

# Discovery of Western European R1b1a2 Y Chromosome Variants in 1000 Genomes Project Data: An Online Community Approach

Richard A. Rocca<sup>1\*</sup>, Gregory Magoon<sup>2</sup>, David F. Reynolds<sup>3</sup>, Thomas Krahn<sup>4</sup>, Vincent O. Tilroe<sup>5</sup>, Peter M. Op den Velde Boots<sup>6</sup>, Andrew J. Grierson<sup>7</sup>

**1** Independent Researcher, Saddle Brook, New Jersey, United States of America, **2** Independent Researcher, Manchester, Connecticut, United States of America, **3** Independent Researcher, Portland, Oregon, United States of America, **4** FamilyTreeDNA Genomics Research Center, Houston, Texas, United States of America, **5** Independent Researcher, Edmonton, Alberta, Canada, **6** Independent Researcher, Amsterdam, North Holland, The Netherlands, **7** Sheffield Institute for Translational Neuroscience, University of Sheffield, Sheffield, United Kingdom

## Abstract

The authors have used an online community approach, and tools that were readily available via the Internet, to discover genealogically and therefore phylogenetically relevant Y-chromosome polymorphisms within core haplogroup R1b1a2-L11/S127 (rs9786076). Presented here is the analysis of 135 unrelated L11 derived samples from the 1000 Genomes Project. We were able to discover new variants and build a much more complex phylogenetic relationship for L11 sub-clades. Many of the variants were further validated using PCR amplification and Sanger sequencing. The identification of these new variants will help further the understanding of population history including patrilineal migrations in Western and Central Europe where R1b1a2 is the most frequent haplogroup. The fine-grained phylogenetic tree we present here will also help to refine historical genetic dating studies. Our findings demonstrate the power of citizen science for analysis of whole genome sequence data.

**Citation:** Rocca RA, Magoon G, Reynolds DF, Krahn T, Tilroe VO, et al. (2012) Discovery of Western European R1b1a2 Y Chromosome Variants in 1000 Genomes Project Data: An Online Community Approach. PLoS ONE 7(7): e41634. doi:10.1371/journal.pone.0041634

**Editor:** Toomas Kivisild, University of Cambridge, United Kingdom

**Received:** March 21, 2012; **Accepted:** June 22, 2012; **Published:** July 24, 2012

**Copyright:** © 2012 Rocca et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** The authors have no support or funding to report.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: rrocca@live.com

## Introduction

Population geneticists are in short supply, and most are involved in the important hunt for genetic factors responsible for susceptibility to disease [1]. Naturally, studies that address genetic susceptibility are prioritized over other types of genetic research. As a result, these studies free up little in the way of both human and monetary resources dedicated to the application of genetics to the history of the human species [2,3]. The study of human phylogenetics is where the practice of citizen science can be a valuable resource to the scientific community. In the name of citizen science, the authors of this study pooled together their resources and volunteered their time to data mine the full genomes of over a thousand samples that were made freely available through the 1000 Genomes Project [4].

The 1000 Genomes Project is the first project to sequence the genomes of a large number of people. The plan for the full project is to sequence about 2,500 samples; however, only the data sets of 1,197 samples from 13 populations were available in early 2011. It is estimated that over 100 million European men belong to haplogroup R1b1a2 (M269) [5], and that greater than 70% of western European men belong to the specific clade defined by SNP L11 [6]. The focus of the work was finding phylogenetic Y-DNA variants in the two largest sub-clades of L11: S116/P312 (rs34276300) and U106/S21 (rs16981293). To illustrate the need for more refined testing, no study to date has tested for L48

(rs13303755), the most frequent variant in the U106 branch of L11 (<http://www.familytreedna.com/public/u106/>), and published studies still show ambiguous S116 (xU152, L21) as the most frequent variant in Spain, Portugal and parts of France [6,7].

Recently, there has been much controversy in the dating and dating techniques used to identify the geographical distribution of R1b1a2 by way of microsatellite variance or diversity [7]. The paucity of haplogroup defining genetic markers has meant that these microsatellite-derived dating calculations have to be conducted without regard to lower level phylogenetic relationships, and therefore erroneously compare populations that may be phylogenetically distant. By identifying the lower level branches of the R1b1a2 phylogenetic tree, more accurate dating of truly related haplogroups will be possible.

## Results

By analysing the 1000 Genomes Project Phase 1 data set of 1,197 individuals, we identified 135 samples bearing the L11 SNP. Excluding Finland, which has a low L11 frequency, approximately 50% of the remaining datasets comprising European populations CEU (CEPH Utah residents with Northern and Western European ancestry), GBR (British in England and Scotland), IBS (Iberian populations in Spain) and TSI (Tuscans in Italy) were derived at L11 [8]. An additional source of L11 derived datasets came from Latin American populations MXL (Mexican Ancestry

in Los Angeles, California), PUR (Puerto Rican in Puerto Rico) and CML (Colombian in Medellin, Colombia) and to a minor extent, the ASW (African Ancestry in Southwest US) population. L11 is divided into two major sub-clades: S116 and U106. A large majority of L11 samples belong to subclade S116 (109 out of 135 or 81%). Using SAMtools and filtering methodology described in the methods section, we identified more than 200 putative non-singleton novel genetic variants in the 135 R1b1a2-L11 samples. The resulting R1b1a2-L11 phylogenetic tree based on Phase 1 1000 Genomes data is presented in Figures 1, 2, 3 and 4.

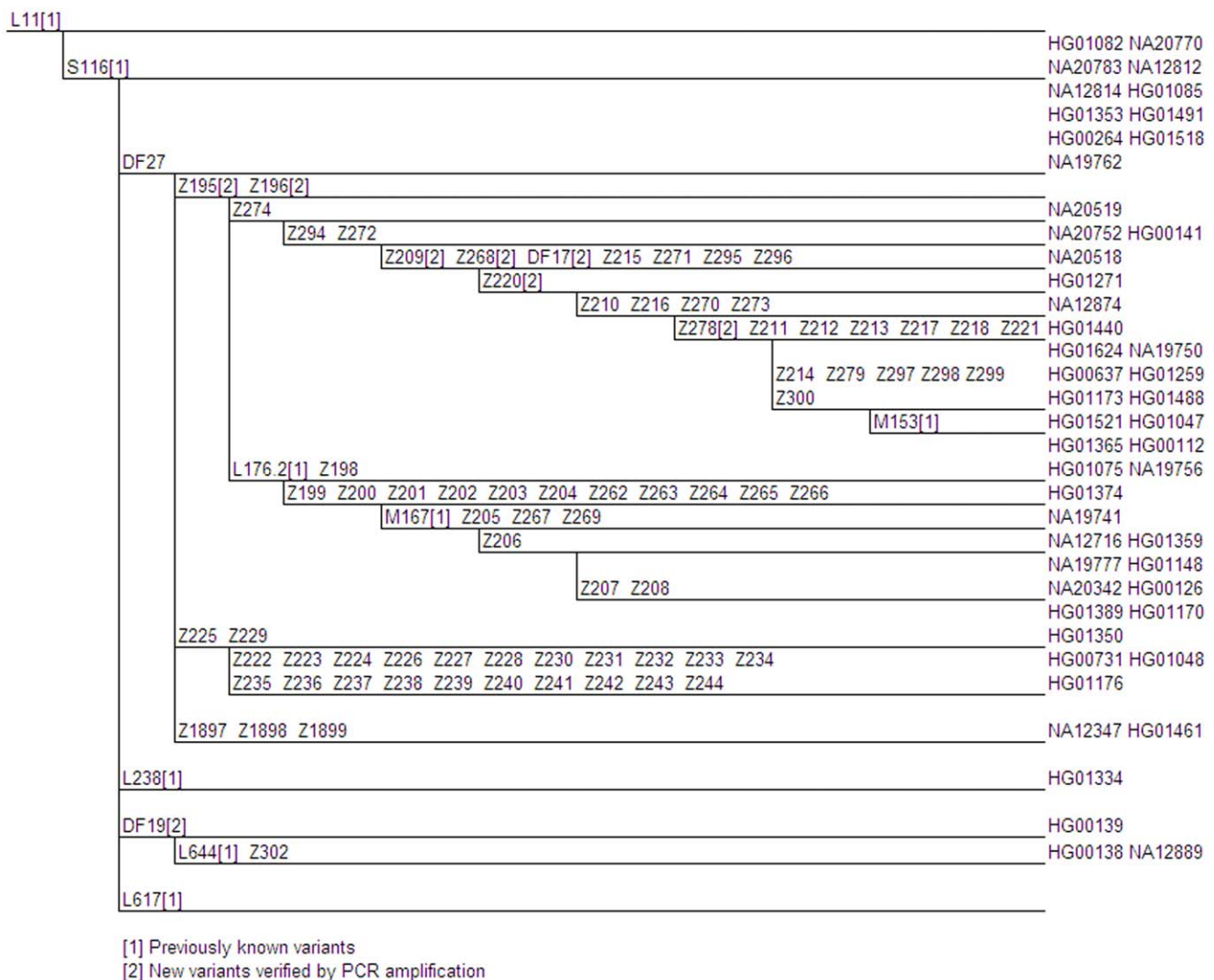
### R1b1a2-S116 Sub-Haplogroups

Forty-two of the 49 samples that would previously have been categorized as belonging to unspecified S116(xU152,L21) were derived for the DF27 SNP (Figure 1). The majority of the DF27 derived samples (71%) were from Latin American/Iberian populations (CLM, IBS, MXL, PUR). This variant is therefore likely to account for the majority of previously unclassified S116(xU152,L21) reported in Iberia and some areas of France [7]. Below DF27, we identified 17 subclades and a total of 73

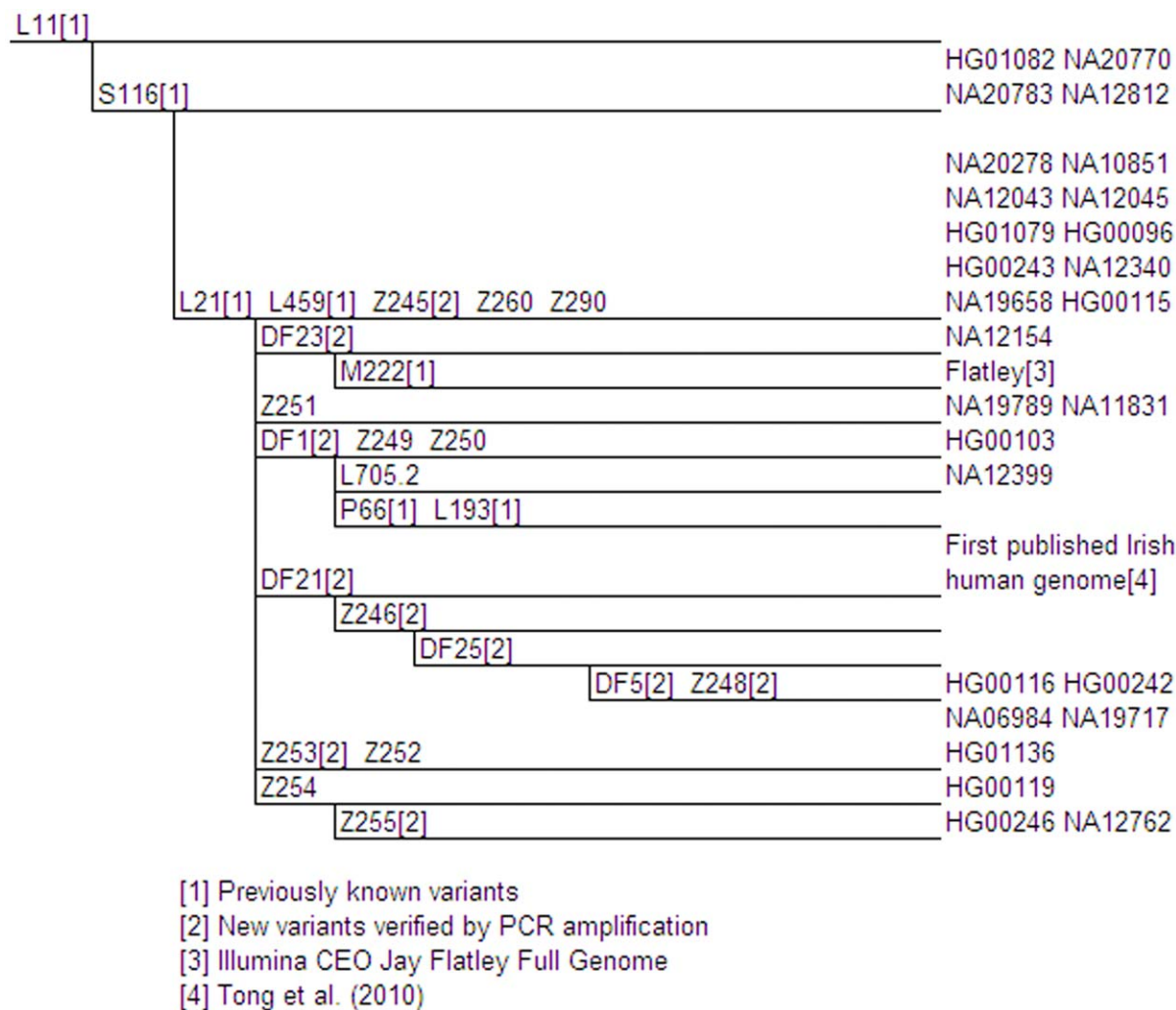
novel mutations. We were also able to place previously known SNP Z278 (rs1469371) into a phylogenetic context and show for the first time a close relationship between markers M153, M167/SRY2627 and L176.2. In the branch of S116 defined by DF27 a total of six new SNPs were validated by PCR amplification and Sanger sequencing. A description of the locations and sequences of all primers developed for validation experiments is given in Table S1.

A smaller subclade of S116(xU152,L21,DF27) consisting of two GBR samples and a single CEU sample was defined by DF19. Another S116(xU152,L21,DF27) GBR sample carried the previously unpublished L238 SNP. The northern European origin of the sample is not surprising given that all previously known L238 samples have been from Scandinavia (<http://www.familytreedna.com/public/atlantic-r1b1c/>).

Marker L21/M529/S127 (rs11799226) is currently perhaps the most clearly geographically localized of the major L11 sub-haplogroups, with a high frequency in the British Isles and Brittany [7]. Twenty-three samples were found to be derived at marker L21 (Figure 2). Unsurprisingly, most L21 derived samples



**Figure 1. Proposed S116 (xU152, L21) Phylogenetic Tree.** Genetic variants are indicated on branches, and branch lengths are not proportional to the number of mutations or the age of the variant. Phylogenetically equivalent markers are shown in alphabetical and numerical order. Full details of these variants are shown in Table S1. The positions of 1000 Genomes samples are given at the tips of the branches. doi:10.1371/journal.pone.0041634.g001



**Figure 2. Proposed L21 Phylogenetic Tree.** Genetic variants are indicated on branches, and branch lengths are not proportional to the number of mutations or the age of the variant. Phylogenetically equivalent markers are shown in alphabetical and numerical order. Full details of these variants are shown in Table S1. The positions of 1000 Genomes samples are given at the tips of the branches.  
 doi:10.1371/journal.pone.0041634.g002

were either from GBR (34.8%) or from the CEU (30.4%) population, which itself is most similar to samples from the Netherlands and the UK [9]. The remainder of the samples were from MXL, CLM, PUR and ASW. We were able to group 13 new variants into 13 subclades of L21. Ten of the new markers have been confirmed by PCR amplification and Sanger sequencing. Some of these were found by comparing L21 derived 1000 Genomes Project samples with two publicly available genomes (see Materials and Methods). The previously known marker M222 was re-positioned within the L21 group below DF23, and new sub-haplogroups were defined by Z251, DF1, DF21 (rs138322855), Z253 and Z255. DF21 and DF1 lie upstream of a number of new and previously identified variants, respectively (Figure 2).

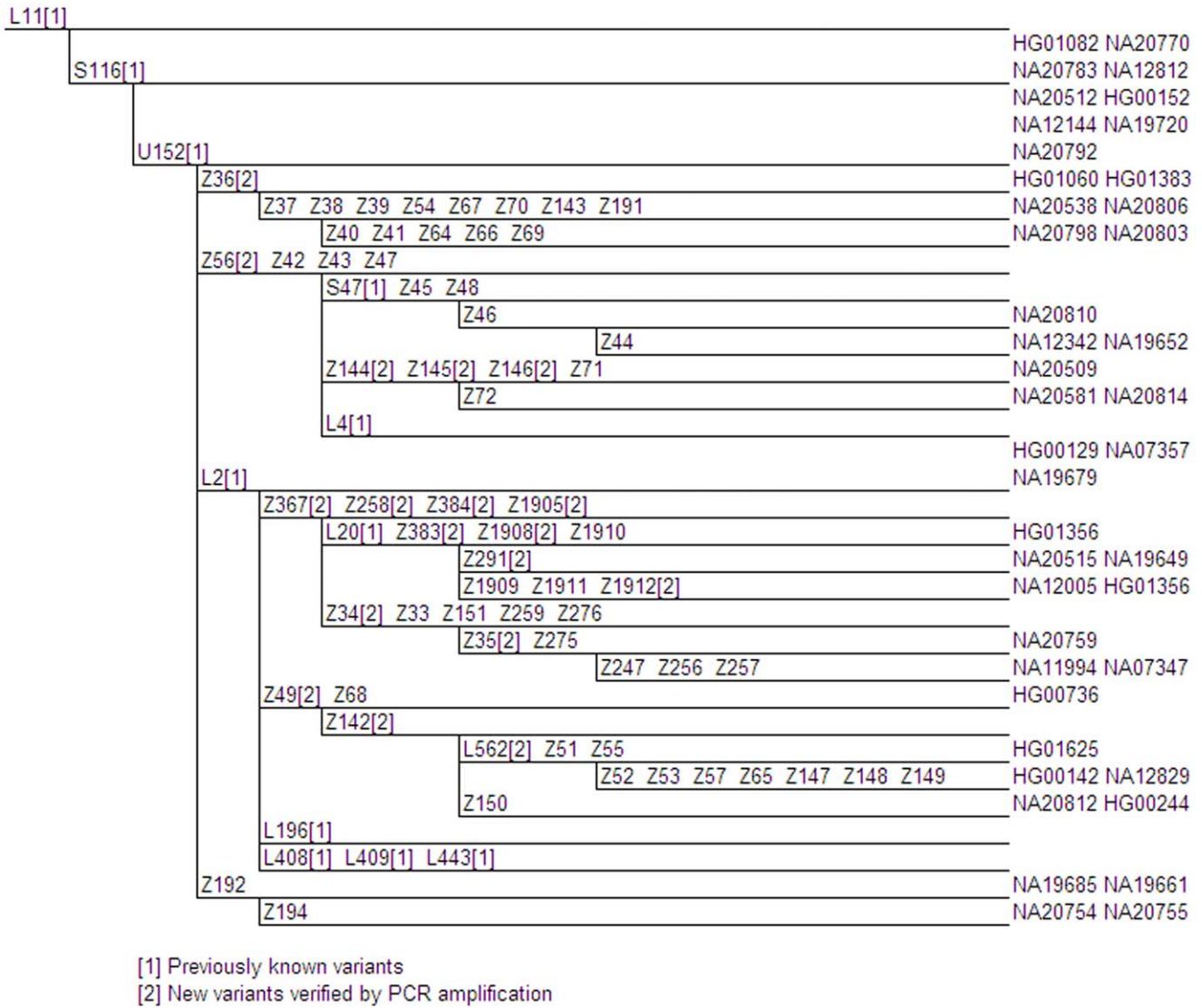
The S116 downstream marker U152/S28 (rs1236440) defines the most common Y chromosome haplogroup in northern and central Italy, Switzerland, eastern France and Corsica [7]. The genomes of 36 U152 derived males were analyzed (Figure 3). Samples from the Tuscany region of Italy were found to have a high frequency of U152 (29.4%). This correlates well with the frequency of 32.4% previously reported in central Italy and the 32.1% found in Corsica [7,10]. We were able to further refine

U152 into 26 subclades. Fifty-four new variants were found in all, 11 of which have been validated by PCR amplification and Sanger sequencing. Additionally, SNPs Z367 (rs7067387), Z258 (rs9785865), Z384 (rs28819996), Z1905 (rs7892878), Z383 (rs34173183), Z1908 (rs11799240) and Z1912 (rs4893798) were placed below U152 in the phylogenetic tree for the first time. To date all U152 samples with DYS492 = 14 appear to be in the Z56 subhaplogroup, including, but not limited to those derived at marker L4 (Family Tree DNA R1b-U152 Project: <http://www.familytreedna.com/public/R1b-U152/default.aspx>).

Of the four remaining S116(xDF19, DF27, L21, L238, L617, U152) samples, three were missing data for DF27 and a fourth had a weak ancestral read.

### R1b1a2-U106 Sub-Haplogroups

Sub-haplogroup U106/S21/M405 (rs16981293) (Figure 4) makes up the other half of the L11 story in Europe. It is the most common R1b1a2 marker in central Europe, and is by far the most frequent SNP in the Netherlands and Belgium [7,11]. Twenty-six samples were found derived at marker U106. As was the case with marker L21, GBR (42.3%) and CEU (42.3%)



**Figure 3. Proposed U152 Phylogenetic Tree.** Genetic variants are indicated on branches, and branch lengths are not proportional to the number of mutations or the age of the variant. Phylogenetically equivalent markers are shown in alphabetical and numerical order. Full details of these variants are shown in Table S1. The positions of 1000 Genomes samples are given at the tips of the branches.  
doi:10.1371/journal.pone.0041634.g003

population samples made up a large percentage of U106 samples. The U106 tree now consists of 30 downstream subclades. We found 49 novel SNPs in this group and were able to place known SNPs Z301 (rs35121273) and Z381 (rs34001725) into the U106 sub-haplogroup. Fifteen of these new variants were confirmed by PCR amplification and Sanger sequencing. The known markers L48 (rs13303755) and U198/S29 (rs17222279) were also relocated to lower branches in the U106 tree. George Church’s genome allowed us to place U198 with respect to other 1000 Genome Project samples (<http://www.personalgenomes.org/pgp10.html>).

**The Validated L11 Phylogeny**

Despite using a number of filtering steps to reduce false discovery rates, it remains formally possible that a number of variants may be artefacts, or are refractory to Sanger sequencing. Thus we also present here a summary of the L11 phylogeny containing all polymorphisms that have been independently

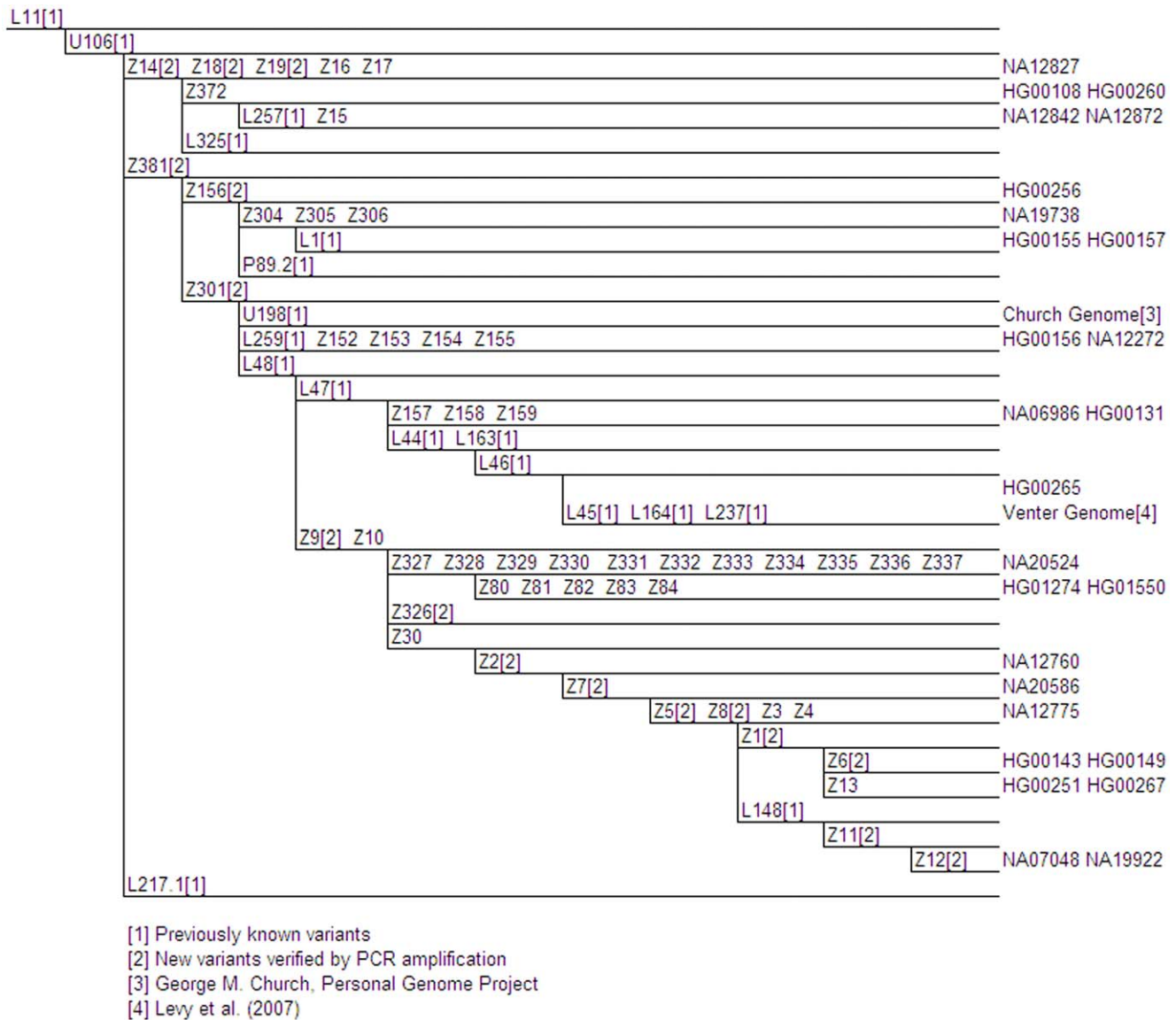
validated to date by PCR amplification and Sanger Sequencing (Figure 5).

**Discussion**

The Y chromosome haplogroup tree has evolved from a collection of 243 unique polymorphisms in 2002, to several thousand markers today [12] (<http://www.isogg.org/tree/>). The availability of high quality whole-chromosome sequence data from thousands of individuals has expedited the discovery of new polymorphisms, and the open access nature of the 1000 Genomes Project has enabled public volunteers to practice citizen science to trail-blaze a new haplogroup tree for R1b1a2-L11.

The R1b1a2-L11 haplogroup is prevalent in Western Europe, and this has led to conflicting opinions about the spread of populations from east to west since the Neolithic [5,6]. Busby et al. stated that coalescence estimates explicitly depend on the STRs that one uses [7], thus the use of relatively small numbers of microsatellite markers in previous studies may have been





**Figure 4. Proposed U106 Phylogenetic Tree.** Genetic variants are indicated on branches, and branch lengths are not proportional to the number of mutations or the age of the variant. Phylogenetically equivalent markers are shown in alphabetical and numerical order. Full details of these variants are shown in Table S1. The positions of 1000 Genomes samples are given at the tips of the branches. doi:10.1371/journal.pone.0041634.g004

problematic, but a more precise sub-classification of R1b1a2-L11 by SNPs will help to resolve these controversies. Better SNP resolution could also answer questions regarding the use of evolutionarily effective and germ line mutation rates [13]. Given the regional affinities of L21, U152, U106 and now DF27, we can already see where studying their downstream variants can help answer questions whose answers have long eluded archaeologists and linguists alike, especially if some of the newly defined haplogroups prove to be geographically clustered.

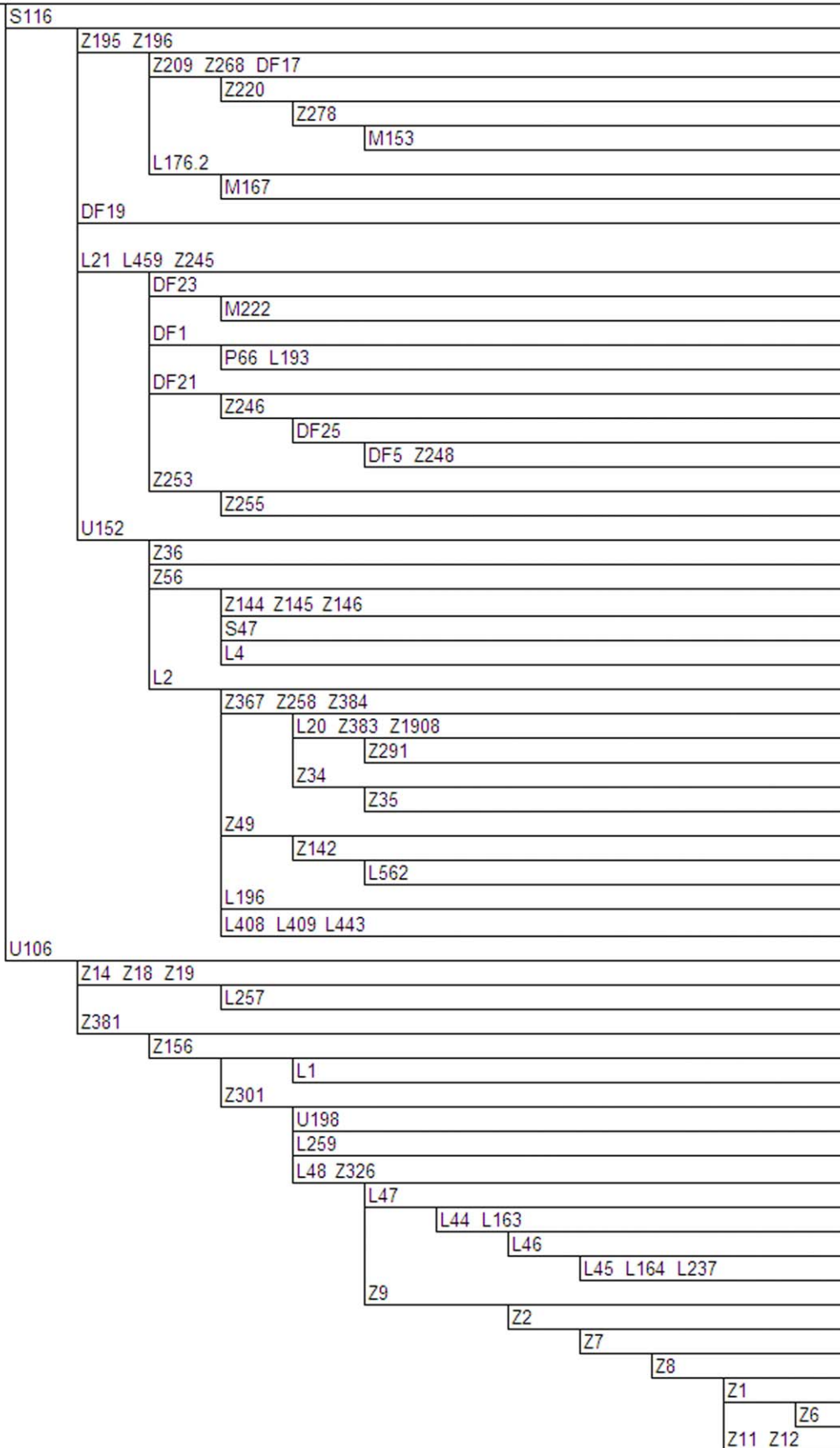
In Iberia, previous data suggests evidence of gene flow between Basques and Catalan speaking populations based on the distribution of M167 (SRY2627) [14,15,16]. Our placement of M153, a marker that has been linked to Basque populations [17], and SRY2627 as downstream variants of DF27 further illustrates this close relationship. The position of M153 many levels down on the phylogenetic tree also shows its relative youth within R1b1a2, as does its low STR diversity [18] (Family Tree DNA M153 Project:

[http://www.familytreedna.com/public/R-M153\\_The\\_Basque\\_Marker/default.aspx](http://www.familytreedna.com/public/R-M153_The_Basque_Marker/default.aspx)).

The concentration of Latin American/Iberian samples derived at DF27 shows the geographical importance of this marker in Iberia. That only Latin American samples were members of the DF27 subclade defined by the mutations Z225 and Z229 further illustrates strong Iberian ties and is consistent with the possibility of colonial era founders in the Americas [19].

Genetic approaches offer unique possibilities to resolve long-standing historical and archaeological dilemmas. Analysis of subclades defined by the new genetic markers reported here could help resolve some of these. As marker L21 has its highest frequencies in areas where insular Celtic languages once dominated, and some areas where they are still spoken today, there is no doubt that understanding its subclades will be instrumental in any debate regarding Celtic origins [20]. Given that U152 is the most frequent marker in northern and central

L11



**Figure 5. Validated L11 Phylogeny.** Genetic variants that have been validated by Sanger sequencing are indicated on branches, and branch lengths are not proportional to the number of mutations or the age of the variant. Phylogenetically equivalent markers are shown in alphabetical and numerical order. Full details of these variants are shown in Table S1. doi:10.1371/journal.pone.0041634.g005

Italy, its new subclades should prove valuable in resolving the fundamental problems of establishing the origin of IE groups in Italy [20,21,22]. Marker U106 and its subclades, so prominent along the Rhine River, can give insights into the spread of Germanic languages. And finally, any future testing of ancient Y-DNA Bell Beaker skeletons should start testing for these markers as Bell Beaker skeletons have already turned up derived at marker R1b-M269 with no downstream markers tested [23]. The discovery of these new markers is only the first step. While commercial testing of some of these markers has commenced, the real benefit will come from typing a large number of L11 derived individuals of phylogeographic background.

We are currently witnessing the resolution of two of the biggest impediments to human population genetics. The first is that of cost, with the \$1,000 full genome sequence seemingly around the corner [24]. The second is the extraction of ancient Y-DNA, which is already becoming a reality [25,26]. The more economically feasible it is to sequence the entire genome of existing humans and the easier it is to apply the novel variants found in this study and compare them to ancient Y-DNA, the quicker our historical, archaeological and linguistic questions will be answered in regard to the populating of Western Europe by R1b1a2.

The progress we have made towards resolving the L11 phylogeny has been significant considering that none of the authors of this manuscript have ever met, nor spoken on the phone. Open source data, open source tools, and open forums enable research collaborations to blossom. In fact, other citizen scientists are currently using similar approaches to find groundbreaking SNPs in other branches of the Y chromosome haplogroup tree.

## Materials and Methods

The main mode of communication was through the DNA-Forums website (DNA-Forums, A Genetic Genealogy Community, <http://dna-forums.org>). It was here that general information on data mining techniques and the discovery of new variants was shared and discussed amongst team members. The ISOGG phylogenetic tree was used as a starting point for phylogenetic placement as it was found to be the most current ([http://www.isogg.org/tree/ISOGG\\_HapgrpR.html](http://www.isogg.org/tree/ISOGG_HapgrpR.html)).

SAMtools [27], and related utilities were used to query low coverage (2–4x) datasets from publicly accessible FTP sites at the European Bioinformatics Institute (<ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/>) and the National Center for Biotechnology Information (<ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/>). Using SAMtools tview, samples were screened for the L11 mutation. BAM index files were then called for each L11 derived sample along with a control group of non-R1b1a2 samples using SAMtools mpileup. The resulting BCF file was then converted into VCF format by using the bcftools vcfutils.pl script. The resulting VCF file was opened in MS Excel for visual identification of potential SNPs downstream of L11.

Novel variants were filtered by verifying that all non-L11 control samples bore the ancestral allele, and by identifying at least

two L11 samples that carried the same derived allele. While hundreds of singletons were found, they were not catalogued. Variants which had heterozygous allele calls were disregarded, as were those with phylogenetic inconsistencies that more than likely arose in duplicated or recombining regions of the Y-chromosome. Once a variant fitting these filtering criteria was found, it was further verified by making individual calls to the respective chromosome Y position using the SAMtools tview. These positions were also cross-referenced against the Family Tree DNA Y Chromosome Browser (<http://ymap.ftdna.com>) to determine whether the variants were novel.

A master list of new variants was kept in a centralized spreadsheet stored in Google Docs (<https://docs.google.com>), a freely available cloud storage service. This made uploading, storing and organising the data easier to manage and readily available to any citizen scientist with access to the Internet.

Primers (see Table S1) were designed for a number of variants of genealogical or phylogenetic importance, using Primer3 [28], or the National Center for Biotechnology Information online Primer-BLAST tool (<http://www.ncbi.nlm.nih.gov/tools/primer-blast/>). Variants were validated using PCR and Sanger sequencing.

Several other publicly available genome data sets were used for the identification of new variants. These included Jay Flatley's genome (<http://aws.amazon.com/datasets/3357>), the first Irish genome [29] and the genomes of Henry Louis Gates Jr and Sr (<http://snp.med.harvard.edu>).

## Supporting Information

**Table S1 Details of all genetic variants identified using 1000 Genomes data details of PCR primers used to validate variants using Sanger Sequencing.** (XLSX)

## Acknowledgments

We thank the 1000 Genomes Project, the European Bioinformatics Institute, and the National Center for Biotechnology Information for making their sample data publicly accessible. We would like to thank Bennett Greenspan along with Astrid-Maria Krahn, Lora Engelhardt, Brent Maning, Arjan Bormans, Ana Bestic and Jennifer Valentine of Family Tree DNA for the rapid development of many of the primers. We would like to thank all other members of the DNA-Forums online community who contributed to this effort as well as George van der Merwe for hosting and administering DNA-Forums. We would like to thank Family Tree DNA project administrators for recruiting end-customers to test for these new markers. Lastly, we wish to give special thanks to the Family Tree DNA customers that tested for these new markers.

## Author Contributions

Conceived and designed the experiments: RAR GM DFR TK VOT PMOdVB AJG. Performed the experiments: RAR GM DFR TK VOT PMOdVB AJG. Analyzed the data: RAR GM DFR TK VOT PMOdVB AJG. Contributed reagents/materials/analysis tools: RAR GM DFR TK VOT PMOdVB AJG. Wrote the paper: RAR GM DFR TK VOT PMOdVB AJG.

## References

1. Russo E (2001) Population Genetics. *Nature* 413: 4–5.
2. Pollack A (2011) A Genome Deluge. *New York Times*. New York. B1.
3. Donnelly P (2011) Genome-sequencing anniversary. Making sense of the data. *Science* 331: 1024–1025.

4. The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467: 1061–1073.
5. Balaresque P, Bowden GR, Adams SM, Leung HY, King TE, et al. (2010) A predominantly neolithic origin for European paternal lineages. *PLoS biology* 8: e1000285.
6. Myres NM, Rootsi S, Lin AA, Jarve M, King RJ, et al. (2011) A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. *European journal of human genetics : EJHG* 19: 95–101.
7. Busby GB, Brisighelli F, Sanchez-Diz P, Ramos-Luis E, Martinez-Cadenas C, et al. (2012) The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. *Proceedings Biological sciences/The Royal Society* 279: 884–892.
8. Lappalainen T, Koivumaki S, Salmela E, Huoponen K, Sistonen P, et al. (2006) Regional differences among the Finns: a Y-chromosomal perspective. *Gene* 376: 207–215.
9. He M, Gitschier J, Zerjal T, de Knijff P, Tyler-Smith C, et al. (2009) Geographical affinities of the HapMap samples. *PLoS One* 4: e4684.
10. Cruciani F, Trombetta B, Antonelli C, Pascone R, Valesini G, et al. (2011) Strong intra- and inter-continental differentiation revealed by Y chromosome SNPs M269, U106 and U152. *Forensic science international Genetics* 5: e49–52.
11. Larmuseau MH, Vanderheyden N, Jacobs M, Coomans M, Larno L, et al. (2011) Micro-geographic distribution of Y-chromosomal variation in the central-western European region Brabant. *Forensic science international Genetics* 5: 95–99.
12. The Y Chromosome Consortium (2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome research* 12: 339–348.
13. Zhivotovsky LA, Underhill PA, Feldman MW (2006) Difference between evolutionarily effective and germ line mutation rate due to stochastically varying haplogroup size. *Molecular biology and evolution* 23: 2268–2270.
14. Hurles ME, Veitia R, Arroyo E, Armenteros M, Bertranpetit J, et al. (1999) Recent male-mediated gene flow over a linguistic barrier in Iberia, suggested by analysis of a Y-chromosomal DNA polymorphism. *Am J Hum Genet* 65: 1437–1448.
15. Lopez-Parra AM, Gusmao L, Tavares L, Baeza C, Amorim A, et al. (2009) In search of the pre- and post-neolithic genetic substrates in Iberia: evidence from Y-chromosome in Pyrenean populations. *Annals of human genetics* 73: 42–53.
16. Adams SM, Bosch E, Balaresque PL, Ballereau SJ, Lee AC, et al. (2008) The genetic legacy of religious diversity and intolerance: paternal lineages of Christians, Jews, and Muslims in the Iberian Peninsula. *Am J Hum Genet* 83: 725–736.
17. Underhill PA, Shen P, Lin AA, Jin L, Passarino G, et al. (2000) Y chromosome sequence variation and the history of human populations. *Nat Genet* 26: 358–361.
18. Alonso S, Flores C, Cabrera V, Alonso A, Martin P, et al. (2005) The place of the Basques in the European Y-chromosome diversity landscape. *European journal of human genetics : EJHG* 13: 1293–1302.
19. Bryc K, Velez C, Karafet T, Moreno-Estrada A, Reynolds A, et al. (2010) Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proceedings of the National Academy of Sciences of the United States of America* 107 Suppl 2: 8954–8961.
20. Mallory JP, Adams DQ (1997) *Encyclopedia of Indo-European Culture*. London and Chicago: Fitzroy-Dearborn.
21. Koch JT (2006) *Celtic Culture: A Historical Encyclopedia*. Anta Barbara: ABC-CLIO.
22. Koch JT (2009) A case for Tartessian as a Celtic language. *Palaeohispanica* 9: 339–351.
23. Lee EJ, Makarewicz C, Renneberg R, Harder M, Krause-Kyora B, et al. (2012) Emerging genetic patterns of the European Neolithic: Perspectives from a late Neolithic bell beaker burial site in Germany. *American journal of physical anthropology*.
24. Defrancesco L (2012) Life Technologies promises \$1,000 genome. *Nature biotechnology* 30: 126.
25. Lacan M, Keyser C, Ricaut FX, Brucato N, Tarras J, et al. (2011) Ancient DNA suggests the leading role played by men in the Neolithic dissemination. *Proceedings of the National Academy of Sciences of the United States of America* 108: 18255–18259.
26. Lacan M, Keyser C, Ricaut FX, Brucato N, Duranthon F, et al. (2011) Ancient DNA reveals male diffusion through the Neolithic Mediterranean route. *Proceedings of the National Academy of Sciences of the United States of America* 108: 9788–9791.
27. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079.
28. Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S, editors. *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Totowa: Humana Press. 365–386. Available: <http://frodo.wi.mit.edu/>, (Accessed 2012 Jan 3).
29. Tong P, Prendergast JG, Lohan AJ, Farrington SM, Cronin S, et al. (2010) Sequencing and analysis of an Irish human genome. *Genome Biology* 11: R91.