View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Digital Commons@Becker

Washington University School of Medicine
Digital Commons@Becker

Open Access Publications

2005

The repetitive landscape of the chicken genome

Thomas Wicker
University of Georgia

Jon S. Robertson
University of Georgia

Stefan R. Schulze
University of Georgia

F. Alex Feltus
University of Georgia

Vincent Magrini
Washington University School of Medicine in St. Louis

See next page for additional authors

Follow this and additional works at: http://digitalcommons.wustl.edu/open_access_pubs

Recommended Citation
Wicker, Thomas; Robertson, Jon S.; Schulze, Stefan R.; Feltus, F. Alex; Magrini, Vincent; Morrison, Jason A.; Mardis, Elaine R.; Wilson, Richard K.; Peterson, Daniel G.; Paterson, Andrew H.; and Ivarie, Robert, ,"The repetitive landscape of the chicken genome." Genome Research.15,. 126-136. (2005).
http://digitalcommons.wustl.edu/open_access_pubs/2082

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact engeszer@wustl.edu.

Authors
Thomas Wicker, Jon S. Robertson, Stefan R. Schulze, F. Alex Feltus, Vincent Magrini, Jason A. Morrison, Elaine R. Mardis, Richard K. Wilson, Daniel G. Peterson, Andrew H. Paterson, and Robert Ivarie

This open access publication is available at Digital Commons@Becker: http://digitalcommons.wustl.edu/open_access_pubs/2082

Downloaded from genome.cshlp.org on January 21, 2014 - Published by Cold Spring Harbor Laboratory Press

GENOME RESEARCH

The repetitive landscape of the chicken genome

Thomas Wicker, Jon S. Robertson, Stefan R. Schulze, et al.

Genome Res. 2005 15: 126-136
Access the most recent version at doi:10.1101/gr.2438004

Supplemental Material
http://genome.cshlp.org/content/suppl/2004/12/10/gr.2438004.DC2.html
http://genome.cshlp.org/content/suppl/2004/10/18/gr.2438004.DC1.html

References
This article cites 42 articles, 14 of which can be accessed free at:
http://genome.cshlp.org/content/15/1/126.full.html#ref-list-1

Creative Commons License
This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at http://creativecommons.org/licenses/by-nc/3.0/.

Email Alerting Service
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on January 21, 2014 - Published by Cold Spring Harbor Laboratory Press

The repetitive landscape of the chicken genome

Thomas Wicker,[1,5] Jon S. Robertson,[1,5] Stefan R. Schulze,[1] F. Alex Feltus,[1] Vincent Magrini,[3] Jason A. Morrison,[3] Elaine R. Mardis,[3] Richard K. Wilson,[3] Daniel G. Peterson,[1,4] Andrew H. Paterson,[1,2,6] and Robert Ivarie[2,6]

[1]*Plant Genome Mapping Laboratory and* [2]*Department of Genetics, University of Georgia, Athens, Georgia 30602, USA;* [3]*Genome Sequencing Center, Washington University Medical Center, Washington University, St. Louis, Missouri 63108, USA;* [4]*Mississippi Genome Exploration Laboratory, Mississippi State University, Starkville, Mississippi 39762, USA*

Cot-based cloning and sequencing (CBCS) is a powerful tool for isolating and characterizing the various repetitive components of any genome, combining the established principles of DNA reassociation kinetics with high-throughput sequencing. CBCS was used to generate sequence libraries representing the high, middle, and low-copy fractions of the chicken genome. Sequencing high-copy DNA of chicken to about 2.7× coverage of its estimated sequence complexity led to the initial identification of several new repeat families, which were then used for a survey of the newly released first draft of the complete chicken genome. The analysis provided insight into the diversity and biology of known repeat structures such as *CRI* and *CNM*, for which only limited sequence data had previously been available. Cot sequence data also resulted in the identification of four novel repeats (*Birddawg*, *Hitchcock*, *Kronos*, and *Soprano*), two new subfamilies of *CRI* repeats, and many elements absent from the chicken genome assembly. Multiple autonomous elements were found for a novel *Mariner*-like transposon, *Galluhop*, in addition to nonautonomous deletion derivatives. Phylogenetic analysis of the high-copy repeats *CRI*, *Galluhop*, and *Birddawg* provided insight into two distinct genome dispersion strategies. This study also exemplifies the power of the CBCS method to create representative databases for the repetitive fractions of genomes for which only limited sequence data is available.

[Supplemental material is available online at www.genome.org and http://plantgenome.agtec.uga.edu/g4g. The sequence data described in this study have been submitted to GenBank under accession nos. CL266240–CL281342. Consensus sequences for the novel repeat families and their major subfamilies were submitted to RepBase.]

The domestic chicken (*Gallus gallus*) provides a major protein source for most human populations throughout the world. Its economic importance has made it the focus of numerous research projects, including a recent effort to sequence the entire chicken genome (http://genome.wustl.edu/projects/chicken). A first draft of the complete chicken genome was made public in March 2004. Approximately 88% of the sequence has been anchored to chromosomes, which include autosomes 1–24, 26–28, and 32, and sex chromosomes W and Z. The remaining unanchored contigs have been concatenated into the virtual chromosome *ChrUn*.

Like most bird species, the chicken has a relatively small genome of ~1200 million base pairs (Mbp), or ~39% of the size of the human genome, and has been shown to contain only ~15% repetitive DNA as measured by reassociation kinetics (Epplen et al. 1978; Schmid et al. 2000). The repetitive DNA fraction, sometimes referred to as "junk DNA," includes short tandem repeats (e.g., telomeric and centromeric repeats) as well as numerous families of interspersed repeats that are often derived from transposable elements. Transposable elements are subdivided on the basis of whether their mobility involves an mRNA intermediate (Class 1) or the DNA itself (Class 2). Class 1 elements are further subdivided on the basis of the presence or absence of long terminal repeats (LTR). Class 1 non-LTR retrotransposons are also referred to as long and short interspersed nuclear elements (LINEs and SINEs), respectively.

It has been speculated that the relatively small genome size of birds in general, and chickens in particular, may reflect selective pressure to optimize metabolism and to minimize the amount of repetitive DNA (Gregory 2002). The most abundant DNA elements known in the chicken genome to date are non-LTR retrotransposons of the *CR1* family, present in an estimated 100,000 copies (Vandergon and Reitman 1994). Six distinct subfamilies of *CR1* (*CR1-A* through *CR1-F*) have been identified previously (Vandergon and Reitman 1994). A complete *CR1* element is ~4.5 kb in length and contains two protein-coding sequences (Haas et al. 1997). ORF-2 encodes a reverse transcriptase as found in many other LINEs in animals and plants, and is responsible for the replication of the element. The exact function of ORF-1 is not yet known. It encodes a protein with a conserved esterase domain and is believed to play a role in a multitude of processes such as protein–protein interactions involved in transcriptional regulation or the horizontal transfer of *CR1* elements (Kapitonov and Jurka 2003). Most chicken *CR1* repeats are truncated at their 5′ ends, so that *CR1* fragments typically contain only a few hundred base pairs. The fact that these 3′ fragments of *CR1* repeats are enriched in the chicken genome may indicate the presence of regulatory (or otherwise advantageous) sequences within this repeat (Stumph et al. 1984). Alternatively, the predominance of 3′ fragments might be a consequence of the premature termination

[5]These two authors contributed equally to this work.
[6]Corresponding authors.
E-mail ivarie@uga.edu; fax (706) 542-3910.
E-mail paterson@uga.edu; fax (706) 583-0160.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.2438005. Article published online before print in July 2004.

15:126–136 ©2005 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/05; www.genome.org

Downloaded from genome.cshlp.org on January 21, 2014 - Published by Cold Spring Harbor Laboratory Press

of reverse transcription, which would result in the integration of an already truncated (dead-on-arrival) element (Petrov and Hartl 1998).

The second main group of class 1 elements (LTR-retrotransposons) represent only a minor fraction of the repetitive DNA in large vertebrate genomes (Lander et al. 2001; Waterston et al. 2002), whereas non-LTR retrotransposons comprise the vast majority of repetitive elements. This is in contrast to the pattern found in plants with large genomes such as maize (SanMiguel et al. 1998) or wheat (Wicker et al. 2001), in which most of the genome can be comprised of LTR retrotransposons. The main characteristic of LTR retrotransposons is that an internal domain containing the coding region for the proteins necessary for replication is flanked by direct repeats of usually several hundred base pairs. These LTR sequences provide promoter and 3′ regions, and are essential in the replication process.

Little is known to date about the contribution to the repetitive elements of the chicken genome made by Class 2 elements (or DNA transposons), which move via excision of the DNA element itself from the genome and integration at a new location through a cut-and-paste mechanism mediated by transposases. Replication of these elements is achieved in a complex process called replicative transposition (see Lewin 1994). To date, 31 chicken repetitive sequences have been deposited into RepBase (Jurka 2000), only two of which represent class 2 transposons (*Mariner1a* and *Mariner1b*).

Another major repeat class populating the chicken genome is short tandem repeats. One example is the telomeric repeat, which consists of tandem arrays of many thousands of copies of the short sequence motif AATGGG. Compared with other vertebrates, chicken chromosomes contain large amounts of telomeric DNA, which by some estimates, comprise 3%–4% of the genome (Delaney et al. 2003). Centromeres and subtelomeric regions also contain multiple classes of larger tandemly repeated units. The most abundant repeat type found in chicken is *CNM* (Matzke et al. 1990), a tandem repeat with a unit size of ~40 bp. *CNM* repeats are members of a large superfamily that also includes other tandem repeats such as *PIR* (Wang et al. 2002) and *XhoI* (Klein and Ellendorf 2000). They show structural and sequence similarity with *CNM* repeats and are therefore believed to share a common ancestor.

The cost of whole-genome sequencing projects in higher eukaryotes combined with the complications introduced by abundant repetitive DNA have motivated the development of a method to fractionate genomes into libraries of DNA components that differ substantially in copy number (Peterson et al. 2001, 2002a,b). Validation of this method by Peterson et al. (2001, 2002a), referred to as Cot-based cloning and sequencing (CBCS), has been followed by its application to exploratory (Yuan et al. 2003) and extensive (Whitelaw et al. 2003) analysis of low-copy sequences from genomes with large repetitive components. Peterson et al. (2002a) also showed the potential value of CBCS for studying repetitive DNA—only 253 sequencing reads from an highly repetitive (HR) Cot library provided representative sequences from DNA families comprising 15% of the sorghum genome. CBCS has also been used on a smaller scale by others for repeat analysis (Ho and Leung 2002).

We report here on the application of a combination of CBCS and bioinformatics to characterize the repetitive landscape of the chicken genome. Sequencing the high-copy DNA of chicken to ~2.7× coverage of its estimated sequence complexity enabled the rapid identification of several new repeat families, including some that are absent from the whole-genome assembly. The repeats identified in this approach were then used in a whole-genome survey. The study of copy number and distribution of these repeats across the chicken genome provided insight into their diversity and biology. In addition, libraries for middle- and low-repetitive sequences were generated for comparison with the high-copy sequences and characterization of low-copy repeats. The relatively low repetitive DNA content of chicken provides an especially stringent test of the ability to fractionate DNA on the basis of copy number by CBCS, and the availability of a whole-genome sequence permits evaluation of what others can expect to learn by applying CBCS to unsequenced genomes.

Results

Three different cot fractions of the chicken genome

A total of 15,103 sequences representing 3.83 Mbp were derived from three Cot fractions as follows: HR (highly repetitive), MR (middle repetitive), and SL (single/low-copy). The fractions were based on previous analyses of the reassociation kinetics of chicken genomic DNA. The Cot curve of the chicken genome is a two-component curve, with a highly repetitive component with a $Cot_{1/2}$ of roughly 1 and a low-copy component with a $Cot_{1/2}$ of ~1000 (Epplen et al. 1978). On the basis of the two Cot decade principle (see Peterson et al. 2002a), 80% of the DNA in a particular component should reassociate within the range 0.1γ to 10γ, where γ is the $Cot_{1/2}$ value of the component. Consequently, HR DNA was prepared by isolating those sequences that renature between 0.1 and 10 M·s; SL DNA was isolated by collecting sequences that do not renature by Cot 100 (i.e., the 0.1γ value for the low-copy component); MR DNA was collected from the small region of the curve not included in the HR and SL components (i.e., Cot 10 to Cot 100), and is thus an ad hoc component of the chicken genome. The fraction obtained at a Cot below 0.1 is thought to be foldback DNA (Britten et al. 1974) and was set aside.

The HR component of the chicken genome represents ~10% of total genomic DNA (Epplen et al. 1978). On the basis of the chicken Cot curve (Epplen et al. 1978), the HR component has a sequence complexity of ~120 kb (see Methods). We sequenced 1520 HR plasmids with a mean insert size of 213 bp for a total of 324 kb, and thus provided (324 kb ÷ 120 kb =) 2.7-fold coverage of the sequence complexity of the HR component.

The SL component should contain 90% of chicken's single-copy sequences. We sequenced only enough SL plasmids (3038 sequences, for a total of 646 kb) to provide a sufficient quantity of sequence data to allow comparison of the SL library with those derived from the repetitive fractions. To make their comparison equitable, the HR and SL components were drawn from the same population of DNA fragments, which had been sheared to a level thought to be sufficient that most fragments would contain only one repetitive element. (We note that if one were to apply the method to complete sequencing of the SL fraction, naturally one would prefer longer fragments than are suitable for repetitive DNA analysis [Peterson et al. 2002a], and, in fact, would best use multiple libraries from DNA populations of different lengths.)

The MR fraction represents sequences reannealing between Cot 10 and Cot 100, thus spanning the gap in the chicken Cot curve between the least repetitive HR sequences and the most repetitive SL. Because the MR fraction is an ad hoc rather than a specific Cot component, it lacks a true $Cot_{1/2}$. To estimate its

Downloaded from genome.cshlp.org on January 21, 2014 - Published by Cold Spring Harbor Laboratory Press

sequence complexity (2900 kb), the arithmetic middle between the borders of the HR and SL fractions (Cot 55) was used in lieu of a $Cot_{1/2}$. Relative to the HR fraction, the MR fraction is expected to contain repeat types that are present at lower copy numbers, as is reflected in its much larger calculated sequence complexity (2900 kb). Our 10,545 sequences cover a total of 2838 kb, and thus provide ~1× coverage of this fraction.

General characteristics of the three cot libraries

A first general characteristic that distinguished the three Cot fractions was their overall GC (G+C) content. The GC content for each Cot fraction was calculated using the average GC content for all individual sequences. The SL fraction showed the lowest GC content of 40.1%, close to the estimate of 40.7% for the entire chicken genome (Epplen et al. 1978). The two repetitive fractions showed significantly higher ($P < 0.01$, as determined by single-factor ANOVA) GC contents of 43.9% (MR) and 51.6% (HR). The GC content of both the SL and MR fractions showed nearly Gaussian distributions, with most sequences having GC content ranging from 20% to 60% and from 30% to 60%, respectively. The HR fraction displays a less homogenous and broader distribution with GC content ranging between 30% and 80%.

To determine the contents of repetitive and low-copy sequences in the respective fractions, all sequences were used in a BLASTN search against the entire chicken genome, and hits to the genome for each sequence were counted. The sequences were then segregated into five repeat classes on the basis of their copy number in the genome: 1–10 copies, 11–100 copies, 101–1000 copies, 1001–10,000 copies, and >10,000 copies. An additional class was formed with sequences that do not match the publicly available genomic sequence. As expected, the content of repetitive sequences was greatest in the HR fraction, as more than 36% of the sequences are present in the genome in more than 10 copies. Interestingly, more than one-third of the HR sequences (536 sequences) are not present in the publicly available genome sequence. Analysis of this set revealed that 33 contain low-complexity sequences (e.g., microsatellites). A group of 21 sequences that clustered together we referred to as *HR_Rep1* (Supplemental material). The rest of the 536 sequences had no obvious characteristics in common.

The MR and SL fractions are very similar to one another with regard to their repeat class content (Fig. 1A). The only notable differences between the two are the slightly higher percentage of sequences with 101–1000 copies in the MR, and the higher percentage of sequences that do not hit the genome in the SL fraction. However, both the MR and SL fraction differ greatly from the HR fraction in their sequence composition, especially with respect to the no hit and the 1–10 copies classes (Fig. 1A).

The gene content of the three fractions was estimated by BLASTX of all sequences against a database containing more than 25,000 predicted proteins for the chicken genome (geneid; http://genome.ucsc.edu). Only hits that were more than 30 acids long and >80% identical were considered. With these stringent criteria, 2.5% of the HR, 4.1% of the MR, and 2.1% of the SL fraction were suggested to contain putative gene sequences.

Cot components are more heterogeneous than is predicted by DNA reassociation models

The 2.7-fold coverage of the sequence complexity of the HR fraction that we have studied would be predicted (on the basis of the Poisson probability distribution function) to provide at least

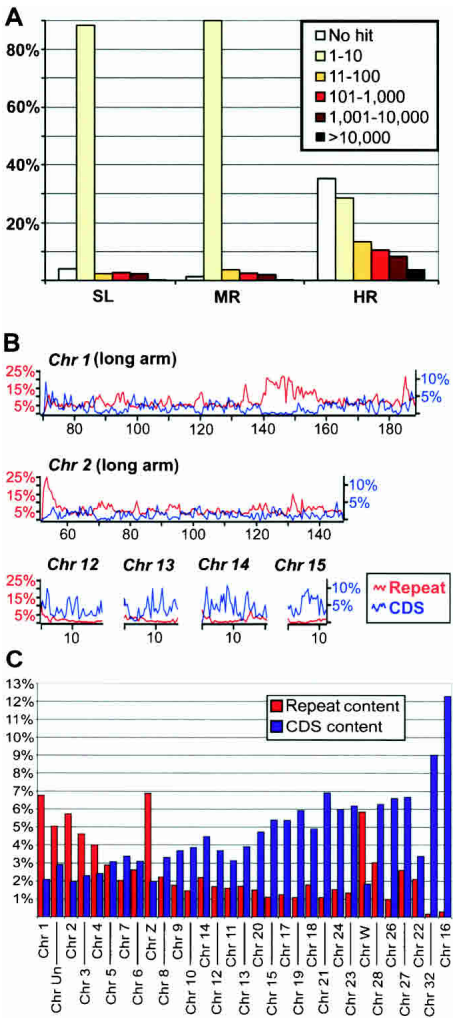

Figure 1. Repetitive DNA content in Cot and genomic sequences. (*A*) Distribution of repeats based on estimated copy numbers for the entire chicken genome (see Results). The *y*-axis indicates the percentage of sequences that contain a respective repeat class for each Cot fraction. The class "No Hit" refers to sequences that had no match in the publicly available genome sequence. (*B*) Repeat and gene coding sequences (CDS) content of chicken chromosomes. The repeat and gene content was calculated in windows of 500 kb along the chromosomes. Note that the scale for the CDS content differs from the one for the repeat content. (*C*) Average repeat and gene content for individual chicken chromosomes. The chromosomes are ordered by size with the largest chromosome (Chromosome 1) on the *left* and the smallest (Chromosome 16) on the *right*. Chromosomes and fragments of chromosomes smaller than 1 Mbp are not included.

one sequence representing 93.3% of repetitive DNA families, and multiple sequences for 75.1% of families, if all DNA families in the HR fraction had the same iteration frequencies. The repeat families found in our sample occurred on an average of 9.12 times (on the basis of a BLASTN search of all HR sequences against themselves), suggesting generally higher abundance than would have been predicted. However, 42% of reads matched only themselves (vs. the expected 18%), suggesting generally lower abundance than predicted. This seeming discrepancy is actually logical, simply reflecting heterogeneity of the HR (or any) fraction. On the basis of the two-Cot-decade principle, the least abundant (slowest reannealing) families in the fraction

Downloaded from genome.cshlp.org on January 21, 2014 - Published by Cold Spring Harbor Laboratory Press

should be 10-fold less abundant than the average, and would thus be expected to be missed entirely in 76.3% of cases (again, on the basis of the Poisson probability distribution function). In contrast, the most abundant elements in the fraction, 10-fold more abundant than average, should be represented by more than 20 reads in 96.8% of cases. Hence, our 2.7-fold coverage is completely satisfactory (even somewhat excessive) for the more abundant elements in the HR fraction, but may have missed rare elements.

Eight types of repeats contribute ~4.3% to the chicken genome

The approaches for identifying novel repetitive elements (see Methods) were used to characterize a set of seven transposable elements (Table 1). Among this group, four, namely *Birddawg*, *Hitchcock*, *Soprano*, and *Kronos* show no similarity to previously described chicken repeats. Both *Galluhop* and *Charlie* are similar to repeats found in RepBase (Mariner1b and Charlie12, respectively). All of the above, plus the publicly available *CR1* sequences, were used for multiple rounds of iterative BLAST searches in order to identify divergent subfamilies for each type. An additional repeat type includes tandem repeats that are generally associated with the telomeric regions. The true telomeric repeat itself, a putative subtelomeric repeat (*PO41*) and the previously described *CNM* (Matzke et al. 1990), *PIR* (Wang et al. 2002), and *XhoI* (Klein and Ellendorf 2000) families were all grouped together as telomeric repeats.

To estimate their copy number and contribution to the total genomic sequence, all of the available DNA sequences for each repeat (novel or previously characterized) were used in a BLASTN search against the chicken genome. The multiple BLAST reports were treated as described below to obtain a total genome count for each repeat type (see Methods). As expected, *CR1* repeats were the most abundant elements in the genome. With 96,230 copies, they contribute ~3.1% to the total genome sequence (Table 1), which agrees well with a previous estimate of 100,000 copies (Vandergon and Reitman 1994). The second most abundant element is the Class 2 transposon *Galluhop* with 13,729 identified copies (Table 1). The other repeats (*Birddawg*, *Charlie*, *Hitchcock*, *Kronos*, and *Soprano*) were found in copy numbers ranging from 1362 to 7404. Telomeric repeats (*CNM*, *XhoI*, *PIR*, and *PO41*) contribute only about 0.1% to the publicly available genome sequence (Table 1), but comprise >11% of the HR cot sequences. The eight repeat types contribute a total of >52 Mbp (4.3%) to the entire genome sequence.

Table 1. Major repeat types in the chicken genome

Name	Class[a]	Copy Number	Total bp
CR1	LINE	96,230	37,160,469 (3.10%)
Galluhop	Class 2/*Mariner*	13,729	6,140,519 (0.51%)
Birddawg	LTR/*gypsy*	7,404	2,697,928 (0.22%)
Kronos	LTR/*gypsy*	4,961	3,021,541 (0.25%)
Hitchcock	LTR?	3,324	811,951 (0.07%)
Charlie	Class 2	2,292	1,203,639 (0.10%)
Soprano	LTR	1,362	768,648 (0.06%)
Telomeric	tandem	362[b]	252,835 (>0.1%)
Total		129,351	52,057,530 (4.31%)

The copy number indicates the number of identified repeat units in the first draft of the publicly available chicken genome sequence. The number in parentheses is the total base-pair count for each repeat in percent of the whole genome.
[a]A question mark indicates that the classification is uncertain.
[b]Copy number for telomeric repeats refers to the number of identified arrays of multiple tandem-repeated units.

Repeats are distributed unevenly across the chicken genome

Positional information for more than 129,000 identified repeat units was used for a graphical illustration of their distribution along the chicken chromosomes and combined with the annotation for predicted genes (geneid, http://genome.ucsc.edu). For the illustration in Figure 1B, the content of repeats and predicted gene-coding sequences was calculated in 500-kb windows along the chromosome. A complete illustration for all chromosomes is available as Supplemental material (Fig. 6, below). The density of identified repeats and predicted genes shows large fluctuations, even at a local level over much of the genome (Figs. 1B and 6, below). One apparent exception to this pattern is a region on the long arm of chromosome 1, which displays high-repeat density and low-gene density over a stretch of ~20 Mbp (positions 140–160 Mbp; Fig. 1B). The average repeat and gene content for all chromosomes was also calculated (Fig. 1C). In general, small chromosomes contain more genes and less repeats relative to their size than the large chromosomes (examples in Fig. 1B). Here, the exceptions are the sex chromosomes W and Z, which show a much higher repeat content than chromosomes of comparable size. These findings are in agreement with previous genetic mapping studies (Schmid et al. 2000). In some cases, repeat density increases near telomeres and centromeres; for example, the repeat content close to the centromere of chromosome 2 is >25% (Fig. 1B).

The most striking irregularity was found in the distribution of tandem repeats. The only tandem repeat types in considerable copy numbers were *CNM* and an element we refer to as *PO41* (Pattern of 41). The latter consists of two motifs of 10 and 11 bp, respectively (Fig. 2A). In most cases, three units of the 10-bp motif are followed by one unit of the 11-bp motif, creating a 41-bp superstructure. Both *CNM* and *PO41* were found in large arrays of dozens of units covering several kilobases of genomic sequences. A total of 95 arrays of *CNM* and 259 arrays of *PO41* were identified. Interestingly, all but one of those arrays were located on the virtual chromosome (*ChrUn*). A dot plot of an array consisting of 21 *CNM* units is depicted in Figure 2B. In two instances, arrays *CNM* and *PO41* were immediately neighboring each other (Fig. 2C), suggesting a possible general association of the two repeat types. As *CNM* has been shown to be associated with telomeric and subtelomeric regions, it is possible that *PO41* is also found mainly in these regions of the genome. Examination of the flanking sequences showed that all but four of the *CNM* arrays, and more than half of the *PO41* arrays were flanked by stretches of N's, at least on one end of the array. This suggests that sequencing the genome was problematic in this region and caused a gap in the assembly.

CR1 repeats

To investigate the relationship between *CR1* subfamilies, sequences of the previously described subfamilies A–F, as well as *CR1-X* and *CR1-Y* (deposited in RepBase) were

Downloaded from genome.cshlp.org on January 21, 2014 - Published by Cold Spring Harbor Laboratory Press

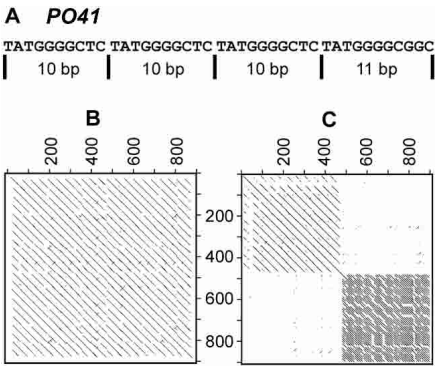

Figure 2. Analysis of tandem repeats. (*A*) Consensus sequence of *PO41*. The basic repeat consists of three subunits of 10 bp followed by one subunit of 11 bp. However, frequent modifications of this 3:1 pattern were observed. (*B*) Dot Plot display of an array of 21 tandem repeated *CNM* units. Most *CNM* repeats have a size of 41 bp, the entire array has a size of 860 bp and is located on the virtual chromosome "ChrUn." (*C*) Arrays of *CNM* and *PO41* immediately adjacent to one another.

used for a BLASTN search against the chicken genome. For this survey, only the 3′ terminal region of ORF-2 was used, because most *CR1* elements are found as short fragments of the 3′ region. The region used corresponds to amino acid positions 818 to 972 (of 980) of the consensus protein for ORF-2 (accession no. AAC60281; Haas et al. 1997). More than 5000 *CR1* elements covering this entire region were identified. Because this set was too large to perform multiple alignments and a phylogenetic analysis, a subset of 500 randomly picked sequences was used for the analysis. Sequences that caused major gaps in the alignment or contained obvious large deletions were removed, resulting in a final set of 335 sequences that were used for the construction of the phylogenetic tree shown in Figure 3A. The 335 *CR1* elements cluster in eight distinct subfamilies. The phylogenetic tree contains multiple members of the previously described subfamilies *A*, *B*, *D*, *F*, *X*, and *Y*. Two additional subfamilies we designated *H* and *I*. Subfamilies *B* and *F* are the largest with 40 and 102 members, respectively (Fig. 3A). No elements similar to the previously described subfamily *E* were found, and a copy of the previously described subfamily *C* is located in a group of 77 elements that are not similar enough to each other or to any other sequence to be resolved into subfamilies (Fig. 3A). However, it is likely that by increasing the sample size, further subfamilies could be resolved in this group.

Very few *CRI* elements are functional

As shown in Figure 3B, the overwhelming majority of *CR1* elements are present as small fragments with a size of <500 bp. The size distribution of *CR1* elements follows an almost perfect hyperbolic distribution with >98% of the identified copies being smaller than 2000 bp (Fig. 3B). A total of 1350 elements were identified that have a size of >2000 bp and only 260 copies larger than 3000 bp. To identify any full-length elements, the consensus protein sequence for ORF-1 (accession no. AAC60280) was used for a TBLASTN search against the chicken genome. Regions that showed similarity were isolated with 4 kb of their flanking regions. In a second step, the resulting slices of genomic DNA sequence were screened for the presence of full-length ORF-2. A total of 63 copies that contain both ORF-1 and ORF-2 were obtained, although the majority of them carry major deletions, and both their ORFs are interrupted by frameshifts or in-frame stop codons. Twenty-five of the full-length elements belong to the *CR1-F* subfamily, whereas the other subfamilies were represented in one to nine copies.

Only one *CR1-F* element with intact ORFs for both proteins was found, on chromosome 6 (positions 661,372–666,220). We consider this element a candidate for a functional "mother" element with the ability to produce functional proteins, and thus, replicate. One additional *CR1-B* element was found to contain an intact ORF-1, but a defective ORF-2. The intact *CR1-F* element (Fig. 3C; Supplemental material) is 4033 bp in length and is flanked by a 5-bp target site duplication. Its borders are in agreement with a *CR1-F* sequence deposited in RepBase. The element itself does not contain any 5′ promoter sequences, as the coding region for ORF-2 starts only 5 bp downstream of the target site duplication. However, the element is inserted into an A/T-rich region that may provide promoter elements. For example, two putative TATA boxes are located 96 and 246 bp upstream of the start codon, respectively (Fig. 3C).

The hypothetical protein encoded by ORF-1 (Supplemental material) is 56% identical with the consensus ORF-1 for *CR1-B* (accession no. AAC60280), and with a size of 358 amino acids shorter than the latter with 420 amino acids. However, a Zinc finger motif ($Cx_2Cx_{14}Cx_2Cx_5Lx_6Lx_6L$) that was previously described in *CR1* elements (Kapitonov and Jurka 2003) was also found to be perfectly conserved. Interestingly, a second full-length, but nonfunctional *CR1-F* element that is 96% identical to the putative mother element was found on chromosome 1. It contains 95 bp of the putative promoter region. We interpret this element as the result of a reverse transcription of almost the entire mRNA molecule of the putative mother element. In that case, the TATA box 246 bp upstream of the start codon (TATA box 1, Fig. 3C) would have served as the initiation point for the RNA transcription (Supplemental material).

LTR retrotransposons

By means of multiple alignments of Cot sequences, we were able to identify four main types of LTR retrotransposons (*Birddawg*, *Hitchcock*, *Kronos*, and *Soprano*). The two most abundant elements are *Birddawg* (7404 identified copies) and *Kronos* (4961 identified copies). The two elements are distantly related members of the *gypsy* family and share virtually no sequence identity at the DNA level. A consensus hypothetical protein sequence of *Kronos* and the hypothetical protein of a *Birddawg*-F element are only 52% similar to one another. Typical features of *gypsy*-type elements are conserved in both proteins, the reverse transcriptase signature (YIDD) in the N terminal half as well as a Zinc finger motif ($HX_4HX_{28}CX_2C$) and a putative DDE motif indicative of the integrase domain in the carboxy-terminal part. *Birddawg* could be classified into seven subfamilies through multiple alignment of 313 LTR sequences (see below). The subfamily *Birddawg*-F appears to be the most recent one, as eight full-length elements could be identified that were >96% identical to one another. In addition, five of the full-length elements contain intact and potentially functional ORFs (Supplemental material).

A second element of particular interest is *Soprano*, for which 1362 copies were identified. The internal domain of *Soprano* encodes a protein that was previously shown to be expressed in embryonic stem cells (*cENS*; Acloque et al. 2001). It was speculated that this gene family may be of retroviral origin. We identified 75 *Soprano* elements that contain an internal domain, and at least one of those contains an intact ORF, whereas the rest of the 1362 copies are solo-LTRs or truncated copies. Thus, our ob-

Downloaded from genome.cshlp.org on January 21, 2014 - Published by Cold Spring Harbor Laboratory Press

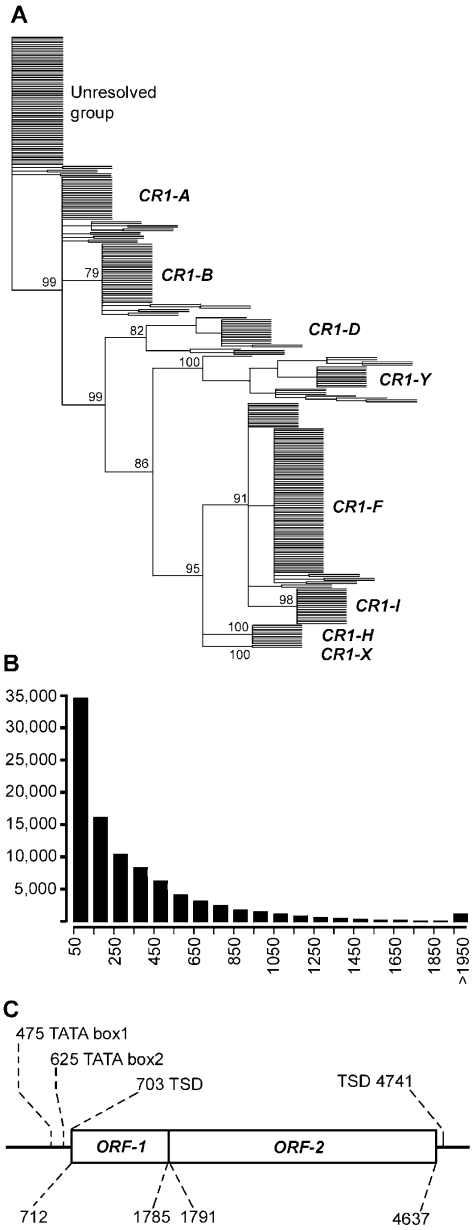

Figure 3. Characterization of *CR1* repeats. (*A*) Phylogenetic analysis of 332 *CR1* sequences. The sequences used cover ~450 bp of the 3′ terminus of the element. Bootstrap values for the major branches are indicated. (*B*) Size distribution of 96,230 *CR1* elements found in the chicken genome. The vast majority of the copies are fragments shorter than 1000 bp. (*C*) Sequence organization of a putative intact *CR1-F* mother element. The element consists of two closely spaced ORFs (ORF-1 and ORF-2), 4 bp of 5′ UTR, and 103 bp of 3′ UTR. The borders are determined by a 5-bp target site duplication (TSD). The element is inserted into an A/T-rich host sequence that provides putative promoter elements (TATA boxes 1 and 2).

servations strongly support, if not prove, the hypothesis of the retroviral origin of the *cENS* gene family.

In general, the number of LTR sequences identified for a particular element is higher than the number of identified internal domains. The excess of LTR sequences is a phenomenon frequently observed both in animals (Benit et al. 1999) and plants (Shirasu et al. 2000), and is generally explained by unequal inter- or intra-element crossing-over that leads to the excision of the internal domain. *Kronos* elements have the highest number of internal domain sequences; of 4961 copies, 1517 contain fragments of the internal domain. For the other types, LTR sequences clearly outnumber the internal domain sequences (*Birddawg*: 7404/894, *Soprano*: 1362/75). In the case of *Hitchcock*, no internal domain could be identified, and the element was classified as LTR mainly because of the conserved termini (TG…CA) and the absence of characteristics typical for SINEs (A-tail, RNA pol III promoter, etc).

Galluhop—A high-copy DNA transposon

The initial copies of *Galluhop* were discovered through multiple sequence alignment of >5000 HR and MR sequences. Iterative BLAST searches as described above identified 102 elements in the Cot sequences and thousands of copies in the chicken genome. A multiple alignment helped to define the borders of the element and led to the identification of 160 full-length elements that were used for a phylogenetic analysis (see below). It also became clear that the chicken genome contains at least two subpopulations of *Galluhop* elements, one with an average size of ~530 bp and the other ~1250 bp in size, with smaller elements outnumbering the larger by ~10-fold.

A BLASTX search against all vertebrate proteins revealed that the larger elements contain a coding sequence for a transposase protein similar to those encoded by *Mariner* in the human genome. As illustrated in Figure 4A, the small elements are derivatives of the large ones, originating by deletion of the coding sequence from the central region. We suggest that the larger, mother elements represent autonomous transposons encoding a transposase that facilitated both their own spread, and that of the smaller *Galluhop*s within the genome. *Galluhop* elements are flanked by 25-bp imperfect terminal inverted repeats, which are conserved in both the nonautonomous and the mother elements. These terminal repeats are similar in sequence to two Mariner elements (*Mariner1a* and *Mariner1b*) from RepBase, consequently making *Galluhop* a member of the *Mariner* superfamily.

Six mother elements (*Galluhop-48*, *Galluhop-56*, *Galluhop-82*, *Galluhop-203*, *Galluhop-241*, and *Galluhop-248*; Supplemental material) containing only minor deletions within their coding regions were used for a multiple alignment to deduce their hypothetical protein sequences. All six coding regions contain multiple frame-shifts and/or in-frame stop codons. A multiple alignment of these six coding regions, however, helped to remove frame-shift mutations and determine the hypothetical start and stop codons. The transposase sequences of human *Mariner* elements (accession nos. AAC52010 and AAC52011) were used as reference sequences in the analysis. The predicted consensus protein (Supplemental material) has a size of 344 amino acids and shares 40%–45% sequence identity with the reference human proteins.

The full-length *Galluhop* elements, as well as the region that is exclusive to the mother elements, were used in a BLASTN search to determine the copy number for both the nonautonomous and the mother elements. A total of 13,729 copies of *Galluhop* were identified, with 1456 being mother elements.

Galluhop has a unique genome dispersion strategy

The large number of full-length elements identified for both *Galluhop* and the numerous LTR sequences for *Birddawg* provided enough data to undertake a phylogenetic analysis for each. Sets of 150 *Galluhop* and 313 *Birddawg* sequences were used to pro-

Downloaded from genome.cshlp.org on January 21, 2014 - Published by Cold Spring Harbor Laboratory Press

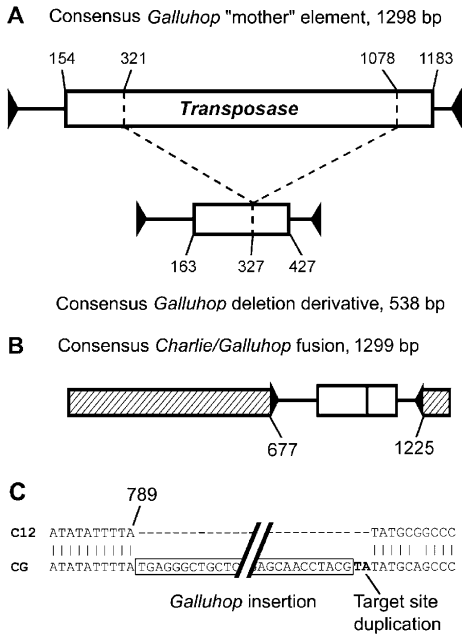

Figure 4. Sequence organization of *Galluhop* and *Charlie*. (*A*) Organization of autonomous and nonautonomous *Galluhop* elements. The nonautonomous elements presumably originated from a deletion within the coding region of the transposase. Start and end positions of the coding sequence, as well as positions of the deleted region refer to consensus sequences for both the mother element and the deletion derivative. Both consensus sequences are available as Supplemental material (*B*) One *Charlie* subfamily carries an insertion of a nonautonomous *Galluhop* element. Positions of the *Galluhop* insertion refer to a *Charlie/Galluhop* fusion consensus sequence. (*C*) Detailed view of the *Galluhop* insertion site as shown by an alignment of the sequence *Charlie12* (C12) from RepBase and the *Charlie/Galluhop* fusion consensus sequence (CG). The *Galluhop* element inserted at the position corresponding to position 790 in *Charlie12* and has created a 2-bp putative target site duplication. The *Charlie/Galluhop* fusion consensus sequence (*Charlie_Galluhop_fusion*) is available as Supplemental material.

duce cladograms for both repeat types (Fig. 5A). The results were striking. *Birddawg* sequences fell into seven subfamilies in both the maximum parsimony and HKY85 distance analyses. The separations between the major subfamilies were supported by bootstrap values above 60 for most branches (Fig. 5A). In contrast, the nonautonomous *Galluhop* elements were all nearly equally closely related to each other, and no major subfamilies could be identified. A few sequences formed small subgroups, whereas the majority grouped together in one main family. This led to the rather unusual cladogram shown in Figure 5B, which is discussed below. A second phylogenetic analysis was performed with 184 copies of the region of the coding sequence that is exclusive for the *Galluhop* mother elements. The results were virtually the same, with most sequences equally closely related to one another (data not shown).

Charlie—An intragenomic vector for *Galluhop*?

The sequence, *Charlie12*, found in RepBase was used as the initial sequence for iterative BLASTN searches that led to the identification of 2292 similar sequences in the chicken genome. The 3′ ends of many of these elements were well conserved, whereas the 5′ ends are highly divergent and often truncated, preventing an unambiguous determination of their full size. A multiple alignment of 100 *Charlie* sequences resulted in a consensus sequence that was 100 bp shorter than *Charlie12,* due to too strong sequence divergence at the 5′ end. The identified *Charlie* elements contained no detectable coding sequence, nor could terminal direct or inverted repeats be identified.

The *Charlie* repeat family is divided into two subpopulations on the basis of an acquisition of a *Galluhop* element. A total of 237 elements were found to contain a full-length copy of a nonautonomous *Galluhop* element at identical positions close to their 3′ ends (Fig. 4B). Deriving a consensus sequence for the *Charlie/Galluhop* fusion element allowed the identification of the precise position of insertion as well as a 2-bp target site duplication flanking the *Galluhop* element. The sequence TA of the target site duplication is typical for *Mariner*-like elements (Plasterk et al. 1999). Because the *Charlie/Galluhop* fusion elements were found in such a high-copy number, the most parsimonious explanation is that a *Galluhop* element inserted into a *Charlie* element that was not disabled by the insertion, and continued to spread, thus forming its own subfamily. The acquisition of foreign sequences by transposable elements has previously been documented (Takahashi et al. 1999; Wicker et al. 2003). However, the *Charlie/Galluhop* fusion is interesting, in that it provides an example of an intragenomic transposable element acquiring another transposable element.

Discussion

Our application of Cot-based cloning and sequencing (CBCS) in a time period when only limited sequence data for chicken was

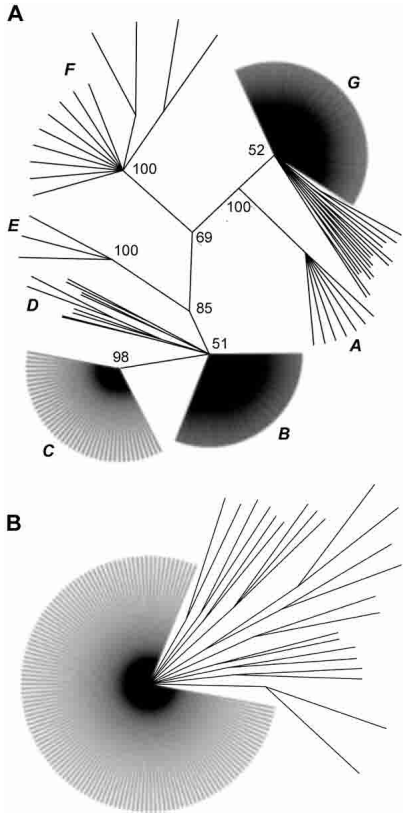

Figure 5. Phylogenetic analysis of two major repeat types. Bootstrap values for the major branches are indicated. (*A*) Cladogram of *Birddawg*. (*B*) Cladogram of *Galluhop*.

Downloaded from genome.cshlp.org on January 21, 2014 - Published by Cold Spring Harbor Laboratory Press

publicly available enabled us to rapidly identify an number of candidates for major repeat types in the chicken genome. The availability of the chicken genome sequence allowed us to immediately expand our analysis to a whole-genome level, also permitting an assessment of what the CBCS approach can provide to researchers working in genomes that are not likely to be fully sequenced soon. The ability to create a high-quality repeat database will be especially valuable for organisms where only limited sequence data is available. It can also provide a reference for repeat sequences in closely related species (Vandergon and Reitman 1994). For example, the discovery of a particular family of CACTA transposons in wheat led to the identification of a large number of similar elements in the rice genome that had been annotated as hypothetical genes (Wicker et al. 2003). CBCS may also produce essential information in organisms in which whole-genome projects are under way, or in the early stages of finishing. The fact that more than one-third of the HR Cot sequences were not found in the publicly available chicken genome sequence demonstrates that CBCS can provide information about regions of the genome that are not easily accessible otherwise.

The relatively low-repeat content of the chicken genome also provided an especially stringent test of the CBCS method. Our results demonstrate that even for genomes with a low level of repetitive DNA, efficient separation of highly repetitive and moderately or low-copy sequences can be achieved by using CBCS. The present application of CBCS to an organism for which a whole-genome sequence is available provides the ultimate validation of the method, as the exact copy number for the majority of the sequences could be determined using BLASTN.

How repetitive is the chicken genome?

Previous studies suggested that the chicken genome contains ~15% repetitive DNA. In the near future, detailed information will become available about the exact repeat content of the chicken genome. Our analysis has focused on a relatively small set of seven main repeat types that contribute ~4.3% to the total genome sequence. It is likely that future quantitative analysis of the genome will discover a number of additional repeat types that are present at lower copy numbers of dozens to hundreds. Eventually, the total content of interspersed repeats in the chicken genome can be expected to be ~6%–8% (W. Warren, pers. comm.). However, large parts of the repetitive landscape of the chicken genome are not represented in the publicly available first draft of the genome sequence. These include the centromeric and telomeric regions that are known to be composed of specific repeat types. Previous studies suggest, for example, that the subtelomeric *CNM* tandem repeats could make up to 10% of the entire genome (Matzke et al. 1990). Also, other types of short tandem repeats such as *PO41* might contribute considerably to the chicken genome, as is indicated by their abundance in the HR Cot sequences. The extremely repetitive nature of these tandem arrays causes a major problem for the final genome assembly, as demonstrated by the fact that almost all identified instances of *CNM* or *PO41* are located on the virtual chromosome *ChrUn* and are immediately flanked by gaps in the genome sequence. It may therefore be difficult to shed light on the true dimensions and exact compositions of these fields of tandem repeated units.

Most chicken repeats are degenerate and nonfunctional

Our results provide insight into the biology of several repeat families and the diversity in the strategies for persistence and spread of selfish DNA in the chicken genome. For example, the identification of an intact *CR1-F* element and the phylogenetic tree for the *CR1* elements suggest that *CR1-F* is the most modern subfamily of *CR1* elements. The finding of only one intact *CR1-F* element and only some dozens of full-length elements is intriguing, as it implies that an overwhelming majority of repetitive sequences are inactive relics of ancient or erroneous transposition events. For other organisms, the number of functional elements can be very limited (Kazazian 1998). Maybe there are a number of mother elements in the chicken genome that remain undiscovered in our analysis, because the genome sequence still contains a large number of gaps, so that additional intact elements were truncated. Our results indicate that certain types of repeats might have become completely silent due to a complete loss of functional elements. Examples are *Galluhop*, for which no intact mother element could be identified and *Hitchcock*, which seems to be present only as a population of solo-LTRs.

Additionally, the question remains unanswered whether the chicken genome has the tendency to selectively remove the 5′ end of *CR1* repeats. The presence of a majority of 3′ fragments of *CR1* elements may suggest a potentially advantageous sequence motif within this region. It could also mean that reverse transcription of the *CR1* elements is simply so erroneous that a vast number of reverse transcription events are terminated prematurely. This would explain the almost perfect hyperbolic shape of the size distribution of the identified *CR1* elements; if reverse transcription is terminated with certain constant probability at any given point during synthesis, a size distribution with the observed shape will be the result.

The chicken genome provides insight into complex patterns of repeat evolution

The different repeat types presented in this study provide fascinating examples of different evolutionary patterns. For example, the presence of the *Charlie/Galluhop* fusion repeat illustrates the symbiosis and coevolution of two originally independent transposable elements. Similar fusions of repetitive elements have been reported previously (e.g., *Alu* dimers, Batzer and Deininger 2002; *B1-dID* elements, Kramerov and Vassetzky 2001). The acquisition of *Galluhop* by *Charlie* founded a rather successful subpopulation of repeats that contribute >10% of the total *Charlie* population in the chicken genome. The fact that the *Charlie* element itself appears to be a nonautonomous element depending on another element(s) for its transposition adds to the complexity of the evolution of these repeats.

It is intriguing that phylogenetic analysis of a high-copy repeat may yield insight into its genome dispersion strategy as illustrated by the cladograms for *Birddawg* and *Galluhop*. After its initial invasion of the genome, *Birddawg* evolved into several subfamilies that were probably active at different times and generated different numbers of copies in the genome. The evolution of *Birddawg*, therefore, produced a phylogenetic pattern similar to the one found for *CR1* or the different subfamilies of *Alu* elements in primates (Batzer and Deininger 2002). This type of pattern might apply to most types of repetitive elements. In contrast, *Galluhop* appears to have spread throughout the chicken genome in one large burst, thus leading to a population of elements that are equally closely related to one another. Additional support for this view is the finding that all nonautonomous *Galluhop* elements show the same deletion within their coding sequence, implying a single deletion event. Independent evolution

Downloaded from genome.cshlp.org on January 21, 2014 - Published by Cold Spring Harbor Laboratory Press

of multiple nonautonomous elements would have led to a variety of deletion derivatives similar to those found for the *Ac/Ds* elements in maize (for review, see Fedoroff 1989). Interestingly, phylogenetic analysis of the *Galluhop* mother elements showed the same pattern, indicating that after the initial invasion, the generation of a nonautonomous derivative and the eventual silencing of the entire population occurred within a relatively short evolutionary time.

A large portion of the chicken genome is neither repetitive nor has obvious gene-encoding capacity

All repetitive elements identified in this study, plus all predicted gene-coding sequences (geneid; http://genome.ucsc.edu) cover a total of 84.7 Mbp or little more than 7% of the genome. A similar situation was found for the Cot sequences, in which 9.2% could be classified either as genes or repeats. It can be expected that the amount of functional DNA-associated genes (promoters, regulatory elements, introns, etc.) significantly exceeds the amount of actual coding sequences. However, even if one assumes that functional DNA constitutes 10-fold the amount of coding sequences, it would only account for ~28% of the genome. Hence, intergenic regions may be low- or single-copy DNA and lacking known function. These sequences might include ancient repeat families or ancient copies of *CR1* elements sufficiently degenerate as to be unrecognizable—several *CR1* elements found in the SL fraction may be well on their way to this fate. The finding that *CR1* elements have diverged into at least eight distinct subfamilies and most copies are heavily truncated, indicates relatively ancient activity. Because repetitive sequences are generally free from selection pressure, their continuing degeneration by mutation results in increasing sequence complexity. Thus, the ongoing production and degradation of repetitive DNA elements could provide a mechanism for a constant turnover of genomic DNA that provides the genetic raw material for the creation and recombination of functional DNA.

Methods

DNA extraction

Heparinized blood collected from adult female chickens via venous puncture (variety Cobb 5 × white leghorn) was centrifuged at 2200 rpm (Beckman GH-3.8 rotor) for 10 min in a 15-mL falcon tube at 4°C. The resulting pellet was resuspended in 100 mL of lysis buffer (10 mM Tris-HCl at pH 7.5, 50 mM EDTA, 1% SDS). Proteinase K (EM Science) was added to a final concentration of 10 µg/mL, and the tube was incubated for 12 h at 55°C. The DNA was then extracted with phenol/chloroform/isoamyl alcohol (25:24:1), precipitated with isopropanol, and resuspended in TE (100 mM Tris-HCl at pH 7.5, 10 mM EDTA). DNA was sheared by sonication in 500-µL aliquots with a Fisher Sonic Dismembrator (Model 300) at 25% of its maximum output for 2 min, and size selected on an agarose gel for a size range of 400–1500 bp fragments.

Cot-based cloning and sequencing (CBCS)

For fractionating DNA on hydroxyapatite (HAP) columns, the protocol of Peterson et al. (2002a) was used. Fractions of single-stranded DNA (ssDNA) and double-stranded DNA (dsDNA) were desalted on Sephadex G-50 columns, precipitated with isopropanol, and cleaned using the QIAquick PCR Purification Kit (QIAGEN). Cot fractions containing dsDNA were digested with mung bean nuclease (New England Biolabs) to create blunt ends. Then an A-overhang was added using *Taq* polymerase (QIAGEN). The resulting products were ligated into the pGEM-T vector (Promega) and cloned in electrocompetent, methylation-insensitive *Escherichia coli* DH10B cells (Invitrogen). For Cot fractions of ssDNA, a second strand was synthesized using random hexanucleotide primers and the Klenow fragment of DNA polymerase I (Promega), and the products were cloned as above.

Plasmid DNA was isolated following a modified alkaline lysis protocol (Marra et al. 1997). Cycle sequencing reactions were performed using the BigDye Terminator Cycle sequencing Kit Version 3.1 (Applied Biosystems) and MJ Research PTC-100 and PTC225 thermocyclers. Sequencing was done on an ABI 3700 automated DNA Analyzer (at UGA) or on a ABI 3730 automated DNA Analyzer as previously described (at WUGSC: Wang et al. 2003). Due to the relatively small insert size, plasmid were sequenced in only one direction.

To calculate the necessary sequence coverage, the sequence complexity for each fraction was calculated using the methods of Peterson et al. (2002a). By definition, the single-copy component has a mean repeat value (R) of 1. The mean repeat values for each repetitive component can be calculated by dividing the $Cot_{1/2}$ of the single-copy component (1000 Ms) by the $Cot_{1/2}$ value for the particular repetitive component. Thus, the HR component has a mean repeat value of [1000 Ms/1 Ms =] 1000. The sequence complexity of individual components was not reported with the chicken Cot curve (Epplen et al. 1978), but can be estimated using the formula (1C × F)/R, where 1C is the genome size in base pairs, F the fraction of genome occupied by each component, and R the mean repeat number of sequences in the component (Goldberg 1978). For the very repetitive (HR) fraction, which constitutes 10% of the chicken genome, the sequence complexity is [(1200 Mbp × 0.1)/1000 =] 120 kb.

Sequence quality control and analysis

Sequence trace data were evaluated using PHRED (Ewing et al. 1998) and vector sequences were masked via CROSSMATCH (http://www.phrap.org). Only inserts with high-quality sequence (Q > 16) larger than 50 bp were used for further analysis. To identify DNA contamination from bacterial, mitochondrial, or human DNA, BLASTN searches (Altschul et al. 1997) were performed against databases containing genome sequences of *E. coli*, cell organelles, and *Homo sapiens*. Putative coding sequences were identified by BLASTX against the nr protein database of GenBank or against local databases containing unpublished hypothetical protein sequences.

Throughout this study, a BLASTN hit was considered significant if two sequences were more than 80% identical over a region of more than 55 bp. A BLASTX hit was considered significant if the two sequences were at least 30% identical over more than 30 amino acids. Multiple sequence alignments were done using CLUSTALW (Thompson et al. 1994), and dot plots for visual comparison of sequences were performed using DOTTER (Sonnhammer and Durbin 1995). For the efficient processing of large numbers of sequences and the evaluation of BLAST outputs, original programs were written in PERL. Phylogenetic analyses by maximum parsimony and Hasegawa-Kishino-Yano (HKY85; Hasegawa et al. 1985) genetic distance with neighbor-joining clustering were done in parallel with PAUP* 4.0b10 (Swofford 2003). The reliability of all trees was characterized by bootstrap resampling (Felsenstein 1985) with 100 replicates.

Identification of repetitive elements and determination of copy numbers

The first approach for the identification of repetitive sequences made use of previously described repeats for which full-length

Downloaded from genome.cshlp.org on January 21, 2014 - Published by Cold Spring Harbor Laboratory Press

elements had been well characterized, including *CR1*, *PIR*, *CNM*, and *XhoI*. These sequences were used as query sequences in BLASTN or TBLASTN searches against sequences from the three Cot fractions, as well as the 25 Mbp of publicly available BAC sequences (this approach predated the publication of the whole-genome sequence). Sequences identified in this way were collected and used for a second round of BLAST searches in order to identify more divergent repeat units. Further iterations of this process were continued until the number of sequences with hits to known repeats did not increase.

The second approach was designed to identify novel repeat elements. Sequences from the Cot fractions were used for multiple alignment via CLUSTALW. For each cluster, defined as at least two sequences with at least 60% similarity, the consensus sequence for the aligned region was considered a candidate repeat and was, in turn, used for a BLASTN search against the Chicken BAC sequences. If multiple loci were hit, these sequences from the BACs plus 200–500 bp from the flanking BAC sequences were isolated. The resulting sequence slices were then compared by multiple alignment and dot plot algorithms to define the actual borders of the repeat element.

The third approach involved all sequences that gave more than two hits on the BAC sequences, but did not match any other sequences in the multiple alignment. This approach was also used for publicly available sequences for which the full size of the element was not known. Full-length elements were isolated from the Chicken BAC sequences in the same manner as described for the second approach.

Copy numbers counts of repeats were performed with initially of up to 50 sequences (e.g., representatives of different subfamilies) for each repeat type that was used in BLASTN searches against the chicken genome. All alignments that showed >80% identity and were longer than 55 bp were considered. Due to the sequence conservation between different subfamilies, most copies were likely to be hit multiple times. In these cases, only the two most extreme positions of such a local accumulation of hits were used to define the borders of an element. In addition, some copies were expected to be divergent to a degree that they do not align with the query sequences over their entire length. Therefore, alignments that were separated by <500 bp were considered to belong to the same copy.

Acknowledgments

This work was funded by National Science Foundation grant number: EHR-0125304 to A.H.P. and R.I.

References

Acloque, H., Risson, V., Birot, A.M., Kunita, R., Pain, B., and Samarut, J. 2001. Identification of a new gene family specifically expressed in chicken embryonic stem cells and early embryo. *Mech. Dev.* **103:** 79–91.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Batzer, M.A. and Deininger, P.L. 2002. *Alu* repeats and human genomic diversity. *Nat. Rev. Genet.* **3:** 370–379.

Benit, L., Lallemand, J.B., Casella, J.F., Philippe, H., and Heidemann, T. 1999. ERV-L elements: A family of endogenous retrovirus-like elements active throughout the evolution of mammals. *J. Virol.* **73:** 3301–3308.

Britten, R.J., Graham, D.E., and Neufeld, B.R. 1974. Analysis of repeating DNA sequences by re-association. *Methods Enzymol.* **29:** 363–418.

Delaney, M.E., Daniels, L.M., Swanberg, S.E., and Taylor, H.A. 2003. Telomeres in the chicken: Genome stability and chromosome ends. *Poult. Sci.* **82:** 917–926.

Epplen, J.T., Leipoldt, M., Engel, W., and Schmidtke, J. 1978. DNA sequence organization in avian genomes. *Chromosoma* **69:** 307–321.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8:** 175–185.

Fedoroff, N.V. 1989. About maize transposable elements and development. *Cell* **56:** 181–191.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39:** 783–791.

Goldberg, R.B. 1978. DNA sequence organization in the soybean plant. *Biochem. Genet.* **16:** 45–68.

Gregory, T.R. 2002. A bird's-eye view of the C-value enigma: Genome size, cell size, and metabolic rate in the class Aves. *Evolution* **56:** 121–130.

Haas, N.B., Grabowski, J.M., Sivitz, A.B., and Burch, J.B. 1997. Chicken repeat 1 (*CR1*) elements, which define an ancient family of vertebrate non-LTR retrotransposons, contain two closely spaced open reading frames. *Gene* **197:** 305–309.

Hasegawa, M., Kishino, H., and Yano, T. 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22:** 160–174.

Ho, I.S. and Leung, F.C. 2002. Isolation and characterization of repetitive DNA sequences from *Panax ginseng*. *Mol. Genet. Genomics* **266:** 951–961.

Jurka, J. 2000. RepBase update: A database and an electronic journal of repetitive elements. *Trends Genet.* **16:** 418–420.

Kapitonov, V.V. and Jurka, J. 2003. The esterase and PHD domains in *CR1*-like non-LTR retrotransposons. *Mol. Biol. Evol.* **20:** 38–46.

Kazazian, H.H. 1998. Mobile elements and disease. *Curr. Opin. Genet. Dev.* **8:** 343–350.

Klein, S. and Ellendorff, F. 2000. Localization of *XhoI* repetitive sequences on autosomes in addition to the *W* chromosome in chickens and its relevance for sex diagnosis. *Anim. Genet.* **31:** 104–109.

Kramerov, D.A. and Vassetzky, N.S. 2001. Structure and origin of a novel dimeric retroposon B1-diD. *J. Mol. Evol.* **52:** 137–143.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Lewin, B. 1994. Replicative and nonreplicative transposition may pass through common intermediates. In *Genes V*, pp.1010–1015. Oxford University Press, Oxford, New York.

Marra, M.A., Kucaba, T.A., Dietrich, N.L., Green, E.D., Brownstein, B., Wilson, R.K., McDonald, K.M., Hillier, L.W., McPherson, J.D., and Waterston, R.H. 1997. High throughput fingerprint analysis of large-insert clones. *Genome Res.* **7:** 1072–1084.

Matzke, M.A., Varga, F., Berger, H., Schernthaner, J., Schweizer, D., Mayr, B., and Matzke, A.J. 1990. A 41–42 bp tandemly repeated sequence isolated from nuclear envelopes of chicken erythrocytes is located predominantly on microchromosomes. *Chromosoma* **99:** 131–137.

Peterson, D.G., Schulze, S.R., Sciara, E.B., Lee, S.A., Bowers, J.E., Nagel, A., Jiang, N., Tibbitts, D.C., Wessler, S.R., and Paterson, A.H. 2001. Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. Accession nos. AZ921847–AZ923007. http://www.ncbi.nlm.nih.gov/entrez.

Peterson, D.G., Schulze, S.R., Sciara, E.B., Lee, S.A., Bowers, J.E., Nagel, A., Jiang, N., Tibbitts, D.C., Wessler, S.R., and Paterson, A.H. 2002a. Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. *Genome Res.* **12:** 795–807.

Peterson, D.G., Wessler, S.R., and Paterson, A.H. 2002b. Efficient capture of unique sequences from eukaryotic genomes. *Trends Genet.* **18:** 547–550.

Petrov, D.A. and Hartl, D.L. 1998. High rate of DNA loss in *Drosophila melanogaster* and *Drosphila virilis* groups. *Mol. Biol. Evol.* **15:** 293–302.

Plasterk, R.H.A, Izsvák, Z., and Ivics, Z. 1999. Resident aliens: The Tc1/mariner superfamily of transposable elements. *Trends Genet.* **15:** 326–332.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J.L. 1998. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20:** 43–45.

Schmid, M., Nanda, I., Guttenbach, M., Steinlein, C., Hoehn, M., Schartl, M., Haaf, T., Weigend, S., Fries, R., Buerstedde, J.M., et al. 2000. First report on chicken genes and chromosomes 2000. *Cytogenet. Cell Genet.* **90:** 169–218.

Shirasu, K., Schulman, A.H., Lahaye, T., and Schulze-Lefert, P. 2000. A

Downloaded from genome.cshlp.org on January 21, 2014 - Published by Cold Spring Harbor Laboratory Press

contiguous 66 kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Res*. **10:** 908–915.
Sonnhammer, E.L. and Durbin, R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167:** GC1–GC10.
Stumph, W.E., Hodgson, C.P., Tsai, M.J., and O'Malley, B.W. 1984. Genomic structure and possible retroviral origin of the chicken *CR1* repetitive DNA sequence family. *Proc. Natl. Acad. Sci*. **81:** 6667–6671.
Swofford, D.L. 2003. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinnauer Associates, Sunderland, MA.
Takahashi, S., Inagaki, Y., Satoh, H., Hoshino, A., and Iida, S. 1999. Capture of a genomic HMG domain sequence by the *En/Spm*-related transposable element Tpn1 in the Japanese morning glory. *Mol. Gen. Genet*. **261:** 447–451.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. **22:** 4673–4680.
Vandergon, T.L. and Reitman, M. 1994. Evolution of chicken repeat 1 (*CR1*) elements: Evidence for ancient subfamilies and multiple progenitors. *Mol. Biol. Evol*. **11:** 886–898.
Wang, D., Urisman, A., Liu, Y.T., Springer, M., Ksiazek, T.G., Erdman, D.D., Mardis, E.R., Hickenbotham, M., Magrini, V., Eldred, J., et al. 2003. Viral discovery and sequence recovery using DNA microarrays. *PLoS Biol*. **1:** E2. Epub 2003 Nov 17.
Wang, X., Li, J., and Leung, F.C. 2002. Partially inverted tandem repeat isolated from the pericentric region of chicken chromosome 8. *Chromosome Res*. **10:** 73–82.
Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Whitelaw, C.A., Barbazuk, W.B., Pertea, G., Chan, A.P., Cheung, F., Lee, Y., Zheng, L., van Heeringen, S., Karamycheva, S., Bennetzen, J.L., et al. 2003. Enrichment of gene-coding sequences in maize by genome filtration. *Science* **302:** 2118–2120.
Wicker, T., Stein, N., Albar, L., Feuillet, C., Schlagenhauf, E., and Keller, B. 2001. Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum L.*) reveals multiple mechanisms of genome evolution. *Plant J*. **26:** 307–316.
Wicker, T., Guyot, R., Yahiaoui, N., and Keller, B. 2003. CACTA transposons in *Triticeae*. A diverse family of high-copy repetitive elements. *Plant Physiol*. **132:** 52–63.
Yuan, Y., SanMiguel, P.J., and Bennetzen, J.L. 2003. High-Cot sequence analysis of the maize genome. *Plant J*. **34:** 249–255.

Web site references

http://genome.ucsc.edu; geneid.
http://genome.wustl.edu/projects/chicken; Chicken Genome Sequencing Project.
http://www.phrap.org; The Phred/Phrap/Consed system home page.
http://plantgenome.agtec.uga/g4g; "Genes for Georgia" Web site.

Received February 9, 2004; accepted in revised form June 2, 2004.