BMC
Bioinformatics

METHODOLOGY ARTICLE

Open Access

YHap: a population model for probabilistic assignment of Y haplogroups from re-sequencing data

Fan Zhang[1,4†], Ruoyan Chen[1†], Dongbing Liu[1], Xiaotian Yao[1], Guoqing Li[1], Yabin Jin[1], Chang Yu[1*], Yingrui Li[1*] and Lachlan JM Coin[1,2,3*]

Abstract

Background: Y haplogroup analyses are an important component of genealogical reconstruction, population genetic analyses, medical genetics and forensics. These fields are increasingly moving towards use of low-coverage, high throughput sequencing. While there have been methods recently proposed for assignment of Y haplogroups on the basis of high-coverage sequence data, assignment on the basis of low-coverage data remains challenging.

Results: We developed a new algorithm, YHap, which uses an imputation framework to jointly predict Y chromosome genotypes and assign Y haplogroups using low coverage population sequence data. We use data from the 1000 genomes project to demonstrate that YHap provides accurate Y haplogroup assignment with less than 2x coverage.

Conclusions: Borrowing information across multiple samples within a population using an imputation framework enables accurate Y haplogroup assignment.

Background

The non-recombining portion of haploid chromosome Y is passed intact from father to son with a mutation rate several times greater than autosomes [1]. As such, patterns of variation in Y are widely used to uncover historical patterns of human migration; are important in genealogical reconstruction and have application in forensic analyses.

The Y Chromosome Consortium (YCC) published a revised Y-chromosome DNA haplogroup tree in 2008, consisting of approximately 600 markers which can be used to characterize 20 major global haplogroups, labeled A-T, as well as sub-classification into a total of 311 haplogroups at the finest level of resolution. Different major haplogroups have been found at high frequencies in different geographical regions, for example the E clade in Africa, and the O clade in Eastern Asia. Particular fine-level haplogroups are found in multiple locations, such as R1a in Eastern Europe, South Asia and Central

Asia, indicating migration of R1a from Eurasian Steppes to the new world. The C3 haplogroup, found at high frequency throughout Asia is commonly interpreted as genealogical remnants of the empire of Genghis Khan [2].

Y haplogroup assignment has traditionally been carried out by targeted genotyping using a combination of short tandem repeat typing, multiplex PCR and mini-sequencing [3,4], often using a hierarchical strategy in order to first refine the major haplogroup, and subsequently genotype markers within that haplogroup which illuminate finer levels of resolution. Such a procedure requires substantial amount of wet-lab analysis, requires stringent replication and quality control to eliminate errors which can arise due to the limited amount of information collected at each step. More recently, personal genetic companies have included specific Y chromosome markers on custom genotyping arrays [5,6]. Nevertheless, the resolution available from genotyping arrays is limited by markers included on the chip.

High coverage high throughput sequencing has the potential to capture all single nucleotide and insertion/deletion variation, and as such provide near-perfect assignment of individuals to Y haplogroups. One recently

* Correspondence: yuchang@genomics.org.cn; liyr@genomics.org.cn;
l.coin@imb.uq.edu.au
†Equal contributors
1BGI-shenzhen, Shenzhen, China
2Institute for Molecular Bioscience, University of Queensland, Queensland, Australia
Full list of author information is available at the end of the article

published method (AMY-tree) demonstrated the effectiveness of assigning Y haplogroups with high coverage sequence data [7]. As high coverage sequencing of large population samples remains expensive, low coverage population sequencing, in which each individual is sequenced at less than 2x haploid coverage is an attractive alternative, but this will not capture all individual-level variation. AMY-tree, for example, found insufficient information in low coverage genomes from 1000 genomes project for confident haplogroup assignment [7].

We hypothesize that, given the sharing of haplogroups within an ethnically homogenous population, it should be possible to borrow information across individuals within a population in order to improve haplogroup assignment. In this manuscript, we present the YHap tool, which has been designed for assigning haplogroups to low-coverage population re-sequencing data. YHap borrows information across all samples to assign samples to haplogroups probabilistically, thus providing an accurate representation of the inference which can be made from the data collected. YHap is a complete solution and can also be applied to high-coverage sequence data, as well as data from genotyping arrays.

Methods
We use the set of haplogroups and mutations defined in [8]. We map the forward and reverse primers described in this manuscript to identify the reported strand of the variation in the GRCh37 reference. After strand correction, we identify whether the mutant allele is the equal to the alternative or reference allele, so that we can subsequently work in reference/alternative allele space on the forward strand, consistent with conventional genotype calling schema. Next, we map each mutation to its position on the pre-defined Y phylogenetic tree T. Finally we create a haplogroup matrix $\mathbf{H}$ of size $N_{ref}*L$ where L is the total number of nodes in T (including leaf and internal nodes) and N_{ref} is the number of predefined Y markers. Each entry $\mathbf{H}_{il} = \{H_{ilg}\}$ is a probability distribution vector expressing the probability that a sampled individual from the clade below node *j* carries allele g (in this case either the reference or alternate allele). At leaf nodes, this probability vector is either *{0,1}* or *{1,0}*, and at internal nodes, it is the proportion of descendant leaf nodes with reference or alternate alleles, respectively.

To assign a sequenced individual to a specific haplogroup, we obtained genotype likelihoods at each putative variant site (inclusive of all markers in H) from chromosome Y VCF files of the 1000 genome project. This results in a matrix G of size N*M where M is the number of sequenced samples, N is the number of putative variants, and $\mathbf{G}_{ij} = \{G_{ijg}\}$ is a vector of genotype likelihoods. We then generated an augmented $\mathbf{H}^*$ matrix by adding in extra sites in $\mathbf{G}$ but not $\mathbf{H}$ with probability

vector $H^*_{il} = \{0.5,0.5\}$. The pipeline is similar for genotype data, except that the genotype likelihoods are taken to be either 1, if $\mathbf{G}_{i,j}=g$ or 0 otherwise. To illustrate the entries of this matrix, we have included a heatmap of both the standard Y chromosome consortium positions which are polymorphic in the CEU, as well as the full $\mathbf{H}^*$ matrix trained on CEU data (Additional file 1: Figure S1).

We can calculate the assignment of each individual using

$$P\left(\mathbf{G}_{.j}|\mathbf{H}^*_{.l}\right) = \prod\nolimits_{i=1..N}\sum\nolimits_{g=\{0,1\}}P\left(\mathbf{G}_{i,j}|g\right)*P\left(g|\mathbf{H}^*_{il}\right) \tag{1}$$

Where $P\left(g|\mathbf{H}^*_{il}\right) = H^*_{ilg}$ and $P\left(\mathbf{G}_{ij}|g\right) = \mathbf{G}_{ijg}$

We can then calculate the posterior probability of each haplogroup amongst a set of haplogroups, where prior haplogroup probability distribution $P(\mathbf{H}^*_{.l})$ is set to the uniform distribution,

$$P\left(\mathbf{H}^*_{.l}|\mathbf{G}_{.j}\right) = \frac{P\left(\mathbf{G}_{.j}|\mathbf{H}^*_{.l}\right)P\left(\mathbf{H}^*_{.l}\right)}{\sum\nolimits_{k=1..L}P\left(\mathbf{G}_{.j}|\mathbf{H}^*_{.k}\right)P\left(\mathbf{H}^*_{.k}\right)} \tag{2}$$

By restricting the set of haplogroups considered in equation (2), YHap can be customized to either only assign to within the major haplogroups (A through to T), or all possible haplogroups at the finest level of classification.

While this model is sufficient for assigning Y haplogroups individually, it does not capture shared information between sequenced samples adequately, particularly for low coverage sequencing. Given that a population sample will share individuals from the same haplogroup, and while none of these individuals may have enough depth at informative Y haplogroup markers, there is enough information across the pooled reads from all samples from the same haplogroup. However, we do not know a-priori which samples can be pooled as coming from the same haplogroup.

In order to pool information between samples, we treat the allele probability distribution *{H*_{ilg}}* at markers present in the sequence data but not present as haplogroup markers, as parameters in our model. We update these parameters using expectation maximization, in which we first calculate the posterior probability assigning each sample j to each haplogroup *l* using equation 2, and then update the *{H*_{ilg}}* to reflect the average of genotypes assigned to haplogroup *l* at position *i*, weighted by this posterior probability of assignment. In this way, the model learns which alleles are characteristic of the pre-defined haplogroups, and is thus able to more accurately assign individuals which may not have good coverage at those sites, but which show similarity to other individuals across the Y chromosome. The probabilities $P(\mathbf{H}^*_{.l})$ are also updated at each step to reflect the proportion of haplogroups assigned in the population.

Results

We applied YHap to low-coverage sequencing data generated in the pilot phase of the 1000 genome consortium which were also part of the Hapmap project, including 19 YRI, 16 JPT, 21 CEU and 9 CHB samples [9]. Major Y haplogroups have been previously assigned to these samples as part of the Hapmap project. The average sequencing depth of these samples is 1.67X as described in 1000 genome Y chromosome analysis report. Compared to haplogroups previously obtained from the Hapmap project [10], YHap showed perfect assignment accuracy (Table 1). We also used YHap on the Hapmap combined phase 1,2,3 Y genotype data and obtained the same assignments previously reported with this data.

In order to investigate the ability of YHap to assign finer-level haplogroups we compared YHap results obtained at complete resolution (i.e. considering all haplogroup leaf nodes on the pre-defined Y phylogenetic tree) on both Hapmap genotype data and also 1000 genomes low coverage sequence data (Additional file 1: Table S1). We see that there is complete concordance at the major haplogroup level, and there is increasing uncertainty in assignment as the resolution of assignment increases, particularly using dense genotype data. We also observe that accuracy remains high amongst those assignments which YHap assigns high confidence.

We compared YHap's performance with AMY-tree on the 1000 genome dataset. Firstly we consider those 1000 genomes samples which were also assigned with AMY-tree based on high coverage Complete Genomics data (Additional file 1: Table S2). YHap achieved greater resolution than AMY-tree relative to this benchmark on 4 of 7 samples (correctly identifying R1b1b2 instead of R1; inferring N instead of NO; D2a instead of D2). In 1 of 7 samples Yhap identified E1b1a7 haplogroup, whereas AMY-hap assigned E1b1a8 using both 1000 genomes and Complete genomics data, although in both cases Yhap assigned the same haplogroup on the basis of Hapmap genotype data.

Next, we use Hapmap genotype data as a validation, using results on 65 samples for which we have Hapmap and 1000 genomes sequence data. Yhap achieved greater resolution than AMY-tree in 30 of 65 samples, whereas AMY-tree never had greater resolution than Yhap. As an example, this included Yhap correctly identifying NA12005 as belonging to R1b1b2 haplogroup vs Root for AMY-tree (Additional file 1: Table S3.) AMY-tree identified a haplogroup inconsistent with Hapmap data in 1 of 65 samples (assigning NA11829 to DE instead of I), and YHap identified an inconsistent haplogroup in 2 of 65 samples (assigning both NA18971 and NA18974 to C2 instead of C3, where AMY-tree only assigned haplogroup C).

Finally, in order to investigate the relationships between sequencing depth and assignment accuracy, we randomly downsampled the original bam files from 1000 genome to 0.6X. For simplicity, we chose JPT to run the test. We see that downsampling increases the uncertainty of assignment (Additional file 1: Table S4), but YHap accuracy remains high amongst those assignments which are made with high posterior probability. This demonstrates that as the underlying amount of information decreases, YHap is still able to extract inference and accurately represent the uncertainty of this inference.

The total complexity for the whole procedure is $O(N^2T)$, conventionally, when using default settings, it will take almost 10 min to locate 10 ~ 20 individuals and approximately 200 Mb memory.

Discussion and conclusions

We have demonstrated the utility of using low-coverage population sequence data to accurately resolve Y haplogroups at high resolution. This can be achieved via efficiently borrowing information between individuals in the population which have a common Y haplogroup using a probabilistic assignment model. Moreover, we have demonstrated that it is possible to accurately quantify the uncertainty in the haplogroup assignment, such that even for very low coverage sequence data (0.6x) it is possible to make inference of Y haplogroups, but only achieve high certainty for a top-level haplogroup assignment.

Moreover, YHap can inform discovery of new haplogroup markers. Essentially the conditional haplogroup allele probabilities H^*_{ij} (which are initialised with probability {0.5,0.5 } which converge to {0,1} or {1,0} represent new mutations which are exclusive to that haplogroup in the population studied, and represent new mutations on the Y haplogroup tree.

YHap currently only incorporates bi-allelic markers, and does not accommodate STR markers, which is a limitation we plan to address in future versions. We also plan to extend Yhap to allow Mitochondrial haplogroup assignment, however, this requires extending the model to incorporate heteroplasmy as an unknown mixture of multiple haplogroups in the same sample.

While Yhap has been designed using low-coverage whole-genome sequence data, given that it only relies on genotype likelihoods, it will also work for capture sequence data. This opens up the possibility of very cost-effective Y chromosome haplogroup analysis on large populations using a custom designed Y chromosome capture array.

Availability

YHap is available from http://www1.imperial.ac.uk/medicine/people/l.coin/.

Table 1 Assignment of individuals included in this study to haplogroups

Pop	ID	Hapmap[1]	1KG[2]	Chip[3]	NGS[4]	Pop	ID	Hapmap[1]	1KG[2]	Chip[3]	NGS[4]	Pop	ID	Hapmap[1]	1KG[2]	Chip[3]	NGS[4]
CHB	NA18558	O	N	N1	N1c1c1	CEU	NA06994	HI	I1	I1	I1b1	YRI	NA18501	E3a	E1b1a8a	E1b1a	E1b1a
CHB	NA18561	O	O2b	O2	O2b	CEU	NA07357	R	R1b1b21	R1b1b2	R1b1b2	YRI	NA18504	E3a7	E1b1a	E1b1a	E1b1a
CHB	NA18562	O	O	O3a	O3a3b1	CEU	NA10851	R	R	R1b1b2	R1b1b2	YRI	NA18507	E3a7	E1b1a	E1b1a	E1b1a
CHB	NA18563	O	O2b	O2	O2b	CEU	NA11829	HI	I1	I1	I1b1	YRI	NA18516	E3a	E1b1a	E1b1a	E1b1a
CHB	NA18572	O	O	O3a	O3a3b1	CEU	NA11831	R	R	R1b1b2	R1b1b2	YRI	NA18522	E3a	E1b1a	E1b1a	E1b1a
CHB	NA18603	O	O	O2	O2a1a	CEU	NA11881	HI	I1	I1	I1b1	YRI	NA18853	E3a	E1b1a8a	E1b1a	E1b1a
CHB	NA18605	O	O	O	O3a3b1	CEU	NA11994	R	R1b1b21	R1b1b2	R1b1b2	YRI	NA18856	E1	E1	E1/E2	E1/E2
CHB	NA18608	O	N	N1	N1c1c	CEU	NA12003	HI	I2b	I2b	I2b	YRI	NA18871	E3a	E1b1a8a	E1b1a	E1b1a
CHB	NA18609	O	O	O3a	O3a3b1	CEU	NA12005	R	R1b1b2	R1b1b2	R1b1b2	YRI	NA19098	E3a	E1b1a	E1b1a	E1b1a
JPT	NA18940	D	D2 × D2b1	D2a	D2a	CEU	NA12043	R	R1	R1b1b2	R1b1b2	YRI	NA19119	E3a	E1b1a8a	E1b1a	E1b1a
JPT	NA18943	O	O2b	O2	O2b	CEU	NA12144	R	R1b1b21	R1b1b2	R1b1b2	YRI	NA19138	E3a	E1b1a8a	E1b1a	E1b1a
JPT	NA18944	D	D2b1	D2a	D2a	CEU	NA12154	R	R1	R1b1b2	R1b1b2	YRI	NA19141	E3a	E1b1a8a	E1b1a	E1b1a
JPT	NA18945	O	O	O3a	O2b	CEU	NA12155	R	R1	R1a1	R1a1c	YRI	NA19144	E3a	E1b1a8a	E1b1a	E1b1a
JPT	NA18948	D	D2b1	D2a	D2a	CEU	NA12716	R	R1	R1b1b2	R1b1b2	YRI	NA19153	E3a	E1b1a8a	E1b1a	E1b1a
JPT	NA18952	D	D	D2a	D2a	CEU	NA12750	HI	I1	I1	I1b1	YRI	NA19160	E3a	E1b1a8a	E1b1a	E1b1a
JPT	NA18953	O	O2b	O2	O2a	CEU	NA1760	R	R	R1b1b2	R1b1b2	YRI	NA19171	E3a7	E1b1a	E1b1a	E1b1a
JPT	NA18959	O	O	O3a	O3a3b1	CEU	NA12762	R	R1	R1b1b2	R1b1b2	YRI	NA19200	E3a7	E1b1a	E1b1a	E1b1a
JPT	NA18960	D	D2 × D2b1	D2a	D2a	CEU	NA12812	R	R1	R1b1b2	R1b1b2	YRI	NA19207	E3a	E1b1a	E1b1a	E1b1a
JPT	NA18961	D	D	D2a	D2a	CEU	NA12814	R	R1	R1b1b2	R1b1b2	YRI	NA19210	E3a7	E1b1a	E1b1a	E1b1a
JPT	NA18965	O	O2b	O2	O2b	CEU	NA12872	R	R1b1b2g	R1b1b2	R1b1b2						
JPT	NA18967	D	D2b1	D2a	D2a	CEU	NA12874	R	R1	R	R1b1b2						
JPT	NA18970	D	D2 × D2b1	D2a	D2a												
JPT	NA18971	C	C1	C1a	C3												
JPT	NA18974	C	C1	C1a	C3												
JPT	NA19005	O	O2b	O2	O2b												

*E1b1a was formerly known as E3a.
[1]Hapmap indicates results from Hapmap consortium.
[2]IKG indicates results form 1000 genomes consortium.
[3]Chip indicates results from YHap applied to Hapmap genotype data.
[4]NGS indicates results from YHap applied to 1000 genomes consortium sequence data.
The resolution reported for YHap is the level at which a single assignment achieved greater than 90% posterior probability.

Additional file

Additional file 1: Table S1. Concordance of Hapmap array data and 1000 genomes sequence data. **Table S2.** Haplogroup assignments comparison using AMY Complete Genomics data based result as golden standard. **Table S3.** Haplogroup assignments comparison using Yhap Hapmap array data based result as golden standard. **Table S4.** Accuracy and certainty of half-coverage. **Table S5.** Accuracy and certainty of downsampling on high-depth sample from 1000 genome project. **Figure S1.** Heatmap representing probability that haplogroup carries non-reference allele on only Y chromosom consortium SNPs. **Figure S2.** Heatmap representing probability that haplogroup carries non-reference allele at all SNPs modelled.

Author details
[1]BGI-shenzhen, Shenzhen, China. [2]Institute for Molecular Bioscience, University of Queensland, Queensland, Australia. [3]Department of Genomics of Complex Disease, School of Public Health, Imperial College, London, UK. [4]Department of Computational Medicine and Bioinformatics, Medical School, University of Michigan, Ann Arbor, USA.

References
1. Xue Y, Wang Q, Long Q, Ng BL, Swerdlow H, Burton J, Skuce C, Taylor R, Abdellah Z, Zhao Y, *et al*: **Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree.** *Curr Biol* 2009, **19**:1453–1457.
2. Zerjal T, Xue Y, Bertorelle G, Wells RS, Bao W, Zhu S, Qamar R, Ayub Q, Mohyuddin A, Fu S, *et al*: **The genetic legacy of the Mongols.** *Am J Hum Genet* 2003, **72**:717–721.
3. Sanchez JJ, Borsting C, Hallenberg C, Buchard A, Hernandez A, Morling N: **Multiplex PCR and minisequencing of SNPs–a model with 35 Y chromosome SNPs.** *Forensic Sci Int* 2003, **137**:74–84.
4. Zerjal T, Wells RS, Yuldasheva N, Ruzibakiev R, Tyler-Smith C: **A genetic landscape reshaped by recent events: Y-chromosomal insights into central Asia.** *Am J Hum Genet* 2002, **71**:466–482.
5. Turner A: **SNPs on chips: a New source of data for Y-chromosome studies.** *Journal of Genetic Genealogy* 2008, **4**(1):iii-iv.
6. Elhaik E, Greenspan E, Staats S, Krahn T, Tyler-Smith C, Xue Y, Tofanelli S, Francalacci P, Cucca F, Pagani L, *et al*: **The GenoChip: a new tool for genetic anthropology.** *Genome Biol Evol* 2013, **5**:1021–1031.
7. Van Geystelen A, Decorte R, Larmuseau MHD: **AMY-tree: an algorithm to use whole genome SNP calling for Y chromosomal phylogenetic applications.** *BMC Genomics* 2013, **14**:101.
8. Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF: **New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree.** *Genome Res* 2008, **18**(5):830–838.
9. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA, The 1000 Genomes Project Consortium: **The 1000 genomes project consortium: a map of human genome variation from population-scale sequencing.** *Nature* 2010, **467**:1061–1073.
10. International HapMap Consortium, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, *et al*: **Integrating common and rare genetic variation in diverse human populations.** *Nature* 2010, **467**:52–58.