Mitochondrial molecular markers for US lineages of *P. infestans*

Brian J. Knaus, Javier F. Tabima and Niklaus J. Grünwald

February 21, 2014

Contents

List of Figures

List of Tables

1 The sample

The sample includes data which was opportunistically gathered from previous publications as well as data which is not yet available to the public (Judelson, unpublished).

Sample	Count	Reference
T30-4	1	[4]
PIC99189 & 90128	2	[8]
13_a2	1	[2]
Yoshida et al.	13	[9]
Martin et al.	3	[7]
Judelson	8	NA
Total	28	

- The sample 'T30-4' was the first sequenced genome and is considered the reference for nuclear work [4]. This genome was assembled prior to high-throughput sequencing (i.e., Illumina and 454 technologies). The data presented here are not the sequences used for the paper but are part of a project by The Broad to resequence this individual using Illumina and 454 technologies.

- Note that both the Yoshida and Martin papers included ancient DNA in their analyses [9, 7]. Here we have omitted those samples and focused on modern samples.

- For enigmatic reasons, not all of the samples from the Yoshida and Martin papers were actually available online. Therefore our numbers here do not match those presented in the papers.

- The Judelson data include a sample of US1 which was sampled at three different time points (us1_1, us1_2 and us1_3). We suspect that these were different samples and not necessarily the same clone. Therefore differences among these samples may either be due to biological or technical factors.

- The Judelson data includes a sample of US8 which has been characterized as having fungicide resistance [3]. This lineage was also sequenced by Martin et al. [7]. These are most likely different samples so differences among these samples may be interpreted as biological.

- The Yoshida data includes one sample of *P. mirabilis* (p7722), this should jump out in the analyses.

- For the mitochondrial data we used the type IIa form [1] because it was the longest sequence and we felt this would provide the best alignment.

We've used the term 'SNP' fairly loosely in this document. The term 'variant' may be more appropriate. Until fairly recently the software tools we've been using could only handle SNPs. They now report short indels as well. We've included both variant types here.

2 Variant discovery

Reads were mapped to the type IIa mitochondrial reference "AY898627.1". Reads were mapped using bowtie2 [5]. Variants were called using SAMtools[6].

2.1 Variant filtering

As a quality control step, the variant files were filtered by quality, read depth and mapping quality (Figures 1,2). For this we used an in-house R package called vcfR. Here, sequencing depth is cumulative over all samples. Quality here is for each variant over all samples and ranges from 1-999.

The genotype caller in Samtools assumes a diploid, bi-allelic model. Because mitochondria are assumed to be haploid we tried to filter out heterozygous calls. Samples which included high quality heterozygote calls (p1362, p6096, p10650, p12204, p10127) were mostly from the Yoshida et al. [9] paper and were among the low sequencing depth samples they included. Because these samples are not among the US lineages we're interested in, and because they are apparently of low sequencing depth, we omitted them for now. However, the sample nl07434 was among the high sequencing depth samples from this paper and is perhaps noteworthy. T30-4 was called as a heterozygote for one variant and is perplexinng.

In an attept to identify high quality variants we employed a filtering strategy. Filtering of the variant panel was based on quality (QUAL=999), cumulative sequencing depth (1st quartile >= DP >= 3rd quartile) and mapping quality (1st quartile >= MQ >= 3rd quartile). This resulted in 37 variants remaining after filtering (Table 1). We have identified a fraction of these as being diagnostic for a small group of samples (Table 2).

The variants remaining after filtering were visualized as a linear chromosome in Figure 3.

```
## gt.m2sfs is commented out
## Before filtering:
## [1] 247
## After filtering:
## [1] 37
```

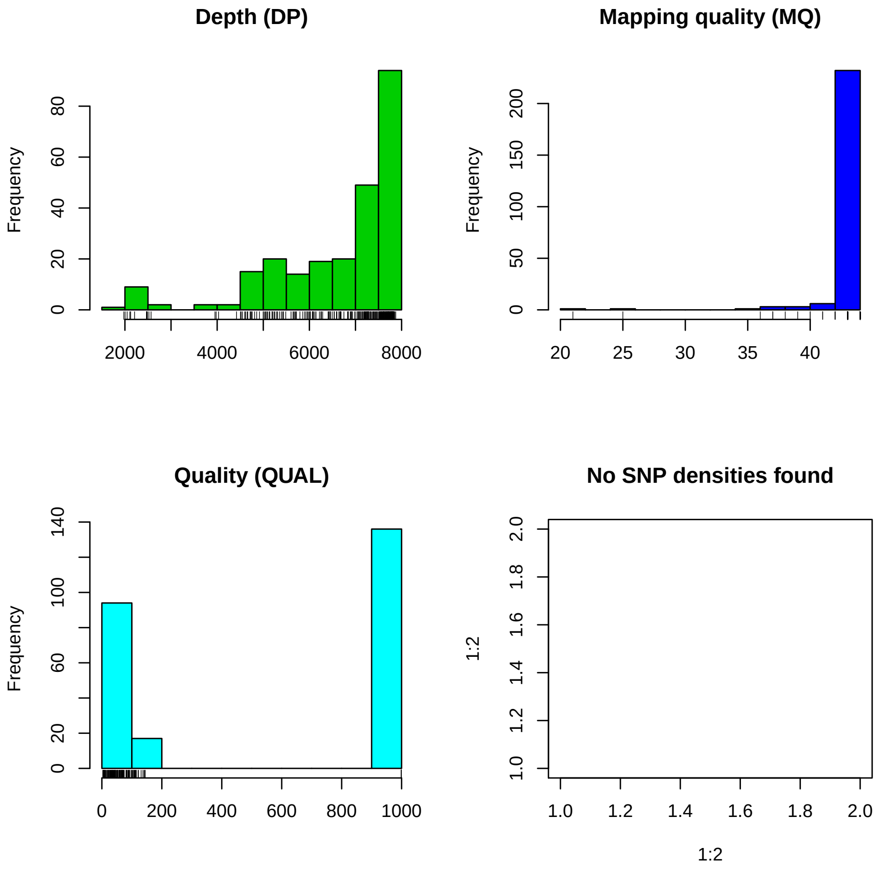


Figure 1: Quality control results for the mtDNA SNP calls before filtering.

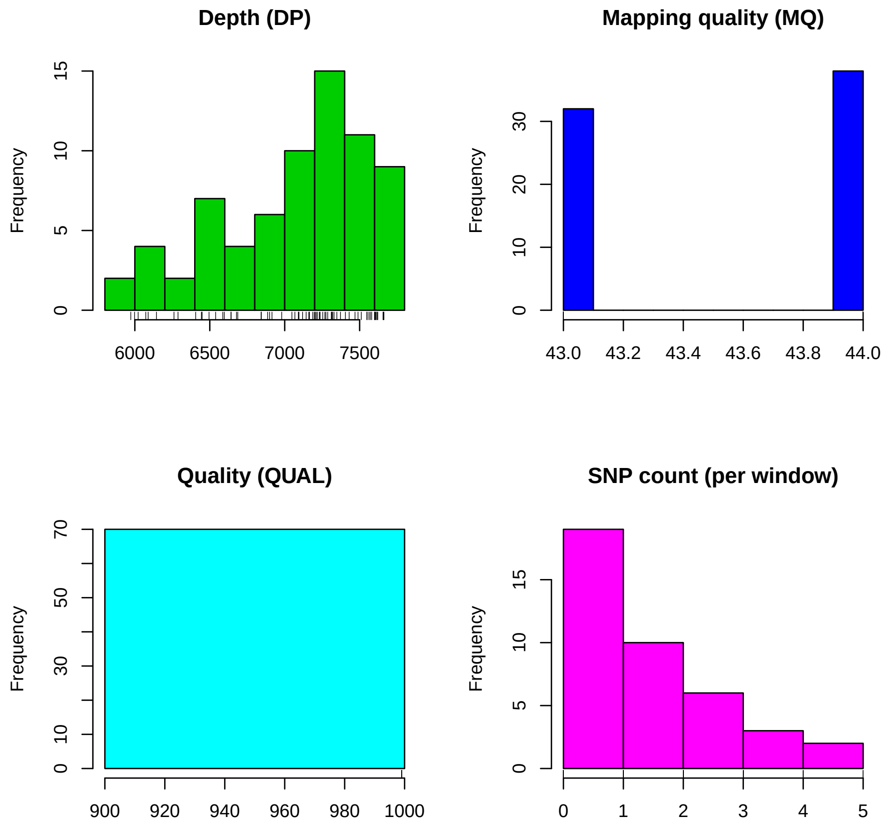


Figure 2: Quality control results for the mtDNA SNP calls after filtering and windowizing variants.

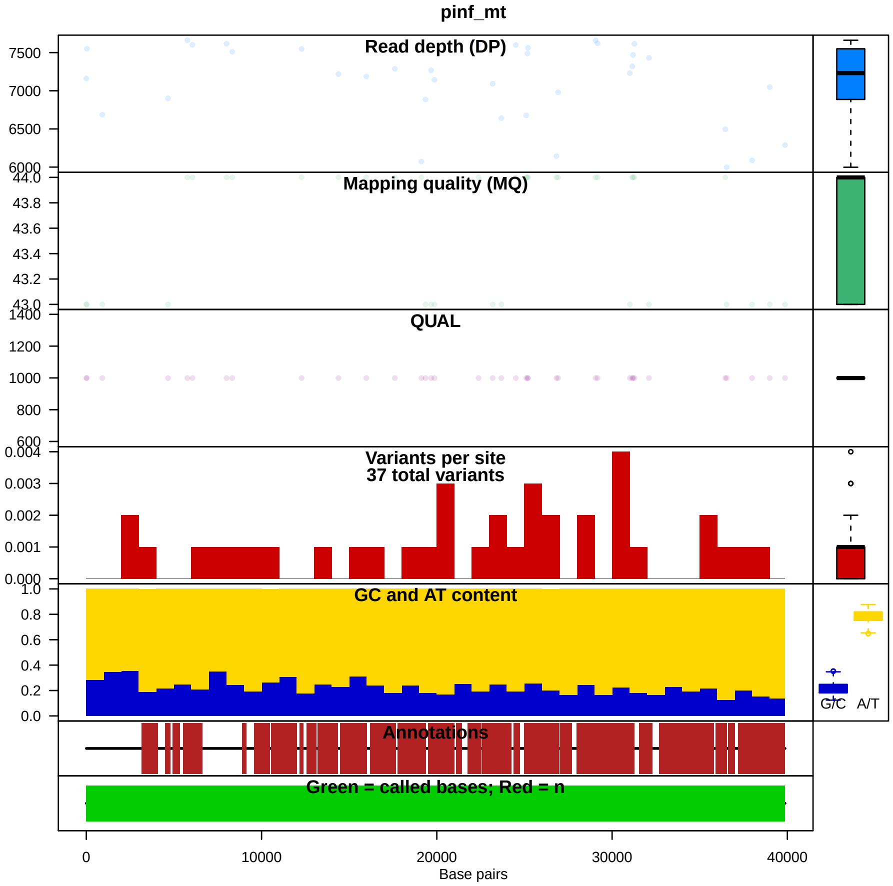


Figure 3: Whole mtDNA genome scan for the P. infestans samples.

3 Variant segregation

In order to visualize how variants segregated among the samples, a phylogeny was inferred. We then used ancestral state reconstruction to map the characters to the tree. At this time we're not trying to say anything bold about phylogeny or character evolution. We're simply using these tools to visualize how the variants segregate.

3.1 Phylogenetic reconstruction

Using the whole genome alignment (28 sequences, 39,870 nucleotides) we performed a whole-genome phylogeny using maximum likelihood (RAxML) and Bayesian inference (BEAST). We used RAxML using no partitions, 1000 bootstrap replicates, a `GTR+I+G` model of nucleotide evolution to obtain a bipartitioned tree with the boostrap values mapped to the branches. For BEAST, we specified `p7722` (*P.mirabilis*) as the outgroup. We used a `HKY+G+I` model of nucleotide substitutions, a strict molecular clock, a constant population size prior, UPGMA starting tree and 10 million Markov chains. The best tree is shown in Figure 4.

3.2 Mapping the SNP's in the BEAST tree

To map the variants found in the mtDNA genome to the coalescent tree, we used `Mesquite`. We did a removal of invariable regions and ancestral state reconstruction for all 37 SNPs using a parsimony reconstruction state (see tree figures).

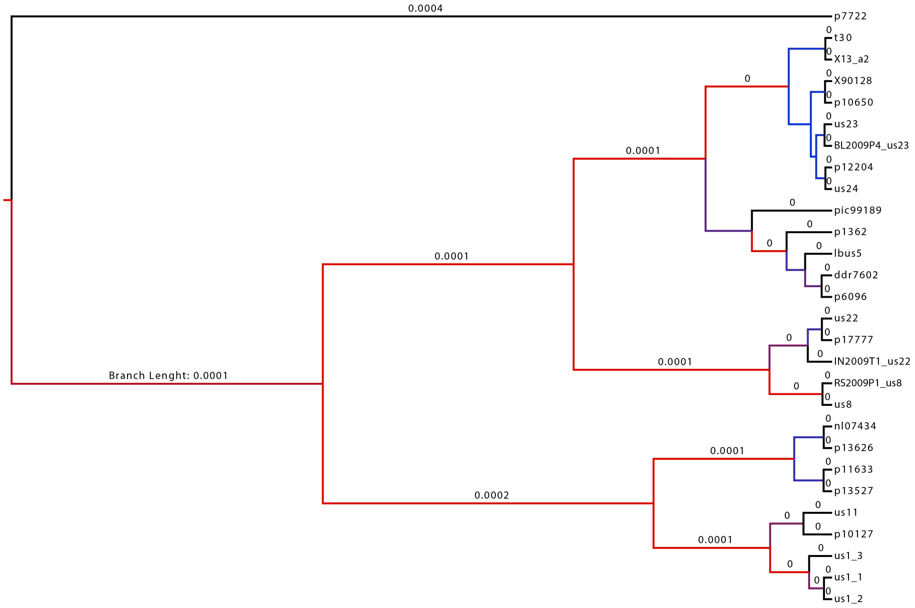
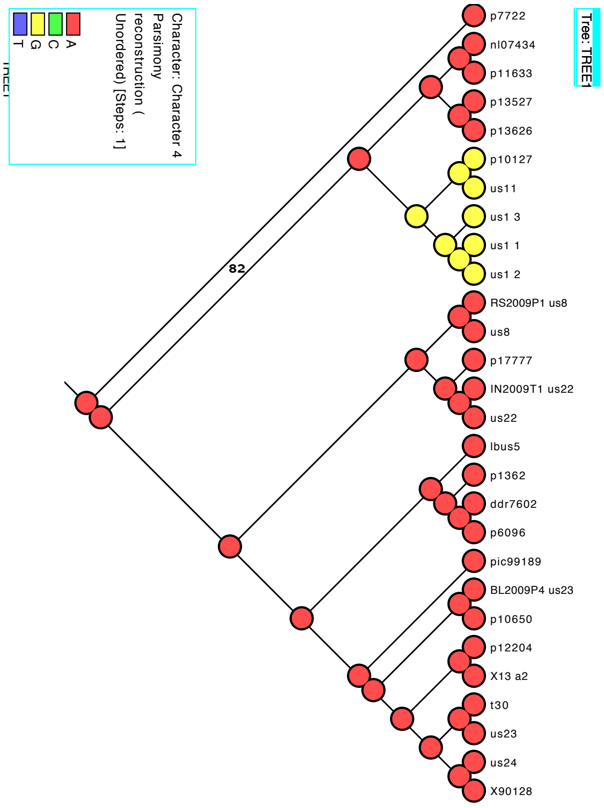
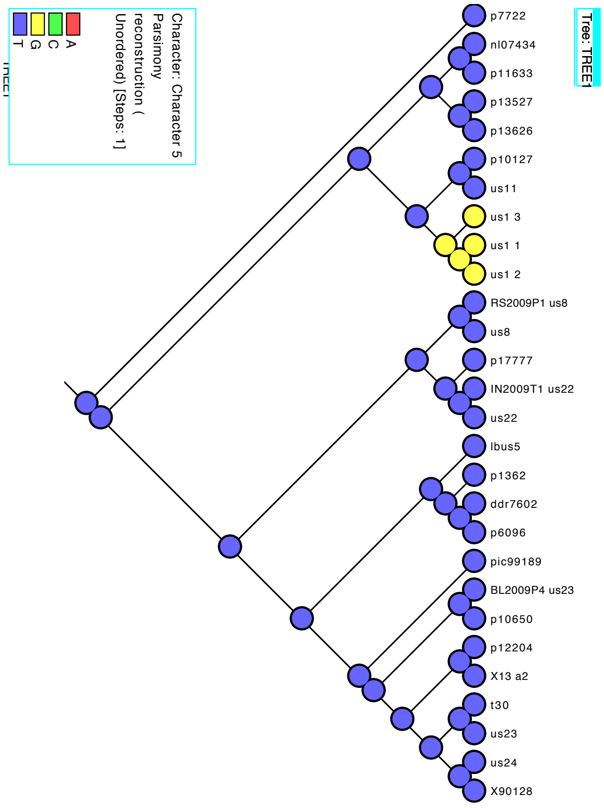
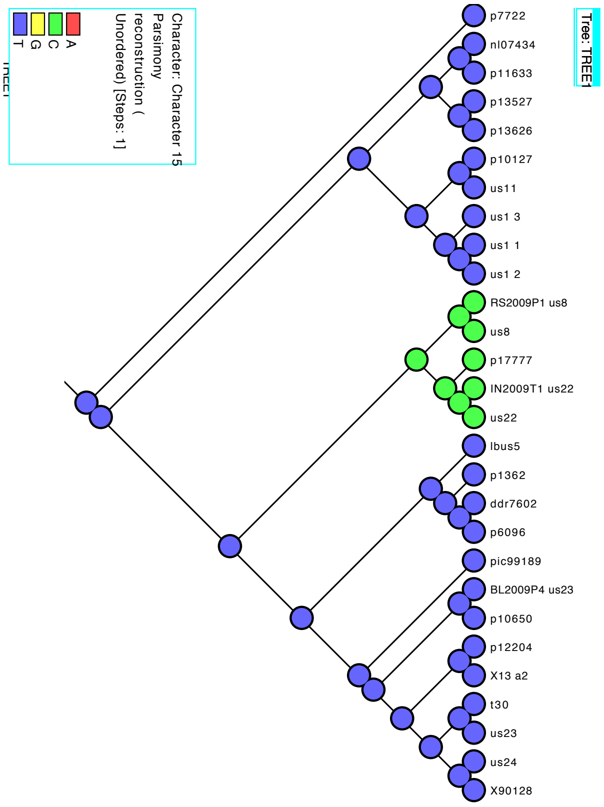
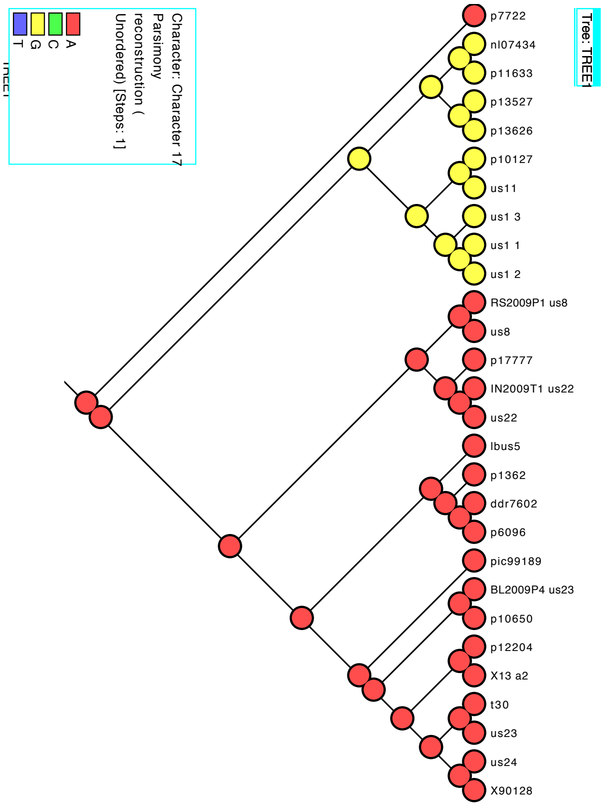


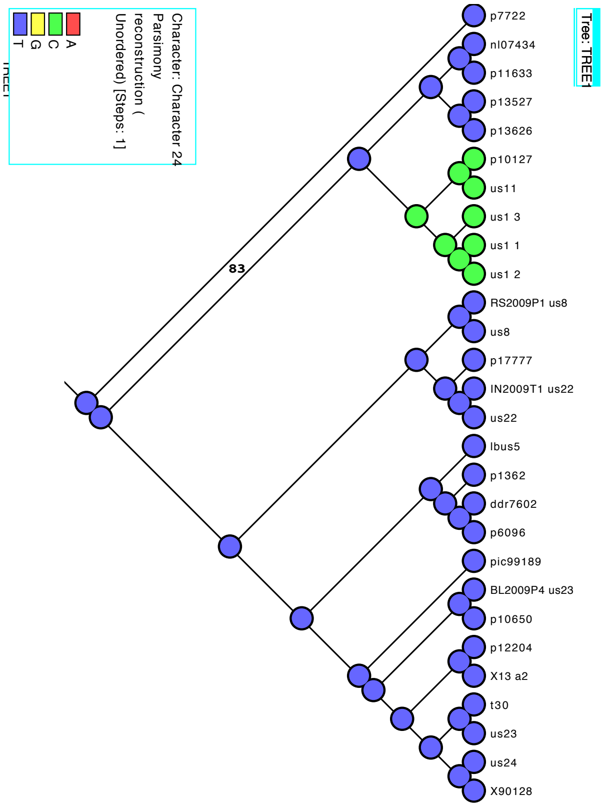
Figure 4: Bayesian coalescent tree of the whole mtDNA genome of *P. infestans* using `BEAST 1.8.0`. Values above branches represent branch lengths (topology is concordant to the bifurcating model of a coalescent reconstruction). Branches are colored based on their posterior probability values (legend indicates color scheme).

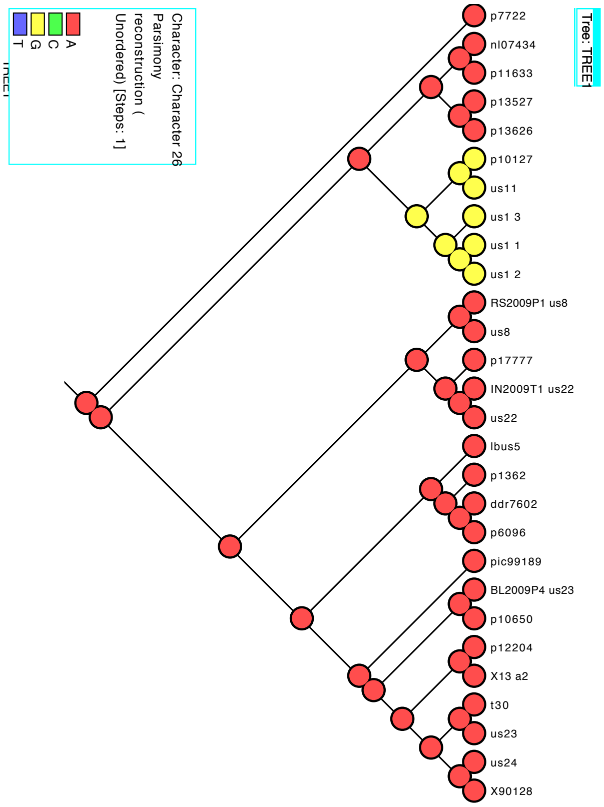

Character: Character 4
Parsimony
reconstruction (
Unordered) [Steps: 1]
A
C
G
T
Tree: TREE1
82
p7722
nl07434
p11633
p13527
p13626
p10127
us11
us1 3
us1 1
us1 2
RS2009P1 us8
us8
p17777
IN2009T1 us22
us22
lbus5
p1362
ddr7602
p6096
pic99189
BL2009P4 us23
p10650
p12204
X13 a2
t30
us23
us24
X90128

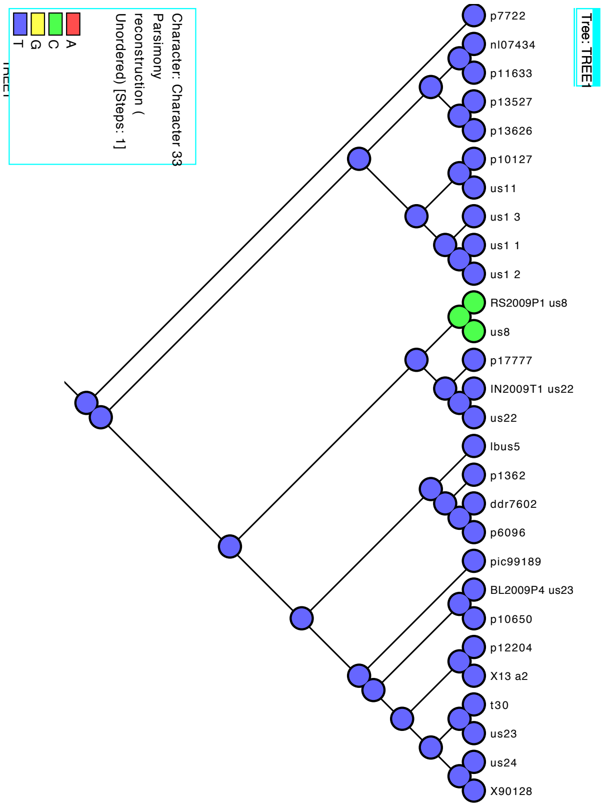

Character: Character 5
Parsimony
reconstruction (
Unordered) [Steps: 1]
A
C
G
T
Tree: TREE1
p7722
nl07434
p11633
p13527
p13626
p10127
us11
us1 3
us1 1
us1 2
RS2009P1 us8
us8
p17777
IN2009T1 us22
us22
lbus5
p1362
ddr7602
p6096
pic99189
BL2009P4 us23
p10650
p12204
X13 a2
t30
us23
us24
X90128


Character: Character 15
Parsimony
reconstruction (
Unordered) [Steps: 1]
A
C
G
T
Tree: TREE1
p7722
nl07434
p11633
p13527
p13626
p10127
us11
us1 3
us1 1
us1 2
RS2009P1 us8
us8
p17777
IN2009T1 us22
us22
lbus5
p1362
ddr7602
p6096
pic99189
BL2009P4 us23
p10650
p12204
X13 a2
t30
us23
us24
X90128


Character: Character 17
Parsimony
reconstruction (
Unordered) [Steps: 1]
A
C
G
T
Tree: TREE1
p7722
nl07434
p11633
p13527
p13626
p10127
us11
us1 3
us1 1
us1 2
RS2009P1 us8
us8
p17777
IN2009T1 us22
us22
lbus5
p1362
ddr7602
p6096
pic99189
BL2009P4 us23
p10650
p12204
X13 a2
t30
us23
us24
X90128


Tree: TREE1
Character: Character 24
Parsimony
reconstruction (
Unordered) [Steps: 1]
A
C
G
T
83
p7722
nl07434
p11633
p13527
p13626
p10127
us11
us1 3
us1 1
us1 2
RS2009P1 us8
us8
p17777
IN2009T1 us22
us22
lbus5
p1362
ddr7602
p6096
pic99189
BL2009P4 us23
p10650
p12204
X13 a2
t30
us23
us24
X90128


Tree: TREE1
Character: Character 26
Parsimony
reconstruction (
Unordered) [Steps: 1]
A
C
G
T
p7722
nl07434
p11633
p13527
p13626
p10127
us11
us1 3
us1 1
us1 2
RS2009P1 us8
us8
p17777
IN2009T1 us22
us22
lbus5
p1362
ddr7602
p6096
pic99189
BL2009P4 us23
p10650
p12204
X13 a2
t30
us23
us24
X90128


Tree: TREE1
Character: Character 33
Parsimony
reconstruction (
Unordered) [Steps: 1]
A
C
G
T
p7722
nl07434
p11633
p13527
p13626
p10127
us11
us1 3
us1 1
us1 2
RS2009P1 us8
us8
p17777
IN2009T1 us22
us22
lbus5
p1362
ddr7602
p6096
pic99189
BL2009P4 us23
p10650
p12204
X13 a2
t30
us23
us24
X90128

4 Session information

```
sessionInfo()

## R version 3.0.2 (2013-09-25)
## Platform: x86_64-apple-darwin10.8.0 (64-bit)
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] xtable_1.7-1 vcfR_0.1     knitr_1.5
##
## loaded via a namespace (and not attached):
## [1] ape_3.0-11      evaluate_0.5.1  formatR_0.10    grid_3.0.2
## [5] lattice_0.20-24 nlme_3.1-113    stringr_0.6.2   tools_3.0.2
```

References

[1] Cruz Avila-Adame, Luis Gómez-Alpizar, Victoria Zismann, Kristine M Jones, C Robin Buell, and Jean Beagle Ristaino. Mitochondrial genome sequences and molecular evolution of the Irish potato famine pathogen, *Phytophthora infestans*. *Current genetics*, 49(1):39–46, 2006.

[2] David EL Cooke, Liliana M Cano, Sylvain Raffaele, Ruairidh A Bain, Louise R Cooke, Graham J Etherington, Kenneth L Deahl, Rhys A Farrer, Eleanor M Gilroy, Erica M Goss, et al. Genome analyses of an aggressive and invasive lineage of the irish potato famine pathogen. *PLoS pathogens*, 8(10):e1002940, 2012.

[3] G Danies, IM Small, K Myers, R Childers, and William E Fry. Phenotypic characterization of recent clonal lineages of *Phytophthora infestans* in the united states. *Plant Disease*, 97(7):873–881, 2013.

[4] Brian J Haas, Sophien Kamoun, Michael C Zody, Rays HY Jiang, Robert E Handsaker, Liliana M Cano, Manfred Grabherr, Chinnappa D Kodira, Sylvain Raffaele, Trudy Torto-Alalibo, et al. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature*, 461(7262):393–398, 2009.

[5] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–359, 2012.

[6] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[7] Michael D Martin, Enrico Cappellini, Jose A Samaniego, M Lisandra Zepeda, Paula F Campos, Andaine Seguin-Orlando, Nathan Wales, Ludovic Orlando, Simon YW Ho, Fred S Dietrich, et al. Reconstructing genome evolution in historic samples of the Irish potato famine pathogen. *Nature communications*, 4, 2013.

[8] Sylvain Raffaele, Joe Win, Liliana M Cano, and Sophien Kamoun. Analyses of genome architecture and gene expression reveal novel candidate virulence factors in the secretome of *Phytophthora infestans*. *BMC genomics*, 11(1):637, 2010.

[9] Kentaro Yoshida, Verena J Schuenemann, Liliana M Cano, Marina Pais, Bagdevi Mishra, Rahul Sharma, Chirsta Lanz, Frank N Martin, Sophien Kamoun, Johannes Krause, et al. Correction: The rise and fall of the *Phytophthora infestans* lineage that triggered the Irish potato famine. *eLife*, 2, 2013.

	CHROM	POS	REF	ALT
6	Supercontig_1.1	2701	C	T
7	Supercontig_1.1	2728	C	A
12	Supercontig_1.1	3519	G	T
36	Supercontig_1.1	6872	accc	acc
42	Supercontig_1.1	7857	A	G
44	Supercontig_1.1	8118	T	G
59	Supercontig_1.1	9857	T	C
62	Supercontig_1.1	10146	A	G
85	Supercontig_1.1	13684	A	G
106	Supercontig_1.1	15563	G	A
114	Supercontig_1.1	16986	G	A
119	Supercontig_1.1	18440	A	T
123	Supercontig_1.1	19793	G	A
124	Supercontig_1.1	20005	C	A
125	Supercontig_1.1	20290	G	T
128	Supercontig_1.1	20464	T	C
141	Supercontig_1.1	22711	G	A
147	Supercontig_1.1	23431	G	A
150	Supercontig_1.1	23872	G	T
154	Supercontig_1.1	24611	C	T
159	Supercontig_1.1	25142	A	C
160	Supercontig_1.1	25204	T	C
161	Supercontig_1.1	25237	G	T
169	Supercontig_1.1	26684	A	G
171	Supercontig_1.1	26767	T	C
186	Supercontig_1.1	28680	C	A
187	Supercontig_1.1	28783	A	G
192	Supercontig_1.1	30427	T	G
193	Supercontig_1.1	30552	A	G
194	Supercontig_1.1	30591	A	G
195	Supercontig_1.1	30660	C	T
201	Supercontig_1.1	31403	T	C
221	Supercontig_1.1	35296	G	T
222	Supercontig_1.1	35367	taaaaaaaaaaa	taaaaaaaaaa
233	Supercontig_1.1	36663	T	C
239	Supercontig_1.1	37562	C	T
242	Supercontig_1.1	38345	C	A

Table 1: Variants remaining after filtering.

Table 2: Diagnostic SNP positions for the mtDNA genome after filtering. The tree number corresponds to the character legend on the tree figures

Position	SNP	Diagnostic for	Tree Number
7857	A/G	p10127, us11 and us1	4
8118	T/G	us1	5
20464	T/C	p17777, us22 and us8	15
22711	G/A	p17777, us22 and us8	17
26767	T/C	p10127, us11 and us1	24
28783	A/G	p10127, us11 and us1	26
36663	T/C	us8	33