Phylogenetic investigation of human tubulin and it's susceptibility to natural selection

Introduction

Microtubules (MTs) are dynamic filaments of the eukaryotic cytoskeleton and highly involved in mitosis, cell motility and intracellular transport. MTs are assembled from the protein tubulin, which is a heterodimer of alpha- and beta-subunits, and the sequence and structure of tubulin subunits is highly conserved among species. Many eukaryotic organisms carry multiple genomic copies of functional α or β tubulin, commonly referred to as isoforms (or isotypes if they are confined to a single organism). In humans for example, 9 genes encode for the α-subunit whereas 10 genes encode the β-subunit which are assembled into functional microtubule polymers. Mutations in different mammalian tubulin proteins have been linked to a wide range of disorders, many of which affect brain development (Moore et al. 2017).

Mutations in specific beta-tubulin isotypes cause severe neuropathies that disrupt axonal transport leading to brain malformations (Huzil et al. 2007). Genetic variations affecting all beta-tubulin genes expressed at high levels in the brain (*TUBB2B, TUBB3*, *TUBB*, *TUBB4A*, and *TUBB2A*) have been linked with malformations of cortical development and disease phenotypes that arise from their disruption, include microcephaly, Lissencephaly and polymicrogyria (Thomas et al. 2014). Beta tubulin is of particular interest since most drugs that counteract these disorders (tubulin related disorders) have their targets directed towards this protein subunit thus making it a good candidate for extensive investigation. In addition, drugs targeting microtubules for cancer chemotherapy bind to beta tubulin (Huzil et al. 2007). Mechanisms of drug action like vinblastine involves destabilizing microtubules and reducing their dynamicity, thus promoting mitotic arrest and eventually apoptosis.

Phylogenetic relationships among the major tubulin proteins groups alpha and beta have not been extensively investigated in a single organism and in humans, for instance alpha tubulin have been well classified but the basis of beta tubulin classification is not well documented and thus not clearly understood. To date, mammalian beta tubulin has been grouped into 9 protein classes, class I, IIa, IIb, III, IVa, IVb, V, VI, VIII (HUGO Gene Nomenclature Committee) which are widely adopted in literature. Despite lack of the consensus phylogeny, it has been shown that class I β-tubulin is the most commonly expressed isotype in humans and as such is also the most common isotype found in cancer cells Alternatively, both classes II and III β-tubulin have been observed at increased levels in human tumors (Ferguson et al. 2005; Mozzetti et al. 2005). There is a high degree of tissue specificity in the expression of some β- tubulins as described and some degree of gene redundancy where the loss of one gene can be compensated by over expression of another as has been shown in yeast (Nsamba et al. unpublished). Here I employ molecular phylogenetics to validate and classify human tubulin isotypes, and at the same time determine the selection pressure and molecular evolutionary relationship between the two major classes of mammalian tubulin.

Methods

Assembling the data

Databases: UniProtKB and NCBI

Protein sequences: The mammalian alpha tubulin protein sequences were similar to those used in Khodiyar et al. 2007 (A revised nomenclature for the human and rodent α-tubulin gene family) whereas the beta tubulin sequences were extracted from UniProtKB database (http://www.uniprot.org/).

Nucleotide sequence: Both α or β-tubulin protein coding gene sequences were extracted from NCBI database using their corresponding gene names. Only the protein-coding portion of each cDNA was used, to prevent differences in length of the UTRs biasing the alignments. cDNA was extracted from (https://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi) a sub-project of NCBI for annotating coding regions.

For phylogenetic analysis, C. elegans α-2 tubulin was used as an outgroup.

Generating the Multiple Sequence Alignment

I employed a MAFFT module available on HPC class to align the sequences. Depending on the type of analysis, I converted the fasta file in different formats that suit a specific phylogenetic analysis for instance, phylip format for Maximum likelihood ad distance analysis.

Computing and Visualizing Phylogenetic Trees

Phylogenetic trees were generated for all data sets using the neighbor joining (NJ) and the maximum likelihood (ML) method. Maximum likelihood was inferred using RAxML7 v.8.0.22 with the GTRGAMMA model of nucleotide substitutions. For Neighbor joining, pairwise distances among taxa was used as input for phylogenetic reconstruction estimated under five different models of amino acid substitutions available in protdist and 4 models of nucleotide substitutions available in dnadist. In all cases, a strict consensus and majority rue trees were built by neighbor joining.

For each analysis, bootstrapping with 1,000 replicates was also performed.

Hypothesis Testing and Detecting Selection with codeml

To determine the rates of evolution in human tubulin genes and ascertain the underlying selection pressure, I used codeml, a PAML (Phylogenetic Analysis by Maximum Likelihood) package, by setting seqtype to 1 and carried out ML analysis using codon substitution models as described in (Goldman and Yang 1994). Both alpha and beta tubulin protein-coding DNA sequences were analyzed separately and dN/dS ratios compared between the two. Omega (ω) ratio ($\omega = d_N/d_S$) is a measure of natural selection acting on the protein and is very informative in understanding natural selection acting on genomes of species

Tree Visualization

Phylogenetic trees used in this study were visualized with Figtree, Newick tree viewer whereas Strict and Majority rule trees used in supplemental data were visualized with Dendrocope.

Results and Discussion

Selection pressure and evolution rates of alpha and beta sequences; beta tubulin is susceptible to negative selection and alpha tubulin positive selection

Consistent with my hypothesis, the omega ($\omega = dN/dS$) between alpha and beta was found to be different (Table 1). Hypothesis testing using codeml revealed that beta tubulin genes are under negative selection pressure ($\omega < 1$) whereas alpha tubulin genes are under positive selection ($\omega > 1$, Table 1). Negative selection refers to removal of deleterious alleles resulting in stabilizing selections through purging of deleterious variants that arise giving them less chances of survival. The magnitude of selection difference between alpha and beta is very strong 61x (Table 1) which be explained in relation to the recent discovery of tubulin disorders (largely known as Tubulinopathies). At least 60% of these disorders are associated with mutations in the beta tubulin genes (Markova et al. 2015). The fact that such mutations are deleterious, they are quickly removed from the gene pool and those that persist cause tubulin diseases. However, such cases are rare as carriers of these harmful mutations have fewer offspring each generation thus reducing the frequency of the mutation in the gene pool.

Furthermore, under Ohta's hypothesis that most amino acid substitutions are deleterious (Gillespie. 1994), purifying selection is more effective in large populations than in small populations, and so differences in population sizes along lineages provide another compatible hypothesis. If amino acid changes are slightly deleterious, we expect them to be removed from the population at a higher rate in a large population than in a small population. As a result, we expect to see a smaller dN/dS ratio in a large population than in a small one, even if there is no difference between the two lineages in selective pressure or gene function. In the context of population sizes, beta tubulin evolved with more protein coding genes (10) than alpha (09) and although the difference is not very significant, it concurs with a smaller dN/dS ratio observed. We thus speculate that although both protein subunits are very essential for MT function, beta tubulin might be conferring specific roles to microtubule function whose perturbation is very penalizing and accumulation of deleterious mutations increases a risk to tubulin diseases.

Table 1. Omega values for alpha and beta tubulin sequences

Nucleotide sequence	dN	dS	w = dN/dS	Ratio of alpha to beta
Beta-genes	0.2460	9.1466	0.02690	65
Alpha-genes	0.9019	0.5141	1.75421	

Gene family evolution of mammalian alpha and beta tubulin genes and the underlying selection pressure

I used Phylogenetic analysis to define the nomenclature of both alpha and beta tubulin protein groups (https://www.genenames.org/, Tables 1 and 2). The basis of alpha tubulin classification has been proposed in literature (Khodiyar et al. 2007) whereas the classification of beta tubulin is not well documented.

Table 2. Classes of beta tubulin genes (https://www.genenames.org/)

Gene name (CDS) Approved symbol	Class (Proposed)	Assigned (New Class)
TUBB	tubulin beta class I	I
TUBB1	tubulin beta 1 class VI	V
TUBB2A	tubulin beta 2A class IIa	II
TUBB2B	tubulin beta 2B class IIb	
TUBB3	tubulin beta 3 class III	III
TUBB4A	tubulin beta 4A class IVa	IV
TUBB4B	tubulin beta 4B class IVb	
TUBB6	tubulin beta 6 class V	III
TUBB8	tubulin beta 8 class VIII	VI
TUBB7P	tubulin beta 7 pseudogene	Not classified

Table 3: Classification of alpha tubulin genes

Approved Symbol	Approved Name	Phylogenetic group (Assigned)
TUBA1A	tubulin alpha 1a	Group 1
TUBA1B	tubulin alpha 1b	
TUBA1C	tubulin alpha 1c	
TUBA3C	tubulin alpha 3c	Group 3
TUBA3D	tubulin alpha 3d	
TUBA3E	tubulin alpha 3e	
TUBA4A	tubulin alpha 4a	Group 2
TUBA4B	tubulin alpha 4b	
TUBA8	tubulin alpha 8	Group 4

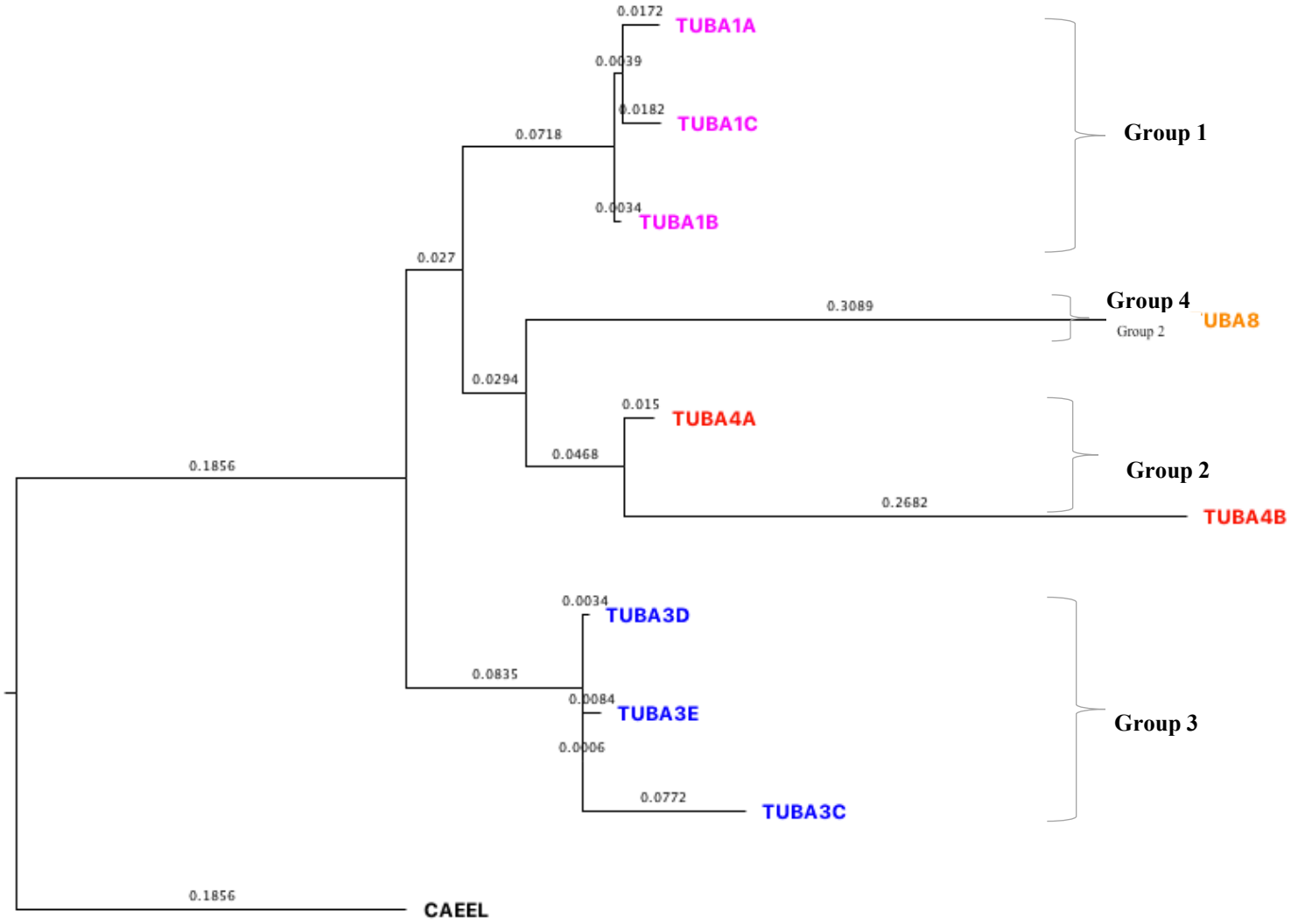
From the phylogenetic analysis of protein coding sequences of alpha tubulin genes, it can be observed that all the 08 alpha protein coding genes are homologous to each other and are clustered into four subgroups (Fig.1A, and Fig.2). This was consistent with the classification proposed by HUGO Gene Nomenclature Committee, (Table 3 and Khodiyar et al. 2007). Beta-tubulin protein coding sequences were also observed to be homologous to each other although they were clustered into 6 protein classes protein instead of 7 classes as reported in literature (Fig. 1B and Fig. 2). *TUBB6* and *TUBB3* are sister to each other on a monophyletic clade and thus belong to the same beta protein class (III) than belonging to classes V and III respectively (Table2) as previously classified.

When the two protein groups were combined together and have their amino acid sequences analyzed by neighbor joining, the resulting tree shows us that the 2 protein groups are paralogs (Fig 2), implying that they arose as a result of gene duplication in the MRCA (Most Recent Common Ancestor). However, within each protein subgroup, all proteins are homologous to each other. After critical investigation, I found out that the tree topologies constructed from the protein coding nucleotide sequences (Fig. 1A and B) and amino acid sequences (Fig 2) were conflicting. Such discrepancy could have resulted from different phylogenetics methods I employed to

construct tree topologies. It should be noted Maximum likelihood (used for protein coding nucleotide sequences) and neighbor joining (used for amino acid sequences) all return unrooted trees which can be rooted anywhere depending on the preferred topology and outgroup included. However, since I included outgroups for phylogenies of alpha (Fig 1A) and beta (Fig 1B), the maximum likelihood tree in this situation provides the most optimum and desirable tree topology. No outgroup was included in Fig 2 since the two proteins were expected to be paralogs.

In the nucleotide-based analysis human *TUBA4B* is sister to *TUBA4B* and all falls within group 2 (Fig 1A, S4), whereas the amino acid-based analysis positions *TUBA4B* outside the outgroup chosen for the phylogenetic analysis (see Fig 2). This is because the human *TUBA4A* protein is 448 amino acids in length, whereas the human *TUBA4B* protein is only 241 amino acids making it have a longer branch length respective of other sequences and with neighbor joining it was likely to be closer to the outgroup. In addition, the alignment of the *TUBA4A* cDNA to the *TUBA4B* genomic sequence provides evidence that another tubulin-related exon containing multiple frameshifts and stop codons is present in TUBA4B as a pseudo gene and thus not transcribed. All these could have led to the observed conflicting topologies in the classification of alpha tubulin protein group. The same applies to *TUBB4A* of the beta tubulin (Fig 1B, S3) where it forms an outgroup for protein classes I and III after the nucleotide-based analysis and thus assigned a distinct class V whereas after the amino acid-based analysis it becomes sister to TUBA4A through a monophyletic clade. (Fig 2, and supplementary figure S1). The reasons are in the follow up discussion below.

Both the nucleotide and protein based phylogenetic analyses are not representative of the tissue-specific expression of the major human b-tubulin isotypes (Antona et al. 2009). Through the use of RT-PCR, Antona et al. 2009 discovered that tissues with the highest beta- tubulin expression were thymus for TUBB, peripheral blood leukocytes for *TUBB1*; brain for *TUBB2A*, *TUBB2B*, *TUBB3*, and *TUBB4* and heart for *TUBB2C* and TUBB6. However, such expression patterns are not representative clades on protein based phylogenetic analysis of human beta tubulin (Fig. 2) and are also not consistent with topology returned by use protein-coding (cDNA) sequence (Fig. 1B). In contrast, the phylogenetic grouping represents the specific gene loci to a greater degree such as *TUBB2A*, *TUBB2B* are both positioned on chromosome 6. However, *TUBB2A* although highly expressed in most tissues than *TUBB2B*, it was not found after examination of the Pan troglodytes (Common chimpanzee) genome which shares a common recent ancestor with humans implying that *TUBB2A* was recently acquired in humans by duplication of *TUBB2B* (Antona et al. 2009).


0.0172
TUBA1A
0.0039
0.0182
TUBA1C
0.0718
Group 1
0.0034
TUBA1B
0.027
0.3089
Group 4
Group 2
UBA8
0.0294
0.015
TUBA4A
0.1856
0.0468
Group 2
0.2682
TUBA4B
0.0034
TUBA3D
0.0835
0.0084
TUBA3E
Group 3
0.0006
0.0772
TUBA3C
0.1856
CAEEL

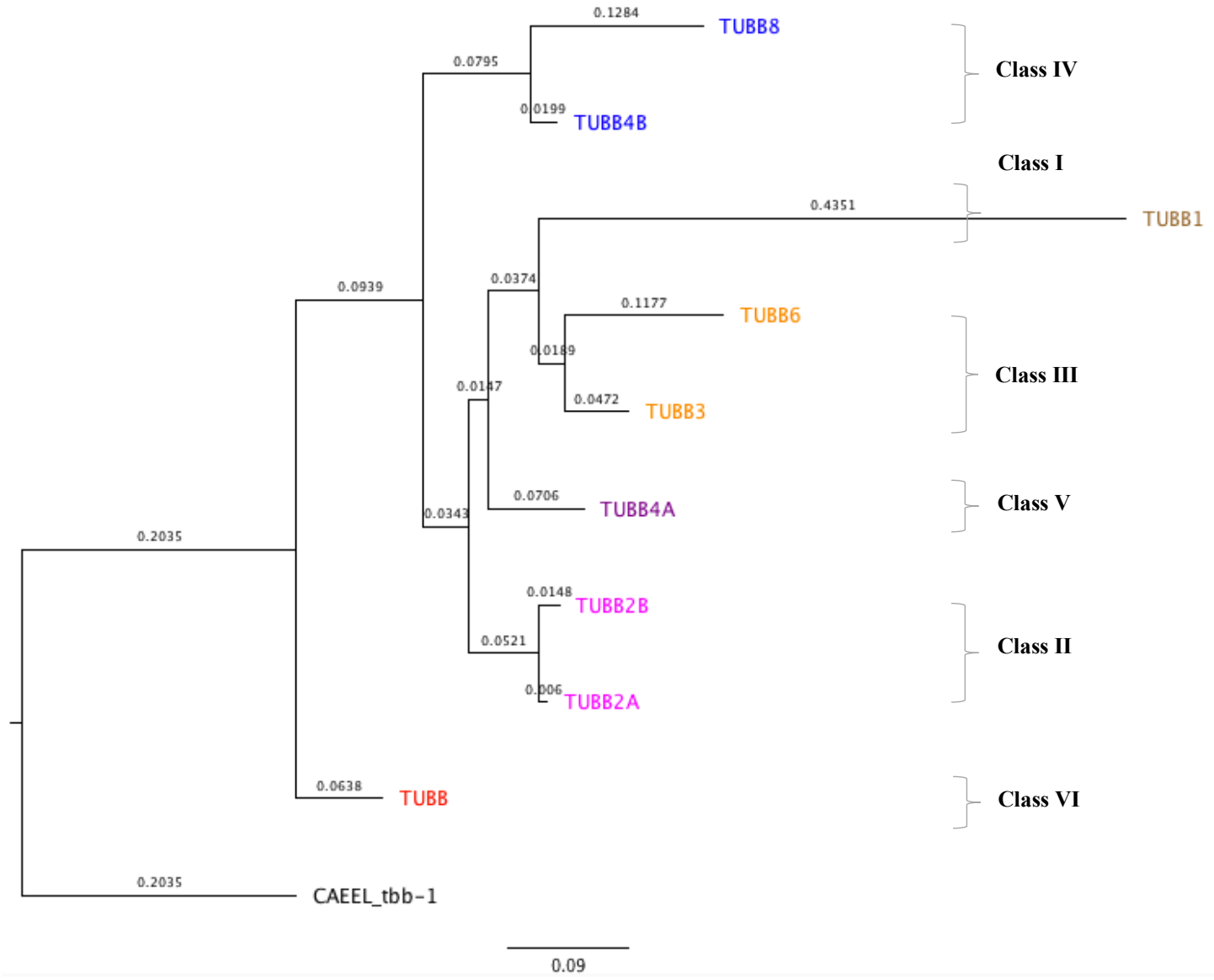


Figure 1. (A) Phylogenetic analysis of human α-tubulin genes (nucleotide coding protein sequences) by Maximum likelihood using the following settings. (1) Method; Maximum likelihood using GTRGAMMA model of nucleotide substitution. (b) substitutions to include: transitions + transversions; (c) rates among sites: uniform rates. (2) Include sites (a) gaps/missing data: complete deletion; (b) codon positions: 1st+2nd+3rd+noncoding. (b) Numbers on the branches refer to the bootstrap values after 1000 replicates. Human sequences, used in this analysis are listed in Table 3 and rooted with *Caenorhabditis elegans* α-2 tubulin CAEEL.
(B). Phylogenetic analysis of human b-tubulin genes (nucleotide coding protein sequences) by Maximum likelihood using similar settings as for a-tubulin. Monophyletic clades clustering beta-tubulin genes as a result of branch specific duplication events are represented as classes. Tubulin protein names/classes used in this analysis are listed in table 2. Numbers on the branches refer to the bootstrap values after 1000 replicates. Phylogenetic trees were visualized by fig tree and represent protein sub-groups that correspond to branch specific gene duplication events and colored for clarity

2A

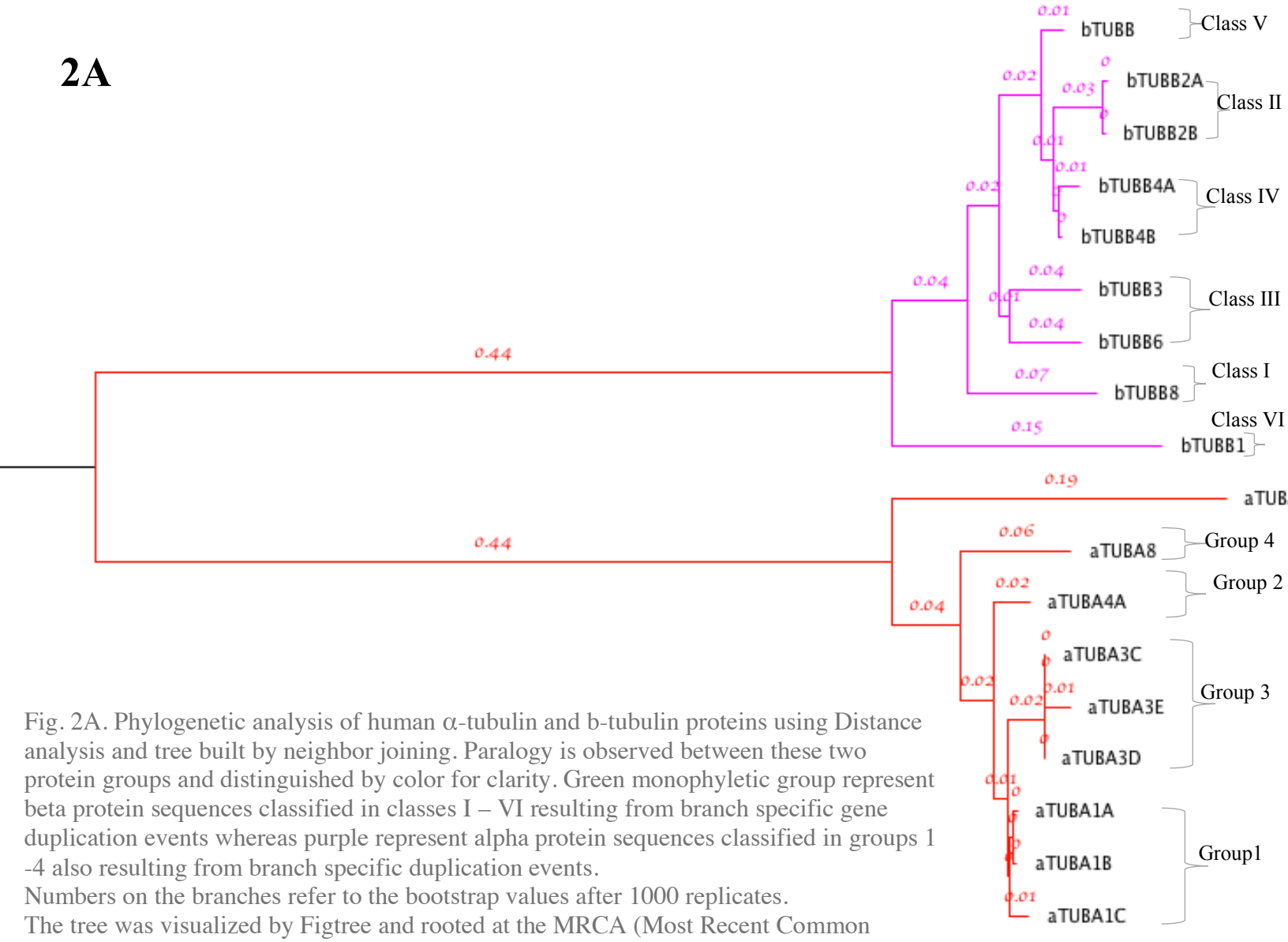


Fig. 2A. Phylogenetic analysis of human α-tubulin and b-tubulin proteins using Distance analysis and tree built by neighbor joining. Paralogy is observed between these two protein groups and distinguished by color for clarity. Green monophyletic group represent beta protein sequences classified in classes I – VI resulting from branch specific gene duplication events whereas purple represent alpha protein sequences classified in groups 1 -4 also resulting from branch specific duplication events.
Numbers on the branches refer to the bootstrap values after 1000 replicates.
The tree was visualized by Figtree and rooted at the MRCA (Most Recent Common Ancestor) of the two protein groups.

Conclusion

There are multiple copies of tubulin genes across the eukaryotic kingdom. Cell biologists working with different species started isolating tubulin genes and giving them randomized names but when they became too many, the naming became complicated resulting in orthologous genes having different nomenclatures. The other discrepancy is due to the fact that these genes are highly similar, co-expressed and sometimes interchangeable (Nsamba et al. un published). For example, on the protein level, human alpha tubulin proteins are 90% identical and 72% identical at the nucleotide level. The HUGO Gene Nomenclature Committee is a web database whose goal is to assign a unique and meaningful name to every human gene.

The previous nomenclature of these proteins is not consistent due to the reasons cited above however, with the advancement in molecular phylogenetics, discrepancy in naming can be resolved and classification updated (Khodiyar et al. 2007). Here, I defined evolutionary relationships of mammalian tubulin and predicted true nomenclature after aligning the sequences first and subsequently building a tree by employing maximum likelihood and distance analysis. Validating the nomenclature of mammalian beta and alpha tubulin protein coding nucleotide sequences resulted into 4 groups for alpha tubulin and 6 classes for beta tubulin than earlier reported. These groups correspond to the prefixes 1, 2, 3 and 4 with the sister members (homologs) assigned with letters A, B, C depending on their number. Beta tubulin protein classes range from I to V1. In addition, hypothesis testing using codeml has revealed to us that alpha and beta tubulin sequences are exposed to different rates of selection, alpha being under positive selection whereas beta negative selection. We thus conclude that phylogenetic analysis is a powerful tool to understand evolutionary relationships among species and on the genome level.

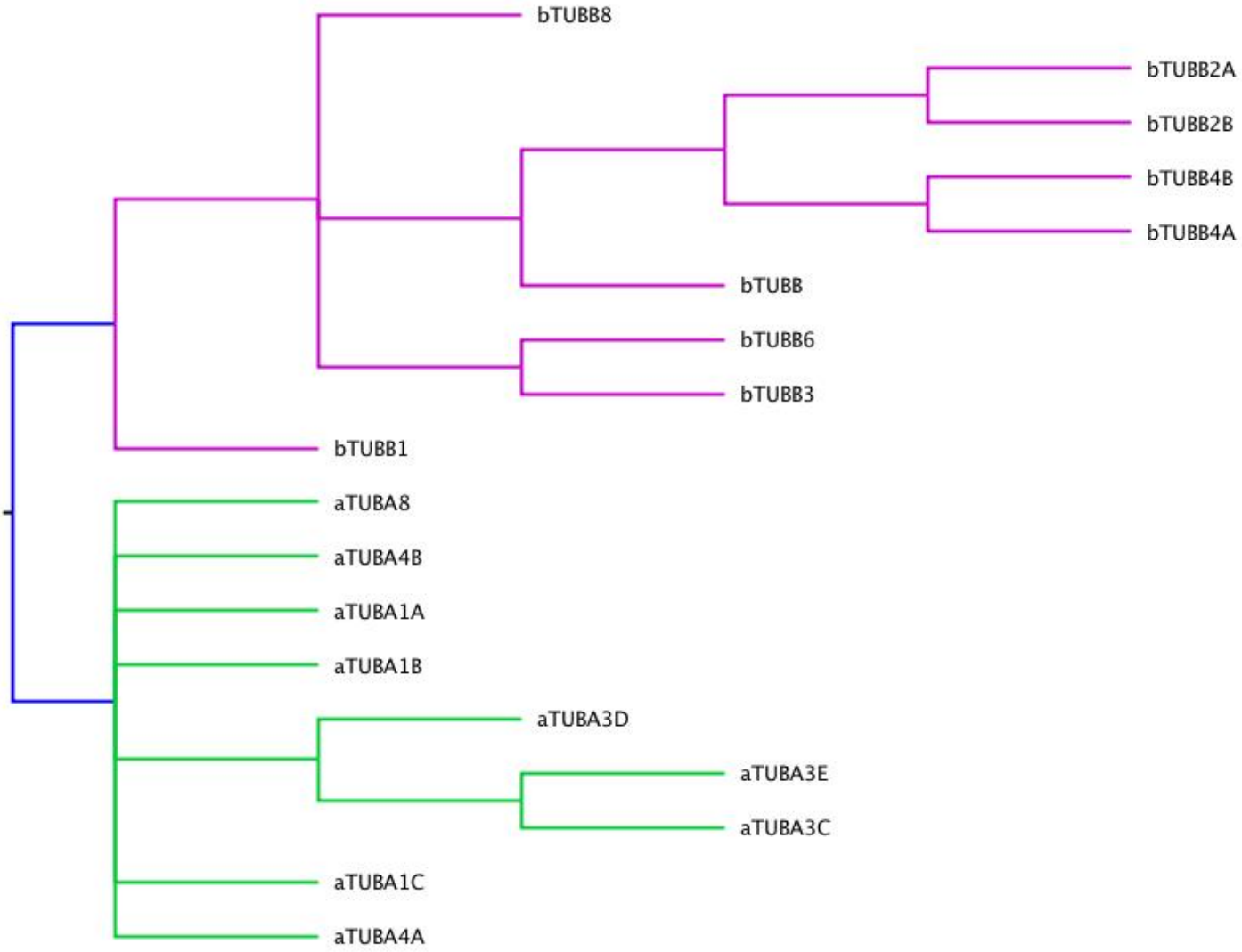
References

1. Huzil et al, 2007, The Roles of β-Tubulin Mutations and Isotype Expression in Acquired Drug Resistance. Cancer Informatics 2007: 3 159–181
2. Huzil et al, 2007, The Roles of β-Tubulin Mutations and Isotype Expression in Acquired Drug Resistance. Cancer Informatics 2007: 3 159–181
3. Thomas D. Cushion et a., 2014, De Novo Mutations in the Beta-Tubulin Gene TUBB2A Cause Simplified Gyral Patterning and Infantile-Onset Epilepsy. The American Journal of Human Genetics 94, 634–641, April 3, 2014
4. Adachi, J., and M. Hasegawa. 1996a. MOLPHY Version 2.3: Programs for molecular phylogenetics based oÂn maximum likelihood. Computer science monographs, 28:1-150. Institute of Statistical Mathematics, Tokyo.
5. Adachi, J., and M. Hasegawa. 1996b. Model of amino acid substitution in proteins encoded by mitochondrial DNA. Journal of Molecular Evolution 42:459-468.[**Entrez**]
6. Markova et al, 2015; Genetic Disorders Affecting Tubulin Cytoskeleton. Journal of Biomedical and Clinical Research
7. http://envgen.nox.ac.uk/bioinformatics/docs/codeml.html
8. http://abacus.gene.ucl.ac.uk/software/pamlFAQs.pdf
9. A. Villasante, *et al..1986*. Six mouse alpha-tubulin mRNAs encode five distinct isotypes: testis-specific expression of two sister genes. Mol. Cell. Biol., 6 (1986), pp. 2409-2419
10. Cristina Rodr ´ıguez-Antona et al 2010. Tumoral and Tissue-Specific Expression of the Major Human b-Tubulin Isotypes. Cytoskeleton, April 2010 67:214–223 (doi: 10.1002/cm.20436)
11. John H. Gillespie, 1995. On Ohta's hypothesis: Most amino acid substitutions are deleterious. J Mol Evol (1995) 40: 64. https://doi.org/10.1007/BF00166596.
12. Jayne Aiken 1, Georgia Buscaglia 2, Emily A. Bates 2 and Jeffrey K. Moore 1, 2017. The a-tubulin gene TUBA1A in brain development: A Key Ingredient in the neuronal isotype blend. J. Dev. Biol. 2017, 5, 8; doi:10.3390/jdb5030008

Supplemental figures

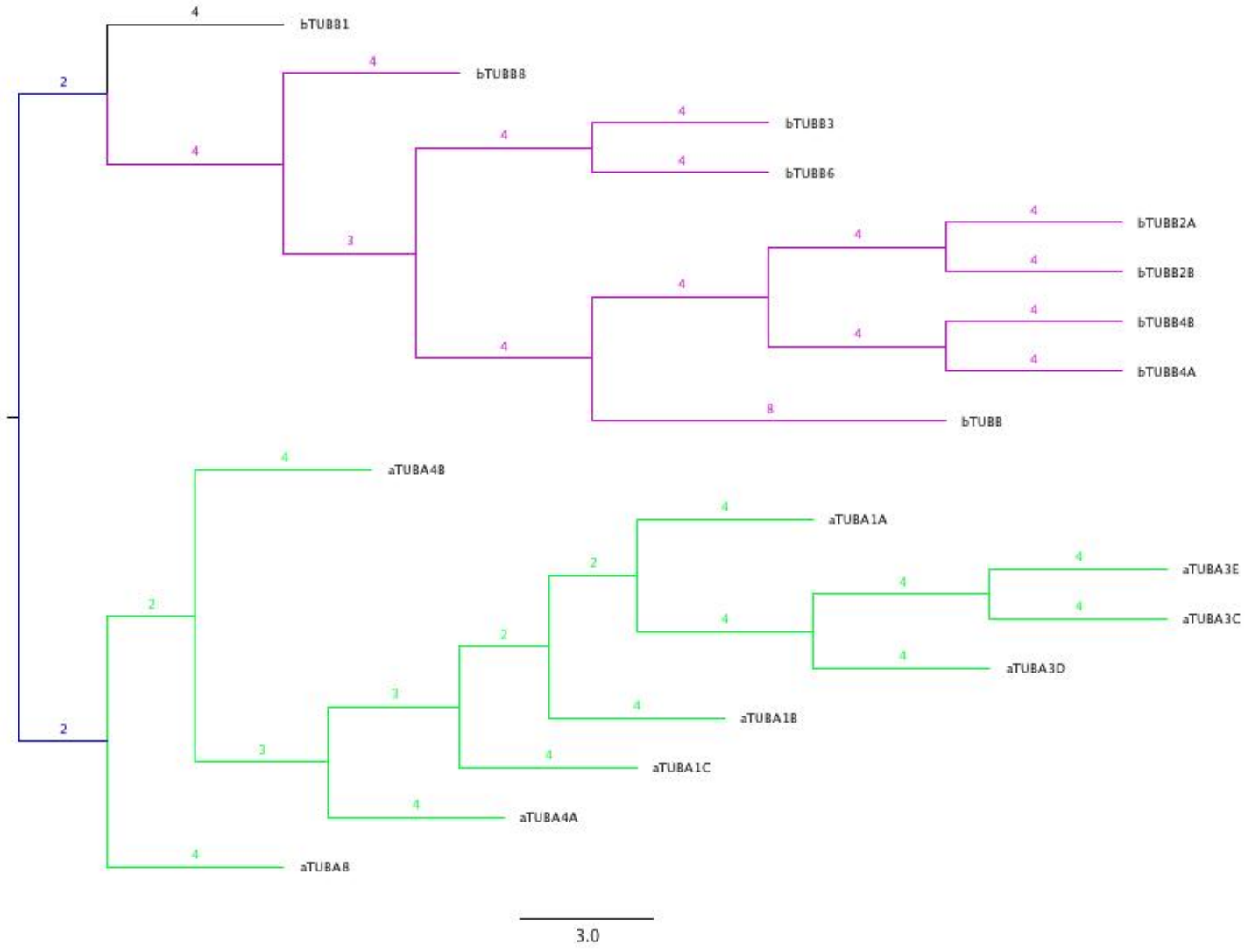
S1

Strict consensus tree for beta-tubulin proteins constructed by neighbor joining method from 4 models of amino acid substitution available in protdist, JTT, PMB, PAM and Kimura. The tree is colored to represent the major tubulin protein groups, purple-beta-tubulin protein isotypes and green-alpha tubulin isotypes.



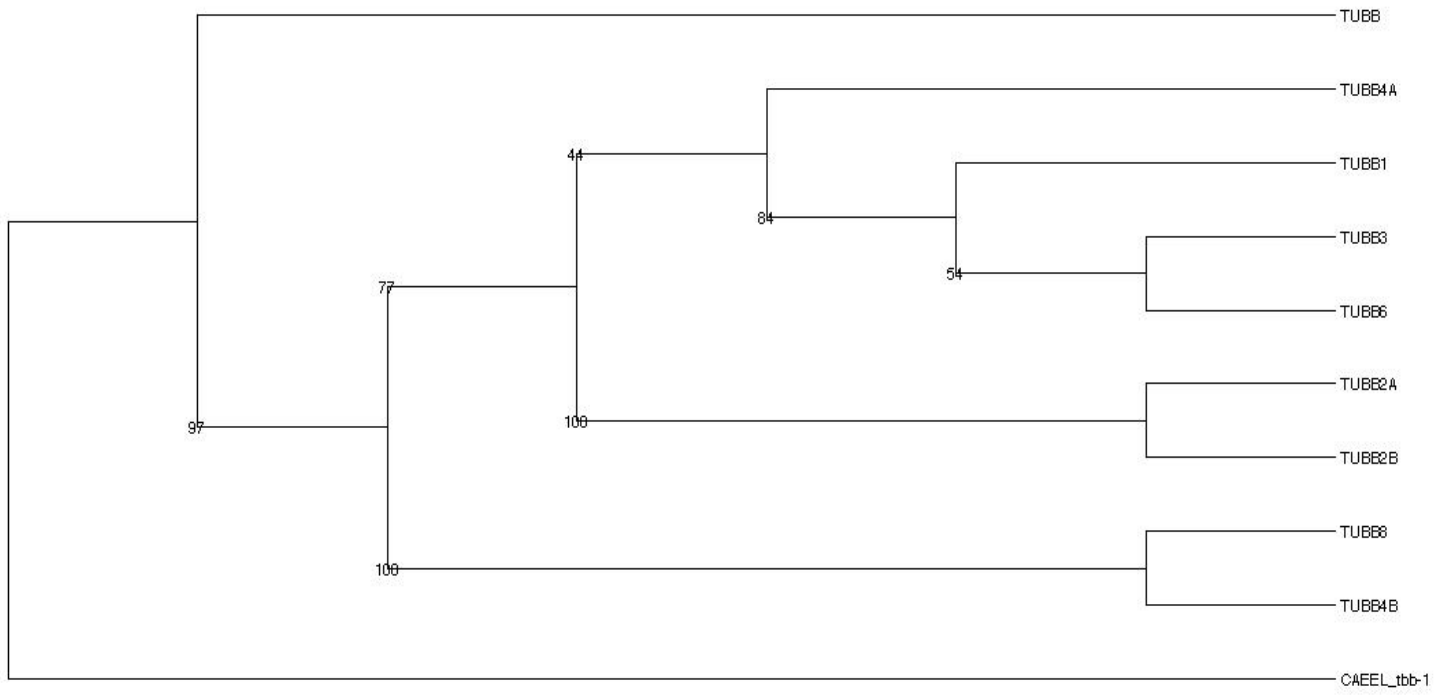
S2

Majority rule tree constructed by neighbor joining method from 4 models of amino acid substitution employed in phylip JTT, PMB, PAM and Kimura. Purple represents beta-tubulin isotypes and green alpha tubulin isotypes.



S3.

Majority rule extended tree of beta tubulin protein coding sequences built by maximum likelihood, model GTRGAMMA and viewed in Dendrocope. Numbers on the branches represent bootstrap support values after 1000 iterations.



S4.

Strict Consensus tree of alpha tubulin protein coding sequences built by maximum likelihood using GTRGAMMA and viewed in Dendrocope.

The clades are clearly distinguished and the numbers represent bootstrap support values after 1000 iterations.