Journal: www.jogg.info
Originally Published: Volume 5, Number 2 (Fall 2009)
Reference Number: 52.010

CLUSTER ANALYSIS AND THE TMRCA PROBLEM: Y-STR MOUNTAINS IN HAPLOSPACE, PART II: APPLICATION TO COMMON POLISH CLADES

Author(s): Peter Gwozdz

Journal of Genetic Genealogy, 5(2):159-185, 2009

Y-STR Mountains in Haplospace, Part II: Application to Common Polish Clades

Peter Gwozdz

Abstract

A number of Y-STR haplotypes are common in Poland and rare elsewhere. The most common of these is called "P type" in this article. This article presents evidence that P type, defined using 36 of 67 STR markers, corresponds to a clade within Haplogroup R1a1, using the "mountain in haplospace" method introduced by the companion article. The "Statistical Background Percent" (SBP), is defined in the companion article. The result for the P type data cluster is SBP = 6.7%. It seems that about 8% of Polish men carry this P type Y-DNA. Estimated time to most recent common ancestor is 2,000 to 3,000 years ago, with significant population expansion less than 1,500 years ago. Several other types are discussed.

Introduction

The first article in this two-part series developed a set of tools for the analysis of Y-STR data from groups of related people (Gwozdz, 2009). The present article applies these tools to Polish Y-STR data. A number of specially defined terms are used in this article and they are defined in the companion article (Part I).

Pawlowski (2002) identified a Y-STR haplotype that is 15 times more common in Poland than in the rest of Europe. That haplotype, determined at nine STR markers, represents 4.6% of Pawlowski's Polish data. This haplotype, called "P type" here, is evidence of a dominant Y-DNA clade in a relatively recent Polish population expansion. One purpose of this article is to add evidence, and to estimate the size and age of the hypothesized P type clade.

Roewer (2005) analyzed European Y-STR data, showing it can supplement single nucleotide polymorphism (SNP) markers in the study of population structure within historical times. Pawlowski's data source is a subset of Roewer's source, Yhrd, a European forensic compilation. Another purpose of this article is to identify common Polish STR haplotypes, as modal haplotype candidates for hypothetical clades, with estimated size and age. Other than brief comments in discussion, comparison to the historical record is beyond the scope of this article.

I use the conventional notation "Haplogroup R1a1" for generally accepted SNP haplogroup clades, and the notation "P type" for types that are introduced here, and "Pc type" for subtypes. A "type" is as defined in Part I of this two-part series of articles. Most of the types are presented as hypothetical subdivisions of Haplogroup R1a1 (R-M17 or R-M198), so the stem haplogroup is R1a1 if not specified. In this article I retain the R1a1 designation for Haplogroup R-M17 in order to be compatible with the reference data. It is the same haplogroup currently designated as R1a1a (ISOGG, 2009). The conventional asterisk (*) as used in the construction "R1a*" indicates those downstream subsets of R1a that are not in any presently defined sub-haplogroups of R1a. "R1a" without an asterisk means all of R1a. For example, R1a includes R1a1; R1a* does not include R1a1.

About half of Polish Y-DNA is in Haplogroup R1a (Wiik, 2008; McDonald, 2005).

"Polish" is difficult to define exactly. Biskupski (2006) published a delightful essay regarding the ambiguity of the word "Polish." In this article "Polish" simply refers to people either born in historical Poland or who have patrilineal lines that go back to historical Poland.

The **Summary** on page 156 of the companion article provides a concise two-page statement of the method of this article, with terms defined in boldface.

Address for correspondence: Peter Gwozdz, pete2g2@comcast.net

Received: June 5, 2009; accepted: September 13, 2009

Materials and Methods

The Y-STR data for this article is available from three Internet sources: Ysearch, Yhrd (Willuweit, 2007), and Family Tree DNA (FTDNA). FTDNA has a number of "projects" that focus on surnames, haplogroups, and geographic areas. The primary source for this article is the Polish Project. See Web Resources at the end of this article for URLs. There is incidental use of data from other projects, which data is available by substitution of the quoted project name for /polish/ in the public address for the Polish Project.

Polish Project data follows the FTDNA standards and marker order, with progressive STR marker sets available, accumulating 12, 25, 37, and 67 markers, plus a few additional rare compound markers, and minus rare missing markers. Data from all available STR markers are used here. The Polish Project has at least the standard 12 markers for all samples, so those 12 markers provide the best statistics, although at lowest resolution.

The Ysearch database consists primarily of Y-STR haplotype data from individual men (herein, "samples"), but also includes some modal haplotypes, fictitious entries for research purposes. The latter records were removed from the data before analysis.

FTDNA projects include family samples. These are edited, as explained in the companion article. Family sets were identified by noting identical family names with identical 67-marker haplotypes, or similar haplotypes with sequential ID codes (or nearly sequential). All but one of each family set were removed, leaving a representative with the most markers. For the Polish Project, where family sets are uncommon, only 13 of 913 (1.4%) data entries were removed. Only two of the R1a1 haplotypes were removed. Those 13 removals are from eight family pairs and one family set of six samples, of which five were removed. That set of six, which had identical haplotypes at 12 markers (four were identical at 25 markers), is an example of how large unedited family sets can distort haplotype frequency data. Additional sets of very similar (zero genetic distance steps at 25 markers, or less than two steps at 37, or less than four steps at 67) haplotypes without identical names, 61 samples (6.8%), were not removed. Insofar as this editing may not remove some family sets without family names, the data are biased slightly toward larger clusters. Insofar as this editing may remove men who actually submitted data independently, data is biased slightly, toward smaller clusters.

Ysearch also has family sets, but Ysearch ID codes are not assigned in sequence, so editing was restricted to identical family names with 67-marker haplotypes at

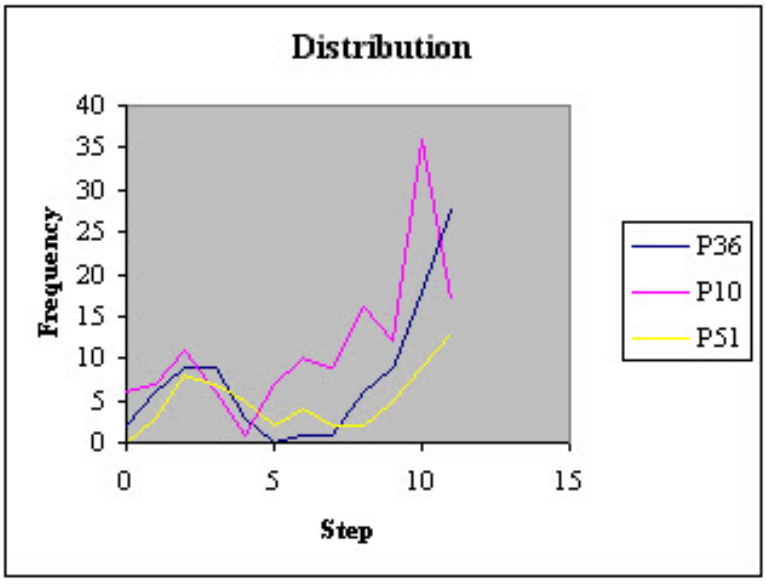

Figure 1. The P type mountain. The definition P36 uses a modal haplotype with 36 markers of the 67 available, for a mountain at cutoff 5 with 29 samples, steps 0 to 4. The gap is steps 5 to 7; 2 samples. Exactly the same 29 samples are captured using a modal haplotype with 13 to 41 markers. Using 51 markers, shown here as P51, also captures the same 29 samples at cutoff 7. Also shown here, P10 (best 10 of 67), captures the same 29 samples plus one more at cutoff 4.

zero or one step separation. Also, Ysearch results were sorted by family name and visually scanned for suspicious sets of samples, for which the detailed data were checked by ID code for "contact person," where the same contact person is assumed to be the collector of a family set; only one such set was identified.

See the companion article, "Y-STR Mountains in Haplospace, Part I: Methods," for the "mountain" method. The word "type" means "mountain" in these articles. Briefly, a "type" is defined using an STR cluster for which a step frequency graph looks like a mountain. Figure 1 is an example. A "signature" is the modal haplotype using markers that best distinguish from other types in the parent haplogroup. A "definition" includes the modal haplotype and "cutoff" that produces the best mountain. The cutoff is the next step, just beyond the mountain. The "background" of samples that may not belong to the type for statistical reasons is estimated based on the samples in the "gap" of steps with low sample frequency just beyond the mountain. The statistical background percent (SBP), a recommended measure of type quality, is the percent of samples in a mountain that may not belong to the type for statistical reasons, at 70% confidence maximum, including objective adjustment for systematic statistical errors. A summary of the mountain method is at the end of the companion article.

A small SBP means an isolated type, taken as evidence, albeit not proof, that the type represents a clade.

Concentration of a type in Poland is taken as evidence, albeit not proof, that the type represents a clade.

The Polish Project is used as representative of Poland. Yhrd is used to verify and update Pawlowski's most common haplotype, here called P type. The Ysearch P type fraction of entries with "Poland" as origin is used as a calibration of the Polish fraction on Ysearch; other types with similar Polish fraction in Ysearch are taken to be similarly Polish.

Age is estimated using Averaged Squared Distance (ASD) in STR data, as explained in the companion article. The time of the most recent common ancestor (TMRCA) is calculated from ASD, although this may be misleading for a number of reasons, some of which are discussed in the companion article. Age is calculated (by various methods) in the *Excel* files, one for each type, in the Supplementary Data. As one simple method, I use Thomas (1998) as a paradigm, with ASD = 0.2226. Age is proportional to ASD. The corresponding Thomas age of 2,650 years may be highly uncertain when applied to other hypothetical clades, so such a simple calibration to Thomas provides little better than a rough estimate of age, calibrated to the published literature. As another method, I use the Chandler (2006) mutation rates to calculate age from ASD. As explained in the companion article, I distinguish two kinds of age: TMRCA vs. time of hypothetical population expansion.

Statistical methods are described in the companion article.

Polish Project At 12 Markers–Ysearch Comparison

The FTDNA Polish Project data were downloaded 15 May 2009. An archive copy of the data is available as "PolishProject15May2009.xls" in the Supplementary Data. The total of 913 samples was edited to 900 as explained above in Methods and Materials. Editing and matching notes are in "Evaluator15May2009.xls" and again in "Results15May2009.xls." All Polish Project samples have values for at least the first 12 markers. Table 1 gives the ten most common 12-marker haplotypes, which are all the haplotypes with more than six samples. Also in Table 1 are four of the eleven haplotypes with six or five samples (discussed below). The file "Polish12.xls" has all 614 haplotypes at 12 markers in a format similar to Table 1.

P type is the most common haplotype at 12 markers in Poland. The 12-marker haplotypes in Table 1 are an introduction; P, K, N, Y, etc. types are defined using the best ranked of all 67 markers with more discussion below.

My "Code Type" labels of Tables 1 and 2 are assigned with hypothetical foresight, anticipating the 67-marker results presented later in this article. A common haplotype, even at only 12 markers, provides a candidate modal haplotype for a type, as explained in the companion article and as demonstrated below. ("Type" means: hypothetical clade / modal haplotype / set of possible haplotypes / associated cluster of data). Most of the haplotypes in Table 1 are such candidates, but the statistical evidence at 12 markers is insufficient by itself to confidently conclude that any of the Table 1 entries is a modal haplotype for a type. Even at 67 markers some of them are highly hypothetical, as discussed below. The fourth most common haplotype in Table 1, eleven samples, which I call PbKa, could also be a candidate type, but I have found no evidence for it. PbKa is one step from P type and one step from K type, so if all three correspond to types, then the 12-marker PbKa data would be a mix of samples from all three, because of mutations.

On 11 July 2009 the top 48 most common Polish haplotypes at 12 markers (those with more than two samples from the Polish Project) were checked on Ysearch for perfect matches at 12 markers. Modal entries were subtracted for the match number, as mentioned above in *Materials*. Family sets were not edited for this 12-marker comparison. The number and percent of Ysearch samples with "Poland" in the male line "Origin" field are all listed in the file "Polish12.xls." Table 1 summarizes this data

Journal of Genetic Genealogy, 5(2):159-185, 2009

Table 1

Common Haplotypes From the Polish Project, Using Only the First 12 FTDNA Markers. The second column has code names for discussion convenience in this article. The two "Polish Pr" columns provide the number of samples with that haplotype, and the percent of the total 900 in the Polish Project. The "Ysearch" columns give the number of samples with that haplotype (matching those 12 markers), and the number and percent of those that give "Poland" as the male-line origin.

Haplo-group	Code Type	Y-STR Haplotype Values, by DYS Location												Polish Pr		Ysearch		
		393	390	19	391	385a	385b	426	388	439	389-1	392	389-2	No.	%	No.	Pol.	%
R1a1	P	13	25	**17**	10	**10**	14	12	12	10	13	11	30	22	2.40	64	27	**42**
R1a1	K	13	25	16	10	11	14	12	12	10	13	11	30	14	1.60	146	18	12
R1a1	Pa	13	25	**17**	10	**10**	14	12	12	10	13	11	**31**	13	1.40	11	6	**55**
R1a1	PbKa	13	25	16	10	**10**	14	12	12	10	13	11	30	11	1.20	19	6	**32**
N1c1	Z	**14**	**23**	**15**	**11**	11	14	11	12	10	13	**14**	29	9	1.00	25	3	12
R1a1	N	13	25	16	10	11	14	12	12	**11**	13	11	**29**	8	0.90	84	12	14
R1a1	KbNd	13	25	16	10	11	14	12	12	**11**	13	11	30	8	0.90	30	5	17
I1	Y	13	**22**	**14**	11	**14**	14	**11**	**14**	**11**	**12**	**12**	**28**	8	0.90	18	9	**50**
R1a1	L	13	25	**15**	**11**	11	14	12	12	10	13	11	30	7	0.80	116	6	5
N	X	**14**	**23**	**15**	**11**	11	**13**	**11**	12	10	14	**14**	30	7	0.80	3	0	0
R1a1	Pc	13	25	**17**	10	**10**	14	12	12	**11**	13	11	30	6	0.70	14	3	21
R1a1	G	13	25	**15**	**11**	**12**	14	12	12	10	13	11	30	5	0.60	16	7	**44**
R1a1	Kd	13	25	16	10	11	**15**	12	12	10	13	11	30	5	0.60	10	3	**30**
I2a	W	13	**24**	16	**11**	**14**	**15**	**11**	**13**	**13**	13	11	32	5	0.60	12	3	25

Notes: In Tables 1 and 2, for R1a1, markers that differ from K are shown boldface.
Types are color coded to facilitate comparisons (see text).

Table 2

Common Haplotypes from Pawlowski (2002). The 2nd column are codes per Table 1. Next column is Pawlowski's code number. The 2 pairs of columns on the right give the number and percent for each haplotype, for Pawlowski and for the Polish Project. The Pawlowski data set is 508 samples at 9 markers. The Polish Project 900 samples were here sorted and counted using only the same 9 markers.

Haplo-group	Code Type	Code Pawl	Y-STR Haplotype Values, by DYS Location												Pawlow-ski		Polish Project	
			393	390	19	391	385a	385b	426	388	439	389-1	392	389-2	No.	%	No.	%
R1a1	P	303	13	25	**17**	10	**10**	14				13	11	30	22	2.4	28	3.1
R1a1	N	229	13	25	16	10	11	14				13	11	**29**	17	3.3	14	1.6
R1a1	K	236	13	25	16	10	11	14				13	11	30	13	2.6	26	2.9
R1a1	PbKa	234	13	25	16	10	**10**	14				13	11	30	10	2.0	12	1.3
I1	Ta	23	13	**22**	**14**	10	**13**	14				**12**	11	**28**	8	1.6	6	0.7
R1a1	L	161	13	25	**15**	**11**	11	14				13	11	30	7	1.4	8	0.9
R1a1	Ne	192	13	**24**	**16**	10	11	14				13	11	**29**	7	1.4	6	0.7
R1a1	Kd	153	13	25	**15**	10	11	14				13	11	30	6	1.2	5	0.6
I2	Wa	212	13	**24**	16	**11**	**14**	**15**				13	11	**31**	5	1.2	11	1.2

for the top haplotypes. The Supplementary Data also has a representative Ysearch 12-marker code for each haplotype, so the reader can quickly update the data on Ysearch.

Although there are a number of reasons why the percent of "Poland" samples at Ysearch may be difficult to interpret as a percent Polish, there is value to compare haplotypes, since the difficulties should be proportional for all haplotypes. Strong evidence is presented below that P type is common in Poland and rare elsewhere. On this basis, the 42% of P type samples on Ysearch with Poland as the paternal origin in Table 1 provides a measure of how well Ysearch evidence for high concentration in Poland can be extended to other haplotypes. In Table 1 boldface is used to indicate a haplotype that has 30% or greater of the samples of Polish origin on Ysearch, as an indication of concentration in Poland. Nine of the top 48 haplotypes passed the 30% test. Table 1 has the top six. Table 1 includes four of the twelve haplotypes with six or five samples, those with more than 20% "Polish Origin " at Ysearch. Sample sizes as small as six or five are, of course, suspected of being just statistical fluctuation.

K type is the most common R1a1 haplotype on Ysearch at 12 markers. K is shown in green in Tables 1 and 2. For other R1a1 haplotypes, markers that differ from K are shown boldface. The most common country for K type on Ysearch is Germany, with only 13%. Scotland, Ireland and England together are only 10%. Ysearch surely has more data for western Europe than for Eastern Europe. From examination of Ysearch results, it appears K type is widespread, concentrated in eastern Europe. There is discussion below that the 12-marker version of K could be a combination of types that may be distantly related.

It is convenient that P and N differ by two markers out of 12 from K, and four markers from each other. P, K, and N show three obvious "mountain" effects even at only 12 markers. Each has fewer samples at every "nearest neighbor" haplotype one-step away out of the 12. Every 12-marker haplotype necessarily has 24 theoretically possible nearest neighbor haplotypes. For example, P type has 38 total samples that differ by one step from the P type modal haplotype, and these occur in only eight of those 24 possible nearest neighbor haplotypes; 12 of the 38 are actually shared, one step from P but also one step from K, for example PbKa with eleven samples. Haplotypes Pa and Pc, two of those 8, have respectively 13 and six of the 38 samples. This is evidence (not proof) that Pa and Pc may be subtypes of P, discussed further below.

Over the past two years, as data accumulated, the general appearance of the data corresponding to Table 1 has been stable, with P always the most common, but details have statistically varied. For example, the third most common, which I call Pa, differs from P by only 1-step. Pa has 13 samples, so the 70% confidence is 9 to 18. Actual rank in Poland may be second to sixth, or even wider. In fact, back in May 2008, Pa was running sixth, and Pc was seventh (now one of five that are tied for eleventh). This is typical when small numbers are involved.

If corresponding P, N, and K clades in fact exist, there is overlap at only 12 markers, with a significant probability of mutation from one haplotype to another. These common haplotypes are only a strong hint of a clade or clades that have corresponding modal haplotypes at 12 markers.

The second most common R1a1 haplotype on Ysearch, L type in Table 1, is more common in western Europe than K type. L has 116 samples on Ysearch but only 5% are Polish on Ysearch, while G type, which differs from L by only one marker out of 12, scores 44% Polish on Ysearch, although this result is highly uncertain because of the small sample.

The most common European haplotype (called "NIST most common" type in the supplementary data file because I originally found it at the National Institute of Standards and Technology site), in haplogroup R1b1, had 1523 samples at 12 markers in Ysearch in January 2008, before Ysearch instituted the time-out failure for long searches. This haplotype has only one in the Polish Project. It is rare in Poland. There are 492 such singleton haplotypes in the Polish Project.

The top 48 haplotypes (more than two samples) at 12 markers had (15 May 2009) 260 samples in the Polish Project. These same haplotypes in Ysearch had (11 July 2009) 232 samples with "Poland" in the "Origin" field (232 / 4255 total = 5.5%) for most distant ancestor. So Ysearch provides another Polish database of about equal size, although it is not fully independent because many men register data at both places. It is very tedious (using on-line data) to identify which samples are from men who registered at both places.

For analysis of common Polish haplotypes, the Polish Project is preferred to Ysearch because the entire Polish Project database is available for immediate conversion to an Excel file, and because in Ysearch the Polish clusters may not be as obvious in a search due to swamping by non-Polish data. Also, "Polish" samples are not well defined in Ysearch; some men use locations (for example "Galicia, Austrian Empire") that are difficult to sort and identify as Polish; many samples have "unknown" origin; some men may prefer not to use the word "Poland" in their data entry for personal reasons. The Polish Project in contrast consists of self identified Polish samples. Nevertheless, as shown in the results below, Ysearch results are close to the Polish Project

results. The Ysearch results are simple to verify and update, because the file "YsearchURLs.xls" in the **Supplementary Data** provides Universal Resource Locator codes for the modal haplotypes discussed in this article.

Discussion; Polish Project is Representative of Poland

"Historical Poland" is a broad definition of Poland and "ethnic Poland" is narrow. Although a definition of "ethnic" would be contentious, for purposes of a Polish Y-STR study, it would be ideal to restrict data to samples from men who have reasonable knowledge of their family genealogy and whose most distant known male line ancestor spoke Polish as a family language and belonged to a family that seemed to be a member of Polish culture. Even Y-DNA data randomly collected in modern Poland is less than ideal in this respect.

The results of the analysis of the Polish Project are presented here as representative of Poland. The Polish Project data is mostly from self selected Polish diaspora. We assume that individuals with an interest in genealogy are more likely to join DNA projects. We assume that an individual is less likely to join the Polish Project Y-DNA section if that individual knows that his most distant known male line ancestor from Poland was actually born elsewhere or was born in Poland but would not be considered ethnic Polish. In that respect, the Polish Project might well be as representative of the Polish ethnic population as modern random data from Poland itself.

However, the Polish Project is relatively open, and welcomes all men from historical Poland, and also welcomes men who do not consider themselves patrilineal ethnic Polish but are searching for Polish Y-DNA connections. So the Polish Project data, although mostly ethnic Polish by any definition, is statistically diluted by non-ethnic-Polish data.

Mayka (2009) provided estimated classification of the R1a section of the Polish Project, based on web data available to the project administrator, along with consideration of surname and given name: of the 433 R1a samples on that day, 300 (69%) seem to be ethnic male line Polish, 27 seem to be German/Prussian from modern Poland, 53 are from historical Poland outside the modern Polish border, 20 seem to be Jewish, 32 are from countries bordering historical Poland, and one is from far away.

Considering the 31% of samples that do not seem to be obviously ethnic Polish, it is not known what fraction of these have knowledge that leads them to believe or to suspect a male line Polish ancestry. It is not known what fraction of the 69% do not consider their male line to be Polish. Acknowledging that "male line ethnic Polish" cannot be defined without contention, it seems to me the R1a1 section of the Polish Project is about 80% male line ethnic Polish, with an estimated confidence interval of 70% to 90%.

Polish R1a are R1a1

About half of eastern European Y-DNA, including Polish, is R1a. Wiik (2008) provides contour plots of R1a, Maps 21 and 22. McDonald's (2005) Figure 2 shows slightly more than 50% R1a for Poland. Sliwinski's (2007) Figure 3 reports the Polish Project R1a fraction as slightly less than 50%. The 15 May 2009 data for the Polish Project has 378 R1a/R1a1 of the 900, for 42.0% (70% confidence interval 39.6% to 44.3%).

All the samples corresponding to the haplogroup labeled R1a1 in **Table 1** are categorized in the Polish Project database as R1a1 "tested" (33%), or R1a "tested" (8%), or R1a "predicted" (59%). "Tested" means assigned to the haplogroup based on the appropriate SNP test. Most, if not all, of the "R1a tested" were not tested for the downstream R1a1 marker. "Predicted" is an FTDNA proprietary method based on STR markers. It is understood that a possible very small fraction of the R1a1 sample counts in **Table 1** may not be R1a1 or indeed may not be R1a. There are a few known downstream SNP markers for R1a1 and commercial tests are available for sub-haplogroups R1a1a through R1a1e, but no one in the Polish project is assigned to one of these sub-haplogroups. Mayka (2008) pointed out that all Polish project R1a tests have been coming out positive for R1a1, and negative for the downstream markers. (Mayka, the administrator, has access to test results by SNP.) On this basis, those common types P, K, N, and G almost certainly correspond to R1a1. For the Polish Project, R1a, R1a1, and R1a1* are essentially equivalent.

Pawlowski Analysis And Verification

Pawlowski (2002) found the most common haplotype, at nine markers, to be 4.6% of his database of 995 Y-STR forensic samples from north Poland, and 15 times more common in Poland than in surrounding countries, 7,752 samples. This result is summarized in Pawlowski's Table III. That haplotype is what I am here calling P type, at nine markers. By email communication in Sep 2007, Richard Pawlowski provided an unpublished version of the database with more samples using up to 19 markers; I analyzed this unpublished version and verified that the results are consistent with both his previous results and my results reported here.

Table II in Pawlowski (2002) lists the number of samples for all 328 haplotypes for a previous version of the database with 508 samples. I list the top nine haplotypes from Pawlowski here in **Table 2**. These are all the haplotypes with more than five samples. Column 2 in **Table 2** has Pawlowki's code number from his Table II.

Column 1 in Table 2 has the type codes used here. The types not associated with R1a1 have the haplogroup hyphenated before the code, although Pawlowski did not determine haplogroup.

The Polish project data were sorted on the basis of those nine markers (900 samples downloaded 15 May 2009, described above). The resulting frequencies by haplotype are displayed in Table 2 along with Pawlowski's results. The file "Polish9.xls" in the Supplementary Data file has details for all 530 haplotypes.

Pawlowski's data is exclusively from northern Poland, near Gdansk. That may explain some of the discrepancies in the frequencies of the common haplotypes; some haplotypes may be more or less common in Gdansk. However, the discrepancies are probably mostly statistical. The 95% confidence interval for six samples is 2.2 to 13, so the bottom two in Table 2, 8th and 9th rank, may be as wide as 3.3 to 19.3 in the Polish Project, which is rank 3rd to 33rd at 95% confidence.

Pawlowski's P type is 4.3%. The one-sigma (70%) confidence interval is 3.4% to 5.2%. The 3.1% P type for the Polish Project at nine markers is just below the one sigma confidence, so the difference may well be statistical.

Reminder: my "Code Type" labels of Table 2 are assigned with hypothetical foresight, as explained for Table 1 above.

The N haplotype (and perhaps the corresponding N type) may be more common near Gdansk.

X type, associated with Haplogroup N, is the 8th most common in the Polish Project at 12 markers. At nine markers there are nine of these. But there is only one of these at nine markers in Pawlowski's Table II; 4 to 14 are expected at 95% confidence in Pawlowski based on the Polish Project, so X seems to be rarer in Gdansk than elsewhere in Poland.

P type shows a remarkable mountain effect even in Pawlowski's 2002 data at nine markers: P type has 22 samples at the modal haplotype. The 18 possible haplotypes at 1-step have 41 samples (36 if the two "between" haplotypes like PbKa are assigned proportionally between P and K). The average is only about two per haplotype. The average random background should be one or less per haplotype at nine markers. This implies that if P type represents one or more clades, the clade or clades are relatively young and total about double the size implied by the modal count of 22. The statistics for only nine markers do not justify a confident conclusion, but Pawlowski (2002) provided an inspiration for the research that led to the present article.

Yhrd Polish Haplotypes

A Yhrd study was done using the database Release 23, dated 15 Jan 2008, with 54,863 haplotype samples in 477 populations. The current (2 May 2009) Yhrd database, Release 27, does not have significantly more data for Poland.

Most of the Yhrd populations are named after cities. There are 13 populations from cities in Poland, with Bialystok divided into five ethnic groups, plus "Southeast Poland" for 18 total Polish populations. All the Polish data have at least the same nine markers as those used by Pawlowski.

Analysis is here at the nine markers. The count analysis by city for Poland and neighboring countries is available in the Supplementary Data.

P type is the most common haplotype in all Polish cities with significant data. (An insignificant example: Zakopane has only seven samples - with one P in a 7-way tie for first place). For all 18 Polish populations, P type has 142 samples out of 3,555 = 4.0%.

In 16 "Border" cities closest to Poland, P type has 50 samples out of 3,997 = 1.3%. In 22 "Neighbor" cities closest to Poland but outside the Border cities, only four cities have 1.4% to 1.7% P type. P type is clearly more common in Poland than in the surrounding countries, according to the Yhrd database.

For the rest of Europe, subtracting the numbers for Poland, P type is 101 out of 23,851 for 0.43%. The ratio of P type in Poland to rest of Europe is 9.5, comparable to Pawlowski's result of 15, considering the same cities are not being compared.

Kayser (2005) found significant difference in Polish vs German populations along the border, using Y-DNA with seven STR markers and ten binary SNP markers, in large part because the Polish - German border marks the western extent of the dominance of the R1a1 haplotype.

However, the "Border" and "Neighbor" cities in this Yhrd analysis includes eastern cities, for example: Vilnius Lithuania (1.3% P type), Hrodna Belarus (1.8% P), Kiev Ukraine (1.5% P). Some cities have high P type without statistical significance (e.g. Thessaly Greece, with one out of 15 samples = 6.7%).

The highest P type non Polish city with reasonable data is Lviv Ukraine (4 out of 105 = 3.8% P). The Ukraine total P type is about half the Polish percentage (8 out of 368 = 2.2% P). This is a very preliminary hint of a genetic link, based on only eight samples at nine markers.

Table 3
Signature Modal Haplotypes for Common Polish Y- STR Types in Haplogroup R1a1. All markers are listed for K type, which is the modal haplotype for Polish R1a1. Blank values are the same as K. Boldface markers are the best signature markers for distinguishing the type cluster from other R1a1 samples in the Polish Project, except Pg, where boldface are best for distinguishing from the parent P type. The Supplementary Data has all 67 marker values for all these types, as well as the best "definition" marker sets.

R1a1	Signature, by DYS Location																				
Type	**19**	**385a**	**439**	**389-2**	**458**	**459b**	**447**	**449**	**464b**	**c**	**e**	**456**	**442**	**578**	**537**	**406**	**413a**	**534**	**481**	**446**	**572**	**565**
P	17	**10**					**23**	31		16			13			**12**			25		**12**	
Pc	17	**10**	11		**16**		23			16						12		**14**	25		**12**	**12**
Pg	17	10	**10**		**16**		23	**31**		16			13			12			**25**		11	**13**
N			**11**	**29**			23		13			17			**11**	12	21		25	**13**		
Na			**11**	**29**			23		**12**		**15**	**17**	13		**11**	12	21		25	**13**		
Nb			**11**	**29**			23		**14**		**16**	**18**	13		**11**	12	21		25	**13**		
Nc	**15**		**11**	**29**			23					17			**11**	12	21		25	**13**		
K	16	**11**	10	30	16	10	**24**	32	15	15	x	16	14	8	**12**	11	22	13	23	**12**	11	13
A					**14**	**11**		**30**	**12**			**14**						14			**10**	
I	17				**14**							**15**		9								

Table 4
Signature Modal Haplotype Examples for Y-STR Types in Other Haplogroups. Y Type is common in Poland according to data in Ysearch, while Z type is not so common. Boldface markers are the best signature markers for distinguishing the type cluster from the Polish Project subset database for the parent haplogroup.

Haplo-group		Signature, by DYS Location																		
	Type	**19**	**391**	**385a**	**385b**	**439**	**389-1**	**392**	**389-2**	**458**	**447**	**449**	**460**	**456**	**442**	**406**	**534**	**444**	**481**	**446**
I1	Y	14	**11**	**14**	14	11	12	**12**	28	15	**24**	**28**			12	9	17		25	14
N1c1	Z	15	11	11	**14**	10	**13**	14	**29**				11	**14**				**13**		16

Berlin, Germany (12 out of 549 = 2.2% P) is known to have a significant Polish immigrant population.

Yhrd data can be combined with the Polish Project at nine markers, because they are independent. (Pawlowski's data is included in Yhrd. Ysearch is not fully independent of the Polish Project because some men join both.) The total for P type is 146 out of 4308 = 3.4% (95% confidence 2.8%-3.9%). If P type corresponds mostly to a single clade, the percent for the clade is surely larger, due to the mountain effect.

Table 5
Number of Men in the Polish Project who are R1a1

Markers Available	Project Totals	R1a1, including those indicated as R1a
67	384	154
37	627	254
25	688	286
12	900	378

Results: Signatures and Definitions of Types

Most of the analysis of the Polish Project, 15 May 2009 download, was done using the 154 R1a1 samples that have all 67 markers. Some analysis was done with fewer than 67 markers–more data, and some analysis was done with other haplogroups. Table 5 shows the R1a1 sample counts available from men in the Polish Project.

There may be bias in 67-marker data toward larger clusters. For example, a man who initially orders the minimum 12-marker panel may be more likely to order the full 67-marker panel for better discrimination if he finds his 12-marker haplotype to be very common. As another example, men with a common haplotype may contact others who closely match and encourage them to order the full 67 panel and to join the Polish Project. At the first 12 markers there are 492 unique singletons (data corresponding to Table 1 - 492 haplotypes with only one sample each in the 15 May 2009 download of the Polish Project). Of these, 207 (42%) have the full 67 markers. So 42% is a measure of the probability that a man with a rare 12-marker haplotype will order the full panel. The top eight haplotypes in Table 1 (eight or more samples each) have a total of 93 samples between them. Of these, 44 (47%) have the full 67 markers. So 47% is a measure of the probability that a man with a common 12-marker haplotype will order the full 67-marker panel. The 70% confidence interval for 44 samples is 37 to 52 (40% to 56%), so the small bias of 42% vs 47% toward larger clusters is not statistically significant in this brief examination. The P type haplotype has 22 samples (Table 1) at 12 markers; eight of these (36%) have 67 samples, within expected statistical fluctuation of 42% or 47%. The lowest result of the top eight is Pa with four of 13 (31%), also within statistical fluctuation. The highest is Y type where all eight samples (Table 1 at 12 markers) have the full 67 panel. This 100% is a surprise, about as likely as eight coin tosses coming up all heads as the highest result (8 coin tosses as one trial with the best result out of eight independent coin flipping experiments). This is a hint that perhaps these Y type men encouraged each other to upgrade to 67 markers and join the Polish Project. Except for the qualifying comments in this paragraph, the analysis below takes the Y-DNA 67-marker data of the Polish Project as representative of Poland.

Table 3 reports best marker signatures (out of the full 67) for the types found in Haplogroup R1a1. Table 4 reports signatures for types from other haplogroups. The Supplementary Data has an *Excel* file with analysis for each type from Tables 3 and 4, and for some hypothetical types not listed in Tables 3 and 4.

Table 6 presents the results for the types with best SBP, and a few examples of speculative types with higher SBP. Selected types are discussed in detail in sections below.

Table 6 includes both full Ysearch calibration and also calibration to Ysearch data with the search restriction "Eastern Europe." The former includes samples with "unknown" origin in the total.

For brevity, Tables 3 and 4 are not the full modal haplotypes. The Supplementary Data has a file "Haplotypes.xls" with full 67-marker best estimate modal haplotypes for all types, including some experimental types not statistically worthy of mention in the tables. That file also has the "definition" modal haplotypes, which usually have fewer than 67 markers. The file "YsearchURLs.xls" has the definition modal haplotypes available in the format of Ysearch Universal Resource Locator codes, for rapid verification.

Full results are highlighted in color by type in file "Results15May2009.xls" in the Supplementary Data. That file should not be misconstrued as assignment of men to clades. Some of the types are highly speculative, while a few types are almost surely clades. Even for likely clades, for example P type in Figure 1, samples from the "foot of the mountain" just short of the "cutoff" are less likely to belong to the hypothetical clade than samples with step zero. The Polish Project web site has my high confidence assignments for R1a1 samples, along with a link to my web page including lower confidence assignments. Those web assignments

are based on cluster definitions that were estimated in 2008, and are scheduled for update after this article is submitted for publication. For use by the Polish Project, links are also available to the Ysearch "User ID" codes for my high confidence definition modal haplotypes, which have been entered into Ysearch, and which are periodically updated.

All blanks in Table 3 are equal to K type at that marker. K is essentially the modal haplotype for Polish Project R1a1. The exceptions are a few bimodal markers such as DYS406 and DYS447, which are close to evenly split in the Polish Project R1a1 data between the K value vs. the P and N value. As a result, the R1a1 modal value for the Polish Project at these markers varies from month to month as data accumulates. The types P, K, and N, however, have stable modal values for these two markers. The current data has the R1a1 modal value DYS406 = 12 but in some previous months the modal value was 11, the value for K.

A few of the very rapid mutators, such as CDY and DYS449, also vary statistically from month to month in modal value, both for R1a1 as a whole in the Polish Project, and for some of the type modal haplotypes. Tables 3 and 4 do not have rapid mutators because they are not useful as signatures.

About half the 67 markers are slow mutators, rarely varying in the Polish Project R1a1, so they are not useful as signature markers, except those few, in Table 3, that happen to correlate as a set, thereby serving as signatures.

Table 3 uses c and e for the third and fifth DYS464 values following the FTDNA format. Four values, a through d, are standard. Most samples do not have a fifth copy value, e. Compound markers are discussed in the companion article, with more detailed explanation in the "Documentation" sheet of the file "Calculator.xls" in the Supplementary Data. As explained there, one or two of the DYS464 values sometimes provide excellent signature markers for a type. These cannot be tested on the Ysearch site because Ysearch ignores all DYS464 markers when any one of the standard four are blank. For compatibility with Ysearch, the modal haplotype definitions use either the entire DYS464 marker set or none of the marker set, although better fit to the data with lower SBP can sometimes be achieved using individual markers. Similarly, DYS389-2 is not used in definitions without DYS389-1.

The 22 markers selected for Table 3 are the most significant markers for the analysis of R1a1 in the Polish Project. Markers not included in Table 3 each have the same value for the modal haplotypes for types P, N, K, and G. In other words, Table 3 has all the markers at which these main Polish Project Y-STR R1a1 types differ from each other (except some rapid mutators). Some of the subtypes in Table 3 differ from the parent type by one or two more markers, discussed below.

The best signature markers for each type are boldface in Table 3.

Signatures should not be entered into Ysearch because they produce too many annoying matches. Ysearch "Freeentry" mode is easier anyway; see "YsearchURLs.xls" for a link. Even definition modal haplotypes should not be permanently entered into Ysearch unless they have high confidence; it is easy to delete experimental modal haplotypes after test searches.

The sequence in Tables 3 and 4, left to right, follows the FTDNA standard sequence.

Note that two of the best signature markers in the case of P type, and three in the case of Y type, are included in the standard 12-marker set. That increased the chance of discovery. Types that happen to have best signature markers not included in the standard 12-marker set may take longer to discover.

The Supplementary Data has an *Excel* file for each type from Tables 3 and 4. Each file includes step frequency data for various combinations of marker choice and cutoff. Modal haplotype, cutoff, SBP, ASD, concentration ranking of markers, and other parameters are automatically calculated. Those files each have data for the parent haplogroup. Each of the types in Tables 3 and 4 were checked using the full Polish Project as a database. In every case, no sample from any haplogroup other than the parent haplogroup came close to the step frequency mountain corresponding to the type, so a copy of the data for the parent haplogroup was used for analysis. Each type was also checked on Ysearch.

Results: P Type

P type clearly represents the most common Y-DNA at 12 markers in Poland. As discussed above, P type (the hypothetical modal haplotype at nine markers) is most common in Pawlowski's (2002) data, in the Polish Project, in Polish cities in Yhrd, and is attributed at 42% to Poland on Ysearch at 12 markers.

P type is my paradigm (best example) for a hypothetical clade using the "Mountains in Haplospace method," as described in the companion article.

The P type analysis at 67 markers is available in the file "PType.xls" in the Supplementary Data. There are 29 samples in the mountain, using the definition P36, which uses 36 of the 67 markers. See Figure 1. The cutoff is step 5, with gap three (steps 5, 6, and 7). SBP = 6.7%. Exactly the same 29 samples are extracted using a modal haplotype from 13 to 41 markers, and again using 51 markers. P36 provides the lowest SBP. Figure 1 com-

Table 6
Examples of Mountain Method Results. P type is a paradigm of the mountain method. A good type has less than 30% SBP, so P type is good at 6.7%. P type seems likely to represent a Polish clade. Pc and Pg Types are examples of SBP results that are marginally good, but both Pc and Pg have small sample size, so the SBP is mostly the high end of the confidence interval due to sampling statistics; these two types may improve as more data is accumulated. "A" type is an example of good results with very few samples, a very isolated type quite likely to represent a clade. A, N, and K types are examples of types that are not concentrated in Poland but common in Eastern Europe. See the text for discussion. Y type seems to be a Polish clade outside R1a1.

Haplo-group	Type Code	Definition			SBP	Number of Samples							
		Modal Markers	Cutoff	Gap		Polish Project		Ysearch			Ysearch-East Europe		
						No.	%	No.	No. Pol.	Pol. %	No.	No. Pol.	Pol. %
R1a1	P	36	5	3	6.7%	29	7.6%	39	22	56%	33	22	67%
R1a1	Pc	47	2	2	61%	5	1.3%	6	3	50%	6	3	50%
R1a1	Pg	21	2	2	68%	5	1.3%	5	2	40%	3	2	67%
R1a1	A	67	9	9	5.9%	6	1.6%	24	2	8%	17	2	12%
R1a1	I	59	8	1	22%	12	3.1%	14	8	57%	11	8	73%
R1a1	N	45	7	2	13%	28	7.3%	57	10	18%	36	12	33%
R1a1	K	34	4	1	26%	54	14%	160	39	24%	93	37	40%
I1	Y	52	3	4	9.2%	8	2.1%	6	6	100%	6	6	100%

Notes:
The full definition modal haplotypes and Ysearch data by step are available in the file "YsearchURLs.xls" for these and other types.
"Polish Project" here refers to the 384 samples with 67 marker data, including all haplogroups.
Polish Project % is taken to be representative of ethnic Poland.
“Modal Markers" is the number of markers out of 67 that provide the best SBP, so these were used for the definition.
The "No." is the total number of samples that match the definition at step count less than the cutoff.
The “Gap” is the number of steps starting at the cutoff with low number of samples, separating the mountain type from the rest of the data.
"Pol." means the Ysearch data that has "Poland" for the male line ancestor.

pares the mountains produced by 10, 36, or 51 markers. The breadth, as defined in the Summary of the companion article (allows <10% mismatch, which is two samples in this case) is three to 51 markers - a good figure of merit. In other words, P type is remarkably insensitive to the choice of modal haplotype over a wide breadth of marker number.

The background is two samples at the gap for this sample set, but, of course, with more data more samples should show up in the gap. That 6.7% is the statistical worst case as automatically calculated by the *Excel* sheet “SBP” within the file “PType.xls” using the formula from the companion article. The SBP of P type has been decreasing for the past two years as data has accumulated in the Polish Project. In a few more months the SBP may well decrease below 5%, the definition for an “island in haplospace” in the companion article.

P type represents 29 samples out of 154 R1a1 at 67 markers = 19%. Among all Polish Project samples that have 67 markers, P type is 29 of 384, or 7.55%. The 70% confidence interval: 23.5 to 35.7 out of 384 is 6.1% to 9.3%.

In the case of P type, the size as automatically calculated with the mountain method equations is 29.3, because the estimated background is 0.7 samples in the mountain while the estimated outliers is 1.0 beyond the mountain. Using 29.3 that 7.55% becomes 7.63%. As mentioned above, the 67-marker data may be slightly biased toward larger clusters, so that 7.6% may be a bit too high for the Polish Project. It is not known if the Polish Project is more or less Polish than Poland itself, introducing more uncertainty in a prediction of the percent size of P type in modern Poland. The most important uncertainty is the extent to which the Polish Project represents “male line ethnic Polish," estimated above at 80%, so the percent of P type among ethnic male line Polish men is likely higher than 7.6%, but more uncertain than indicated by the statistical confidence interval in the previous paragraph.

Conclusion: About 8% of ethnic Polish men belong to the hypothetical P type clade.

Of the 22 P type at 12 markers (Table 1) eight of these have 67 markers available. Of these eight, seven (88%) are in the P type cluster using the 36-marker definition, so a 12-marker match to P is a good but not foolproof indicator for P type in the Polish Project. On a personal note, that 8th sample that matches P type at 12 markers fits well in I type (below) at 67 markers, and is the data for my maternal grandfather. That was the sample that motivated me to study R1a1 and led me to the Pawlowski (2002) article.

At 12 markers, modal haplotype P12(12), still using the data with all 67 samples, eleven samples are at step 1. Nine of these are P type (defined by P36). So missing the 12-marker modal haplotype by only one step is also a good indicator of P type. At two steps, only four of 19 samples are P type; this is expected because these include the samples that match K type at 12 markers.

Of the eleven PbKa type at 12 markers (Table 1) four of these have 67 markers available. Only two come out in the P type cluster. This is as expected; PbKa at 12 markers is between P type and K type, so at 67 markers these should be partly P partly K, and perhaps partly neither. This is consistent with (not a proof of) the mountain method evidence for a P type clade even at only 12 markers (discussed above).

P type samples can be recognized using only the FTDNA first 25 markers. This analysis is included in the file "PType.xls" in the Supplementary Data. The definition modal haplotype uses 18 of the first 25 markers, modal haplotype P18(25), with a cutoff 3. SBP = 37%, so the P18(25) mountain is not convincing by itself. It was calibrated to the 67-marker data. Of the 29 P type samples (P36 using all 67 markers), all but one are captured by the P18(25) mountain plus three foreigners are captured, for a mismatch of 4/29 = 14%. In fact using 15 to 18 automatically ranked markers (out of 25) captures the same samples. That 14% does not mean 86% accuracy for a new sample at 25 markers, because of sampling statistics. Also, one foreigner is at step two and two are at step three (see "PType.xls"). The 17 samples at steps zero and one (using P18(25)) are all P type (using P36). A new sample with only 25 markers available that matches P type at step zero or one (using P18(25)) is likely to be P type at 67 markers (the probability is difficult to estimate, but might be better than 90%). Additional samples in the database with only 25 or 37 available markers were identified as probably P type, but this prediction was not used for further analysis in this article.

Using the full 67 markers, modal haplotype P67, SBP = 8.8%, which is a respectable P type mountain. The data and graph are available in the sheet "P67 SBP" within "PType.xls." Correlation is possible. There are 25 of the 29 P type (the 29 from the P36 definition) captured in this P67 mountain type, plus two foreigners (foreigners according to P36). The four that are missed by P67 are one each at steps 2 and 3 and two from step 4 (according to P36, cutoff 5). That P67 mountain cutoff is 16 steps and those two foreigners are at steps 14 and 15. Leaving out the nine samples at steps 14 and 15 (P67), the best 19 (P67) have steps 5 through 13 (P67) and have steps 0 to 3 (P36). This means there is a good correlation at 67 markers compared to the 36-marker definition, except at the foot of the mountain, at step 4 of P36, just before the cutoff at 5. Although there are 0% foreigners (foreigners as defined by P36) in the P67 mountain below step 14, the confidence in that 0% is very difficult to calculate because of the selection bias (I chose 14 because this particular database has no foreigner below 14). Surely the confident future correlation prediction, P36 vs P67, is close to 100% at low step count, and decreases with step count. Conclusion for this paragraph: The full 67-marker data adds credibility to the P36 definition of P type; the two definitions correlate well.

However, it is not reasonable to use the 67 markers as a definition. The SBP, and the samples included in the mountain, varies significantly with the number of markers above 51. This is due to those markers that mutate rapidly. On the other hand, P36 is a good definition because, as mentioned above, the samples in the mountain are identical using 13 to 41 markers, chosen automatically by rank, with SBP 6.7% to 20%. Figure 1 shows P51, which captures the same 29 samples as P36, SBP = 20%, although some of the modal haplotypes with markers in the 40's miss by one sample, and P52 misses by three samples. With careful manual selection of markers (instead of automatic selection by calculated rank) it is possible to achieve slightly better fit with more than 51 markers, but not with SBP below 10%. Reminder: the mountain method does not automatically vary all parameters to calculate a minimum SBP; these results are my results after a long but reasonable search effort.

Those best three markers are DYS385a, DYS572, and DYS406. Those three markers capture all 29 of the P36 data, plus the two samples from the gap at P36, and no others. The data is available in a column within the file "PType.xls." SBP = 67% for the modal haplotype using only these three markers. That does not mean we expect 67% foreigners. The SBP method is misleading above 50% because the background is estimated by the data count at the gap, and the gap by necessity has many samples when only three markers are used for the modal haplotype. However, the fact that P type as defined by P36 can be verified using P3 with only three markers adds credibility to the P36 definition. Also, those three markers are credible candidates as slowly mutating

markers that just happened to mutate about the time when the hypothetical P type ancestor lived.

DYS389-1 does not rank well but it is forced into the P36 modal haplotype to be compatible with Ysearch. (Ysearch cannot include the well ranked DYS389-2 without DYS389-1, but an *Excel* spread sheet can.) In previous months, excluding DYS389-1 provided a lower SBP, but with this particular download excluding DYS389-1 increases the SBP from 6.7% to 7.2%, a small difference.

DYS464 comments: P36 includes all four of the DYS464 markers "a" through "d." Genetic distance mutation step count for DYS464 is calculated following the method used at Ysearch, "infinite alleles," including markers DYS464 "e" and "f" that appear in some samples. The DYS464 set works well in P type for two reasons. First, DYS464c differs from K type and N type, so "c" ranks well by itself. The *Excel* files use a "mask" row for rapid editing of individual markers. Using only the "c" from DYS464 as an individual marker changes P36 to a P33 modal haplotype, capturing exactly the same 29 samples, with exactly the same SBP of 6.7%, but with a lower cutoff at four steps. This is a coincidence. In previous months the SBP using only DYS464c came out lower than with the full DYS464 set. I follow the Ysearch method by default because some readers may be skeptical of using a single DYS464 marker in a modal haplotype. The second reason the full set works well: Many N type samples have values at DYS464e and DYS464f, and some of these fall in the P type gap when the DYS464 set is not used; DYS464 gives these samples higher mutation counts to better distinguish from P type. Usually, DYS464 mutates too rapidly to be of value. The value of DYS464 is evidence that P type represents a very young clade, as discussed further below. For readers uncomfortable with using DYS464 in a modal haplotype, P32 with all four DYS464 markers removed (but still including DYS389-1) provides SBP of 12% and captures one more sample for a mountain with 30, so it is not very different from P36.

DYS413a has two samples with mutations of step 2, apparent double mutations, so it was excluded from the modal haplotype for P type, although it ranks 33rd. (P36 is defined using all markers up to rank 33, but there are two markers at rank 33, and for consistency with Ysearch the poorly ranked DYS389-1 and two poorly ranked DYS464 markers are included.) With the 37-marker haplotype including DYS413a, SBP comes out 10% (vs 6.7% without it) and the cutoff is 6 (vs 5), but the same 29 samples are captured by this 37-marker modal haplotype. The breadth is not as wide using DYS413a. DYS413b is not in the modal haplotype because it does not rank that well.

The file "YsearchURLs.xls" has the step frequency data from the international Ysearch database using a free entry URL for P type, using the P36 modal haplotype with 36 markers from the Polish Project. The total is in Table 5. 56% of the data in the mountain is Polish ("Poland" indicated as origin). On Ysearch the percent Polish is higher for the P type modal haplotype using a 36-marker definition than the percent Polish using only the first 12 markers for P type (42% in Table 1). Although the statistics are not convincing, this may mean there are proportionally more non-Polish samples (in Ysearch compared to the Polish Project) outside Poland that match P type at only 12 markers but clearly do not belong to the type with all 67 markers.

The Ysearch P type data cluster, using P36 from the Polish Project, has SBP = 9.2%, compared to 6.7% for the Polish Project. Restricting to Ysearch Polish, SBP = 12%. These SBP differences may well be mostly due to sampling statistics, but a higher SBP in Ysearch is expected because the identification of Polish samples is not complete, as discussed above. A Ysearch data download (described below, in the K type section) was analyzed using the file "PTypeYsearch.xls," available in the Supplementary Data. The ranking of the markers came out slightly differently, but the approximate rank of markers was the same as in the Polish Project. The analysis proceeded similarly to the analysis of the Polish Project data. I did not find a Ysearch P type modal haplotype with lower SPB than the P36 fit to the Polish Project, although I spent much more time searching for a fit in the Polish Project data.

There are 599 samples in the FTDNA project "R1aY-Haplogroup" open to all R1a (24 April 2009 download). There are ten P type at 12 markers, 1.67%, less than the Polish Project at 2.4%. Of the ten , five are "Poland," one is "Ukraine" and four don't say. As expected, P type is predominantly Polish.

There are four small FTDNA projects worth monitoring from Slavic countries next to Poland. These were checked on 12 May 2009: The "UkraineBlackSea" project has no P type at 12 markers out of 20 Y samples. The "LithuanianDNA" project has no P type at 12 markers out of 50 Y samples. The "Czech" project has three P type at 12 markers out of 180 Y samples. The "Slovakia" project has three P type at 12 markers out of 74 samples. All these are all lower percent P type at 12 markers than the Polish Project.

Two Jewish FTDNA projects are mentioned below, in the A type discussion. Neither of the these have any P type at 12 markers.

Results: P Subtypes

Pc. As shown in Table 3, the Pc type signature, in addition to sharing the signature of P type, differs from P at four markers (plus three rapid mutators not shown in Table 3, that have little effect). DYS442 is useful to

distinguish Pc from P, but not as useful for distinction from R1a1 as a whole, because the Pc value at DYS442 is the modal value for R1a1. The "PcType.xls" file in the Supplementary Data has details.

Pc type is very small, with only five samples at 67 markers. All five samples are also P type. The breadth of Pc is 16 to 50; the same five samples segregate into a well defined Pc mountain using 16 to 50 markers. Exactly the same mountain, with SBP = 61%, is produced from the 15 May 2009 data using 17 to 47 markers. Accordingly, the definition of Pc49 uses 47 markers, with cutoff 2.

The background of Pc actually comes out zero. The sample size is small, so the mountain method produces that SBP = 61% in order to account for confidence in the statistical expectation. The 70% confidence interval on that size = 5 for Pc type is 3 to 9. As more data accumulates, the SBP value may come down. Although that SBP = 61% value is too high to conclude that Pc type is likely a clade, Pc is discussed here as an example of how the mountain method works for small types. Pc is also discussed in the next section, on age.

Of the six Pc type at 12 markers in Table 1, only two have 67-marker data, and both fall into Pc type at 67 markers; in the future with more data not all would, of course. None of the eight P type at 12 markers with 67-marker data fall into the Pc type at 67 markers; with more data some would, of course.

Pc also produces a similar mountain on Ysearch, see Table 6 and Supplementary Data. Two of the six Ysearch Pc type samples are clearly the same as two of the Polish Project Pc type samples, with the same ancestor name. The other four are different. Three of the six indicate Poland as origin, two are Russia, and two are Ukraine.

The Ysearch Pc mountain differs from the Polish Project by only one more sample, six total. The gap is identical. For Ysearch Pc, SBP = 47% (vs 61% for the Polish Project); which demonstrates how SBP is sensitive to just one more sample in the mountain, as it should be for small samples.

Pa. Pa is an instructive example, because Pa in Table 1, with signature DYS389=(13,31) looks twice as good as Pc with signature DYS439=11. However, with all 67 markers, that DYS389=(13,31) signature fails to form a mountain, while 439=11 correlates with three other slowly mutating markers to produce Pc type.

Pa, signature DYS389-2=31, is actually a value of 18 for the larger of the DYS389 pair. The R1a1 modal value, and the value for P type, is DYS389=(13,30), for a value of 17 at that larger one. There are eight samples in the 67-marker R1a1 database with DYS389-2=31, but half of those are DYS389=(14,31), which has the modal value of 17.

The other half are only four samples with 67-marker data. Pa is unlucky in that only four of the 13 samples (Table 1) have all 67 markers for analysis. Anyway, DYS389 is not a particularly useful marker in this case; it does not form a type with the 67 markers available at this time. SBP comes out >200%.

Pg. Although highly speculative with only five samples at this time, Pg type is an instructive example of a type that is not visible at 12 markers because it is primarily distinguished by one marker, DYS572=11. SBP = 68%, which is not bad for such a small type, in view of the large confidence interval automatically calculated for small types in the mountain method. Maybe the value might come down as more data accumulates. Breadth is 18 to 51 markers. The definition uses 21 markers. See "PgType.xls."

Pg type has two of those four samples with the DYS389 = (13,31) for Pa in Table 1. However, the DYS389 marker is variable in the Pg cluster, so that marker does not rank well, and is not used for the definition. This explains in part why Pa does not produce a mountain type with reasonable SBP.

Table 6 has data for Pg, but this is just a very preliminary hint, with so little data. Like Pc, Pg also has only preliminary data on Ysearch at this time.

Pg type can also be extracted from a database with only the 29 P type samples, because that DYS572 marker works all by itself. Only the five Pg type samples have that value within P type. 572 is not effective extracting from the full R1a1, because the Pg value DYS572=11 is modal for R1a1. The samples in the gap next to the Pg mountain are all P type, so the extraction from P type produces the same SBP = 68%.

Pg differs from P by only that 572 marker (and the inconsequential CDa). The reason Pg separates from P is because Pg type has no variation (in the five samples) at a few markers that are quite variable in P type (remaining 24 samples with Pg extracted). The four best of these markers (along with 572) are boldface in Table 3.

In other words, Pg looks like a subtype within P. A subtype is expected to have less variation, because the parent type has additional variation at those markers where other subtypes differ (population structure). Pg and 572 are further discussed in the next section.

Other P Subtypes. I found no further credible mountain subtypes in P type. At a glance, two other markers look

good as cluster candidates: Out of the 29 P samples at 67 markers, DYS458=17 has twelve samples, only two less than the modal value 16 for P type. DYS456=15 has ten samples. These individually look like subtypes, because of the bimodal distributions. They do not correlate with each other. There are some pairs of markers that are weakly correlated, but none worthy of mention at this time. Perhaps as more data accumulates in the Polish Project more P subtypes will score well for SBP. No doubt if more than 67 markers become available some will correlate to produce more P subtype candidates.

Age of P Type

Age of P type is estimated three ways, using the data from the definition modal haplotype with 36 markers: A: Cutoff 5: The best choice mountain data, all 29 P type samples at steps 0 through 4. B: Cutoff 3: This choice leaves out the twelve samples at steps 3 and 4, yielding the 17 samples that best fit the modal haplotype. These are anticipated to provide a younger age. These are likely to have less background from foreign outlier samples. C: Cutoff 9: P type plus the additional two samples in the gap at steps 5 to 7 plus six samples at step 8 beyond the gap, yielding 37 samples. These are anticipated to provide an older age. These are more likely to have all the P type outliers, but with more background. In the Supplementary Data, file PType.xls, each of these three ways uses two sheets, one for the data and one for the ASD calculations. These three ways demonstrate how much the ASD age depends on the choice of cutoff.

Using the same five markers as Thomas (1998), the P type ASD = 0.115, smaller than the Thomas value. The age comes out 1,365 years using the Thomas method, compared to 2,650 years found by Thomas for his data. There are only 16 total mutations in the 29 samples using these five markers.

Conclusion by comparison to Thomas: P type, if a valid clade, is surely very young.

With the Thomas method, the 17 sample test comes out 1,252 years and the 37 sample test comes out 1,242 years. That latter number was expected older but came out younger because those extra 20 samples just happen to have fewer mutations at the Thomas markers. The Thomas method age is not very sensitive to how P type is extracted (exact cluster definition), because all the R1a1 samples in the Polish Project with step counts not far from the P type modal haplotype have few mutations using the Thomas markers.

The age using Chandler's rates for the five Thomas markers varies widely; DYS392 has no mutations for zero age and DYS19 with eight of the 29 samples mutated comes out oldest at 3,307 years. This may be sampling statistics, since the sample size is not large. On the other hand, this may be a clue that DYS19 harbors a subtype. Indeed DYS19 is bimodal, with eight samples at 16, the modal value for R1a1, and with 21 samples at 17, the modal value for P type, and no samples at any other value.

With all 67 markers and the Chandler rates, the age of P type comes out 1,905 years. Those best 17 samples come out 1,242 years, and those 37 samples come out 2,108 years.

The ASD sheet has a tool for sorting markers by apparent age, and for masking out selected markers.

The ASD sheet has a default mask with 59 markers, masking out those compound markers that seem to mutate most rapidly by recLOH (the CDY pair, the DYS385 pair, and the DYS464 quartet). The recLOH issue is reviewed in the Supplementary Data. With this mask, P type comes out 1,645 years. Those best 17 samples come out 1,521 years, and those 37 samples come out 1,869 years. These ages are not much younger than using all 67 markers, despite the obvious recLOH objections to using compound markers as individuals in the 67-marker average. It seems for P type that the high recLOH counts for these compound markers are mostly compensated by the high mutation rates for these markers.

DYS385b is the oldest marker - 21,835 years. This is obviously due to the four samples with recLOH mutation from DYS385=(10,14) to DYS385=(10,10). The DYS385 pair is excluded from the age calculation with that 59-marker mask. Although DYS385b should be excluded from age calculations, DYS385a can be included as an independent marker with no recLOH in the P type data.

DYS464a is 7th oldest - 4,695 years, also due to obvious recLOH. All four of the DYS464 marker copies are excluded in that 59-marker mask.

The CDY pair of markers are the two fastest mutating markers according to Chandler's rates. For the 29 P type samples the raw ASD age comes out 2,416 years (CDYb - 15th oldest marker) and 1,078 years (CDYa - 34th oldest). The age should be slightly younger, because the 29 samples have two obvious recLOH copy mutations, both low, one at the lowest value (32,32) vs. modal value (34,39), which contributes too much to the ASD variance and thus to the age. The CDY pair are excluded by that 59-marker mask, but CDY makes little difference in the averaged age for P type.

DYS413a is excluded from the modal haplotype because of two double step mutations, discussed above. Indeed DYS413a is 9th oldest at 3,738. No recLOH are expected for the DYS413 pair because it is modal (18,18). In this situation, the "a" marker gets the step down muta-

tions from both, and the "b" marker gets the step up mutations from both. There happen to be few step up mutations, so DYS413b comes out only 412 years. To be fair both should be excluded, with a small effect on the average.

The DYS459 pair has modal value (9,10) so recLOH produces only one step. Similarly, the 395 pair has modal value (17,17). YCAII has modal value (19,23) but there are no recLOH in P type. There is no reason to exclude these pairs. That concludes the consideration of all compound markers.

Best raw ASD age estimate: A 58-marker mask is in-place in the three ASD calculation sheets when the file "PType.xls" is opened; a backup copy is in the "Mask" sheet. The nine markers removed: the CDY pair, the DYS413 pair, DYS385b, and the DYS464 quartet. Age of P type using this 58-marker mask is **1,601 years**. Age of the other two sets: 1,490 and 1,823 years.

DYS426 is the second oldest marker - 9,248 years. This age is due to only one single step mutation in this very slowly mutating marker. This anomaly is not removed, because it is statistically balanced by the 19 markers that have no mutations and zero age.

DYS565 is the third oldest - 6,697 years. This is probably because 565 is a marker for Pc type. See Table 3. The same is true for 442, 4th oldest - 6,055 years. This subtype variation is discussed further below.

The far right of the ASD sheet has all the markers sorted by age, with notes.

Fast mutating markers tend to be good individual indicators of TMRCA for very young clades, as explained in the companion article. Chandler rates are available in row 5 of each copy of the "ASD" sheet in the Excel files. The Chandler rates are ranked in the "Mask" sheet in each copy of the "Type" files, as part of a tool for constructing a mask, for ease in restricting calculations by mutation rate. A mask was constructed for the 15 Chandler fastest mutating markers, excluding the nine markers that were excluded for the ASD age above. This mask is available at the bottom of the "Mask" sheet. Using this mask the raw ASD age for the 29 P type samples is 1,793 years, which is not very different than the 58-marker average above.

Since P type is a significant fraction of the Polish population, a young age means that P type went through a rapid population growth from a small founder population.

TMRCA summary: As explained in the companion article, time of the most recent common ancestor (TMRCA) may be up to a factor of four older than the raw ASD age of 1,601 years (due to "effective mutation rate factor" as small as 1/4). Since P type grew rapidly, however, that factor should be closer to 1, not likely as large as 2. TMRCA may perhaps be from 2,000 to 3,000 years ago.

As explained in the companion article, time of population expansion for P type is likely younger than the raw ASD age of 1,601 years.

Apparent age by marker comes out in a broad distribution, with 19 markers at zero age due to no mutations, and with 14 markers older than 2,000 years (19 including five markers that are among the nine excluded above).

Most of the age variation by marker is no doubt data sampling statistics, which gets averaged in that 1,601 year result.

Some of the age variation may be imprecision in the Chandler (2006) rates, due to sampling statistics and biases in both the mutation data and the calibration data used to obtain the rates, as discussed by Chandler. Again, the old and the young results are averaged.

But at least some of the apparently old markers are no doubt due to population structure, as explained in the companion article. If P type went through a relatively recent rapid population expansion, those markers that were relatively homogeneous in the small founder population would come out with young raw ASD age pointing to the time of population expansion, or a range of times if the expansion happened in waves. Those markers that just happened to vary in the founder population would come out older.

If the founder population could be subdivided into subtypes based on those markers that varied (in theory one subtype for each man from the founding population who has male-line descendants in the database), each subtype would come out with a raw ASD age equal to both the TMRCA and the time of population expansion (for a rapid population expansion, with a large enough database for statistical precision).

The signature marker for subtypes (those markers that differed in the founding population) should produce an ASD age older than other markers (given enough data). Specifically, if Pc and Pg are valid subclades of P, the signature markers (Table 3) should come out older. As mentioned above, 565 and 442 indeed do. As mentioned above, DYS19 is an example of a marker with a bimodal distribution and an old ASD age that does not produce a subtype, but that may just mean that a DYS19 clade does not significantly differ from the parent P type at other slowly mutating markers in the standard 67-marker set. All the markers that come out old are candidates for hypothetical subtypes, but many if not most of them are no doubt just statistical fluctuations due to limited data sampling.

The Supplementary Data automatically provides an ASD sheet for Pc Type and for Pg type. Pg comes out quite young. Pc comes out about the same age as P. As mentioned in the Pg section above, Pg has little variation. With only five samples each, any discussion regarding these subtypes is highly speculative. I provide here a speculative discussion as an example of how hints may be gleaned using the mountain method for very small types.

Pc type is the about same age as P type - 1,507 years using the same 58 markers. See "PcType.xls." DYS442, one of the markers for Pc type, has a lot variation, so after separating Pc from P, the age of Pc does not get reduced by much, but the age of P type does get reduced. This may mean that Pc is a best fit of multiple subtypes defined by the 442 markers. More data should clarify this issue.

Using the Thomas markers, the Pc age comes out old, 3,429 years, because many of the samples with high mutation counts segregate into Pc. That must be at least partly sampling statistics because of the small sample size, but a high age for Pc is consistent with the previous paragraph suggestion that Pc may be a best fit of multiple subtypes.

Pg comes out young - 898 years using the same 57 markers. The ASD sheet has notes. Thomas age is 762 years. This may be the first example of a valid subtype within P type descended from a man or tribe of men who were identical for the slow mutating markers in the standard 67-marker set. This conjecture is uncertain until better statistics accumulate.

I checked all the other old markers for evidence of subtypes. There is no further evidence statistically worthy of mention, but as more data accumulates in the Polish project some of these "old marker" hints may provide subtypes with low SBP and with relatively young raw ASD age.

Old markers were not excluded in the analysis, but clearly if some of these are due to population structure before population expansion, then they should be excluded to estimate the time of expansion. The trouble is, we do not which to exclude for structure, and which to retain for statistical averaging.

However, those 19 markers with zero mutations certainly suggest that many if not most of them were homogeneous in a hypothetical small founder population for P type.

DYS572 justifies three paragraphs further discussion: DYS572=11 is the modal value for R1a1 in the Polish Project. P modal value is 12, which is the second best marker for distinguishing P type. Pg has the 11 value. Nevertheless, all Pg samples segregate well into P type because other markers correlate very well. All five P type samples with the 11 value segregate into the Pg subtype. The gap at step two has no samples.

Outside P type, only three samples from the Polish Project R1a1 population have DYS572=12, and those three have high mutation step distance from P type. There are no borderline cases; there are no DYS572=12 at or near the gap of the P type mountain. DYS572=10 is a signature of A type, and only three non-A, non-P samples have the 10 value.

The Chandler rates (as extended to 67 markers on the web) do not provide a very low mutation rate for 572; it is 40th, ranking by slowest to fastest rates. It seems surprising that it works so well in distinguishing P type. Perhaps the true rate, at least within R1a1, is slower than the posted rate. It is not clear if the Pg = 11 value for 572 is a back mutation from P type to the haplogroup modal value, or if Pg represents a subset (perhaps just a few closely related men) of the founder population who descend from near the node of P type, after the prime signature DYS385a = 10 mutated, but before the 572 mutation. Either way, even if Pg type represents a very old branch, Pg type should come out with a young raw ASD age in Poland, equal to the time of population expansion, if 572 is the only marker of the 67 available that distinguishes it. Of course, that 898-year age is highly uncertain by at least a factor of 2. In the interests of full disclosure, I must admit that another explanation is that all this is a false positive error due to my intense study of the P type data. These paragraphs are provided as an example of how far the mountain method can be pushed to glean a reasonable, if speculative hypothesis. Time and more data, perhaps a future SNP, will either falsify these speculative paragraphs about Pg and 572, or justify the mountain method.

This concludes the speculative discussion of hypothetical subtypes.

Total mutated markers for P type (infinite alleles, disregarding step, treating compound markers as one, all markers) is 257, so sampling statistics is not a big issue for averaging many markers here. The ASD sheet automatically calculates this total mutation number, with the infinite alleles limit set higher than 50%.

Summary; population expansion of P type: Population expansion time should be less than the raw ASD age, as explained in the companion article. For P type, that means less than 1,601 years. Although there is no statistically compelling evidence for subtypes, those markers with age over 3,000 years, and all those markers with zero age, are evidence of subtypes at the time of population expansion. P type population expansion appears to

have happened less than 1500 years ago. How much "less than" we cannot say until more data accumulates.

P Type Summary

If the hypothetical P type is a valid clade, about 8% of Polish men belong to this clade, with confidence interval 6.4% to 9.6%. TMRCA is probably more than the raw ASD age of 1,601 years, perhaps more like 2,000 to 3,000 years ago. This is quite young for such a large clade, so P type must have significantly expanded in population, less than 1,601 years ago, perhaps 1,000 to 1,500 years ago. Because P type is isolated in Polish Y-DNA, it may represent an immigration from elsewhere, or it may represent an older population that almost went extinct. Either way, there seems to have been a small closely related founder population before the expansion. It has not escaped my attention that Poland as a nation appears in written history a little more than 1,000 years ago. It is fascinating to speculate that a rapid expansion of population, including P type, occurred shortly before the appearance of Poland as a nation, although further discussion is beyond the scope of this article.

A Type; Ashkenazi

A type is small in the Polish Project, but it is a very well defined mountain.

The same six samples segregate very well into a mountain using any number of makers for the modal haplotype, from one to 67 breadth. Accordingly, the definition uses all 67, with a cutoff 9. The gap is 9; step count 9 to 17; there are no samples in that gap. See "AType.xls" for the graph. The six samples have step count 2 - 8. In the future, the cutoff can be 10 or more if outliers show up, with still an excellent mountain effect.

SBP = 5.9%. As more data accumulates A type will surely qualify as an "island in haplospace," defined as SBP < 5% in Part 1 of this two-article series.

Using only the first 25 standard markers, with the 67-marker data, the same six samples are captured, again using any number of markers from one to 25. The best SBP is 24% using 24 markers. This is because five of the top seven markers for A type just happen to be in the 13 - 25 panel. The standard 37-marker set is even better, including the best marker for A type. This means A type can be confidently extracted using as few as 25 markers. Accordingly, twelve total samples from the Polish Project are assigned to A type. At 12 markers A type is identical to K type, the modal haplotype for R1a1 in the Polish Project, so samples with only 12 markers cannot be assigned.

The best marker for A type is 456=14. In R1a1, all eleven A type samples with at least 37 markers have this value, and none of the other samples (254 total with at least 37 markers) have it. No doubt as more data accumulates, a few exceptions should show up, so this one marker should not continue to distinguish A type by itself. The 14 value is common in other haplogroups.

Another good marker for the A type signature is DYS459b=11. However DYS459b is bimodal, with near 50-50 split vs the R1a1 modal value of 10. As data accumulates, 10 is usually the modal value for A, but sometimes alternates with 11. This is an example of a situation where it is advantageous to set the modal haplotype at a value that is not clearly modal, because the 10 value does not distinguish A type from the rest of R1a1. Using DYS459b=11 in the modal haplotype for A type yields a very good ranking. Currently, out of the 286 samples in R1a1 with this marker, four of the twelve A type have DYS459b=11 and only two other samples have this value, and those samples are clearly not A type because they have high step count for other markers. Only two of the non-R1a1 in the Polish Project samples have the 11 value.

This bimodal DYS459b distribution may indicate a subtype. No other marker of the other 66 seems to be correlated with this one in the A type data. The highly tentative subtype Ab, with DYS459b = 10, is not listed in Table 3.

For all twelve A type, and for only 23 other samples in R1a1 at 25 markers, DYS464 = (12,12,15,15) indicates a specific kind recLOH event, where the 12 and 15 chains copy together onto the other two. For this reason, DYS464b works well as a signature marker. None of the twelve A type have more than four values for DYS464, so this mutation is not an expansion to more copies like the mutation in N type (below). On the other hand, all but one of those other 23 samples have values for DYS464e and DYS464f, indicating an expansion, which is a different kind of mutation from the A type mutation. So the full DYS464 set by itself is also an excellent signature for A type in the Polish Project R1a1. These results are easily viewed using the *Excel* "Filter" tool with the file "Results15May2009.xls" in the Supplementary Data.

Any two of the five A type boldface markers in Table 3 (aside from DYS456, which works by itself) are sufficient to cleanly separate the A type samples as a mountain in the Polish Project.

A type is identical to K type at 12 markers, but at 67 markers only some of the K(12) resolve into A type. Of the 14 K type in Table 1, twelve have 25 markers available, but only six are A type, slightly less than half.

Mayka (2008) posted this type in Ysearch with 67 markers. I have since updated that Ysearch entry. The Ysearch code and the 67 markers are available in the Supplementary Data.

A type seems to be Ashkenazi. Ten of the twelve A type have ancestor names supplied. Most if not all of these are Jewish family names. Jewish affiliation is not indicated in the database of the Polish Project.

Ysearch shows a very well defined mountain effect for A type. Almost the same samples come up regardless of the number of markers used. The result using A67, the modal haplotype from the Polish Project, is in the file "YsearchURLs.xls" in the Supplementary Data. Most of the names below step 9 seem Jewish, and progressively lower percent Jewish names appear at higher step counts. More than half are listed as Eastern European origin in the A type mountain on Ysearch. See Table 6. Only two of the 24 A type on Ysearch have "Origin" Poland. Using the Polish Project definition, on Ysearch the A type SBP is 8.4%.

The modal haplotype comes out the same on Ysearch as for the Polish Project, except for that bimodal 459b marker discussed above, which comes out 10 here; 15 samples at 10 and nine samples at 11. Changing this one marker to 10 reduces cutoff to 8, the sample count to 23 and the SBP to 7.7%. However, the Ysearch data has samples in that wide cutoff, and a better SBP can be found. In Ysearch, the best A type SBP that I found is 4.6%, using 60 to 65 markers, 26 samples, cutoff 9, gap 2. Ysearch conclusion: the A type Modal Haplotype at Ysearch is almost the same as in the Polish Project, with almost the same SBP, well within statistical expectations. But A type is not concentrated in Poland (except insofar as Ashkenazim were concentrated in historical Poland).

Nebel (2005) provides the 14 most common Ashkenazi haplotypes, in a supplementary file. Nebel uses six of the standard first 12 markers. Nebel's most common type, with 25 of his 56 samples, matches A type (and also K type). The remainder differ by one step (seven types), two steps (three types), or three steps (one type). In other words, samples that match any of Nebel's 14 most common Ashkenazi haplotypes may or may not fall into the A type mountain, depending on the STR values beyond the 6. Behar (2004) used ten markers for his Ashkenazi study but again these are all from the standard 12, which do not distinguish A type within K type.

The "JewishDNAProject" at FTDNA has 687 samples (24 April 2009 download). Of these, 166 have all 67 markers. A type forms an excellent isolated mountain type in this data, seven samples, 4.2%. This compares to six / 384 = 1.6% in the Polish Project (all haplogroups at 67 markers). At 37 markers there are 16 A type in a nice mountain. Regarding that DYS456=14 marker that by itself distinguishes A type in the R1a1 37-marker data in the Polish Project, it works again in the JewishDNAProject. All 16 have this value and none of the other R1a1 have this value.

The "Jewish_Ukraine_West" project at FTDNA has 220 samples (24 April 2009 download). Of these, 80 have all 67 markers. A type forms an extremely isolated mountain type in this data at 67 markers, three samples, 3.8%. Only one of these three kit numbers matches (same man) an A type Kit number from the JewishDNAProject (above). At 37 markers there are eight A type and again 456=14 value perfectly distinguishes them from all the other R1a1. The Jewish_Ukraine_West map shows the place of family origin of more than half the project samples as the west half of The Ukraine, with most of the remainder spread throughout Eastern Europe.

Using the A modal haplotype with the Thomas six markers (six in the data, only five used for ASD) in the JewishDNAProject, eight A type samples match all five markers and seven non-A type match, for close to 50-50. Ten more samples with 12 markers that match are not included. All are R1a1. That compares to the Polish Project (data with 25 or more markers) where eight A type match and 60 of the non-A type match. In other words A type is a small fraction of K type (at six markers) in Poland but is half of the common K type (at six markers) in the JewishDNAProject.

Back to the Polish Project, age of A type can be compared to Thomas (1998), which used Jewish Cohanim data. All twelve A type samples can be used at 25 or more markers because these have the five Thomas markers. Perhaps there are more A type in the 12-marker data, but these cannot be distinguished in the Polish Project. ASD for the twelve samples using only the five Thomas markers is 0.056, compared to the Thomas value of 0.2226. This means A type is about 1/4 the age of the Thomas data set. A type is only 661 years old following the Thomas method. A type seems to be a particularly young but common subtype within the Ashkenazim. Thomas could not separate A type from other types with the same 5-marker modal haplotype, so it makes sense that this subtype of the Ashkenazi population comes out younger than the full population. Thomas uses a single mutation rate which is close to (23% greater than) the Chandler rate average for those five markers, so no population factor is included; and perhaps not needed for this population which obviously grew fast. Using the Chandler rates for the same five markers the result is 812 years for A type.

Not surprisingly, that bimodal DYS459b marker comes out as seeming to be 6,708 years old. Oldest is DYS578 at 69,444 years. Third oldest is DYS531- 9,384 years,

because of a mutation from 11 to 12 at this slow mutating marker. Those three markers do not correlate with each other in the A type data.

It appears that those three markers define three different subtypes. This is highly speculative, with only twelve samples available. A type is an example of how examination of ASD age by marker provides (tentative) population structure hints, even with minimal data.

It is likely that at least one of those three, perhaps all three, are statistical flukes. If that be the case, then A type is an excellent example of how flukes in data cause huge effects in ASD age.

That said, averaging all 67 markers, the age is 626 years; 49 of the 67 markers have zero mutations, providing age of zero for averaging. A population structure factor is probably not appropriate for this very young type that obviously grew very fast (if it is a valid clade). There are no recLOH issues in the A type data. Those three old markers probably act as a sufficient accounting of population structure. Even using an additional population structure factor of 1/2 only doubles the age to 1252 years. More important here is the possibility of false positive bias. I am discussing A type at length, to serve as an example of a type that seems to represent a young fast growing clade that provides a distinct mountain even with very little data. If A type did not come out that way I would have chosen another type to discuss.

A type summary: No precise statement can be made about the exact size and age of A type until more data accumulates. The good isolation of A type, quantified by the SBP value of 5.9% in the Polish Project, is good preliminary evidence that A type represents a very young clade within R1a1. The data from Ysearch and the two Jewish projects adds evidence. Although present in Poland, A type does not seem to be concentrated in Poland.

I Type

I Type does not appear in Table 1 because there are only three samples at 12 markers. At 67 markers, the I type mountain has twelve samples. SBP = 22%. The modal haplotype has 59 markers, cutoff 8, gap 1. The analysis is available in "IType.xls."

The breadth of I Type is 18 to 64 markers. Within that breadth, marker range 19 to 59 produce exactly the same twelve-sample mountain with a one-step gap with zero samples. The SBP equation (companion article) provides SBP of 22.4% for a twelve-sample mountain with a one-step gap of zero, even though the cutoff value is larger for more markers. So the definition uses the largest of those equivalent modal haplotypes. The narrower marker range 42 to 55 actually provides an SBP of 22.1%, but that marginal improvement is because the last step of the mountain in those cases has a single sample (a sample that is always in the last step of the mountain, for all marker sets), and the SBP for an 11-sample mountain with a two-step gap containing one sample is 22.1%. The mountain method allows human judgment to intervene in situations like this, where it seems silly to exclude a sample for a marginally lower SBP. That one sample may or may not actually belong to a true clade corresponding to I type.

The I type mountain is not monolithic. It is more like a set of rolling hills. A number of subtypes can be tentatively identified, but all have SBP > 30% because of the small sample sizes.

Ysearch has 14 I type samples, SBP 19%. The Polish Project modal haplotype I59, and the cutoff at 8, gap 1, fits well on Ysearch. The percent Polish, Table 6, is very similar to P type, so on this basis I type also seems to be concentrated in Poland. The confidence that I type is in fact a clade is good, based on SBP, and based on concentration in Poland, although the confidence is not as good as for P type.

The age of I type is analyzed in the file "IType.xls," ASD sheet, using the twelve I type samples. Age using the five Thomas markers comes out 3,406 years, largely due to variation in markers 390 and 19. This is much older than P type. In contrast, age using all 67 markers comes out 1,483 years, slightly younger than P type using 67 markers, largely because there are no recLOH issues in the twelve data samples for I type. The oldest marker is 578, at 69,444 years, due to four identical mutations. Second is 436 at 10,610 years, due to only one mutation. Third is 537 at 8,224 years, due to three identical mutations. These look like subtypes, but I see no obvious correlations, so there may be multiple subtypes each with only one marker that differs from the modal value.

Eight markers give ages older than 5,000 years, 33 markers have no mutations for zero age, and 12 markers give ages less than 1,000 years (range 337 to 850). There are no obvious problems for those old markers in the twelve samples; these are just markers that have many mutations in this data, or few mutations with a low Chandler rate.

The I type range of ages by markers is very suggestive of subtypes, consistent with the "rolling hills" comment above. One simple hypothetical explanation: a few individuals (or a few sets of closely related males) from an I type clade were in a founding population of relatively recent rapid population expansion. The individuals had Y chromosomes that varied at a few markers - those markers that come out very old now in the Polish Project. Those individuals were homogeneous at most markers, which now produce very young ages. It is tempting to

speculate that I type and P type participated in the same Polish population expansion. Other than such speculation, the I type data is too few and too diverse to make confident statements concerning TMRCA or time of population expansion. Perhaps as data accumulates in the Polish Project I type will be cleanly resolved into a number of smaller subtypes with better statistics.

The A type samples are just beyond the gap of the I type mountain. I type and A type samples are color coded in "IType.xls." There is a dip in the step frequency curve for I type, at step 14, suggestive of a broader type that seems to include the I samples, all the A samples, and several others. This is probably because of K type, discussed below.

Although there are three samples that match I type at the first 12 markers, only one of those has 67-marker data, and it falls into the I type cluster. As mentioned above, the I type cluster also has one of the samples that match P type at the first 12 markers (one of 22 samples, Table 1). None of the K type or N type from Table 1 fall into I type at 67 markers. We expect that as data accumulates there will be more examples demonstrating that the 12-marker data is a good but not foolproof indicator for assigning samples to subtypes within Polish R1a1.

N Type

The N type mountain using 67 markers has 28 samples. SBP = 13.3%. The modal haplotype has 45 markers, cutoff 7, gap 2. The analysis is available in "NType.xls."

The breadth of N type is 10 to 47 markers (same data within two samples). But the lowest SBP of 13.3% is achieved only using 44 or 45 markers. The cutoff is somewhat arbitrary, because within that breadth, SBP varies only slightly as the cutoff and gap are adjusted by one step. The three samples in the gap and the four samples at step 6 in the mountain can easily be adjusted into or out of a reasonable mountain definition with SBP less than 20%. However the 24 samples below step 6 fall into the mountain data cluster for any reasonable cutoff within the breadth.

Most but not all N type samples have DYS464 "e" and "f" data, due to obvious recLOH copy mutation. Any modal set of values for DYS464 produces high mutation count for many samples in N type. The single marker DYS464a, however, ranks well. A 44-marker definition using only the "a" marker from DYS464 provides an SBP of 10.0%. For compatibility with Ysearch, however, all DYS464 markers were masked out of the analysis for N type.

The recLOH copy mutations seem to imply two subtypes, Na and Nb, in Table 3, but these are very speculative, because they do not separate well and do not produce low SBP. These subtypes may not separate even with more data; more markers may be needed to distinguish them, if they are valid.

N type is quite common on Ysearch, 55 samples (17 Jul 2009), SBP 20%, using the Polish Project modal haplotype N45, and the cutoff at 7, gap 2. However, changing the gap to 1 changed SBP to 18%. In Ysearch as in the Polish Project, various marker sets produce SBP less than 25% for N type, again sensitive to the exact cutoff definition. The modal values for the markers on Ysearch are the same as in the Polish Project (except for a few rapid mutators that do not rank well and are not used for the definition).

N type is not concentrated in Poland, compared to P type and I type, as evidenced in the Polish percent, Tables 1 and 6. N type, if valid as a clade, seems to be a common Slavic subtype of R1a1. Ysearch should be a better database to study N type. The advantage to studying N type in the Polish Project lies in the possibility that a small Polish subtype may be discovered in the future as more data accumulates. N type may well be concentrated in a region other than Poland, and such concentration would be additional evidence of validity. The concentration and age of N type is better studied in Ysearch, beyond the scope of this article.

K Type

K type has the largest cluster identified by this article, 54 samples in the Polish Project using 67 markers. The analysis is in the Supplementary Data, file "KType.xls." The definition modal haplotype, K34, uses 34 markers; the cutoff is at genetic distance step 4, with a gap of 1. A, P, I, N, and K types are color coded in the "KType.xls" analysis file for ease of comparison. SBP = 26%.

The other types in this article have SBP dominated by the statistical confidence associated with a small number of samples in the gap. K type, by contrast, provides a statistical background (12) only 50% higher than the estimated background (8), so the SBP will not reduce as much as more data accumulates in the future. In other words, K type, if a valid clade, probably has 10% to 26% background in the associated data cluster in the Polish Project.

K type distinguishes from P and N types without overlap. The P67 modal haplotype (P type at 67 markers) is at step 5 (using K34, cutoff 4). Five P type samples are also at step 5, the remainder are at steps 6 to 10. The P36 definition modal haplotype uses different markers, so it is not surprising that the P type samples are at or beyond the step of the P modal haplotype. Also, a modal haplotype drifts slowly with time because it is a best fit of many samples, so the K vs P modal haplotypes are not expected to be as far apart as most samples from

the two types. The N67 modal haplotype is at step 8 in the K34 mountain curve, and the closest N type sample is at step 7. The 28 samples on the far side of the mountain (K34 steps 9 to 13, the bottom of the Calculator sheet) include four P type, 22 N type, and two others. N is clearly farther away than P from K.

Conclusion: if N, P, and K are indeed valid clades, the node for N type is older than the nodes for P and K types

As additional evidence of validity, the types P and K reinforce each other. As noted above, some samples of the K34 mountain at step 5 fit P type well, so the gap for the K type mountain may be visualized as actually lower and wider than indicated by the SBP calculation, if P type samples are removed. Similarly, the gap for P type may be visualized as lower by removing the samples at or near the P gap that fit K type well. This consideration of the interaction between types is mentioned in the companion article, but I do not propose a quantitative adjustment of SBP along these lines, keeping the SBP definition simple. This paragraph is additional qualitative evidence for the consistency of P and K types.

The K modal haplotype can alternatively be interpreted as the modal haplotype for all of the 67-marker Polish Project R1a1. See the discussion above, in *Results; Signatures and Definitions of Types*. In fact, without effort to find a low SBP, most trial modal haplotypes using the K type marker values (different combinations of markers, but the same STR value for each marker) produce the full R1a1 mountain. The file "KTypeR1a1Example.xls" has all the Polish Project 67-marker data, using the K34 modal haplotype. The step frequency curve has two minima: one at step 4 with SBP = 26% reproducing the K34 result, and a second minima at step 14, gap 2, zero samples in the gap, SBP = 0.7%, cleanly separating a mountain with all the R1a1 from the R1b just beyond the mountain There is only one exception, a single R1b sample at the last step (13) of the R1a1 mountain.

K type seems to be the main R1a1 Y-DNA tree trunk. If this is indeed true in Poland and in the Polish Project data, then alternate K type modal haplotypes can be defined younger or older (fewer or more samples) than K34 by judicious choice of markers, and the SBP of each alternate provides an estimate of the background. I tried that. Most but not all alternates have high SBP. Each alternate is a cut off the top of this main tree trunk, and the background is mostly samples that are outliers from the branches of the main trunk just above (older than) the cut. K34 is a cut made in a region of the trunk with few branches.

K type SBP less than 10% can be found by selecting markers for a much larger cluster with more than 90 samples. These results are probably good candidates for older versions of K type. However, the low SBP is misleading. The mountain method is really intended for types that are a small fraction of the parent haplogroup. If K type is defined as more than half the R1a1 database, the cutoff comes on the down slope of the R1a1 mountain, and that down slope reduces the SBP artificially. I found a number of such solutions. One valuable comment from these solutions: they all have many but not all of the P type samples (background from the P branch), but few or none of the N type. This adds evidence that the P node is higher up the trunk (younger than N), providing background for alternates defined just above (younger than) the P node, or providing outliers in the gap for alternates defined at or below (older than) the P node.

Lower SBP can also be achieved by selectively removing markers. This is selection bias, so it should not be done except for markers that obviously have problems. I found no need to remove markers in the K type analysis. DYS DYS464 does not rank well for K type, so it automatically does not come up for the definitions.

The K type analysis does not converge: when markers are chosen by rank to form the definition modal haplotype, a new cluster is extracted; that new cluster provides a new ranking of markers and with the new ranking a slightly different set of markers provides the definition, from which yet another cluster is extracted, different near the cutoff. To minimize this effect, I used only the best fit 36 samples from a previous fit for the ranking to select the markers for the K34 modal haplotype. Basically, I stopped the iteration procedure with the best fit after a few hours of searching. This K34 fit has a breadth of 14 to 35 markers, and produces an identical mountain curve for 32 to 34 markers.

The samples at low step count (33 samples less than step 3) consistently come out in K type for just about any choice of well ranked markers. (This is the same comment as for N type above, but the variation in N type is less than in K type.) The values of the markers do not vary; the various trial modal haplotypes differ in the marker ranking, but marker value varies only for the poorly ranked markers, which are never used for a definition. All this variation is in the choice of which marginally ranked markers to use to distinguish K type from the rest of the Polish Project R1a1.

All six of the A type samples are in K type, at step zero or one. A type is clearly a subtype of K type. (This analysis is hypothetical, valid if the types A, I, and K are indeed clades.) All twelve of the I type samples are in K type, one at step 3, one at 2, and the other ten at steps 0 or 1. I type is a subtype of K type using 67 markers, although the I modal haplotype differs from the K modal haplotype at two of the first 12 standard markers.

Although the K type cluster has 54 samples, the mountain method estimates outliers as fewer (4) than the estimated foreign samples in the mountain (8), so the estimated size of K type is 50 samples, confidence interval 42 to 59.

The file "KTypeYsearch.xls" has the 67-marker data for all of R1a from Ysearch, 17 July 2009. Downloading from Ysearch is quite tedious; explanation notes are in the sheet "Download." Briefly, this file includes all Ysearch samples within 42 steps of K type at 67 markers, including R1* and some R1b outliers, and including many samples of "unknown" haplogroup. After editing for modal haplotypes and family sets, there are 702 samples. This *Excel* file was used for Ysearch analysis of all the types discussed in this article, because analysis is tedious using the Ysearch on-line tool. The results were verified using the Ysearch on-line tool as reported in Table 6 and reported with more detail in the file "YsearchURLs.xls." During on-line verification, comparison to "KTypeYsearch.xls" easily identifies new samples, modal haplotypes, and family sets. The K type Ysearch results in Table 6 are the 7 Aug 2009 on-line verification, with more details in "YsearchURLs.xls."

The K34 modal haplotype produces a mountain on Ysearch when restricted to Polish data, SBP = 33%. However, for all the Y search data, there is no K mountain. One reason is due to I type, which is part of K type, but concentrated in Poland. There may well be other Polish subtypes within K, as yet unidentified. There is another reason for no K34 mountain on Ysearch, below.

Table 6 compares the full Ysearch result to the result restricted to Eastern Europe. N type is slightly more concentrated in Eastern Europe (36/57 = 63%) than K type (93/160 = 58%), but the distinction is not statistically significant at this time.

As mentioned above, the K67 modal haploltype is essentially the modal haplotype for R1a1 in the Polish Project. K67 is similarly essentially the modal haplotype for the database in "KTypeYsearch.xls." "Essentially" means there are a few markers that differ, but these are rapid mutators or markers with bimodal distributions. The full comparison is in the "Haplotypes" sheet of "KTypeYsearch.xls."

Modal haplotypes for R1a1 had been entered into Ysearch by others. These are similar to K, differing only at some of those "non-essential" markers.

K type is not concentrated in Poland, so it is better studied using Ysearch data, including age, same as N type (comment above). Study of K type in the Polish Project is valuable for this article because I type seems to be a Polish subtype of K; more Polish subtypes will probably be discovered in K in the future.

L type is mentioned above as the second most common on Ysearch, with 116 samples, at 12 markers. Fourteen of these have 67-marker data, and using the K34 modal haplotype these 14 fall at steps 3 to 8; four of the 14 are at step 4, which is the K34 cutoff in the Polish Project. This explains why K type does not form a mountain outside Poland using the K34 modal haplotype. There is overlap with L type, which is rare in Poland. Further study of L type is beyond the scope of this article.

Many of the Polish Project R1a1 samples (men) are distant from each type. These are rare haplotypes individually, but they add up to a significant fraction of the database. No doubt most of these belong to small clades that branched from R1a1 before the nodes for P, N, and K. On the web site that I use to classify the Polish Project samples: http://www.gwozdz.org/PolishClades.html, I classify these remainder as "R," which is not a type. As data accumulates some small types may be discovered within R. Samples without enough markers for assignment, and samples that fall within the overlap between types, are unassigned, so I classify them as "U" in the Polish Project R1a1.

Comment added 4 Nov 2009: The web site mentioned in the previous paragraph now has a new small hypothetical subtype of K, dubbed B type. SBP = 14%. B type falls at the outer down slope of the K mountain, at step 3 (K cutoff is 4). The 10 B type samples have K step values of 3 and 5. This means the K type gap is better than it seems, in light B type, because many of the samples near the K cutoff are accounted for as solid members of B, and therefore could be removed in a new calculation of SBP. Similarly, B type, which has zero samples at the cutoff step 10 and zero at step 11, has exclusively K type samples at steps 12 through 16. In other words, two types can reinforce each other in apparent validity, if the solid members in one type are marginal members or outliers in the other type. In this case of B type, only one B type marginal sample (B step 9, K step 6) is not accounted for as a solid member of K type. This reinforcement does not depend upon the two types being closely related, because such statistical reinforcement to the validity of each will apply even if the two types are in fact distantly related, with similar STR values by coincidence.

Y Type

This is a good Polish mountain type, from Haplogroup I1. SBP = 9.2%. The eight samples from Table 1 form the type; they all have 67 markers. They are very well isolated from the 27 samples in Haplogroup I1 with 67-marker data; the breadth is 1 to 67; any number of markers can separate these eight samples from the I1 data. The best modal haplotype uses 52 markers with a cutoff 3, zero samples in the gap from 3 to 6.

That marker that resolves Y type by itself is DYS392=12, but no doubt with more data exceptions will turn up.

Ysearch also has a Polish "mountain" for Y type, but this is not an independent result, because the seven matches are the same men as six of these eight in the Polish Project (one man entered two samples). That Ysearch Polish mountain has a wide gap with zero samples; beyond the gap are more than 100 samples including only 2% Polish.

After I isolated the Y type data cluster from the Polish Project, Mayka (2009) pointed out that this cluster had been independently discovered before. One of the seven Ysearch matches seems to be a modal haplotype by Marek Skarbek-Kozietulski. See "YsearchURLs.xls" for the Ysearch codes. This cluster has been analyzed and posted on the web by Nordtvedt (2008), with the code name M253-P. For more data, see the FTDNA project "Normans-CE."

Y type seems to have selection bias, as discussed above, in *Results; Signatures and Definitions of Types*. During review of this article, I determined by email communication with Marek Skarbek-Kozietulski that the Y type men tested independently, but he convinced most of them to upgrade to all 67 samples. This means selection bias in the size of Y type in the 67-marker data of the Polish Project. Y type is not as large in the Polish population as it seems from the Polish Project results. The low SBP result is misleading in this case. Y type may nevertheless correspond to a small Polish clade. Regardless of validity, Y types serves as an example of how cluster analysis can be misleading when there is selection bias.

Z Type

Z type is speculative, but it is another example of how the mountain method can be used to glean hints from minimal data - only four well isolated samples in haplogroup N1c1. SBP = 37%. A "Notes" sheet is available with discussion in "ZType.xls." Z type seems concentrated in Lithuania.

G Type

G type looks just as Polish as P type in view of the Ysearch data at 12 markers (Table 1 - 43%). However, only two of those five in Table 1 have all 67 markers, and those two are far apart. Wait for more data.

Median Joining network

Mayka (2007) produced a median joining network for the Polish Project data with 37 or more markers in March 2007. The P type data does form a deep branch in that network, and the A type data falls into a small branch high in the tree. N type (as defined here) is not resolved as a distinct branch in that 2007 network.

Russian R1a1 Web Sites

The web site dnatree.ru, no longer active, in Russian Cyrillic, offered in 2008 hypothetical R1a1 subdivision trees and modal haplotypes for clusters. Most of the data appeared to be from the FTDNA web site. P type is not there. Haplotypes very similar K type and N type are listed.

Rodstvo.ru has recently included R1a1 trees and modal haplotypes, also in Russian Cyrillic.

Polish Nobles

It is obvious to wonder if the common Polish DNA types are more common in the nobility or less common. All kinds of archeological investigations come to mind, beyond the scope of this paper.

The FTDNA project "DNA-stia" is a "Central-European Nobility DNA Project" for self selected descendants of gentry from Poland and near-by countries. The download on 24 April 2009 has only 65 members, predominantly with "Poland" in the field for most distant ancestor. There are four P type at 12 markers, for 6.15%. (Those are four of 35 R1a1, 12.9%.) All four are Polish. Two have 67 markers, both with that DYS572=12 value for P type. The 70% confidence interval for those four samples is 2.0 to 7.3, 3.1% to 11.2%, greater than the Table 1 value of 2.4% in the Polish Project. The percent is higher if data is restricted to "Poland" samples. This very small preliminary sample seems to hint that P type was even more common in the Polish gentry than in Poland as a whole.

R1a1 in India

Ploski (2002) found that Polish Y-DNA is particularly homogeneous, based on Y-DNA STR variation. Homogeneity is evidence of founder effects and relatively recent Polish population expansion. This is consistent with the findings here that there are relatively young Y-DNA types in Poland.

As mentioned above, at this time most R1a are R1a1, and very few R1a1 have the known downstream SNP markers, so R1a is effectively comprised of R1a* plus R1a1*. R1a* seems to be absent in Poland.

Sharma (2009) found R1a1 to be relatively inhomogeneous in India. Sharma also reports for the first time finding populations where R1a* is relatively common. Sharma takes this as evidence that R1a1 may have originated in India. For STR analysis, Sharma uses six

markers, reporting R1a1 ASD ranges from 0.24 to 0.52 for various castes and tribes. For India as a whole, 509 samples, R1a1 ASD is 0.38.

The types in this article have much smaller ASD, meaning they are much younger, but these types are selected for STR correlation. File "R1a1Type.xls" in the **Supplementary Data** considers all 378 R1a1 from the Polish Project because all Polish Project data have Sharma's markers. The calculations of this section are highlighted in red on the "ASD" sheet. Average ASD using Sharma's six markers is 0.40.

So the Polish R1a1 seems to be slightly less homogeneous than Indian R1a1, with ASD 0.40 slightly larger than Sharma's 0.38, and therefore ASD age slightly older. This seems like a solid statistical result, because the number of samples and the number of total mutations is large. For all 67 markers, the December 2008 Polish Project average ASD was 0.388, and the current (May 2009) ASD is 0.383. However, the average Polish Project ASD using only the six Sharma markers decreased from December to May, 0.45 to 0.40. The biggest change: two samples appeared with DYS392 = 13, a value absent last December, raising ASD for that marker from 0.013 to 0.032.

However, any single ASD value may be misleading due to population structure. For these six markers in the Polish Project the ASD varies from 0.032 (DYS392 with six mutations) to 1.23 (DYS19 with 191 mutations). The corresponding ages using Sharma's rate come out from 1,146 years (459 years last December) to 44,671 years. That high value for DYS19 is expected, because DYS19 is a prime signature marker for P type, which produces high ASD at that marker. As explained in the companion article, "raw ASD age" is influenced very strongly by population structure. This application of ASD age is an example of strong population structure.

In other words, although Polish DNA is particularly homogeneous, Polish R1a1 comes out more inhomogeneous than Indian R1a1 using Sharma's six markers, because the most common type differs at one of these markers, and all six are relatively slowly mutating markers.

If there is a large, young R1a1 clade in India, it may also make Sharma's R1a1 reported ages too old due to population structure. On the other hand, if there are no R1a1 dominant young types in India, with STR distributions unusually close to random, Sharma's R1a1 reported ages may be too young.

Sharma uses a mutation rate of 0.00069 per 25 year generation, giving Zhivotovsky as a reference. Zhivotovsky, and the "factor" used for "equivalent mutation rate" are discussed in the companion article. The average father-son rate from Chandler for Sharma's six markers is 0.00201. This means a relative factor of 1 / 2.9 for Sharma, well within the range traditionally used. In other words, Sharma uses a typical mutation rate adjustment factor to account for population structure, a reasonable estimate. It would take more than six markers to improve Sharma's age estimate for R1a1 in India.

Applying Sharma's 1/2.9 factor to my spread sheet using Chandler's rates, the ages for all 67 markers in the Polish Project (file "R1a1Type.xls") range from 0 years (DYS472, DYS450,and DYS617 with no mutations in the respective 155, 154, and 155 samples because of a sample with a 13-marker panel not finished) to 79,214 years (DYS578 with 15 mutations out of 155 samples). DYS19 comes out second oldest, 59,197 years, this way because it has a low Chandler rate, and it defines P type.

DYS578 is the second slowest mutator of the 67 markers (actually less than 67, because of compound markers), according to the current on-line extension of Chandler's rates. There are only 15 mutated values of 578 in the Polish Project R1a1 data, and all 15 of those mutated values at DYS578 are 9, compared to 8 for the other 140 samples. There are no other values in the Polish Project R1a1. Two of the six A type have the 9 value, as do eight of the twelve I type, and five other samples. It seems the Polish Project data includes few independent mutations in this marker, perhaps only three. It is not surprising they are all plus 1, because markers with low STR count have a tendency to mutate up more than down, as explained in the companion article. It is too soon to judge the validity of my I type, which seems to be a type that experienced rapid population expansion in Poland. A type is likely valid, also with rapid population expansion. It seems that the Polish Project data includes few independent mutations at DYS578, but at least two of the mutations happened in types that experienced rapid population expansion carrying DYS578 = 9, contributing many mutated values in the data, creating an artificially old "raw ASD age" for this marker in R1a1 as a whole in the Polish Project. Or maybe there was migration with DYS578 = 9 from elsewhere, including A type. This paragraph is a simple example of the point made in the companion article, that in using ASD age calculations it is important to look at the individual markers for signs of extreme population structure. To summarize this discussion: inhomogeneity of STR data is very sensitive to which markers are chosen, because of population structure; for Polish R1a1 DYS578 comes out quite inhomogeneous, but that should be considered a hint of a subtypes with rapid population expansion in this case.

Sharma's data has been requested for further analysis. It will be interesting to see if any particular caste or tribe has a significant percent of P type (at those six markers).

The R1a Y-Haplogroup Project, mentioned at the end of the P type section above, has 46 samples segregated as "Indian subcontinent." All are R1a or R1a1, as expected in this project. None are P type at 12 markers. In fact, none are P type using only the first four markers. This is lacking evidence, not disproof, of significant P type in India.

Origin of R1a1

Although in the previous section I point out caveats in the use of ASD as a measure of diversity, Sharma's (2009) high diversity of R1a1 in India, plus his new finding of R1a* in India are good circumstantial, if not conclusive, evidence that R1a1 may have originated in India. Someone may find a pocket of even higher R1a* in a tribe in the Eurasian Steppes, saving the traditional opinion that R1a1 originated there. Or not. Time will tell.

Regardless, there is no reason to doubt the traditional view that R1a1 men migrated into Poland from elsewhere. The low ASD and therefore young ages of the Polish types reported here adds evidence of founder effects, which could be rapid population growth, or immigration, or both.

Polish Population Expansion

As shown above, it seems P type, from Haplogroup R1a1, went through a rapid population expansion somewhat less than 1,500 years ago in the area that is now Poland. Y type is a type from Haplogroup I1 that is also young, perhaps younger than P type, and also concentrated in Poland. Table 1 provides hints that more Polish types will be identified soon, as more data accumulates in the Polish Project.

It makes sense that a population expansion would not be confined to a single type, but might include other types and indeed other haplogroups, according to the population mix in the population experiencing the expansion. It is tempting to anticipate that additional data will point to multiple Polish types, of various sizes, with about the same population expansion time. If the data comes out that way, it will suggest the time of the growth of the Polish nation.

Conclusions: Polish Y-STR Haplotypes

Based on data from the Polish Project, Ysearch, Pawlowski (2002), and Yhrd, there is very high confidence that the P type haplotype at nine to twelve markers is concentrated in Poland. The concentration decreases with distance from Poland.

P type forms a very well isolated type using all 67 markers available in the Polish Project. The SBP is 6.7%. The implication is that P type is likely a Polish clade within R1a1. About 8% of Polish men belong to the hypothetical P type clade, confidence interval 6.4% to 9.6%. The TMCA is probably more than 2,000 years, but it seems P type experienced a significant population expansion somewhat less than 1,500 years ago, although there is uncertainty in those age estimates. Other young types seem to be concentrated in Poland, although there is not enough data yet to be confident these are valid clades.

Note Added in Proof: The New M458 SNP

A new SNP, M458, was recently announced by Underhill et. al. (2009) that splits R1a-M17. The corresponding STR data from the Underhill study is available. The Polish Project P type and N type match well to the M458 clade, dubbed R1a1a7. I find in the Underhill data respectable mountains for both P and N, although the number of markers is not sufficient for a low SBP. R1a1a*, the part of R1a1a that is negative for M458, is widespread in Eurasia, according to Underhill. I find that R1a1a* clearly contains K type and other non-Polish types not discussed in detail in this article. At this time, there are 29 results of M458 tests in the Polish Project (Larry Mayka, private communication), and all are consistent with (albeit not a proof of) the types introduced here: Of the 18 M458 who were found to be positive, eight had been assigned to P type, one had been noted as P borderline in the P type gap, seven were N type, and two uncertain. Of the eleven found negative, four had been assigned K type, five K borderline, and two were uncertain. Two of the uncertain samples, one positive and one negative for M458, had been assigned as "Remainder," meaning confidence that they are not part of one of the types identified so far. The web site www.gwozdz.org/R1aClades will provide updates on SNP results plus detailed analysis of the Underhill data.

Underhill reports that the highest coalescent time (age) for R1a1a7 is among Polish, 10,700 years. This calculation uses an average mutation rate 0.00069 per 25 years, which includes a factor of 1/3 to account for the stochastic reduction of variance in slowly growing populations, as demonstrated by Zhivotovsky (2006); see Part I section "Mutation Rates." Zhivotovski is a co-author of the Underhill paper. Zhivotovsky (2006) shows that for large populations, or for rapid growth, the mutation rate factor approaches one, which is to say that the coalescent time comes out up to 3 times younger. The coalescence time for R1a1a7 including P and N types may well be as much as 10,700 years ago. This is because P type and N type have quite different STR values. But many of the modern carriers of M458 in Poland come from two population expansions that are much more recent, because each type is much less diverse (lower ASD) than the total. Since both are large young types, the simplest explanation is that each grew recently from a small founding population. On the other hand, the two types may represent immigra-

tion of two tribes from distinct regions. Or the situation may be more complicated, with any number of immigrations and any number of population expansions, at different times. The important observation from this example: a clade may be composed of one or more daughter clades that are much younger than the parent. Indeed this is no surprise, since M458 is much younger than the parent R1a1a (R1a-M17) clade.

Underhill tentatively identifies M458 as a mutation from the Mesolithic, a reasonable conclusion. However, the corresponding haplogroup may not have grown much at first, or may have grown and then dwindled over the millenia. It appears there was a recent resurgence, from two or more small founding populations. It will be interesting to identify the cultures of those founders.

Supplementary Data.

In the Supplementary Data file for Part I there is an index with links to other on-line files. Those at the JoGG web site do not change, but support the article at the time of its publication. The directory includes tools, data, analysis, and detailed results for this article, and for the companion article. For similar information that is updated, see the author's web site:

http://www.gwozdz.org/PolishCladesUpdate.

Web Resources

Family Tree DNA Web Site
http://www.familytreedna.com

Polish DNA Project
http://www.familytreedna.com/public/polish/

YHRD Y-STR Haplotype Reference Database
http://www.yhrd.org

Ysearch Y-STR Database
http://www.ysearch.org

References

Behar DM, Garrigan D, Kaplan ME, Mobasher Z, Rosengarten D, Karafet TM, Quintana-Murci L, Ostrer H, Skorecki K, Hammer MF (2004) Contrasting patterns of Y chromosome variation in Ashkenazi Jewish and host non-Jewish European populations. *Hum Genet*, 114:354-365.

Biskupski M (2006) Who is a Pole and Where is Poland? *Rodziny - The Journal of the Polish Genealogical Society of America* (PGSA), Summer 2006:5-12.

Chandler JF (2006) Estimating Per-Locus Mutation Rates. *J Genet Geneal*, 2:27-33. See also the on-line extension to 67 markers .

Gwozdz P (2009) Y-STR mountains in haplospace, Part I: Methods. *J Genet Geneal*, 5:137-158.

Kayser M, Roewer K, Ploski R, et. al. - 32 authors total (2005) Significant genetic differentiation between Poland and Germany follows present-day borders, as revealed by Y-chromosome analysis. *Human Genetics*, 117:1432-1203.

Mayka L (2007) median-joining network. If that link does not work, look for update at the Polish Project, results page: www.familytreedna.com/public/polish/

Mayka L (2008) email communication. Mayka is administrator of the Polish Project.

Mayka L (2009) email communications. Mayka provided many helpful suggestions for this and the companion article, Polish Project data, literature references, and web references.

McDonald D (2005) World haplogroups map.

Nebel A, Filon D, Faerman M, Soodyall H, Oppenheim A (2005) Y chromosome evidence for a founder effect in Ashkenazi Jews. *Eur J Hum Genet*, 13:388-391.

Nordtvedt K (2008) Founders' haplotypes for Y Haplogroup I varieties and clades.

Pawlowski R, Dettlaff-Kakol A, Maciejewska A, Paszkowska R, Reichert M, Jezierski G (2002) Population genetics of 9 Y-chromosome STR loci w Northern Poland. *Arch Med Sadowej Kryminol*, 52(4):261-277.

Ploski R, Wozniak M, Pawlowski R, Monies DM, Branicki W, Kupiec T, Kloosterman A, Dobosz T, Bosch E, Nowak M, Lessig R, Jobling MA, Roewer L, Kayser M (2002) Homogeneity and distinctiveness of Polish paternal lineages revealed by Y chromosome microsatellite haplotype analysis. *Human Genetics*,10:592-600.

Roewer L, Croucher PJP, Willuweit S, Lu TT, Kayser M, Lessig R, de Knijff P, Jobling MA, Tyler-Smith C, Krawczak M (2005) Signature of recent historical events in the European Y-chromosomal STR haplotype distribution. *Hum Genet*, 116:279-291

Sharma S, Rai E, Sharma P, Jena M, Singh S, Darvishi K, Bhat AK, Bhanwer AJS, Tiwari PK, Bamezai RNK (2009) The Indian origin of paternal Haplogroup R1a1* substantiates the autochthonous origin of Brahmins and the caste system. *J Hum Genet*, 54:47-55.

Sliwinski RP (2007) Genetic Genealogy - A Polish-American Perspective of Y-DNA Testing. *Rodziny - The Journal of the Polish Genealogical Society of America* (PGSA) Spring 2007:5-12.

Thomas MG, Skorecki K, Parfitt T, Bradman N, Goldstein DB (1998) Origins of Old Testament Priests. *Nature*, 394:138-140.

Underhill PA, et. al, 34 authors (2009) Separating the post-glacial coancestry of European and Asian Y chromosomes within Haplogroup R1a. *Eur J Hum Genet*, online publication 4 November 2009. The STR data may be obtained at the following URL: http://www.nature.com/ejhg/journal/vaop/ncurrent/extref/ejhg2009194x4.pdf

Wiik K (2008) Where did European men come from? *J Genet Geneal*, 4:35-85.

Willuweit S, Roewer L (2007) Y chromosome haplotype reference database (YHRD): Update. *Foren Sci Int: Genetics*, 1, 83-87. See YHRD database URL in Web Resources.