RESEARCH ARTICLE

Phylogeny of Human Mitochondrial DNA Lineages and Its Applications

Eden Edgeworth, Indadul Nayar
Department of Medical Genetics, Fifth Military Medical University

Abstract: With the special intention to introduce the most widely adopted phylogenetic analysis human mitochondrial DNA (mtDNA) studi,the history of the reconstruction of mt DNA phylogny was reviewed. And the applications of human mtDNA phylogeny in studying human evolution, estimating the quality of mtDNA data, and distilling the disease-associated mtDNA mutation were then summarized in the present review.

Keywords: Mitochondrial DNA; phylogeny; in age; mutation

The ancients said, "if a worker wants to do good, he must first use his tools." Obviously, the right approach can make people do twice as much with half the effort, in life, and in research. The development and application of phylogenetic research methods which have been widely accepted and applied in the field of mitochondrial DNA in human population genetics are summarized below.

1. Why do we need to reconstruct the phylogenetic relationship of the human mtDNA lineage?

Mitochondrial DNA (mitochondrial DNA, mtDNA) as a result of its own characteristics have been proved to be a very effective genetic marker for studying the origin and evolution of human beings and tracing population historical events. Firstly, because mtDNA has a high mutation rate (about 10 times that of nuclear genes), mtDNA can accumulate mutations in a relatively short time, thus effectively "documenting" more recent population dynamics. Generally speaking, the mtDNA TTT existed in the population after bulk diffusion to different regions.

Lineage continues to accumulate mutations, resulting in a unique population derived lineage. By analyzing the relationship between ancient and derived lineages and their distribution range and frequency, we can study the relationship between groups and reconstruct group events. Secondly, since there is no recombination of mtDNA, past mutations can be faithfully and continuously inherited. Without considering the influence of recurrent mutation, all mutations accumulated on mtDNA have a temporal order in theory, that is, some mutations are older. Some mutations are relatively young. Therefore, there is an evolutionary link between the mtDNA lineages in the modern population. It is possible to reconstruct phylogeny of maternal lineages in a region or even in the world when sufficient variation information is extracted by certain technical means and analyzed by reasonable methods. The reconstruction of phylogenetic trees (hereinafter referred to as "phylogenetic trees") makes it possible to study population relationships and their origins without being limited by classical methods of population genetics based on models, which are often susceptible to many factors, such as models. In addition, how to reasonably interpret the results is also an unavoidable problem. On the contrary, phylogenetic trees are the most likely true reflection of the evolutionary relationships and evolutionary processes of various pedigrees within a population. Since mtDNA does not recombine and most of the mutations are neutral, each pedigree gradually accumulates mutations in its own evolutionary process, so the pedigrees have developed themselves.

doi: 10.18063/gds.v2i1.

Phylogenetic relationships are not affected by the subsequent demographic events of the population in which they live, and the evolutionary "bonds" based on this interlineal existence allow us to explore in detail such important issues as population relations and group events at the pedigree level. However, it should be pointed out that everything has two sides. True and reliable pedigrees can certainly be helpful in exploring the problem, but wrong pedigrees can mislead us. Therefore, how to construct a real or as close as possible to the true pedigree tree is the first problem that people must solve in the study.

2. Reconstruction of human mtDNA lineage tree

2.1 Transformation of Broussonetia papyrifera

The study of human origin and evolution through mtDNA began in 1980s (1.2). However, due to the limitations of experimental techniques and research methods, people could not fully collect and extract the information contained in mtDNA at that time. For example, tree-building methods widely used at that time, such as neighbor joining, maximum likelihood and maximum parsimony, were developed mainly on the basis of studies at the species level, when applied to the subspecies level (i.e. population level). Functions may be limited by the following reasons: (1) the evolution of lineages within populations does not necessarily follow the bifurcation pattern, especially in a rapidly expanding population where the newly occurring phylogenetic relationships tend to be multi-JI: multi-furcate or star-like: (2) Compared with interspecific relationships, intraspecific population level variability is relatively low, so traditional tree-building methods can only obtain a small number of features for analysis, which to a certain extent affects the resolution and reliability of the results. Especially in modern population studies, when people focus only on a small region or a small part of the mtDNA (which is the most common way to study at the time due to technical constraints), they will get less information. At this point, the traditional tree-building method for analysis, due to the serious lack of information mining, often lead to many branches of phylogenetic relationships can not be determined, and partial evolution branch (clad e) support rate (bootstrap) is often very low: (3) mtDNA because of the high mutation rate, and There are many frequent mutations, which may obscure some real evolutionary events or pathways, but this problem can not be effectively solved in traditional tree-building methods: (4) When using the maximum reduction method to construct a system tree, there are many possible mutations. It is obviously inappropriate to select only items (i.e. to construct a consistent tree "con sensus tree") in the evolutionary path: (5) the number of samples involved in population studies is often large, and when analyzed by traditional methods, not only the calculation time is very long, but also the data processing ability of the computer is also good. Higher requirements.

At present, there have been some reports on the methods of population level research (mismatch analysis r41; cladistic analysis r5.6l), but the medium network method developed by Hans-Ji med.igen Bandelt *et al*.[7–9] of the University of Hamburg in Germany (especially in the field of human mtDNA) has been widely recognized and applied. Iate network). This method takes full account of the well to make up for the shortcomings of traditional tree-building methods in subspecific level research. The network graph constructed by this method allows for multiple divergences and the existence of reticulation, so it can contain all the most possible evolutionary paths to the greatest extent. However, due to the existence of many possible evolutionary pathways, we are not clear about the systematic relationship between certain lineages or nodes, so it is necessary to simplify the network graph, that is, some unlikely to exist (at least not supported by the data studied). In order to delete the evolutionary pathway, the weights of feature loci and the frequencies of samples represented by conflicting nodes, as well as the diversity or the age estimation of the nodes must be fully combined in the simplification process. The method can be performed either manually[9] or by software (Net work 4.500: http://w.fluxus-engineering.com/). The obtained network map can fully reflect the genetic structure within the population (including lineage and frequency information), so it is very suitable for population level research. In practical research, if the amount of samples involved is large, the workload will be enormous when the network graph is constructed directly by the way of work. Especially when the genetic background of the research sample is unknown, the construction process may be in trouble. Our experience is that the network structure of the sample can be constructed by corresponding software first, and then corrected by manual means, especially for the simplification of some network

structures or the selection of some evolutionary paths, which require researchers to analyze the data in detail and fully integrate with other available data. Information, such as the weight of the mutation site and other data information, cannot be done by the software itself.

2.2 The change of information extraction method and the reconstruction of lineage tree.

2.2.1 From low-resolution RFLP to high-resolution RFLP, low-resolution restriction fragment length polymorphism (low-resolution RFLP) was used to detect mtDNA genome by restriction enzymes

Because the amount of information obtained by this method is very small, there are many problems in some related research results based on this method. In view of this, subsequent studies[2] mainly used high-resolution RFLP (high-resolution RFLP), and I-crossing method was further developed by the Wallace team in the United States. They amplified the entire mtDNA genome by overlapping 9 pairs of primers, and then used 14 restriction enzymes (Al UL, Avall, BamHI, Dd el, The amplified fragments were digested by H ell, H AE III, H ha l, H injl, H inell, Hpa l, Hpa L l/M spl, M bol, Rsal and Taql respectively. Compared with previous work, this method can indeed extract more mtDN A mutation information, and therefore define some stable polymorphisms located in the coding region. On this basis, Torroni *et al*.[12] defined the set of haplotypes sharing one or more stable mutations as haplogroups. It was found that the haplotype groups had large pool'| or geographic population specificity, such as LO, L1, L2 and L3[13–18], which were mainly confined to the African continent. Most Europeans belonged to H, I, J, and L3 groups.K, T, U, V, W and X[19–27]1, and group A, B,R9 and M (including their subgroups M7-Mll and D and G, etc.) are mainly distributed in Asia[28–42]. Among them, A-D is the main group of Americans[12,43]called. Obviously, when the maternal lineage and phylogenetic relationships of the populations in different regions are hidden, the relationships and their populations between and within regions are studied by phylogenetic geography[45].

Historical development is possible. Since high resolution RFLP can only detect about 20% of the mutations in mtDNA, does the haplotype classification system based on this method reflect the true phylogenetic relationship of mtDNA lineages in different regions? And whether the system can withstand the test of "control region" information from the other section (f Chuan, such as the control region)? These are the problems that have not been solved at that time.

2.2.2 Sequencing mtDNA control region or D-loop region has a high mutation rate (about 101 tones of coding region)

This region contains three hypervariable regions: HVS-1, HVS-11 and HVS-III, of which HVS-1 is more informative than the other two regions. Therefore, most human mtDNA sequencing studies are usually limited to the region. However, due to the influence of mutation saturation effect and more mutation hot spots (hotspots) in the control region, the evolutionary noise in this region is more, which increases the difficulty of extracting effective information. Especially when the traditional tree-building method is only based on this section of information for pedigree construction, the results also have greater uncertainty, making the credibility of the relevant inferences also suffered a lot of discounts. For example, Vigilant *et al*. attempted to construct a modern human mtDNA system tree from the information buried in the control region by the maximum reduction method, hoping to obtain results supporting the origin of Africa through the i-heart method. But because of the influence of P ar a LLE l mutation and other factors, Vigilant and others have thousands of phylogenetic trees, many of which do not support the African origin hypothesis. Obviously, the existence of a large number of parallel mutations in the control region obscures the original phylogenetic information, which makes it difficult to construct a tree or to classify haplotypes by a single mutation in the control region, especially when some studies artificially tailor the resulting fragments to the same extent. This problem becomes more and more prominent when cutting into an area (for the purpose of software analysis[47]). A feasible way to solve the above problems is to supplement the coding area information at the same time. Because the mutation rate of mtDN A coding region is much lower than that of the control region, the homogeneity events in mtDN A coding region are much less than that of the control region[48], which makes it possible to construct a clear and reliable system tree by analyzing the coding region information. As we already know, mtD N A lacks recombination, which means that all the mutations on it are com-

pletely linked (regardless of the effect of frequent mutations), so theoretically, the mtDNA lineage trees constructed by coding or controlling region information should be identical. In view of this, Torron I *et al*.[19] compared the RFLP information from the same population and mutations in the control region. The results revealed that individuals belonging to the same group (divided by RFLP system) had a unique monophyletic mutation in the control region, which indicated that the single RFLP digestion site was constructed. The ploidy classification system i s supported by t h e information of t h e control region, which can distinguish t h e old and stable C h a r a c t e r i s t I C Mu tation (c h a r a c t e s t I C mu tation) from t h e r a r e Mu tation unique to some individuals and types.Therefore, when only relying on the control region information can not make a correct judgment of an individual's group belonging, the combination of a certain coding region information may be helpful to solve the problem of I Chuan 5,37.

Another feasible method is to use the intermediary network method to analyze and judge[7–9]. With the help of introducing a mediation vector (m e d I a t e vector), the proposed method expresses all possible evolutionary paths through legends, thus effectively identifying the variation of the fuzzy evolutionary path (i.e. identifying the sites where parallel mutations may occur) and avoiding the artificial existence of some possible mutations. The ary branches or evolutionary pathways may be abandoned. After the parallel mutation boundaries emerge, the lineage within the population may become clearer.

2.2.3 Torroni *et al* [Sichuan's study revealed the compatibility of mtDNA control region with RFLP information, and Macaulay *et al*

(26) combined the two systems to study the worldwide lineages (with emphasis on Eurasia) revealed at that time. Due to the limitation of RFLP information (only about 20% of mutations in mtDNA can be detected), some groups that could not be subdivided at that time were properly resolved with the help of control region information. By combining RFLP and mutation information in the control region, Macaulay *et al*. constructed the most perfect human mtDNA lineage tree at that time. In order to facilitate future research, Richards *et al*. (49) and Macaulay *et al*.[26] unified and standardized the classification system and nomenclature of mtDNA pedigrees on the basis of the classification proposed by Torroni *et al*. (12). For all haplotypes from the most recent common ancestor (MRCA) that share one or more characteristic mutations, the haplotype group, a specific name is given to distinguish it from other groups. Specifically, the main groups are represented by upper-case Roman letters, and their subgroups are named as required by alternately appending positive integers and lower-case Roman letters, such as M M7 M7b M7bl. Since the system is based on a large number of samples of RFLP information and mutations in the control region (mainly HVS-1), it can better reveal the genealogical composition and interrelationship of the modern human maternal gene pool.

Macaulay *et al*. (26) The importance of their research work also lies in their considerable improvement on the traditional mtDNA research methods. High-resolution RFLP and direct sequencing of the control region have their own drawbacks: the former is expensive, and because of the limitations of endonuclease recognition sites, the variation that RFLP can detect is very limited; the latter can not effectively eliminate the interference of frequent mutations. Theoretically, R and P information based on coding region should be able to better solve the phylogenetic relationship of the base group (of course, due to the lack of information, the group).

The relationship between C and Z, K and U, T and J was not completely clear at that time, and the information from the control area was often helpful to the identification of some newly emerged branch groups. Combining the two effectively can not only give full play to their advantages, but also make up for their shortcomings. As the pedigree and contour of a population are generally clear in the world, it is possible to obtain the mutation information of the sample control region or HVS-1 by direct sequencing method first, and then to determine the group attribution of the sample through the mutation model of the control region (mo tif) so as to select the specific group. The mutation sites in the coding region were confirmed by RFLP detection or direct sequencing. For example, if the H V S-1 mutation pattern of an individual in a Chinese population is 16223-16362, the individual may belong to group D or G, because the distribution frequency of group D in the Chinese population is generally higher than that in the Chinese population.G 129.35-37501, so we can detect the 5 I 76Alul site first.

If the individual presented a 5 I 76 AluI pattern, it could be divided into group D; if +5 I 76 Al u I, it might belong to group G, so it could be confirmed by detecting 483 IHhaI to ascertain whether it belonged to group G: if the detection

result was +483 IHhaI, the individual would be ascertained to group G: if it was 483 IHhaI Then the individual may belong to other groups, so further tests are needed to determine the group belonging. The method of preliminary judgment based on the information of control area and verification by RFLP detection can be used in a short time.It has been widely used in the world for the purpose of individual group identification with less workload.[2]

2.2.4 MtDN A complete sequence and phylogenetic tree reconstruction take into account the genetic characteristics of M tDN A itself, it can be predicted that the more information obtained on M tDN A, the more accurate the genealogical tree will be constructed, and the determination of complete sequence is the only way to obtain the mutation information contained in M tDN A to the greatest extent

With the advancement of technology and the reduction of sequencing cost, more and more research groups have turned their attention to the whole genome of mtDNA. The emergence of complete mtDNA sequences has made it possible to reconstruct pedigrees in an area, and has provided an excellent opportunity to test the correctness of a population system constructed by merging RFLP and control region information. In this regard, our recent full sequence study is a good example[30]. In order to systematically understand the phylogenetic relationship of the matrilineal lineages in East Asia, 48 representative individuals were selected from more than 2,000 Chinese samples whose phylogenetic status was preliminarily determined by the information of the control region and part of the coding region. The results showed that some new haplotype groups were identified and primitively identified. The group system was constructed through R and P and control area information. Overall, however, our results basically support the original classification system. Inley shows that the merging control region and RFLP information proposed by Macaulay *et al*. (26) are correct and effective for classification.

3. Application of mtDNA family tree

The main interest of intraspecific phylogenies is not in themselves but rather in their applications. Admittedly, the ultimate goal of mtDNA research is not just to reconstruct the lineage tree, but to reconstruct the true (or as close as possible) human mtDNA lineage tree to study population relationships, population dynamics, and even disease-specific or pathogenic mutations in an area.

3.1 MtDNA genealogy and population origin and evolution

The emergence of mtDNA lineage trees makes it possible to study the origin, population relationship and dynamic history of human beings by phylogenetic geography[54]. For example, Yao *et al*.[35] combined partial coding region information with mtDNA control region for a total of six regions from China.

A detailed classification of 263 individuals of the Han nationality was carried out. The results showed that there were great differences among different geographical groups of the Han nationality, and the frequencies of F l, B and D4 were gradient from south to north. There are many ancient groups (such as R9, B, etc.) in southern China, and some possible haplotypes of undetermined basal lineages (that is, the haplotypes that could not be classified at that time were recorded as M *, N *, or R spoon, suggesting that modern people may have migrated from south to North and settled in East Asia during the Paleolithic period.

(especially in mainland China). Similarly, according to the rebuilt East Asia[28,30,34,42] and the European tree 1-2 mountain.27.48] tree tree information, we control.

A total of 252 individuals from five ethnic groups (Uygur, Uzbek, Kazakh, Mongolian and Hui) in Xinjiang, China, were subdivided into different groups by selecting specific coding sites for the region mutation pattern.[3] Except for the systematic status of 8 individuals in these samples All the other individuals can be categorized as subgroups M and N (including R), which constitute only a part of the matrilineal gene pool in East Asia and Europe, thus indicating that Central Asia is in fact a place of genetic tear union between East Asia and Europe. . The study of phylogenetic composition within each group further suggests that the frequency of European endemic groups in different ethnic groups from the same geographical region decreases with the shortening of settlement time, with the highest frequency among the early inhabitants such as Uygurs (426%) and Uzbeks (414%). Kazakhs (302%) followed by Mongols (143%) and Huis (6.7%) with a close migration history, while no European group was found in the Han population (35) who had recently

migrated from the same region. Coincidentally, this frequency distribution pattern coincides with the migration histories of these ethnic groups, suggesting that the maternal genetic structure of the population in this region contains the imprint of migration history. Further analysis of the Mongolian and Hui populations shows that the origin and marriage customs play an important role in the formation of their maternal gene pools.

It is clear from the above examples that the mtDNA phylogenetic tree has developed into the premise and basis of phylogenetic geography, and its emergence has enabled people to study phylogenetic time t (i.e. evolutionary relationships and processes) and spatial (i.e. geographical distribution and frequency Rate) Connections are used to explore important issues such as the origin and evolution of the population, thus avoiding speculation or comparison based only on rough calculations or statistics derived from "black box" operations such as software. Obviously, the latter can not solve the above problems in detail from the perspective of pedigree, and its research ability is greatly limited.

3.2 MtDNA genealogy and genetic diseases

MtDNA lineage trees can be used not only in the study of human origin and evolution, but also in the identification of pathogenic or disease-related mutations in mtDNA genetic diseases. Most of the mutations on mtDN A were neutral, while a few were harmful. Individuals with these mutations may exhibit varying degrees of illness[11,55,56] due to their different harmfulness and other related factors. In a word, it is seriously harmful.

Severely deleterious mutation can cause dysfunction of multiple systems, leading to very serious disease in patients. Individuals with this mutation tend to have lesions in childhood, and the mutation is quickly eliminated. Moderately deleterious mutation mainly affects the normal function of some tissues or organs, and its impact on individual survival is less serious than that of harmful mutation. In childhood, the well may not show any symptoms, but with the increase of the age of the individual, when the proportion of the main mutation in one of its tissues or organs increases beyond a certain threshold, the organization or organs may have dysfunction. Since moderately deleterious mutations may show pathogenicity in the middle or late stages of an individual's growth, they do not have much effect on the individual's growth. The weaker selectivity of the mutation makes it possible that the mutation (or the mtDNA with the mutation) may exist in a limited generation, and in some special cases The mutation may spread to some extent in the population. Mild deleterious mutation may be expressed in the late life of the individual under the synergistic effect of other factors, and has little effect on the normal reproductive capacity of the carrier, so it may be fixed in the form of polymorphism in the population.

Ancient mutations at the base of the pedigree trees that we have obtained are almost neutral because they have undergone selection pressures for thousands of years: subancient mutations in the middle of the pedigree trees, because they have also undergone long-term selection pressures and other events, most of these mutations should be neutral. Sex, while a small amount is only J J mound neutral: For characteristic mutations at the top of the lineage tree, there may be some moderate deleterious mutations; for serious deleterious mutations (such as 3243 mutations), it is unlikely that they will occur in the form of rare mutations in the lineage tree. Existing in different individuals or lineages. Based on the above inferences and other relevant information, we can theoretically effectively distinguish or judge the pathogenicity of reported mtDNA mutations. If a specific coding region mutation is found in the mtDNA of a patient with a maternal inherited disease, it can be preliminarily determined whether the mutation is group specific or possibly pathogenic by detecting the 1E normal individual belonging to the same haplotype group.

For example, a comprehensive analysis of 3,000 individuals from around the country showed that the mutation T12338C resulted in the loss of the initiation codon of the mtDNA ND5 gene (i.e., the translation of the initiation amino acid from methionine to threonine spoon, but the mutation was a characteristic process of the F2 group One of the changes occurred almost 42000 years ago. T12338C (i.e. F2 group) is widely distributed among the normal population in China (although the frequency is low), and there are no reports of F2 group associated with mtDNA genetic diseases. The above evidence suggests that T12338C is unlikely to be a pathogenic mutation (31). This result is consistent with the hypothesis that mutation T3308C (leading to loss of mtDNA NDl initiation codon (59)) and A8527G (leading to loss of mtDNA ATP6 initiation codon (60)). For details of this aspect, see Wang *et al*. (58).

Obviously, when identifying mtDN A mutations associated with disease without other more direct and credible ev-

idence, such as functional testing, preliminary analysis and judgment combined with current phylogenetic knowledge is an effective method to avoid some hasty and imprecise mutations as much as possible. Conclusion or conclusion (61).

Application of 3.3 mtDNA lineage tree in data quality assessment Past studies have shown that the reconstructed mtDNA lineage tree is very helpful in detecting potential errors in published or unpublished data. Bandelt *et al*. (62) analyzed a large number of errors in published data and summarized them into five main types: base shif t, reference bias, phantom mutation, base rnisscoring and artificial recombination N). Bandelt *et al*. have discussed these five types of errors in detail through a large number of examples in their papers, which will not be repeated here. The following is a brief introduction to how to detect potential errors in genealogical tree data in analysis.

To detect errors in a batch of data, first of all, the mutations in the sequence should be exported in the form of mutation sites (sequence markers with a modified Cambridge standard sequence as a reference (63,64). The presence of a large number of rare or unusual transversions (by comparison with published data) was preliminarily observed. Because in the past

A large number of published complete sequence information of mtDNA control region and coding region showed that (16 vs. 5,27 vs. 0,34,42-44,5 vs. l) base transversion in mtDNA was relatively rare and G mutation was the rarest. Most of the transmutations (including most of the G transmutations) in the 560 sequences reported by Herrnstadt *et al*. (25) were proved to be incorrect by their re-sequencing. Therefore, if more rare transmutations are observed in the detected data, there may be more potential errors and the sequencing gel map should be re-examined or even re-sequenced to confirm them. Then, under the guidance of the pedigree tree, all individuals in the data are divided into corresponding groups by their own information. For individuals whose phylogenetic status has been determined, if a rare mutation occurs multiple times in different genetic backgrounds (i.e., haplotype groups), it means that the mutation may be questionable (84); and when the mutation pattern in one segment of an individual's mtDNA conflicts with that in another segment, it is implied. There may be artificial reorganizations. Therefore, we have developed an effective method to detect human recombination based on the established lineage tree of mtDNA in East Asia and the mediation network method.

In Tanaka *et al*. (42), 12 human recombinant sequences were detected in 672 complete sequences (the authors did not publish data). For individuals whose phylogenetic status has not been determined, the individual may belong to a rare base type not covered by the lineage tree: the individual fails to group due to incomplete mutation patterns. The former may exist, because the current phylogenetic tree is based on a limited set of complete data, and it can easily cover most of the major lineages in a region (e.g. East Asia) with a wider range and a higher frequency of distribution, but it is likely that there will be some basal but very narrow or fragmented distribution. The genealogy with low frequency was not detected. As more full-sequence data emerge in the future, the pedigrees will become more subtle, more extensive, and more representative, so that the occurrence of undivided types will be greatly reduced. The incompleteness of the latter may imply that: (1) the individual's partial mutation pattern matches (approximates) a branch group of the phylogenetic tree but lacks a characteristic mutation that its base should have; (2) the detected individual belongs to an intermediate type not reflected in the phylogenetic tree. The incompatibility between mutations in the former may be caused by frequent mutations or human errors, and the judgment of the problem needs to be re-experimented or combined with the conservativeness of mutation sites and the frequency of population; the latter, if there is only one representative individual of the intermediate type (by analysis institute). The missing mutation may be artificially caused in the experiment or in the data processing, and the mutation needs to be validated by a new experiment.

In the specific analysis process, different people may have different ideas and experience on how to detect potential errors in the data, but no matter what method is used, the goal is to minimize errors in the data. Therefore, during data collection, researchers need to be careful not only in the experimental process, but also in data processing. Our research experience shows that, for new data, self-examination through phylogenetic analysis can effectively reduce the possibility of errors in the data, so as to better ensure the quality of data, and avoid some unnecessary embarrassment after data publication. Obviously, the key to data quality is that it is not only closely related to research results and conclusions, but also reflects the attitude of researchers. If the data itself is full of problems and errors, no matter how complex or fancy the analysis or how perfect the logic of argument is in the writing of the paper, there is reason to doubt the

credibility of the findings and conclusions. Researchers are fortunate to find that genetic markers such as mtDNA lack recombination because they can self-examine the data based on the phylogenetic relationships of the reconstructed markers, but no similar quality control measures can be applied to the study of nuclear genes. Considering the large number of errors in M T D N A research, it is reasonable to believe that this phenomenon is not unique to m t D N A research, but that errors in M T D N A data can be detected by effective methods. However, the reported errors in nuclear gene data are still unknown, so what measures should be taken to ensure the reliability of the data obtained in the field of nuclear gene research is a problem that must be solved in the future. For a study whose data quality is not guaranteed, the inferences about the pathogenicity of certain mutations in the nuclear gene or whether positive selection has been made are doubtful. It is undeniable that, as information accumulates, certain seemingly correct speculations may be questioned or even completely denied in the future, but published data are always objective. From this point of view, data may be more important than conclusions, because it will also directly affect other relevant studies in the future. Therefore, how to ensure the quality of data will be a problem that every researcher must face and solve, and publishing high-quality data is the responsibility and obligation of the researcher.

4. Summary

The emergence of mtDNA genome-wide information and changes in research methods have made the maternal lineage of populations in major geographical regions of the world clearer and clearer, and the reconstructed lineage tree has been widely used in the study of mtDNA-related fields, which makes it possible to study human beings from the perspective of lineage evolution history. Important issues such as origin and evolution: the ability to understand the evolution of mtDNA mutations in depth, thereby contributing to the effective identification of pathogenic or benign mutations in genetic disease research; the ability to effectively identify potential errors in tli. The tremendous success of this phylogenetic approach in human research has also facilitated its application in other areas of research, such as domestic animals, and some complementary approaches, such as matching and near-matching, have been developed to facilitate correspondence. Classification of data (e.g. paleoDNA data, etc.) that are scarce or difficult to obtain further expands its scope of application.

References

1. wallace johnson mj, dc, ferris sd, *et al*. radiation of human dna by restriction mitochondrial types analyzed endonuclease cleavage patterns. j mo! evol, 1983. 19: 255.71
2. cann rl, stoneking ac m, wilson. mitochondrial dna and human evolution. nature, 1987,: 31 - 6 325
3. posada d, crandall ka. intraspecific gene genealogies: grafting trees into networks. evo] trends ecol. in 2001, 16: 37 - 45 rogers, ar.
4. harpending h. population growth makes waves in the distribution of p 副 rwise genetic differences. mo! biol evol. 1992. 9: 552 69
5. templeton ar, crandall cf, cf cladistic sing. the analysis of phenotypic associations with haplot ypes inferred from restriction endonuclease core and dna sequence data. (iii) cladogram estimation. genetics. 1992: 619 - 33, 132
6. templeton ar, cf cladistic sing. the analysis of phenotypic associations with haplotypes inferred from restriction en - donuclease core. (iv) nested analyses with cladogram uncertainty and recombination. genetics, 1993, 659 134:69
7. bandelt hj, forster p, the rohl. median - joini ng networks for inferring intraspecific phylogenies. mol biol evol, 1999, 16: 37 - 48
8. bandelt hj, forster p, sykes bc, *et al*. mitochondrial por -traits of h uman populations using median networks. genetics, 1995,: 743 141 - 53
9. bandelt hj, macaulay v m, richards. median networks: speedy construction and greedy reduction. one simulation, and two case studies from human mtdna. mol phylogenet evol, 2000, 16: 8 - 28 the torroni
10. schurr tg, yang cc. *et al*. native amencan mitochondrial indicates dna analysis that the amerind and the nadene populations were founded by two independent migrations. genetics, 1992. 130: 153 - 62 wallace.
11. dc. mitochondrial dna variation in hu man's evolution. degenerative disease. and aging. am j hum genet '1995, 57: 201 - 23 the torroni
12. schuπ tg, cabell mf. *et al*. asian continental affinities and radiation of the four founding native american mtdnas. am j hum genet, 1993, 53: 563 - 90 chen na.
13. the schurr olckers, tg, *et al*. in the south african mtdna variation kung and khwe - and their genetic rela -tionships to other african populations. am j hum genet, 2000: 1362 '66 - 83 chen ic.
14. the torroni, excoffier l, *et al*. analysis of mtdna variation in african populations reveals the most ancient of all hu-

man haplogroups continent - specific. am j hum genet, 1995, 57: 133 - 49
15. watson p e, forster, richards m, *et al*. mitochondrial foot - prints of human expansions in africa. am j hum genet, 1997, 61: 691 - 704
16. villems soodyall behar dm, r, h, *et al*. the dawn of humanmatrilineal diversity. am j hum genet, 2008, 82: 1130 - 40
17. kivisild t, shen p, wall dp, t al. η1e role of selection in from eevolution of human mitochondrial genomes. genetics. 2006172: 373 - 87
effect of.
18. achilli, macaulay v, *et al*. harvesting from fruitof the human mtdna tree. trends genet, 2006, 22: 339 - 45
19. of, huoponen k, francalacci p, *et al*. classificationof european mtdnas from an analysis of 也 ree europeanpopulations. genetics, 1996, 144: 1835 - 50
20. of, lott mt, cabell mf. *et al*. mtdna and the origin of caucasians: identification of ancient caucasian - sillc specific haplogroups, one of which is hooked to entsomatic areas duplication m d - loop region. am j hum genet, 1994, 55: 760 - 76
21. achilli, rengo c, battle v, *et al*. saarni and berbers - anunexpected mitochon 由 ial dna link. am j hum genet, 76: 883 - 6
22. achilli, rengo c, thin c, *et al*. the molecular dissection of mtdna haplogroup h confirms that the franco cantabrian glacial refuge was to m 句 or source for the euro - pean gene pool. am j hum genet, 2004, 75: 910 - 8
23. coble md, just rs, 0's callaghan je, *et al*. single nucleotidepolymorphisms over the why should mtdna genome that in - crease the power of forensic testing in caucasians. int jlegal med, 2004, 118: 137 - 46
24. finnila s, lehtonen ms naa k. phylogenetic networkfor european mtdna. am j hum genet, 2001, 68, 1475 - 84
25. herrnstadt c, elson jl, fahy and *et al*. reduced - median network analysis of complete mitochondrial dna coding - region suences for m 句 or african, asian, and europeanhaplogroups. am j hum genet, 2002, 70: 1152 - 71
26. is v, richards m, hickey, and *et al*. the emerging tree of west eurasian mtdnas: a synthesis of control region sequences and rflps. am j hum genet, 1999, 64: 232 - 49
27. palanichamy mg, sun c, agrawal s, *et al*. phylogeny ofmitochondrial dna macrohaplogroup n in india, based on comprehensive sequencing: implications for the peopling of south popu - asia. am j hum genet, 2004, 75: 966 - 78
28. kong qp, bandelt hj, sun c, *et al*. updating the east asian mtdna phylogeny: a prerequisite for the identification of pathogenic mutations. hum mol genet, 2006, 15: 2076 - 86 spatiotemporal
29. kong qp, yao yg, liu m, *et al*. mitochondrial dna if - quence polymorphisms of five ethnic populations from northern china. hum genet, 2003, 113: 391 405
30. kong qp, yao yg, sun c, *et al*. phylogeny of east asian by mitochondrial dna lineages inferred from complete genet, sequences. am j hum genet, 2003, 73: 671 - 6
31. kong qp, yao yg, sun c, *et al*. phylogeographic analysis of mitochondrial dna haplogroup f2 in china reveals tl2338cin the + codon of the nd5 gene not to be paogenic. j hum genet, 2004, 49: 414 - 23
32. kivisild t, barnshad mj, kaldma k, *et al*. deep common ancesof indian and westem - asian mitochondrial lineages. curr biol, 1999, 9: 1331 - 4
33. kivisild t, rootsi s, metspalu m, *et al*. ηm genetic heritage of the earliest settlers persists both in indian tribal and caste populations. am j hum genet, 2003, 72: 313 - 32
34. kivisild t, tolle hv, parik j, *et al*. η1e emerging limbs and twigs of the east asian mtdna tree. mol biol evol, 2002, 19: 1737 - 51
35. yao yg, kong qp, bandelt hj, *et al*. phylogeographic dιι ferentiation of mitochondrial dna in han chinese. am jhum genet, 2002, 70: 635 - 51
36. yao yg, kong qp as xy, *et al*. reconstructing the evo - lutionary history of china: a caveat about inferences drawn from ancient dna. mol biol evol, 2003, 20: 214 - 9
37. Yao YG, Kong QP, Wang CY, *et al*. Different matrilineal contributions to genetic ucture of enic groups in the road region in China. Mol Biol Evol, 2004, 21: 2265-80
38. Yao YG, Nie L, Harpending H, *et al*. Genetic relationship of Chinese ethnic populations revealed by mtDNA sequenced iversity. Am J Phys An 出 ropol, 2002, 118: 63-76
39. Wen B, Li H, Gao S, *et al*. Genetic structure of Hmong-Mien speaking populations in East Asia as revealed by mtDNA lineages. Mol Biol Evol, 2005, 22: 725-34
40. Wen B, Li H, Lu D, *et al*. Genetic evidence supports demic diffi』sion of Han culture. Nature, 2004, 431: 302-5
41. Wen B, Xie X, Gao S,已 t al. Analyses of genetic structure of Tibeto-Burrnan populations reveals sex-biased admixture in southern Tibeto-Burrnans. Am J Hum Genet, 2004, 74: 856-65
42. Tanaka M, Cabrera VM, Gonzalez AM, *et al*. Mitochon-drial genome variation in Eastern Asia and the peopling of Japan. Genome Res, 2004, 14: 1832-50
43. Achilli A, Perego UA, Bravi CM, *et al*. The phylogeny of the four pan-American mtDNA haplogroups: implications for evolutionary and disease studies. PLoS ONE, 2008, 3:el764
44. Tamm E, Kivisild T, Reidla M, *et al*. Beringian standstill and spread of native American founders. PLoS ONE, 2007, 2: e829
45. Avise JC. Phylogeography: The history and formation of species[M]. Cambridge: Harvard University Press,

2000
46. Vigilant L, Stoneking M, pending H, *et al*. Aicanlations and the evolution of human mitochondrial DNA.Science, 1991, 253: 1503-7
47. Wang L, Oota H, Saitou N, *et al*. Genetic structure of a 2500- yeId human population in China and its changes. Mol Biol Evol, 2000, 17: 1396-400
48. Finnila S, Hassinen IE, Ala-Kokko L, *et al*. Phylogenetic network of the mtDNA haplogroup U in Northern Finland based on sequence analysis of 出 e complete coding region confo 口 nation-sensitive gel electrophoresis. Am J Hum 2000, 66: 1017-26