Studying the Evolution of Histone Variants Using Phylogeny

Antoine Molaro, Ines Drinnenberg

Studying the evolution of histone variants using phylogeny

Antoine Molaro[1] and Ines A. Drinnenberg[2]

[1]Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, United States.

[2]Institut Curie, Paris Sciences et Lettres Research University, Centre National de la Recherche Scientifique UMR 3664, F-75005 Paris, France

Correspondence: amolaro@fredhutch.org and ines.drinnenberg@curie.fr

Running head: evolutionary analysis of histone variants

ABSTRACT

Histones wrap DNA to form nucleosomes that package eukaryotic genomes. Histone variants have evolved for diverse functions including gene expression, DNA repair, epigenetic silencing and chromosome segregation. With the rapid increase of newly sequenced genomes the repertoire of histone variants significantly expands demonstrating a great diversification of this protein family across eukaryotes. In this chapter, we are providing guidelines for the computational characterization and annotation of histone variants. We describe methods to predict the characteristic histone fold domain and list features specific to known histone variants that can be used to categorize newly identified histone fold proteins. We continue describing procedures and sources to retrieve additional related histone variants for comparative sequence analyses and phylogenetic reconstructions to refine the annotation and to determine the evolutionary trajectories of the variant in question.

Keywords: Histone variants, H2A, H2B, H3, rapid evolution, homology predictions, multiple sequence alignments, phylogenetic inference, annotation

1. INTRODUCTION

The fundamental repeating unit of chromatin in eukaryotes is the nucleosome, in which DNA is wrapped around an octameric histone core complex [1]. The majority of nucleosomes in the cell are assembled by canonical histones and consist of 2 copies of each of the canonical histones H2A, H2B, H3 and H4. Histones all share a common fold, the histone fold domain (HFD), composed of three helices that mediate both protein dimerization and DNA-binding in the context of the nucleosome [2]. Canonical histones provide the primary structural scaffold for DNA wrapping during whole genome replication. As such, canonical histone expression and synthesis peaks during S-phase and their deposition occurs genome-wide [3]. In addition to canonical histones, unique histone variants can be incorporated into nucleosomes to confer specialized functions in specific genomic regions. Such functions include processes as diverse as transcription, chromosome segregation, DNA repair and recombination, chromatin remodeling, germline-specific DNA packaging, and even extra-nuclear acrosome formation [4]. To accomplish their specific chromatin functions, variants differ from canonical histones in their primary amino acid sequence. In addition to variation in the HFD, most variants have evolved N- or C-terminal amino acid extensions with specific sequence motifs or acquired entirely new domains. These sequence features are key to categorize variants accomplishing similar functions between organisms.

The evolutionary trajectories of histone variants are also unique from their canonical counterparts. First, unlike most canonical histones, which are encoded by genomic clusters containing the four core histone genes, “stand-alone” genes encode variants [3]. Second, while most canonical histones have an old evolutionary history that can be traced back to the

last common ancestor to all Eukaryotes, variants have diverse evolutionary origins [4]. Some variants are broadly distributed throughout the eukaryote phylogeny, and likely arose early in eukaryotic evolution, while others originated from recent lineage-specific events [5]. In fact, with the increasing number of sequenced genomes from diverse eukaryotes, the repertoire of new histone variants or diverse homologs of existing variants continues to expand [6]. As some of these newly-sequenced organisms are not easily accessible to genetic manipulations enabling experimental investigation, histone variants characterization solely relies on *in silico* comparisons to known and well-annotated histone variants. In this context, the phylogenetic investigation of variants provides a unique insight into the origins and selective constrains driving their specific biological functions.

In this chapter, we will provide guidelines to enable the systematic characterization and phylogenetic study of histone variants with an emphasis on H2A and H3 variants that are most abundant in model organisms. First, we will describe approaches used to identify histone variant candidates from publically available curated or non-curated genomic sequences. We will then provide instructions for the application of phylogenetic analyses to help distinguish them from their canonical counterparts. Because the combination of distinct sequence and phylogenetic features are strong predictors of their structure and function, we also provide a list of features characteristic of known histone variants that can be used to survey candidates.

2. Materials

Sequence sources: curated histone sequences can be found in browsable and searchable public databases such as Histone DB 2.0 [7], NCBI [8] and UCSC [9]. User provided sequences can also be used.

Analysis tools: all of the analysis presented here can be performed on a desktop computer. Alignment and other sequence analysis software referenced here are open source software that can be freely downloaded from the web or run online.

3. METHODS

Evolution-guided prediction of histone variant proteins.

In the following section, we provide general guidelines to perform sequence homology searches against publically available genome, transcriptome, and protein databases (3.1). Following up on those, we review characteristic features and evolutionary relationships of all known histone variants to help identify and classify HFD containing proteins (3.2). Once the histone variant has been identified with some confidence, we provide the general guidelines to extend searches to closely related and more distantly related species (3.3). Finally, we explain how to apply phylogenetic analyses to determine their evolutionary trajectories (3.4). The general workflow outline to study the evolution of histone variants is illustrated in Figure 1.

3.1 Commonly used search tools and databases

With an unknown query sequence in hand; the first step is to establish its relationship, or its identity, to known histone fold containing proteins. Two strategies can be applied depending on the degree of divergence of the query sequence:

(i) Search algorithms like BLAST or PSI-BLAST that test for direct sequence homology

(ii) Hidden-Markov-Model-based algorithms capable of identifying remote homologies

BLAST and PSI-BLAST

1. BLAST (basic local alignment search tool, see Note 1) [10] is a commonly used and efficient suite of tools to discover the evolutionary relationship of a query sequence with, searchable, publicly available sequences or between multiple user-provided input sequences. There two types of query that are supported by BLAST: nucleotide and amino acid sequences. When running BLAST on nucleotide sequences (BLASTn) it is advised to start with a spliced mRNA or coding sequence (CDS); as intron and regulatory regions might complicate searches since they are less evolutionarily constrained. A successful BLAST search should result in a list of sequence hits with additional attributes (i.e. genome of origin, chromosome location, protein ID…). These hits can be downloaded and further processed outside of the BLAST environment.
2. Three parameters are used to describe a match: similarity, coverage and expect value (*E* value). As general guidelines, a minimum of 30% similarity and at least 70% coverage of the histone fold domain between the query and the template are necessary for unambiguous alignment and potential structural models. The *E* value is a statistical estimate that describes the number of matches that can be expected by chance in a given database of a particular size. The lower the *E* value, the more "significant" the given match is. Generally, *E* value less than 0.001 are considered to be significant to conclude homology.
3. When using protein queries, a more sensitive strategy than BLASTp (Protein BLAST) to detect more distant evolutionary relationships is position-specific iterated-BLAST (PSI-BLAST) [11]. Starting from a single sequence, a sequence profile or position-specific-scoring matrix (PSSM) of related proteins above a certain threshold (based on the *E* value) found using BLASTp is created. The PSSM is then used to search the protein database iteratively for several rounds. The matches are included to create another PSSM and the procedure is repeated until no additional matches are found. By including all related proteins in the search, PSI-BLAST helps to uncover more distant homologs than ordinary BLASTp.
4. tBLASTn allows the user to search a nucleotide database (e.g. a whole genome) that has been *in silico* translated using a protein sequence query. This has the advantage of finding protein-coding nucleotide sequences that have no annotated protein, missed by BLASTp, or are highly diverged at the nucleotide level, missed by BLASTn. In addition, this allow the characterization of pseudogene sequences that have early stop codons or frame-shifts but still have significant homology to your query.
5. Finally, for highly diverged sequences, it might be better to run the search using the canonical histone sequence as a bait and focus on divergent matches (e.g. using H3 for cenH3 searches).

Hidden Markov Model-based predictions

While most histones can be detected using searches for direct sequence homology, some histone-fold proteins exhibit low levels of sequence conservation requiring alternative prediction tools to detect remote homologies [12]. In fact, it has been shown that two proteins can share a high degree of structural similarity despite the lack of detectable sequence similarity. Hidden Markov Model (HMM) based techniques implemented in servers such as HHpred [13] are currently the most sensitive tools detecting remote protein homology.

1. In the first step, HHpred builds a multiple sequence alignment from the query sequence by multiple iterations of PSI-BLAST against the nr database from NCBI.
2. In the next step, a profile HMM is generated from this alignment that includes the information about predicted secondary structure.

3. This is then compared with a database of HMMs representing proteins with known structure (such as PDB [14]) and protein families (such as PFAM [15]). The result is a list of matches with pairwise alignments.
4. For the interpretation of the results, the most relevant statistical measure is the probability in percent that the database match is indeed homologous. This value takes the score of the secondary structure into account that helps to distinguish homologs from chance hits (see Note 2, for a discussion on false positives).

3.2 Characteristic features used to distinguish histone variants and their evolutionary trajectories

Histone variants are subject to strong selective pressures to perform their unique function. Consequently, many sequence features affecting both the HFD and additional, variant-specific, domains can be used to discriminate each variant from their canonical counterpart across broad phylogenetic distributions. It is important to keep in mind that canonical histones also display some degree of divergence across eukaryotes and, thus, only "conserved" substitutions should be used to specifically annotate variants. In table 1 we provide an overview of the evolutionary origins and distinguishing features of histone variants.

3.3 Finding orthologous sequences

One key step in studying the evolution of histone variants, or any protein, is to compare it to homologous sequences across species to determine their evolutionary trajectories/origins.

To do so, a set of homologous sequences first needs to be collected keeping in mind that the phylogenetic distribution of the corresponding species used can greatly influence downstream analyses (see 3.4). In addition to surveying the group of organisms where orthologs of the corresponding histone variant are expected to be found, it is advised to include an homolog encoded by a more distantly related species in order to orient the comparison (i.e. root the phylogeny). To collect homologous sequences to the variant of interest, collection from public histone databases can be used as starting points. As an alternative and to further extend the set, BLAST searches can be performed to obtain homologs of additional, more closely related species (using the tools described in 3.1). We briefly describe both approaches.

HistoneDB 2.0:

HistoneDB 2.0 [7] is a database classifying histone protein sequences by type and variant. It collects canonical histones and histone variants, provides protein sequence, structural and functional annotations to facilitate the performance of comparative analyses across organisms. In addition, one can input any histone protein to this database to find its most closely related histone annotation. HistoneDB 2.0 consists of two complementary parts.

1. First, each histone clade (H2A, H2B, H3 and H4) can be explored through a set of manually curated variants with annotations of their specific features and functions. This set is used to construct Hidden Markov Models in order to automatically extract collections of related proteins from the non-redundant database of protein sequence maintained by NCBI (nr).
2. Results for this search constitute the second part of this database. From these 2 sets of proteins HistoneDB also provides an estimated phylogenetic distribution for the variant.
3. This enables browsing through histone variant entries to analyze their characteristic features and predict their evolutionary trajectories (also see Note 3). HistoneDB 2.0 therefore represents a useful resource to obtain sets of histone variants for constructing multiple sequences alignments (MSA) and phylogenetic analyses as described in the following section.

BLAST searches:

The most thorough way of finding homologous sequences across species is to use BLAST searches using NCBI genomes and databases. One can also run BLAST on user complied datasets (genomes, transcriptomes…). Through the GeneBank portal (https://www.ncbi.nlm.nih.gov/genbank/), one can explore the available datasets and their phylogenetic relationships.

1. We advise to begin by looking at the phylogeny (user determined or published) of species closely related to the query species – separated by a few My – and using this set of species to extend the search over distant clades (< 100My). With a few related sequences in hand, identifying distant orthologs becomes easier where conserved sequences features starting to emerge.
2. As mentioned in 3.1 one should carefully decide between using amino acid or nucleotide sequences as input – generally, we advise to use tBLASTn and work with *in silico* translated protein sequences.
3. Finally, we advise to always retrieve the genetic context in which the homologs are found. Indeed, depending on the biology of the species as well as the evolutionary time span, certain histone variants are found within identical genomic context and this information can be crucial to help finding the correct phylogenetic relationship between sequences (3.4). More generally if the gene/protein is annotated in any assembly, then BLAST searches can be combined with synteny searches using a comparative genomics online platform named Genomicus (http://www.genomicus.biologie.ens.fr/genomicus-88.01/).

3.4 Sequence alignments and phylogenetic reconstructions

When studying the evolution of histone variants several questions can arise: when did this histone variant appear in genomes and how is it related to another variant? For example, while H2A.Z variants are expected to have deep roots in the most recent common ancestor of eukaryotes, the phylogeny of cenH3 variants is currently unresolved. It was proposed that cenH3 variants have evolved several times independently in multiple eukaryotic lineages from H3 [5], yet with limited statistical support for this hypothesis. Instead, the more parsimonious model of a single origin for cenH3 in an early eukaryotic ancestor appears more likely due to its conserved presence at the eukaryotic centromere. Are there new sequence features that arose in a group of species? For example, H2A.W acquired a unique sequence motif that is unique to plants, while the motif in H2A.X might have arisen several times during eukaryotic evolution. How diverged is this variant compared to other histone genes? While canonical variants as well as H2A.Z and H3.3 are deeply conserved, the protein sequences of other H2A and cenH3 variants are divergent among species.

Phylogenetic analyses use information derived from sequence alignments to infer the possible evolutionary path that led to extant gene/protein sequences. It allows the reconstruction of ancestral sequences and the assessment of orthologies between sequences. Finally, phylogenetic trees convey more information than sequences alignments alone. Following up on the retrieval of histone variants homolog sequences, we provide general guidelines to perform multiple sequence alignments (MSA) and to build meaningful phylogenies. However, since these are general tools that are applied outside of the study of histone variants, we refer to individual chapters dedicated to both alignments and phylogenies [16,17].

3.4.1. Methods for multiple sequence alignments

Since all of the input sequences are histone-related genes/proteins one should attempt to anchor the alignment around the HFD, possibly using the appropriate canonical histone as a

reference. Although current algorithms are very accurate, alignments should be inspected visually, especially when dealing with distantly related variants. At first glance the MSA can inform on highly constrained residues in the alignment or, to the contrary, reveal unexpected regions of variability. In addition, MSA will allow the user to trim potentially retained non-coding intronic sequences when using genomic sequences as input or to identify annotations errors such as mis-annotated start codons (see Note 4).

In the following, we review a few MSA tools:

1. **MUSCLE**: MUSCLE stands for stands for **MU**ltiple **S**equence **C**omparison by **L**og-**E**xpectation [18,19]. This software that can be run both on an online server (http://www.ebi.ac.uk/Tools/msa/muscle/) and locally (the source code is freely available under http://www.drive5.com/muscle). One of the advantages of MUSCLE is that it can handle both DNA and protein sequences and is suitable for large datasets (>100 sequences).
2. **Clustal Omega**: Clustal Omega (http://www.clustal.org) is part of the widely used Clustal alignment suite [20]. Clustal Omega also achieves fast execution times for large datsasets by implementing the co-called mBed algorithm [21] for the fast construction of guide trees to produce sequence alignments.
3. **DNA versus protein alignments***:* Protein phylogenies are very well suited to inform on the patterns and rates of substitution that occurred between sequences, which will help determining the origin and relatedness of the variant. On the other hand, nucleotide phylogenies can be more robust since they contain ~3 times as many positions to be considered in the MSA than proteins (3 nucleotides per amino acid). In addition, nucleotide phylogenies will better inform on the nature of the selective forces that have shaped the histone variant, since one can measure and compare the rates of synonymous (silent) and non-synonymous codon substitutions across the phylogeny [22]. In contrast to protein alignments that can be used when studying histone variants over deep evolutionary time scales (e.g. >100My for mammals), alignments between nucleotide sequences are restricted to closely related species (e.g. <50My for primates), but it largely depends on the mutation rate and generation time of the species studied since the total number of synonymous substitutions can quickly saturate and become uninformative.

3.4.2. Inferring the phylogenies of histone variants

The final and last step to study histone variant evolution is the construction and visualization of a phylogeny. There are 3 major methods to infer phylogenies from sequence alignments:

1. **Maximum parsimony*:*** Phylogenies based on parsimony will treat each aligned position as a separate “character” and will attempt to build a relationship tree that minimizes the total number of steps required to explain the distribution of the “characters” in the MSA. A lot of platforms support parsimony tree building, one of the most commonly used is PAUP* (Phylogenetic Analysis Using Parsimony * and other methods) [23]. However, maximum parsimony methods are computationally heavy since they explore all possible trees before assessing the “most parsimonious” one, so they are best suited for a small number of sequences that are not too distantly related (e.g. all carnivores H2A.Zs).
2. **Neighbor joining:** This method is widely used since it is suitable for large datasets, and runs quickly locally or online. Neighbor joining phylogenies use identities in the MSA to assign distances between sequences. These distances become the branch length in the tree and the algorithm finds the “best phylogeny” by minimizing the total branch length of the tree [24]. This method can also be used with bootstrapping – which applies random sampling of branches to assign statistical significance to the tree

topology. However, since this method assumes a constant rate of substitution across the MSA it will not resolve very deep nodes in the phylogeny, e.g. the base nodes of all land plants.

3. **Maximum Likelihood:** This is the most statistically robust method to build phylogenies. Computationally, building ML phylogenies is as intense as building maximum parsimony phylogenies, however it has the advantage of using a substitution model that evaluates the probability of mutations across the MSA. Since varying evolutionary rates are permitted along the tree, ML phylogenies are really efficient at dealing with deeply branching phylogenies and should be the preferred method when studying old histone variants. In addition, ML phylogenies should be preferred over other methods when attempting to measure selective pressure that acted upon a set of histone variants as it will provide a robust assessment of topology (substitution pattern) and branch length (distance between sequences). Substitution models vary between sequence types and overall divergence [25] and many online tools exist to evaluate the best model to use – such as ModelOMatic [26] or prottest [27]. To run ML phylogenies the reference software is phyML [28] which is an open access software and can be run online on a multitude of platforms.

Visualization of phylogenetic trees

The most commonly used output file formats for phylogenetic trees include nexus, phylip and newick. These formats are suitable to the visualization of trees using a number of platforms as most contain information relative to the tree length, bootstrap support of topology, substitution rate (usually the branch scale), and additional information regarding the alignment.

1. When visualizing phylogenies, the user should decide if the tree should be rooted on a sequence, or a group of sequences, which is known to be the most “distantly related” histone in the MSA.
2. Rooted trees are preferred when attempting to resolve the evolutionary trajectories of the histone variant from its origins to its patterns of diversification. When unsure about which sequence to use, one should start with canonical histones from a similar set of species as well as from a true outgroup species for which there are no histone variant sequences being used. For example, when studying H2A.1 (known to be restricted to mammals) one can use mammalian canonical H2As and/or bird canonical H2As as outgroups. The assumption being that canonical H2A in these species arose before H2A.1.
3. Unrooted trees are most useful when studying deep phylogenies with many distantly related histone variants. These will allow one to infer the monophyletic (one origin or polyphyletic (multiple independent origins) nature of the studied variant (e.g. a phylogeny with representative H3s from most eukaryotes will group canonical sequences together and different variants as separated groups, see next section).

3.4.3. Interpretation of the obtained phylogeny – the evolution of histone variants

With these phylogenies in hand, it is now possible to trace back the evolutionary origin of the variant, assess its evolutionary rates within specific lineages and even determine the selective forces that shaped its evolution. In this last section, we discuss specific histone variant examples to illustrate possible interpretations.

If the studied histone variant sequences all share a common ancestor they are expected to group together with a single node separating them from other, e.g. canonical, histone sequences. In most cases, histone variants are derived from their canonical counterpart via

duplication or retroposition so these should constitute their closest relative in the tree. However, while most previously studied histone variants arose once and thus follow this topology, H3.3 and H2A.X variants have proposed to have multiple evolutionary origins. These 2 variants, diverge from their canonical counterparts several times in a phylogenetic tree consistent with either multiple independent origins of the variant or of canonical H2A or H3, or both [5,4]. For H2A.X, considering its deep eukaryotic origin, it is still up for debate if H2A.X could have preceded the evolution of canonical H2As. Overall, these observations further highlight the necessity of phylogenetic approaches to raise important yet unresolved questions about histone ancestry.

Additional information that can be gained from the phylogeny is the rate at which the variant sequences accumulate mutations (evolutionary rate). This is an indicator of the selective forces shaping its evolution and can provide insight into functional constraints that structural analysis alone cannot uncover. The length of branches in a phylogenetic tree can convey such information. While some trees have branch lengths proportional to the total length of the tree (indicative of relative distances between sequences), others illustrate actual mutation rates in the form of substitution per site (either amino acids or nucleotides).

Across a wide phylogenetic distribution, variants such as macroH2As, H2A.Z, H2A.1, H3.3 as well as canonical histones display very short branch lengths indicative of the strong purifying selection acting on these genes to maintain their function. However, branches connecting these histone clades can be long and are indicative of the major evolutionary transitions that have shaped the variant's sequence/function (see Note 5 for caveats associated with long branches in phylogenies). In addition, these transitions can also be limited to a group of species within the variant phylogeny and indicate lineage specific innovations that can be investigated further (e.g. H2A.M or H2A.J) On the other hand, histone variants such as cenH3 and short H2As have extremely long branches and their phylogeny does not necessarily match the expected species tree (see Note 6). In these cases, rapid sequence changes are hypothesised to be the result of diversifying/adaptive selection related to their special function.

Signatures of selection (purifying or adaptive) can be further determined when analysing phylogenies based on nucleotide sequences. Nucleotide substitutions can be categorized into synonymous (S) or non-synonymous (N) codon substitutions. Assuming that an appropriate substitution model and maximum likelihood phylogeny was used, one can compare the rates of dS and dN along the tree. A small dN is usually the sign of purifying selection while dN>dS can indicate diversifying selection. Equal dN and dS rates indicate that the sequences are not subject to selection at all. There are excellent online tools to perform such analysis and provide some statistical assessment of the selective forces at play (e.g. the Datamonkey server http://www.datamonkey.org/).

4. NOTES

1. *BLAST and related tools:* BLAST is a free software, if one wants to run jobs locally, or can be run remotely through its host at the NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi). When using BLAST one can specify the database that should be searched. The non-redundant protein/nucleotide sequence databases (nr database) are the most comprehensive and include both curated and non-curated sequences [29]. For comparative structural modeling of protein queries, the protein databank [14] that contains sequences associated with experimental structures is appropriate. BLASTp and BLASTn retrieve sequences based on similarity to the input sequence. This means that the alignment between the query and the hit must be sufficient to be reported, and works well with closely related sequences, but has clear limitations with increased divergence. Because protein databases used for BLASTp can be incomplete with annotations of certain variants missing and coding sequences can accumulate

synonymous substitutions that limit BLASTn searches, these search algorithms might not return the true orthologs of the query. Instead, given that related histone genes are also often present in the genomes, BLASTp and BLASTn might return hits that only look closely related leading to wrong conclusions. We advise to systematically run translated nucleotide BLAST (tBLASTn) to take into account these limitations.

2. False-positives in HMM-based predictions: In contrast to the *E* value in the BLAST output, no general rule exists which probability value is considered to be sufficiently high to conclude true homology. Instead, the alignment needs to be manually inspected using criteria like similarity of the secondary structure, hydrophobicity profiles and potential gaps.

While these algorithms are currently the most sensitive approaches to reveal remote homologies, the false-positive rate among hits can be notably high. Therefore, additional criteria are often needed to evaluate putative histone-fold candidates. These can include short sequence motifs specific to certain variants and the overall protein domain architecture. Still, putative homologues should generally be experimentally verified.

3. Histone DB: As mentioned by the curators of this resource, not all variants might be represented even though new variants are recurrently added.

4. Informative MSA: As the parameters and algorithms differ, it is generally advised to use at least two different programs to construct the MSAs. Also, when possible, one should try to perform and compare both DNA and protein sequence alignment since they inform on different level of selective pressure (codons *vs.* amino acids). In order to pick the "right" number and type of sequences to use, one should consider the overall amount of diversity present in the alignment. Having too many nearly identical sequences will not provide enough information to discriminate between alternative evolutionary scenarios. The same holds true if too little identity exist in the alignment.

5. Long-branch attractions: When dealing with highly diverged sequences, as it is the case for cenH3 and short H2As or over large evolutionary span, one should pay special attention to the branching patterns of the longest branches. Indeed, a classical caveat to phylogenetic trees, is that sequences that have accumulated many mutations share little homology to the rest of the alignment, and are thus artificially brought together during tree building. They may wrongfully appear as sharing a common ancestry. Using an appropriate substitution model can greatly improve this issue. One should always try to use alternative methods (synteny, species phylogeny…) to infer the best relationship of the sequences subject to long branch attraction.

6. Gene conversion: When interpreting the topology of a phylogeny, concerted evolution (the fact that paralogs share more identity than their true orthologs) can create wrong assumptions. In such case, duplicates might cluster together by species suggesting independent origins instead of forming separated clades suggestive of independent evolution from an ancestral duplication. Mammalian short H2As and, ironically, canonical histones display such complex phylogenies. While canonical histone genes seem to be subject to a "birth-and-death" mode of evolution [30], short H2As duplicates are undergoing gene conversion, through non-homologous recombination, which homogenizes their sequences within lineages (Molaro et al., 2017, in press). When confronted to such topologies, synteny can greatly inform on the ancestry of the duplication and help resolve confusing phylogenies and uncover interesting modes of evolution like the ones mentioned above.

Figure 1: Workflow to annotate and characterize new histone variants

Table 1: A summary table of histone variants and their characteristics

Note that frequently not all characteristics are common to all homologs of a particular variant and should therefore only be considered as general guidelines. Experimental validations are often needed to confirm the function of candidates. Furthermore, conservation across a specific group of organisms (e.g. Ophistokonts) does not imply that every eukaryote encodes this variant.

Acknowledgments

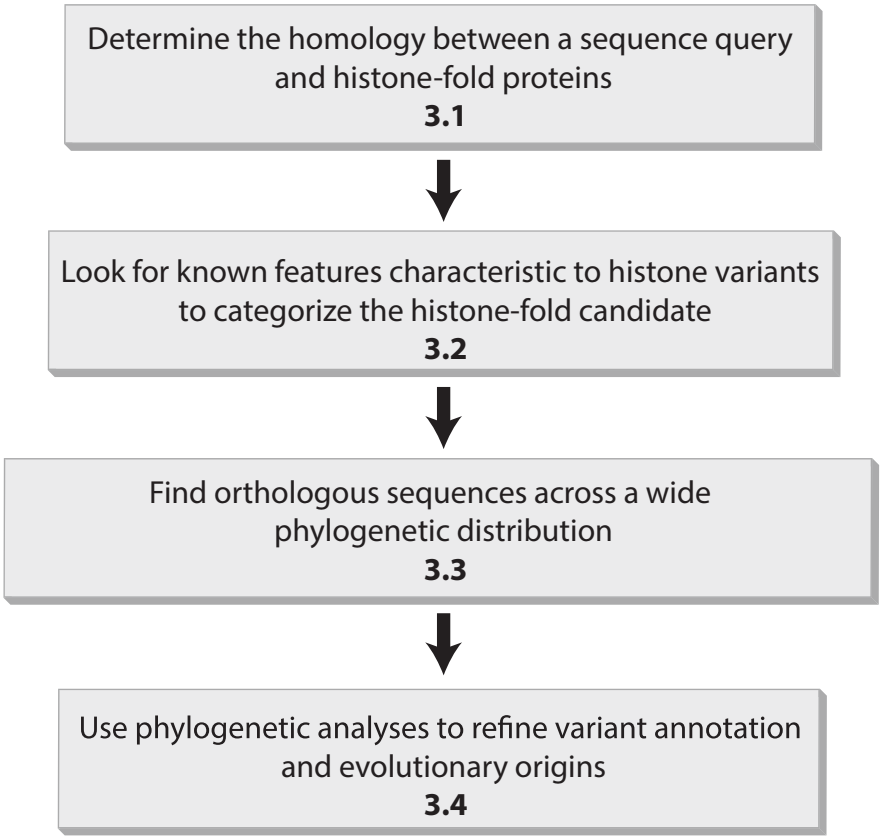
We would like to thank Paul Talbert and members of the Drinnenberg lab for comments on the manuscript. I.A.D. receives salary support from the CNRS. This work is supported by the Labex DEEP ANR-11-LABX-0044 part of the IDEX Idex PSL ANR-10-IDEX-0001-02 PSL, the Institut Curie and funds from the Atip Avenir 2015 program. A. M. was supported by a Damon Runyon Cancer research foundation (DRG:2192-14).

1. Kornberg RD (1974) Chromatin structure: a repeating unit of histones and DNA. Science 184 (4139):868-871
2. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 389 (6648):251-260. doi:10.1038/38444
3. Marzluff WF, Wagner EJ, Duronio RJ (2008) Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. Nature reviews Genetics 9 (11):843-854. doi:10.1038/nrg2438
4. Talbert PB, Henikoff S (2010) Histone variants--ancient wrap artists of the epigenome. Nature reviews Molecular cell biology 11 (4):264-275. doi:10.1038/nrm2861
5. Malik HS, Henikoff S (2003) Phylogenomics of the nucleosome. Nature structural biology 10 (11):882-891. doi:10.1038/nsb996
6. Kawashima T, Lorkovic ZJ, Nishihama R, Ishizaki K, Axelsson E, Yelagandula R, Kohchi T, Berger F (2015) Diversification of histone H2A variants during plant evolution. Trends in plant science 20 (7):419-425. doi:10.1016/j.tplants.2015.04.005
7. Draizen EJ, Shaytan AK, Marino-Ramirez L, Talbert PB, Landsman D, Panchenko AR (2016) HistoneDB 2.0: a histone database with variants--an integrated resource to explore histones and their variants. Database : the journal of biological databases and curation 2016. doi:10.1093/database/baw014
8. Coordinators NR (2016) Database resources of the National Center for Biotechnology Information. Nucleic acids research 44 (D1):D7-19. doi:10.1093/nar/gkv1290
9. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome research 12 (6):996-1006. doi:10.1101/gr.229102. Article published online before print in May 2002
10. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of molecular biology 215 (3):403-410. doi:10.1016/S0022-2836(05)80360-2
11. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research 25 (17):3389-3402
12. Schleiffer A, Maier M, Litos G, Lampert F, Hornung P, Mechtler K, Westermann S (2012) CENP-T proteins are conserved centromere receptors of the Ndc80 complex. Nature cell biology 14 (6):604-613. doi:10.1038/ncb2493
13. Soding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic acids research 33 (Web Server issue):W244-248. doi:10.1093/nar/gki408
14. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic acids research 28 (1):235-242
15. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A (2010) The Pfam protein families database. Nucleic acids research 38 (Database issue):D211-222. doi:10.1093/nar/gkp985
16. Whelan S, Morrison DA (2017) Inferring Trees. Methods in molecular biology 1525:349-377. doi:10.1007/978-1-4939-6622-6_14
17. Bawono P, Dijkstra M, Pirovano W, Feenstra A, Abeln S, Heringa J (2017) Multiple Sequence Alignment. Methods in molecular biology 1525:167-189. doi:10.1007/978-1-4939-6622-6_8
18. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC bioinformatics 5:113. doi:10.1186/1471-2105-5-113
19. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research 32 (5):1792-1797. doi:10.1093/nar/gkh340
20. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology 7:539. doi:10.1038/msb.2011.75
21. Blackshields G, Sievers F, Shi W, Wilm A, Higgins DG (2010) Sequence embedding for fast construction of guide trees for multiple sequence alignment. Algorithms for molecular biology : AMB 5:21. doi:10.1186/1748-7188-5-21
22. Yang Z (2005) The power of phylogenetic comparison in revealing protein function. Proceedings of the National Academy of Sciences of the United States of America 102 (9):3179-3180. doi:10.1073/pnas.0500371102
23. Wilgenbusch JC, Swofford D (2003) Inferring evolutionary trees with PAUP*. Curr Protoc Bioinformatics Chapter 6:Unit 6 4. doi:10.1002/0471250953.bi0604s00
24. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution 4 (4):406-425
25. Shapiro B, Rambaut A, Drummond AJ (2006) Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Molecular biology and evolution 23 (1):7-9. doi:10.1093/molbev/msj021
26. Whelan S, Allen JE, Blackburne BP, Talavera D (2015) ModelOMatic: fast and automated model selection between RY, nucleotide, amino acid, and codon substitution models. Syst Biol 64 (1):42-55. doi:10.1093/sysbio/syu062
27. Darriba D, Taboada GL, Doallo R, Posada D (2011) ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27 (8):1164-1165. doi:10.1093/bioinformatics/btr088

28. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59 (3):307-321. doi:10.1093/sysbio/syq010
29. Coordinators NR (2017) Database Resources of the National Center for Biotechnology Information. Nucleic acids research 45 (D1):D12-D17. doi:10.1093/nar/gkw1071
30. Rooney AP, Piontkivska H, Nei M (2002) Molecular evolution of the nontandemly repeated genes of the histone 3 multigene family. Molecular biology and evolution 19 (1):68-75
31. Contrepois K, Coudereau C, Benayoun BA, Schuler N, Roux PF, Bischof O, Courbeyrette R, Carvalho C, Thuret JY, Ma Z, Derbois C, Nevers MC, Volland H, Redon CE, Bonner WM, Deleuze JF, Wiel C, Bernard D, Snyder MP, Rube CE, Olaso R, Fenaille F, Mann C (2017) Histone variant H2A.J accumulates in senescent cells and promotes inflammatory gene expression. Nature communications 8:14995. doi:10.1038/ncomms14995
32. Marzluff WF, Gongidi P, Woods KR, Jin J, Maltais LJ (2002) The human and mouse replication-dependent histone genes. Genomics 80 (5):487-498
33. Churikov D, Siino J, Svetlova M, Zhang K, Gineitis A, Morton Bradbury E, Zalensky A (2004) Novel human testis-specific histone H2B encoded by the interrupted gene on the X chromosome. Genomics 84 (4):745-756. doi:10.1016/j.ygeno.2004.06.001
34. Boulard M, Gautier T, Mbele GO, Gerson V, Hamiche A, Angelov D, Bouvet P, Dimitrov S (2006) The NH2 tail of the novel histone variant H2BFWT exhibits properties distinct from conventional H2B with respect to the assembly of mitotic chromosomes. Molecular and cellular biology 26 (4):1518-1526. doi:10.1128/MCB.26.4.1518-1526.2006
35. Dalmasso MC, Echeverria PC, Zappia MP, Hellman U, Dubremetz JF, Angel SO (2006) Toxoplasma gondii has two lineages of histones 2b (H2B) with different expression profiles. Mol Biochem Parasitol 148 (1):103-107. doi:10.1016/j.molbiopara.2006.03.005
36. Dalmasso MC, Sullivan WJ, Jr., Angel SO (2011) Canonical and variant histones of protozoan parasites. Front Biosci (Landmark Ed) 16:2086-2105
37. Santoro SW, Dulac C (2012) The activity-dependent histone variant H2BE modulates the life span of olfactory neurons. eLife 1:e00070. doi:10.7554/eLife.00070
38. Waterborg JH (2012) Evolution of histone H3: emergence of variants and conservation of post-translational modification sites. Biochemistry and cell biology = Biochimie et biologie cellulaire 90 (1):79-95. doi:10.1139/o11-036
39. Postberg J, Forcob S, Chang WJ, Lipps HJ (2010) The evolutionary history of histone H3 suggests a deep eukaryotic root of chromatin modifying mechanisms. BMC evolutionary biology 10:259. doi:10.1186/1471-2148-10-259
40. Schenk R, Jenke A, Zilbauer M, Wirth S, Postberg J (2011) H3.5 is a novel hominid-specific histone H3 variant that is specifically expressed in the seminiferous tubules of human testes. Chromosoma 120 (3):275-285. doi:10.1007/s00412-011-0310-4
41. Urahama T, Harada A, Maehara K, Horikoshi N, Sato K, Sato Y, Shiraishi K, Sugino N, Osakabe A, Tachiwana H, Kagawa W, Kimura H, Ohkawa Y, Kurumizaka H (2016) Histone H3.5 forms an unstable nucleosome and accumulates around transcription start sites in human testis. Epigenetics & chromatin 9:2. doi:10.1186/s13072-016-0051-y
42. Wiedemann SM, Mildner SN, Bonisch C, Israel L, Maiser A, Matheisl S, Straub T, Merkl R, Leonhardt H, Kremmer E, Schermelleh L, Hake SB (2010) Identification and characterization of two novel primate-specific histone H3 variants, H3.X and H3.Y. The Journal of cell biology 190 (5):777-791. doi:10.1083/jcb.201002043
43. Witt O, Albig W, Doenecke D (1996) Testis-specific expression of a novel human H3 histone gene. Experimental cell research 229 (2):301-306. doi:10.1006/excr.1996.0375
44. Nishino T, Takeuchi K, Gascoigne KE, Suzuki A, Hori T, Oyama T, Morikawa K, Cheeseman IM, Fukagawa T (2012) CENP-T-W-S-X forms a unique centromeric chromatin structure with a histone-like fold. Cell 148 (3):487-501. doi:10.1016/j.cell.2011.11.061
45. Albig W, Ebentheuer J, Klobeck G, Kunz J, Doenecke D (1996) A solitary human H3 histone gene on chromosome 1. Human genetics 97 (4):486-491
46. Chakravarthy S, Bao Y, Roberts VA, Tremethick D, Luger K (2004) Structural characterization of histone H2A variants. Cold Spring Harb Symp Quant Biol 69:227-234. doi:10.1101/sqb.2004.69.227
47. Eirin-Lopez JM, Gonzalez-Romero R, Dryhurst D, Ishibashi T, Ausio J (2009) The evolutionary differentiation of two histone H2A.Z variants in chordates (H2A.Z-1 and H2A.Z-2) is mediated by a stepwise mutation process that affects three amino acid residues. BMC evolutionary biology 9:31. doi:10.1186/1471-2148-9-31
48. Faast R, Thonglairoam V, Schulz TC, Beall J, Wells JR, Taylor H, Matthaei K, Rathjen PD, Tremethick DJ, Lyons I (2001) Histone variant H2A.Z is required for early mammalian development. Current biology : CB 11 (15):1183-1187
49. Pehrson JR, Fried VA (1992) MacroH2A, a core histone containing a large nonhistone region. Science 257 (5075):1398-1400
50. Pehrson JR, Fuji RN (1998) Evolutionary conservation of histone macroH2A subtypes and domains. Nucleic acids research 26 (12):2837-2842
51. Rivera-Casas C, Gonzalez-Romero R, Cheema MS, Ausio J, Eirin-Lopez JM (2016) The characterization of macroH2A beyond vertebrates supports an ancestral origin and conserved role for histone variants in chromatin. Epigenetics 11 (6):415-425. doi:10.1080/15592294.2016.1172161
52. Eirin-Lopez JM, Ishibashi T, Ausio J (2008) H2A.Bbd: a quickly evolving hypervariable mammalian histone that destabilizes nucleosomes in an acetylation-independent way. FASEB J 22 (1):316-326. doi:10.1096/fj.07-9255com
53. Govin J, Escoffier E, Rousseaux S, Kuhn L, Ferro M, Thevenon J, Catena R, Davidson I, Garin J, Khochbin S, Caron C (2007) Pericentric heterochromatin reprogramming by new histone variants during mouse spermiogenesis. The Journal of cell biology 176 (3):283-294. doi:10.1083/jcb.200604141
54. Ferguson L, Ellis PJ, Affara NA (2009) Two novel mouse genes mapped to chromosome Yp are expressed specifically in spermatids. Mamm Genome 20 (4):193-206. doi:10.1007/s00335-009-9175-8
55. Shaytan AK, Landsman D, Panchenko AR (2015) Nucleosome adaptability conferred by sequence and structural variations in histone H2A-H2B dimers. Curr Opin Struct Biol 32:48-57. doi:10.1016/j.sbi.2015.02.004

Variant	Conservation	Origin	Characteristics to canonical form *	References
H2A.Z	Eukaryotes	monophyletic	60% protein sequence identity to H2A Under strong purifying selection Divergent C-terminal end - one a.a ** insertion in loop L1 7 diverged residues within loop L1/α2 "DEELD" motif in the acidic patch - truncation of αC. Duplication in vertebrates: H2A.Z.1 and .2 - 3 divergent a.a ** : T/A15, S/T39 and V/A128.	3, 4, 43-45
H2A.X	Eukaryotes	polyphyletic	95% protein sequence identity to H2A along the HFD C-term SQ(E/D)-phi motif (phi: hydrophobic residue - usually, Ile, Tyr or Phe).	4, 13
macroH2A	Metazoa	monophyletic	~60% protein sequence identity to H2A along the HFD Extended C-term "macro domain" (Pfam:PF01661) Duplication in vertebrates: macroH2A.1 and macroH2A.2	46-48
Short H2As	Placental mammals	monophyletic	Rapidly evolving, <50% identity with canonical H2A Diverged N-term residues and loss of acidic patch Truncated C-term tail with no Lys at position 119	49-52
H2A.W	Seed Bearing plants	monophyletic	Paralogs have >80% protein sequence identity Diverged loop 1 - "RY-S/A-K/Q" C-term motif "KSPK-K/S-A/K"	5
H2A.M	Moss, Liverwort, Lycophyte, Angiosperms	monophyletic	Similar to H2A.W Additional "KSPK" C-term motif Shorter N-term than H2A.W, at least 6 a.a **	5
H2A.J	Mammals, n.d.	likely monophyletic	90% protein sequence identity to H2A Val at position 11 C-term "SQK" motif variable position 4 a.a. ** from the stop codon.	28
H2A.1	Mammals, n.d.	monophyletic	Ile at position 31 and 44, Ser at position 71 C-term "A/V/S/T-Q-S/A/T" motif. germline restricted	3, 13, 29
H2B.W	Mammals, n.d.	likely monophyletic	<50% protein sequence identity to H2B 30 a.a. extension of N-term sperm restricted	30, 31
H2B.1	Mammals, n.d.	monophyletic	Ile at position 43 many N-terminal substitutions germline restricted	3, 13, 52
H2B.E	Mouse	n.d.	5 diverged residues olfactory neuron restricted	34
H2B.Z	Apicomplexa	likely monophyletic	>90% protein sequence identity to H2B Shorter N-term tail Variable residues within α2	13, 32, 33
H3.3	Eukaryotes	polyphyletic	Under strong purifying selection Ser/Thr at position 31 Often additional differences at position 87, 89, and 90	3, 35, 36
cenH3	Eukaryotes	possibly monophyletic (unclear)	50-60% protein sequence identity to H3 Rapidly evolving extended N-terminal tail (20-200 aa) and loop1 region Lack of Gln at position 68; Trp at position 84 (instead of Phe); Ala, Cys or Ser at position 107 (instead of Thr)	3, 4
H3.5	Hominids	monophyletic	Lack of Lys at position 37	37, 38
H3.T	Mammals	monophyletic	testis-restricted expression Val, Met, Ser and Val at position 24, 72, 98, and 111	40, 42
H3.X	Primates	monophyletic	Several mutations to canonical H3, almost identical to H3.Y	39
H3.Y	Primates	monophyletic	Several mutations to canonical H3, almost identical to H3.X	39
CENP-T	Ophistokonts	likely monophyletic	Rapidly evolving Large N-terminal domains with HFD at the C-terminus C-terminal 2-helical extension	10, 41
CENP-W	Ophistokonts	likely monophyletic	Rapidly evolving	10, 41
CENP-S	Eukaryotes	likely monophyletic	n.d.	10, 41
CENP-X	Eukaryotes	likely monophyletic	n.d.	10, 41
subH2B	Mammals, n.d.	likely monophyletic	Bipartite nuclear localization motif at N-term and C-term. of the Less than 50% identities with canonical H2B	58

* Positions are counted according to the canonical form
** amino acid


Determine the homology between a sequence query and histone-fold proteins
3.1
Look for known features characteristic to histone variants to categorize the histone-fold candidate
3.2
Find orthologous sequences across a wide phylogenetic distribution
3.3
Use phylogenetic analyses to refine variant annotation and evolutionary origins
3.4