, C.; Ward, P. N.; Wickstead, B.; Wortman, J.; White, O.; Fraser, C. M.; Stuart, K. D.; Andersson, B. *Science* **2005**, *309*, 409.

(27) *C. elegans* Sequencing Consortium. *Science* **1998**, *282*, 5396.

(28) Stein, L. D.; Bao, Z.; Blasiar, D.; Blumenthal, T.; Brent, M. R.; Chen, N.; Chinwalla, A.; Clarke, L.; Clee, C.; Coghlan, A.; Coulson, A.; Eustachio, P.; Fitch, D. H. A.; Fulton, L. A.; Fulton, R. E.; Griffiths-Jones, S.; Harris, T. W.; Hillier, L. W.; Kamath, R.; Kuwabara, P. E.; Mardis, E. R.; Marra, M. A.; Miner, T. L.; Minx, P.; Mullikin, J. C.; Plumb, R. W.; Rogers, J.; Schein, J. E.; Sohrmann, M.; Spieth, J.; Stajich, J. E.; Wei, C.; Willey, D.; Wilson, R. K.; Durbin, R.; Waterston, R. H. *PLoS Biol.* **2003**, *1*, e45.

(29) Adams, M. D.; Celniker, S. E.; Holt, R. A.; Evans, C. A.; Gocayne, J. D.; Amanatides, P. G.; Scherer, S. E.; Li, P. W.; Hoskins, R. A.; Galle, R. F.; George, R. A.; Lewis, S. E.; Richards, S.; Ashburner, M.; Henderson, S. N.; Sutton, G. G.; Wortman, J. R.; Yandell, M. D.; Zhang, Q.; Chen, L. X.; Brandon, R. C.; Rogers, Y.-H. C.; Blazej, R. G.; Champe, M.; Pfeiffer, B. D.; Wan, K. H.; Doyle, C.; Baxter, E. G.; Helt, G.; Nelson, C. R.; Gabor Miklos, G. L.; Abril, J. F.; Agbayani, A.; An, H.-J.; Andrews-Pfannkoch, C.; Baldwin, D.; Ballew, R. M.; Basu, A.; Baxendale, J.; Bayraktaroglu, L.; Beasley, E. M.; Beeson, K. Y.; Benos, P. V.; Berman, B. P.; Bhandari, D.; Bolshakov, S.; Borkova, D.; Botchan, M. R.; Bouck, J.; Brokstein, P.; Brottier, P.; Burtis, K. C.; Busam, D. A.; Butler, H.; Cadieu, E.; Center, A.; Chandra, I.; Cherry, J. M.; Cawley, S.; Dahlke, C.; Davenport, L. B.; Davies, P.; de Pablos, B.; Delcher, A.; Deng, Z.; Mays, A. D.; Dew, I.; Dietz, S. M.; Dodson, K.; Doup, L. E.; Downes, M.; Dugan-Rocha, S.; Dunkov, B. C.; Dunn, P.; Durbin, K. J.; Evangelista, C. C.; Ferraz, C.; Ferriera, S.; Fleischmann, W.; Fosler, C.; Gabrielian, A. E.; Garg, N. S.; Gelbart, W. M.; Glasser, K.; Glodek, A.; Gong, F.; Gorrell, J. H.; Gu, Z.; Guan, P.; Harris, M.; Harris, N. L.; Harvey, D.; Heiman, T. J.; Hernandez, J. R.; Houck, J.; Hostin, D.; Houston, K. A.; Howland, T. J.; Wei, M.-H.; Ibegwam, C. *Science* **2000**, *287*, 2185.

(30) Xia, Q.; Zhou, Z.; Lu, C.; Cheng, D.; Dai, F.; Li, B.; Zhao, P.; Zha, X.; Cheng, T.; Chai, C.; Pan, G.; Xu, J.; Liu, C.; Lin, Y.; Qian, J.; Hou, Y.; Wu, Z.; Li, G.; Pan, M.; Li, C.; Shen, Y.; Lan, X.; Yuan, L.; Li, T.; Xu, H.; Yang, G.; Wan, Y.; Zhu, Y.; Yu, M.; Shen, W.; Wu, D.; Xiang, Z.; Genome analysis, g.; Yu, J.; Wang, J.; Li, R.; Shi, J.; Li, H.; Li, G.; Su, J.; Wang, X.; Li, G.; Zhang, Z.; Wu, Q.; Li, J.; Zhang, Q.; Wei, N.; Xu, J.; Sun, H.; Dong, L.; Liu, D.; Zhao, S.; Zhao, X.; Meng, Q.; Lan, F.; Huang, X.; Li, Y.; Fang, L.; Li, C.; Li, D.; Sun, Y.; Zhang, Z.; Yang, Z.; Huang, Y.; Xi, Y.; Qi, Q.; He, D.; Huang, H.; Zhang, X.; Wang, Z.; Li, W.; Cao, Y.; Yu, Y.; Yu, H.; Li, J.; Ye, J.; Chen, H.; Zhou, Y.; Liu, B.; Wang, J.; Ye, J.; Ji, H.; Li, S.; Ni, P.; Zhang, J.; Zhang, Y.; Zheng, H.; Mao, B.; Wang, W.; Ye, C.; Li, S.; Wang, J.; Wong, G. K.-S.; Yang, H. *Science* **2004**, *306*, 1937.

(31) Holt, R. A.; Subramanian, G. M.; Halpern, A.; Sutton, G. G.; Charlab, R.; Nusskern, D. R.; Wincker, P.; Clark, A. G.; Ribeiro, J. M. C.; Wides, R.; Salzberg, S. L.; Loftus, B.; Yandell, M.; Majoros, W. H.; Rusch, D. B.; Lai, Z.; Kraft, C. L.; Abril, J. F.; Anthouard, V.; Arensburger, P.; Atkinson, P. W.; Baden, H.; de Berardinis, V.; Baldwin, D.; Benes, V.; Biedler, J.; Blass, C.; Bolanos, R.; Boscus, D.; Barnstead, M.; Cai, S.; Center, A.; Chatuverdi, K.; Christophides, G. K.; Chrystal, M. A.; Clamp, M.; Cravchik, A.; Curwen, V.; Dana, A.; Delcher, A.; Dew, I.; Evans, C. A.; Flanigan, M.; Grundschober-Freimoser, A.; Friedli, L.; Gu, Z.; Guan, P.; Guigo, R.; Hillenmeyer, M. E.; Hladun, S. L.; Hogan, J. R.; Hong, Y. S.; Hoover, J.; Jaillon, O.; Ke, Z.; Kodira, C.; Kokoza, E.; Koutsos, A.; Letunic, I.; Levitsky, A.; Liang, Y.; Lin, J.-J.; Lobo, N. F.; Lopez, J. R.; Malek, J. A.; McIntosh, T. C.; Meister, S.; Miller, J.; Mobarry, C.; Mongin, E.; Murphy, S. D.; O'Brochta, D. A.; Pfannkoch, C.; Qi, R.; Regier, M. A.; Remington, K.; Shao, H.; Sharakhova, M. V.; Sitter, C. D.; Shetty, J.; Smith, T. J.; Strong, R.; Sun, J.; Thomasova, D.; Ton, L. Q.; Topalis, P.; Tu, Z.; Unger, M. F.; Walenz, B.; Wang, A.; Wang, J.; Wang, M.; Wang, X.; Woodford, K. J.; Wortman, J. R.; Wu, M.; Yao, A.; Zdobnov, E. M.; Zhang, H.; Zhao, Q. *Science* **2002**, *298*, 129.

(32) Dehal, P.; Satou, Y.; Campbell, R. K.; Chapman, J.; Degnan, B.; De Tomaso, A.; Davidson, B.; Di Gregorio, A.; Gelpke, M.; Goodstein, D. M.; Harafuji, N.; Hastings, K. E. M.; Ho, I.; Hotta, K.; Huang, W.; Kawashima, T.; Lemaire, P.; Martinez, D.; Meinertzhagen, I. A.; Necula, S.; Nonaka, M.; Putnam, N.; Rash, S.; Saiga, H.; Satake, M.; Terry, A.; Yamada, L.; Wang, H.-G.; Awazu, S.; Azumi, K.; Boore, J.; Branno, M.; Chin-bow, S.; DeSantis, R.; Doyle, S.; Francino, P.; Keys, D. N.; Haga, S.; Hayashi, H.; Hino, K.; Imai, K. S.; Inaba, K.; Kano, S.; Kobayashi, K.; Kobayashi, M.; Lee, B.-I.; Makabe, K. W.; Manohar, C.; Matassi, G.; Medina, M.; Mochizuki, Y.; Mount, S.; Morishita, T.; Miura, S.; Nakayama, A.; Nishizaka, S.; Nomoto, H.; Ohta, F.; Oishi, K.; Rigoutsos, I.; Sano, M.; Sasaki, A.; Sasakura, Y.; Shoguchi, E.; Shin-i, T.; Spagnuolo, A.; Stainier, D.; Suzuki, M. M.; Tassy, O.; Takatori, N.; Tokuoka, M.; Yagi, K.; Yoshizaki, F.; Wada, S.; Zhang, C.; Hyatt, P. D.; Larimer, F.; Detter, C.; Doggett, N.; Glavina, T.; Hawkins, T.; Richardson, P.; Lucas, S.; Kohara, Y.; Levine, M.; Satoh, N.; Rokhsar, D. S. *Science* **2002**, *298*, 2157.

(33) Aparicio, S.; Chapman, J.; Stupka, E.; Putnam, N.; Chia, J.-m.; Dehal, P.; Christoffels, A.; Rash, S.; Hoon, S.; Smit, A.; Gelpke, M. D. S.; Roach, J.; Oh, T.; Ho, I. Y.; Wong, M.; Detter, C.; Verhoef, F.; Predki, P.; Tay, A.; Lucas, S.; Richardson, P.; Smith, S. F.; Clark, M. S.; Edwards, Y. J. K.; Doggett, N.; Zharkikh, A.; Tavtigian, S. V.; Pruss, D.; Barnstead, M.; Evans, C.; Baden, H.; Powell, J.; Glusman, G.; Rowen, L.; Hood, L.; Tan, Y. H.; Elgar, G.; Hawkins, T.; Venkatesh, B.; Rokhsar, D.; Brenner, S. *Science* **2002**, *297*, 1301.

(34) Jaillon, O.; Aury, J.-M.; Brunet, F.; Petit, J.-L.; Stange-Thomann, N.; Mauceli, E.; Bouneau, L.; Fischer, C.; Ozouf-Costaz, C.; Bernot, A.; Nicaud, S.; Jaffe, D.; Fisher, S.; Lutfalla, G.; Dossat, C.; Segurens, B.; Dasilva, C.; Salanoubat, M.; Levy, M.; Boudet, N.; Castellano, S.; Anthouard, V.; Jubin, C.; Castelli, V.; Katinka, M.; Vacherie, B.; Biemont, C.; Skalli, Z.; Cattolico, L.; Poulain, J.; de Berardinis, V.; Cruaud, C.; Duprat, S.; Brottier, P.; Coutanceau, J.-P.; Gouzy, J.; Parra, G.; Lardier, G.; Chapple, C.; McKernan, K. J.; McEwan, P.; Bosak, S.; Kellis, M.; Volff, J.-N.; Guigo, R.; Zody, M. C.; Mesirov, J.; Lindblad-Toh, K.; Birren, B.; Nusbaum, C.; Kahn, D.; Robinson-Rechavi, M.; Laudet, V.; Schachter, V.; Quetier, F.; Saurin, W.; Scarpelli, C.; Wincker, P.; Lander, E. S.; Weissenbach, J.; Roest Crollius, H. *Nature* **2004**, *431*, 946.

(35) International Chicken Genome Sequencing Consortium. *Nature* **2004**, *432*, 695.

(36) Lindblad-Toh, K.; Wade, C. M.; Mikkelsen, T. S.; Karlsson, E. K.; Jaffe, D. B.; Kamal, M.; Clamp, M.; Chang, J. L.; Kulbokas, E. J.; Zody, M. C.; Mauceli, E.; Xie, X.; Breen, M.; Wayne, R. K.; Ostrander, E. A.; Ponting, C. P.; Galibert, F.; Smith, D. R.; deJong, P. J.; Kirkness, E.; Alvarez, P.; Biagi, T.; Brockman, W.; Butler, J.; Chin, C.-W.; Cook, A.; Cuff, J.; Daly, M. J.; DeCaprio, D.; Gnerre, S.; Grabherr, M.; Kellis, M.; Kleber, M.; Bardeleben, C.; Goodstadt, L.; Heger, A.; Hitte, C.; Kim, L.; Koepfli, K.-P.; Parker, H. G.; Pollinger, J. P.; Searle, S. M. J.; Sutter, N. B.; Thomas, R.; Webber, C.; Lander, E. S. *Nature* **2005**, *438*, 803.

(37) International Human Genome Sequencing Consortium. *Nature* **2001**, *409*, 860.

(38) Mouse Genome Sequencing Consortium. *Nature* **2002**, *420*, 520.

(39) The Chimpanzee Sequencing and Analysis Consortium. *Nature* **2005**, *437*, 69.

(40) Rat Genome Sequencing Consortium. *Nature* **2004**, *428*, 493.

(41) Arabidopsis Genome Initiative. *Nature* **2000**, *408*, 796.

(42) Yu, J.; Hu, S.; Wang, J.; Wong, G. K.-S.; Li, S.; Liu, B.; Deng, Y.; Dai, L.; Zhou, Y.; Zhang, X.; Cao, M.; Liu, J.; Sun, J.; Tang, J.; Chen, Y.; Huang, X.; Lin, W.; Ye, C.; Tong, W.; Cong, L.; Geng, J.; Han, Y.; Li, L.; Li, W.; Hu, G.; Huang, X.; Li, W.; Li, J.; Liu, Z.; Li, L.; Liu, J.; Qi, Q.; Liu, J.; Li, L.; Li, T.; Wang, X.; Lu, H.; Wu, T.; Zhu, M.; Ni, P.; Han, H.; Dong, W.; Ren, X.; Feng, X.; Cui, P.; Li, X.; Wang, H.; Xu, X.; Zhai, W.; Xu, Z.; Zhang, J.; He, S.; Zhang, J.; Xu, J.; Zhang, K.; Zheng, X.; Dong, J.; Zeng, W.; Tao, L.; Ye, J.; Tan, J.; Ren, X.; Chen, X.; He, J.; Liu, D.; Tian, W.; Tian, C.; Xia, H.; Bao, Q.; Li, G.; Gao, H.; Cao, T.; Wang, J.; Zhao, W.; Li, P.; Chen, W.; Wang, X.; Zhang, Y.; Hu, J.; Wang, J.; Liu, S.; Yang, J.; Zhang, G.; Xiong, Y.; Li, Z.; Mao, L.; Zhou, C.; Zhu, Z.; Chen, R.; Hao, B.; Zheng, W.; Chen, S.; Guo, W.; Li, G.; Liu, S.; Tao, M.; Wang, J.; Zhu, L.; Yuan, L.; Yang, H. *Science* **2002**, *296*, 79.

(43) Tuskan, G. A.; DiFazio, S.; Jansson, S.; Bohlmann, J.; Grigoriev, I.; Hellsten, U.; Putnam, N.; Ralph, S.; Rombauts, S.; Salamov, A.; Schein, J.; Sterck, L.; Aerts, A.; Bhalerao, R. R.; Bhalerao, R. P.; Blaudez, D.; Boerjan, W.; Brun, A.; Brunner, A.; Busov, V.; Campbell, M.; Carlson, J.; Chalot, M.; Chapman, J.; Chen, G. L.; Cooper, D.; Coutinho, P. M.; Couturier, J.; Covert, S.; Cronk, Q.; Cunningham, R.; Davis, J.; Degroeve, S.; Dejardin, A.; dePamphilis, C.; Detter, J.; Dirks, B.; Dubchak, I.; Duplessis, S.; Ehlting, J.; Ellis, B.; Gendler, K.; Goodstein, D.; Gribskov, M.; Grimwood, J.; Groover, A.; Gunter, L.; Hamberger, B.; Heinze, B.; Helariutta, Y.; Henrissat, B.; Holligan, D.; Holt, R.; Huang, W.; Islam-Faridi, N.; Jones, S.; Jones-Rhoades, M.; Jorgensen, R.; Joshi, C.; Kangasjarvi, J.; Karlsson, J.; Kelleher, C.; Kirkpatrick, R.; Kirst, M.; Kohler, A.; Kalluri, U.; Larimer, F.; Leebens-Mack, J.; Leple, J. C.; Locascio, P.; Lou, Y.; Lucas, S.; Martin, F.; Montanini, B.; Napoli, C.; Nelson, D. R.; Nelson, C.; Nieminen, K.; Nilsson, O.; Pereda, V.; Peter, G.; Philippe, R.; Pilate, G.; Poliakov, A.; Razumovskaya, J.; Richardson, P.; Rinaldi, C.; Ritland, K.; Rouze, P.; Ryaboy, D.; Schmutz, J.; Schrader, J.; Segerman, B.; Shin, H.; Siddiqui, A.; Sterky, F.; Terry, A.; Tsai, C. J.; Uberbacher, E.; Unneberg, P. *Science* **2006**, *313*, 1596.

(44) Mirsky, A. E.; Ris, H. *J. Gen. Physiol.* **1951**, *34*, 451.

(45) Gregory, T. R.; Herbert, P. D. *Genome Res.* **1999**, *9*, 317.

(46) Thomas, C. A. *Annu. Rev. Genet.* **1971**, *5*.

(47) Selker, E. U. *Annu. Rev. Genet.* **1990**, *24*, 579.

(48) Smit, A. F. *Curr. Opin. Genet. Dev.* **1999**, *9*, 657.

(49) Finnegan, D. J. *Trends Genet.* **1989**, *5*, 103.

(50) Brookfield, J. F. Y. In *Mobile Genetic Elements*; Sherratt, D. J., Ed.; IRL Press: Oxford, 1995.

(51) Kapitonov, V. V.; Jurka, J. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 8714.

(52) Feschotte, C.; Mouches, C. *Mol. Biol. Evol.* **2000**, *17*, 730.

(53) Xiong, Y.; Eickbush, T. H. *EMBO J.* **1990**, *9*, 3353.

(54) Eickbush, T. H. In *The Evolutionary Biology of Viruses*; Morse, S. S., Ed.; Raven Press: New York, 1994.

(55) Malik, H. S.; Eickbush, T. H. *Genome Res.* **2001**, *11*, 1187.

(56) Eichler, E. E. *Trends Genet.* **2001**, *17*, 661.

(57) Samonte, R. V.; Eichler, E. E. *Nat. Rev. Genet.* **2002**, *3*, 65.

(58) Bailey, J. A.; Yavor, A. M.; Massa, H. F.; Trask, B. J.; Eichler, E. E. *Genome Res.* **2001**, *11* (6), 1005.

(59) Bailey, J. A.; Church, D. M.; Ventura, M.; Rocchi, M.; Eichler, E. E. *Genome Res.* **2004**, *14*, 789.

(60) Tuzun, E.; Bailey, J. A.; Eichler, E. E. *Genome Res.* **2004**, *14*, 493.

(61) She, X.; Horvath, J. E.; Jiang, Z.; Liu, G.; Furey, T. S.; Christ, L.; Clark, R.; Graves, T.; Gulden, C. L.; Alkan, C.; Bailey, J. A.; Sahinalp, C.; Rocchi, M.; Haussler, D.; Wilson, R. K.; Miller, W.; Schwartz, S.; Eichler, E. E. *Nature* **2004**, *430*, 857.

(62) Johnson, M. E.; Viggiano, L.; Bailey, J. A.; Abdul-Rauf, M.; Goodwin, G.; Rocchi, M.; Eichler, E. E. *Nature* **2001**, *413*, 514.

(63) Paulding, C. A.; Ruvolo, M.; Haber, D. A. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 2507.

(64) Bailey, J. A.; Eichler, E. E. *Nat. Rev. Genet.* **2006**, *7*, 552.

(65) Horvath, J. E.; Schwartz, S.; Eichler, E. E. *Genome Res.* **2000**, *10*, 839.

(66) She, X.; Jiang, Z.; Clark, R. A.; Liu, G.; Cheng, Z.; Tuzun, E.; Church, D. M.; Sutton, G.; Halpern, A. L.; Eichler, E. E. *Nature* **2004**, *431*, 927.

(67) Jiang, N.; Bao, Z.; Zhang, X.; Hirochika, H.; Eddy, S. R.; McCouch, S. R.; Wessler, S. R. *Nature* **2003**, *421*, 163.

(68) Cheng, Z.; Ventura, M.; She, X.; Khaitovich, P.; Graves, T.; Osoegawa, K.; Church, D.; DeJong, P.; Wilson, R. K.; Paabo, S.; Rocchi, M.; Eichler, E. E. *Nature* **2005**, *437*, 88.

(69) Glockner, G.; Szafranski, K.; Winckler, T.; Dingermann, T.; Quail, M. A.; Cox, E.; Eichinger, L.; Noegel, A. A.; Rosenthal, A. *Genome Res.* **2001**, *11*, 585.

(70) Quesneville, H.; Bergman, C. M.; Andrieu, O.; Autard, D.; Nouaud, D.; Ashburner, M.; Anxolabehere, D. *PLoS Comput. Biol.* **2005**, *1*, 166.

(71) International Rice Genome Sequencing Project. *Nature* **2005**, *436*, 793.

(72) Kurtz, S.; Choudhuri, J. V.; Ohlebusch, E.; Schleiermacher, C.; Stoye, J.; Giegerich, R.; Kurtz, S. *Nucleic Acids Res.* **2001**, *29*, 4633.

(73) Benson, G. *Nucleic Acids Res.* **1999**, *27*.

(74) Parisi, V.; De Fonzo, V.; Aluffi-Pentini, F. *Bioinformatics* **2003**, *19*, 1733.

(75) Kolpakov, R.; Bana, G.; Kucherov, G. *Nucleic Acids Res.* **2003**, *31*, 3672.

(76) Castelo, A. T.; Martins, W.; Gao, G. R. *Bioinformatics* **2002**, *18*, 634.

(77) Jurka, J. *Trends Genet.* **2000**, *9*, 418.

(78) Jurka, J.; Klonowski, P.; Dagman, V.; Pelton, P. *Comput. Chem.* **1996**, *20*, 119.

(79) Bedell, J. A.; Korf, I.; Gish, W. *Bioinformatics* **2000**, *16*, 1040.

(80) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. *J. Mol. Biol.* **1990**, *215*, 403.

(81) Quesneville, H.; Nouaud, D.; Anxolabehere, D. *J. Mol. Evol.* **2003**, S50.

(82) Andrieu, O.; Fiston, A. S.; Anxolabehere, D.; Quesneville, H. *BMC Bioinf.* **2004**, *5*, 94.

(83) McCarthy, E. M.; McDonald, J. F. *Bioinformatics* **2003**, *19*, 362.

(84) Massimiliano, M. R.; Caizzi, R. *Gene* **2005**, *357*, 115.

(85) McCarthy, E. M.; McDonald, J. F. *Genome Biol.* **2004**, *5*, R14.

(86) McCarthy, E. M.; Liu, J.; Lizhi, G.; McDonald, J. F. *Genome Biol.* **2002**, *3*, R53.

(87) Kalyanaraman, A.; Aluru, S. *J. Bioinform. Comput. Biol.* **2006**, *4*, 197.

(88) Volfovsky, N.; Haas, B. J.; Salzberg, S. L. *Genome Biol.* **2001**, *2*, 27.

(89) Bao, Z.; Eddy, S. R. *Genome Res.* **2002**, *12*.

(90) Pevzner, P. A.; Tang, H.; Tesler, G. *Genome Res.* **2004**, *14*, 1786.

(91) Price, A. L.; Eskin, E.; Pevzner, P. A. *Genome Res.* **2004**, *14*, 2245.

(92) Edgar, R. C.; Myers, E. W. *Bioinformatics* **2005**, *21*, 152.

(93) Price, A. L.; Jones, N. C.; Pevzner, P. A. *Bioinformatics* **2005**, *21*, i351.

(94) Chimpanzee Sequencing and Analysis Consortium. *Nature* **2005**, *437*, 69.

(95) Caspi, A.; Pachter, L. *Genome Res.* **2006**, *16*, 260.

(96) Bailey, J. A.; Gu, Z.; Clark, R. A.; Reinert, K.; Samonte, R. V.; Schwartz, S.; Adams, M. D.; Myers, E. W.; Li, P. W.; Eichler, E. E. *Science* **2002**, *297*, 1003.

(97) Bailey, J. A.; Yavor, A. M.; Viggiano, L.; Misceo, D.; Horvath, J. E.; Archidiacono, N.; Schwartz, S.; Rocchi, M.; Eichler, E. E. *Am. J. Hum. Genet.* **2002**, *70*, 83.

(98) Negroni, M.; Buc, H. *Annu. Rev. Genet.* **2001**, *35*, 275.

(99) Zhi, D.; Raphael, B. J.; Price, A. L.; Tang, H.; Pevzner, P. A. *Genome Biol.* **2006**, *7*, R7.

(100) Sanger, F.; Donelson, J. E.; Coulson, A. R.; Kossel, H.; Fischer, D. *J. Mol. Biol.* **1974**, *90*, 315.

(101) Sanger, F.; Coulson, A. R.; Hong, G. F.; Hill, D. F.; Petersen, G. B. *J. Mol. Biol.* **1982**, *162*, 729.

(102) Fleischmann, R. D.; Adams, M. D.; White, O.; Clayton, R. A.; Kirkness, E. F.; Kerlavage, A. R.; Bult, C. J.; Tomb, J. F.; Dougherty, B. A.; Merrick, J. M.; Venter, J. *Science* **1995**, *269*, 296.

(103) Green, P. http://www.genome.washington.edu/UWGC/analysis-tools/phrap.htm, 1994.

(104) Bonfield, J. K.; Smith, K. F.; Staden, R. *Nucleic Acids Res.* **1995**, *23*, 4992.

(105) Kececioglu, J.; Myers, E. W. *Algorithmica* **1995**, *13*, 7.

(106) Myers, E. W. *J. Comput. Biol.* **1995**, *2*, 275.

(107) Huang, X.; Madan, A. *Genome Res.* **1999**, *9*, 868.

(108) Kim, S.; Segre, A. M. *J. Comput. Biol.* **1999**, *6*, 163.

(109) Chen, T.; Skiena, S. S. *Bioinformatics* **2000**, *16*, 494.

(110) Lander, E. S.; Waterman, M. S. *Genomics* **1988**, *2*, 231.

(111) Pop, M.; Salzber, S. L.; Shumway, M. *IEEE Comput.* **2002**, *35*, 47.

(112) Zhi, D.; Keich, U.; Pevzner, P. A.; Heber, S.; Tang, H. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2007**, *4*, 54.

(113) Weber, J. L.; Myers, E. W. *Genome Res.* **1997**, *7*, 401.

(114) Green, P. *Genome Res.* **1997**, *7*, 410.

(115) Kent, W. J.; Haussler, D. *Genome Res.* **2001**, *11*, 1541.

(116) Myers, E. W.; Sutton, G. G.; Delcher, A. L.; Dew, I. M.; Fasulo, D. P.; Flanigan, M. J.; Kravitz, S. A.; Mobarry, C. M.; Reinert, K. H.; nbsp; J.; Remington, K. A.; Anson, E. L.; Bolanos, R. A.; Chou, H.-H.; Jordan, C. M.; Halpern, A. L.; Lonardi, S.; Beasley, E. M.; Brandon, R. C.; Chen, L.; Dunn, P. J.; Lai, Z.; Liang, Y.; Nusskern, D. R.; Zhan, M.; Zhang, Q.; Zheng, X.; Rubin, G. M.; Adams, M. D.; Venter, J. C. *Science* **2000**, *287*, 2196.

(117) Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J.; Sutton, G. G.; Smith, H. O.; Yandell, M.; Evans, C. A.; Holt, R. A.; Gocayne, J. D.; Amanatides, P.; Ballew, R. M.; Huson, D. H.; Wortman, J. R.; Zhang, Q.; Kodira, C. D.; Zheng, X. H.; Chen, L.; Skupski, M.; Subramanian, G.; Thomas, P. D.; Zhang, J.; Gabor Miklos, G. L.; Nelson, C.; Broder, S.; Clark, A. G.; Nadeau, J.; McKusick, V. A.; Zinder, N.; Levine, A. J.; Roberts, R. J.; Simon,

M.; Slayman, C.; Hunkapiller, M.; Bolanos, R.; Delcher, A.; Dew, I.; Fasulo, D.; Flanigan, M.; Florea, L.; Halpern, A.; Hannenhalli, S.; Kravitz, S.; Levy, S.; Mobarry, C.; Reinert, K.; Remington, K.; Abu-Threideh, J.; Beasley, E.; Biddick, K.; Bonazzi, V.; Brandon, R.; Cargill, M.; Chandramouliswaran, I.; Charlab, R.; Chaturvedi, K.; Deng, Z.; Francesco, V. D.; Dunn, P.; Eilbeck, K.; Evangelista, C.; Gabrielian, A. E.; Gan, W.; Ge, W.; Gong, F.; Gu, Z.; Guan, P.; Heiman, T. J.; Higgins, M. E.; Ji, R.-R.; Ke, Z.; Ketchum, K. A.; Lai, Z.; Lei, Y.; Li, Z.; Li, J.; Liang, Y.; Lin, X.; Lu, F.; Merkulov, G. V.; Milshina, N.; Moore, H. M.; Naik, A. K.; Narayan, V. A.; Neelam, B.; Nusskern, D.; Rusch, D. B.; Salzberg, S.; Shao, W.; Shue, B.; Sun, J.; Wang, Z. Y.; Wang, A.; Wang, X.; Wang, J.; Wei, M.-H.; Wides, R.; Xiao, C.; Yan, C. *Science* **2001**, *291*, 1304.

(118) Li, X.; Waterman, M. S. *Genome Res.* **2003**, *13*, 1916.

(119) Wang, J.; Wong, G. K.; Ni, P.; Han, Y.; Huang, X.; Zhang, J.; Ye, C.; Zhang, Y.; Hu, J.; Zhang, K.; Xu, X.; Cong, L.; Lu, H.; Ren, X.; Ren, X.; He, J.; Tao, L.; Passey, D. A.; Wang, J.; Yang, H.; Yu, J.; Li, S. *Genome Res.* **2002**, *12*, 824.

(120) Batzoglou, S.; Jaffe, D. B.; Stanley, K.; Butler, J.; Gnerre, S.; Mauceli, E.; Berger, B.; Mesirov, J. P.; Lander, E. S. *Genome Res.* **2002**, *12*, 177.

(121) Jaffe, D. B.; Butler, J.; Gnerre, S.; Mauceli, E.; Lindblad-Toh, K.; Mesirov, J. P.; Zody, M. C.; Lander, E. S. *Genome Res.* **2003**, *13*, 91.

(122) Havlak, P.; Chen, R.; Durbin, K. J.; Egan, A.; Ren, Y.; Song, X. Z.; Weinstock, G. M.; Gibbs, R. A. *Genome Res.* **2004**, *14*, 721.

(123) Mullikin, J. C.; Ning, Z. *Genome Res.* **2003**, *13*, 81.

(124) Huson, D. H.; Reinert, K.; Kravitz, S. A.; Remington, K. A.; Delcher, A. L.; Dew, I. M.; Flanigan, M.; Halpern, A. L.; Lai, Z.; Mobarry, C. M.; Sutton, G. G.; Myers, E. W. *Bioinformatics* **2001**, *17*, S132.

(125) Pop, M.; Kosack, D. S.; Salzberg, S. L. *Genome Res.* **2004**, *14*, 149.

(126) Pevzner, P. A.; Tang, H.; Waterman, M. S. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 9748.

(127) Kececioglu, J.; Yu, J. Annual Conference on Research in Computational Molecular Biology (RECOMB'01), Montreal, Canada, 2001; p 176.

(128) Tammi, M. T.; Arner, E.; Britton, T.; Andersson, B. *Bioinformatics* **2002**, *18*, 379.

(129) Sankoff, D.; El-Mabrouk, N. In *Current Topics in Computational Biology*; Jiang, T., Xu, Y., Zhang, M., Eds.; MIT Press: Cambridge, MA, 2002.

(130) Sturtevant, A. H. *Publ. Carnegie Inst. Washington* **1931**, *421*, 1.

(131) Sturtevant, A. H.; Dobzhansky, T. *Proc. Natl. Acad. Sci. U.S.A.* **1936**, *22*, 448.

(132) Dobzhansky, T. H.; Sturtevant, A. H. *Genetics* **1938**, *23*, 28.

(133) O'Brien, S. J.; Menotti-Raymond, M.; Murphy, W. J.; Nash, W. G.; Wienberg, J.; Stanyon, R.; Copeland, N. G.; Jenkins, N. A.; Womack, J. E.; Graves, J. A. *Science* **1999**, *286*, 458.

(134) Nadeau, J. H.; Taylor, B. A. *Proc. Natl. Acad. Sci. U.S.A.* **1984**, *81*, 814.

(135) Mural, R. J.; Adams, M. D.; Myers, E. W.; Smith, H. O.; Miklos, G. L.; Wides, R.; Halpern, A.; Li, P. W.; Sutton, G. G.; Nadeau, J. *Science* **2002**, *296*, 1661.

(136) Pevzner, P. A.; Tesler, G. P. *Genome Res.* **2003**, *13*, 37.

(137) Kent, W. J.; Baertsch, R.; Hinrichs, A.; Miller, W.; Haussler, D. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 11484.

(138) Bourque, G.; Pevzner, P. A.; Tesler, G. *Genome Res.* **2004**, *14*, 507.

(139) Palmer, J. D.; Herbon, L. A. *J. Mol. Evol.* **1988**, *27*.

(140) Kececioglu, J.; Sankoff, D. *Algorithmica* **1995**, *13*, 180.

(141) Bafna, V.; Pevzner, P. A. the 34th Annual IEEE Symposium on Foundations of Computer Science, 1993; p 148.

(142) Bafna, V.; Pevzner, P. A. *Mol. Biol. Evol.* **1995**, *12*, 239.

(143) Bafna, V.; Pevzner, P. A. *SIAM J. Comput.* **1996**, *25*, 272.

(144) Hannenhalli, S.; Pevzner, P. A. 27th Annual ACM Symposium on the Theory of Computing, 1995; p 178.

(145) Hannenhalli, S.; Pevzner, P. A. *J. Assoc. Comput. Mach.* **1999**, *46*, 1.

(146) Hannenhalli, S.; Pevzner, P. A. The 36th Annual IEEE Symposium on Foundations of Computer Science, 1995; p 581.

(147) Tesler, G. *J. Comp. Syst. Sci.* **2002**, *65*, 587.

(148) Copeland, N. G.; Jenkins, N. A.; Gilbert, D. J.; Eppig, J. T.; Maltais, L. J.; Miller, J. C.; Dietrich, W. F.; Weaver, A.; Lincoln, S. E.; Steen, R. G.; Stein, L. D.; Nadeau, J. H.; Lander, E. S. *Science* **1993**, *262*, 57.

(149) Pevzner, P. A.; Tesler, G. P. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 7672.

(150) Sankoff, D.; Trinh, P. *J. Comput. Biol.* **2005**, *12*, 812.

(151) Peng, Q.; Pevzner, P. A.; Tesler, G. *PLoS Comput. Biol.* **2006**, *2*, e14.

(152) Stankiewicz, P.; Lupski, J. R. *Trends Genet.* **2002**, *18*, 74.

(153) Shaw, C. J.; Lupski, J. R. *Hum. Mol. Genet.* **2004**, *13*, R57.

(154) Myers, S. R.; McCarroll, S. A. *Nat. Genet.* **2006**, *38*, 1363.

(155) Lupski, J. R. *Am. J. Hum. Genet.* **2003**, *72*, 246.

(156) Bailey, J. A.; Liu, G.; Eichler, E. E. *Am. J. Hum. Genet.* **2003**, *73*, 823.

CR0683008