Review

Eur J Hum Genet 1993;1:3–18

L.L. Cavalli-Sforza[a]
A. Piazza[b]

[a] Department of Genetics, Stanford University School of Medicine, Stanford, Calif., USA, and
[b] Dipartimento di Genetica, Biologia e Chimica Medica, Centro CNR per l'Immunogenetica e l'Istocompatibilità, Università di Torino, Italia

Human Genomic Diversity in Europe: A Summary of Recent Research and Prospects for the Future

Key Words
Genome
Evolution
Europe
Diversity
Blood groups
DNA polymorphisms
Neolithic
Agriculture
Linguistics

Abstract

Gene frequencies in Europe are intermediate with respect to those of other continents. A phylogenetic tree reconstructed from 95 gene frequencies tested on 26 European samples shows some deviant populations (Lapps, Sardinians, Greeks, Yugoslavs, Basques, Icelanders and Finns) and other weakly structured populations. This behavior may have a simple interpretation: Europeans have not evolved according to a tree of descent probably because of the major role played by migrations in prehistorical and historical times. The leading component of the European genetic landscape is a gradient that originates in the Middle East and is directed to the northwest. According to the hypothesis by Ammerman and Cavalli-Sforza this gradient was generated by a migration of Neolithic farmers from Anatolia followed by continuous, partial admixture of the expanding farmers with local hunter-gatherers. Other leading components of the gene frequencies in Europe show correlations with possible movements of Uralic-speaking people and pastoral nomads from a region north of the Caucasus and Black Sea, which according to Gimbutas is the area of origin of Indo-European speakers. This analysis is based on classical pre-DNA genetic markers. The prospect of future research using DNA polymorphisms is discussed in the context of the Human Genome Project.

Received: May 26, 1992
Revision received: August 2, 1992
Accepted: August 17, 1992

Professor Alberto Piazza
Università di Torino
Dipartimento di Genetica
Via Santena 19
I-10126 Torino (Italy)

Introduction

The study of individual variation is one of the principal areas of interest for geneticists. Variation is the key to evolution. In the eighties, genetic research shifted much of its attention from gene products to the genes themselves. With the introduction of techniques that allow the direct study of nucleic acids, and with continuous improvements to these techniques, it has become possible to study variation at the DNA level, but this avenue of research has been explored by the new techniques to only a very limited extent. In fact, at its inception, the *Human Genome Project* showed very little interest in variation [1].

In 1991 a project for testing human genomic diversity was proposed [2], and a HUGO committee was appointed by its President, Sir Walter Bodmer, with the aims of (a) planning a systematic study of genetic differences in suitably chosen samples of our species, and (b) saving, for future analysis, the DNA of a significant proportion of individuals representing current human diversity. The program will in part focus on populations that are currently vanishing either owing to a combination of low fertility and high mortality or because of a loss of identity due to acculturation, urbanization and migration. The human species is moving toward increasingly intensive amalgamation. The genetic differences between extant populations which are rapidly disappearing are an irreplaceable source of information for understanding our evolution. Genetic information is being obtained from fossil records [3], but if we have no bona fide modern data with which to compare it, this knowledge will be of little use.

In this paper we summarize the genetic knowledge about Europe derived from classical (pre-DNA) polymorphisms, and discuss some of its specific problems in the frame-

Table 1. Number of samples (A), mean gene frequencies × 1,000 (B), F_{ST} values × 1,000,000 (C), for the genetic loci and alleles in five continents

Locus	Allele	Europe		
		A	B	C
ABO	A	2,650	267	3,725
	A1	337	202	4,243
	A2	337	74	10,168
	B	2,650	84	11,230
	O	2,650	650	7,639
ACP1	B	182	620	6,951
	C	182	60	4,753
ADA	1	139	932	4,516
AK1	1	140	965	2,786
PI	M	85	947	73,215
AG	X	55	253	32,575
LPA	A	42	192	27,942
CHE1	U	54	981	4,135
CHE2	+	41	40	7,269
C3	S	58	810	5,233
FY	A	193	418	8,463
ESD	1	65	880	5,133
G6PD	def	125	81	57,250
GPT	1	37	531	2,484
BF	S	35	711	40,291
	F	35	218	9,422
	F1	35	57	132,245
	SO.7	35	12	3,375
GLO1	1	54	413	5,621
GC	1	257	724	9,981
	1F	42	146	4,642
HP	1	410	380	3,463
	1S	28	231	18,643
HLA-A	1	131	141	8,345
	2	131	280	3,764
	3	131	134	8,434
	9	131	121	5,724
	10	131	57	6,500
	11	131	61	2,493
	19	75	126	24,401
	23	42	22	8,032
	25	44	20	10,284
	28	131	38	3,200
	29	109	40	11,338
	30	59	33	27,456
	31	62	24	25,523
	32	102	37	8,342
	33	55	15	10,292
HLA-B	5	132	82	16,728
	7	132	109	16,877
	8	132	89	13,576
	12	132	127	10,862
	13	132	29	5,624

Africa			Asia			America			Australia		
A	B	C	A	B	C	A	B	C	A	B	C
733	176	17,499	1,453	205	9,569	393	81	174,785	39	220	74,608
189	119	23,404	294	168	19,378	243	65	205,618	12	235	96,080
189	52	8,932	294	29	18,688	243	3	20,489	12	0	0
733	129	11,486	1,453	198	4,183	393	23	72,334	39	22	78,091
733	694	14,523	1,453	596	5,289	393	896	216,358	39	758	55,022
94	783	63,020	238	702	64,956	111	674	273,201	18	978	7,785
94	3	19,131	234	6	22,331	111	1	9,586	18	0	0
24	995	10,230	120	913	34,547	38	998	14,222	11	988	29,004
67	981	37,016	185	953	42,209	63	997	21,305	12	1,000	0
13	968	12,561	32	984	12,872	3	954	74,222	1	1,000	0
8	99	4,516	7	704	90	7	380	214,838	3	551	36,715
3	353	164,782	3	175	97,021	5	91	27,563			
17	985	12,901	25	985	88,403	27	982	25,197	2	997	5,003
11	18	24,585	40	36	7,657	19	32	16,473			
7	872	46,979	49	884	68,228	6	971	30,024			
92	110	215,173	172	596	284,041	234	700	87,083	7	988	84,739
29	873	49,799	118	742	50,382	78	797	107,538	8	913	66,690
199	104	73,161	195	85	127,361	54	2	9,503	6	0	0
22	806	53,576	52	534	35,331	17	531	44,969	11	795	63,311
16	380	73,728	27	725	72,921	7	964	37,524			
16	534	95,581	27	255	33,782	7	28	29,214			
16	44	4,582	24	4	9,353	7	1	1,608			
16	39	20,208	25	18	93,207	7	5	2,286			
28	292	13,377	52	237	57,542	31	280	72,272	9	25	39,311
65	872	46,444	184	774	40,863	107	749	116,445	10	832	51,698
19	602	173,144	46	409	220,289	18	336	35,128	3	346	21,556
174	549	70,111	383	247	28,696	261	512	137,767	16	275	47,345
2	213	27,181	8	186	28,599	1	239		2	229	36,599
18	59	47,851	55	50	56,942	34	8	20,091	6	0	0
18	166	10,586	55	213	30,661	34	370	112,400	6	157	4,677
18	79	19,402	55	46	48,433	34	6	9,835	6	0	0
18	109	6,616	55	257	93,352	34	308	208,748	6	267	2,696
17	63	27,556	55	75	38,491	34	5	35,255	6	306	149,871
17	28	5,018	55	125	73,419	34	7	43,443	6	16	75,580
17	104	17,624	49	30	27,445	33	103	124,307	6	0	0
17	53	20,174	43	13	28,439	27	1	3,699	3	0	0
12	137	37,223	25	22	28,374	28	11	42,393	3	5	4,854
13	50	77,014	23	38	23,739	27	149	131,384	3	1	4,441
16	33	57,185	37	20	35,285	27	11	45,833	3	0	0
12	57	26,058	26	56	43,064	27	18	65,425	3	0	0
17	65	28,361	55	140	35,997	34	123	64,437	6	0	4,549
18	101	33,282	55	44	15,653	34	11	50,628	6	0	0
18	46	17,848	55	19	27,138	34	5	14,522	6	0	0
18	92	21,537	55	52	19,175	34	7	15,891	6	0	0
17	15	8,427	55	39	24,867	34	1	7,299	6	84	89,928

work of the human genome diversity program. Europeans form more than one-tenth of the world's population [4] and have a different social structure compared to much of the rest of the world.

A Genetic Picture of Europe

We review here the main conclusions of research into the genetic history of Europe which are part of a forthcoming book [5], where detailed references will be found.

Europeans Are Genetically More Homogeneous

Compared with the aborigines of other continents, Europeans are more homogeneous. Genetic data collected from the available literature are summarized in table 1: genetic loci, alleles, number of collected samples, mean gene frequencies, F_{ST} values [6] in Europe and, for comparison, in Africa, Asia, America and Australia are listed. The mean gene frequencies in Europe are intermediate with respect to those of other continents, but this may be in part an artifact, because almost all polymorphisms were first detected in Europeans. F_{ST} values are the variances of gene frequencies among populations divided by the variances due to sampling [6]: they measure standardized between-population genetic differences. They are lower in Europe than in other parts of the world, and this is initial evidence for the genetic homogeneity of Europeans when compared with the populations of other continents.

On the basis of the genetic loci and alleles listed in table 1 and by eliminating several less tested populations, genetic distances between 26 selected populations were calculated according to the coancestry coefficient of Reynolds et al. [7]. The populations were Austrian, Basque, Belgian, Czechoslovakian,

Table 1 (continued)

Locus	Allele	Europe		
		A	B	C
HLA-B	14	132	33	7,072
(continued)	15	132	55	12,284
	16	117	37	11,496
	17	132	42	5,736
	18	129	67	27,053
	21	129	34	12,815
	22	130	21	1,700
	27	132	39	6,270
	35	132	109	15,289
	37	57	11	1,114
	40	131	51	10,273
	41	37	10	7,903
IGHG1G3	za;g	157	183	10,352
	zax;g	156	79	15,219
	f;b	157	727	15,318
	za;b	11	13	11,923
	za;b0stb3b5	29	12	17,655
	fa;b	94	10	91,731
IGKC(KM)	1&1,2	108	81	5,503
KEL	K	315	42	3,310
JK	A	60	499	4,983
LE	Le	73	663	42,747
	Le(a+)	25	196	
LU	A	70	26	5,297
MNS, GYPA	M	577	573	4,466
	S	172	329	5,742
	MS	161	247	5,183
	Ms	161	325	7,538
	NS	161	83	9,925
	Ns	161	344	10,322
P1	1	201	493	14,146
PTC	T	139	511	41,851
PGM1	1	207	749	8,878
PGD	A	94	972	7,521
RH	D	1,287	617	7,822
	C	362	472	18,566
	E	362	142	6,001
	Du	52	15	12,723
	CDE	362	4	15,272
	CDe	362	451	18,107
	Cde	362	16	5,634
	cDE	362	131	6,504
	cDe	362	29	7,713
	cdE	362	6	5,301
	cde	362	362	13,515
SE, FUT2	Se	122	539	17,937
TF	C	118	991	5,994
Mean				14,200
Median				8,389

Table 1 (continued)

Africa			Asia			America			Australia		
A	B	C	A	B	C	A	B	C	A	B	C
17	29	13,604	55	12	19,208	34	8	53,551	6	0	0
17	26	22,314	49	102	64,740	34	115	159,809	6	95	46,752
12	31	19,909	45	47	43,135	32	132	124,119	3	5	4,137
17	155	65,138	49	58	30,914	34	3	9,858	6	0	0
17	37	10,325	49	23	30,850	34	3	15,215	6	0	0
17	51	80,804	50	25	65,012	32	15	76,330	6	0	0
17	15	11,724	49	37	23,561	34	7	74,224	6	249	76,047
17	15	23,497	55	21	12,032	32	38	86,104	6	0	0
18	52	25,453	49	96	28,258	34	203	173,500	6	0	4,549
5	18	17,417	22	13	13,196	12	1	4,785			
17	51	43,964	55	138	75,627	34	187	147,099	6	421	55,271
6	5	1,869	11	7	28,977	9	0	0			
97	85	133,132	185	303	195,003	111	744	108,952	26	668	68,002
45	17	16,545	185	93	54,941	111	167	162,105	26	240	26,657
75	117	371,466	93	386	313,057	61	25	125,390			
104	601	271,964	27	122	271,403	16	19	33,950	26	87	311,238
			100	117	139,346	84	70	114,955			
31	2	1,720	159	367	547,734	101	15	27,082			
79	340	16,385	129	188	65,224	80	353	77,228	14	295	49,616
94	21	38,865	163	29	58,951	190	6	29,391	6	1	724
29	662	104,754	70	474	58,823	182	457	65,665	2	515	132,244
18	435	140,755	23	557	160,445	124	550	84,270	3	692	37,298
57	42	29,557	57	14	34,410	56	3	36,749	2	0	0
249	521	22,229	517	611	33,678	342	712	66,907	40	274	58,163
110	185	30,151	204	238	124,046	236	284	92,321	27	0	2,272
108	125	44,930	195	159	83,311	233	229	78,508	24	0	0
108	404	27,051	195	445	56,185	233	481	68,443	24	314	51,141
108	60	15,123	195	76	61,427	233	57	52,773	24	0	0
108	379	24,073	195	320	91,038	233	233	80,394	24	686	51,141
106	666	66,376	211	351	111,996	242	444	120,046	22	209	60,691
38	679	86,195	178	519	116,670	45	744	124,014	1	298	
92	820	48,221	52	534	35,331	124	851	56,124	18	853	73,394
116	946	42,152	211	944	24,193	80	974	62,365	13	950	13,734
392	778	40,085	624	824	42,858	324	960	113,586	30	996	124,437
258	135	166,084	385	637	75,227	284	571	87,260	30	698	41,152
258	74	28,378	385	159	73,725	284	403	74,576	30	252	69,032
143	78	54,045	27	46	56,133	2	70	78,215			
258	2	22,322	385	10	41,713	284	42	59,682	30	54	29,592
258	121	183,769	385	608	76,524	284	524	87,718	30	639	32,931
258	12	19,396	385	19	33,862	284	4	75,252	30	4	124,437
258	69	24,642	385	141	68,219	284	357	75,992	30	198	76,197
258	592	275,073	385	63	60,491	284	46	70,212	30	105	56,317
258	3	21,162	385	7	57,252	284	3	43,564	30	0	0
257	200	54,047	385	151	117,369	284	23	95,046	30	0	0
23	536	27,463	129	501	48,645	65	926	357,815	2	890	181,863
111	961	48,691	284	985	25,920	118	977	94,433	18	875	36,036
		52,020			66,800			77,540			39,290
		27,320			43,060			69,330			29,000

Table 2. Genetic distances (× 10,000) of 26 European populations (lower triangle), and their SEs (upper triangle)

	Bas.	Lap.	Sar.	Aus.	Cze.	Fre.	Ger.	Pol.	Rus.	Swi.	Bel.	Dan.	Dut.
Basque		101	68	39	32	22	37	27	29	37	22	33	22
Lapp	629		105	55	92	54	49	56	56	64	66	53	55
Sardinian	261	667		70	77	64	72	74	52	81	71	78	66
Austrian	195	308	294		14	9	10	39	22	6	4	8	20
Czechoslovakian	159	470	327	36		18	15	24	25	8	13	11	17
French	93	350	283	38	72		5	13	11	6	7	8	7
German	169	314	331	19	52	27		9	16	2	4	4	7
Polish	146	395	282	72	64	66	47		8	15	14	16	12
Russian	140	323	266	64	75	59	60	30		13	13	12	11
Swiss	165	375	353	12	31	23	10	60	78		5	3	4
Belgian	107	333	256	16	43	32	15	40	51	14		5	3
Danish	184	334	348	27	54	43	16	69	80	19	21		2
Dutch	118	341	307	38	66	32	16	54	57	16	12	9	
English	119	404	340	55	60	24	22	70	79	28	15	21	17
Icelandic	221	494	396	153	173	146	106	144	169	115	78	88	101
Irish	145	557	393	115	117	93	84	150	160	86	75	68	76
Norwegian	195	317	424	61	76	56	21	58	90	33	24	19	21
Scottish	146	447	357	74	104	62	53	121	129	59	59	40	48
Swedish	168	333	371	80	90	78	39	82	110	55	34	36	41
Greek	231	308	190	86	126	131	144	177	161	148	103	191	199
Italian	141	339	221	43	77	34	38	64	75	44	30	72	64
Portuguese	145	324	340	48	46	48	51	65	98	53	31	77	60
Spanish	104	452	295	69	65	39	69	117	122	43	42	80	76
Yugoslavian	176	565	294	110	101	124	118	137	170	120	50	157	136
Finnish	236	210	334	77	175	107	77	139	153	112	63	96	123
Hungarian	153	338	279	40	69	70	46	25	30	57	52	78	71

Danish, Dutch, English, Finnish, French, German, Greek, Hungarian, Icelandic, Irish, Italian, Lapp, Norwegian, Polish, Portuguese, Russian, Sardinian, Scottish, Spanish, Swedish, Swiss, Yugoslavian. The matrix of genetic distances is given in table 2, and a phylogenetic tree reconstructed by average linkage [8] from the distances in table 2 is shown in figure 1.

By applying to this tree the bootstrap technique [9], one finds that Lapps are outliers in 76% of the resampled trees; they are replaced as the most extreme outlier by Sardinians 18% of the time; Sardinia is the next outlier in 63% of the bootstrapped trees in which Lapps are first.

There is a group of five other extreme outliers: Greeks, Yugloslavs, Basques, Icelanders and Finns. All of them are separated from the other populations of the tree in $> 50\%$ of the bootstrapped trees. The remaining populations form a series of small groups, all of which are geographically close neighbors which appear clustered in the same configuration in $< 50\%$ of the bootstrapped trees. If one draws a consensus tree among all the bootstrapped trees by applying what is commonly called the 'majority rule' (i.e. an agree-

prehistorical

Eng.	Ice.	Iri.	Nor.	Sco.	Swe.	Gre.	Ita.	Por.	Spa.	Yug.	Fin.	Hun.	
22	33	27	41	26	25	48	26	30	15	53	38	30	Basque
68	112	109	55	85	63	50	50	59	83	230	36	52	Lapp
70	66	87	88	80	76	30	54	79	58	86	73	69	Sardinian
26	37	28	34	16	32	13	11	15	17	105	13	9	Austrian
15	36	30	22	22	24	19	16	10	13	71	41	22	Czechoslovakian
5	24	20	10	13	16	28	6	14	9	90	17	12	French
6	24	21	6	10	18	31	6	15	17	95	11	8	German
15	36	32	16	21	29	75	11	15	20	125	31	7	Polish
16	47	40	18	23	30	49	24	28	29	148	49	9	Russian
9	24	18	10	13	14	27	7	16	13	91	19	14	Swiss
6	18	19	6	12	11	27	6	7	11	16	13	17	Belgian
6	20	15	9	9	15	29	14	16	14	99	25	17	Danish
5	26	18	6	10	16	42	16	18	19	93	31	15	Dutch
	16	7	6	6	10	56	8	9	9	99	21	16	English
76		24	18	25	26	68	27	27	34	136	41	44	Icelandic
30	99		22	6	21	57	26	27	22	91	50	35	Irish
25	74	79		14	12	62	15	17	18	106	19	18	Norwegian
27	111	29	58		15	42	16	18	18	116	31	26	Scottish
37	106	94	18	74		57	18	16	15	107	16	31	Swedish
204	288	289	235	253	230		29	25	33	133	32	16	Greek
51	143	132	88	112	95	77		12	15	88	19	11	Italian
46	149	115	73	97	78	103	44		17	87	19	11	Portuguese
47	163	113	97	100	99	162	61	48		84	27	27	Spanish
160	317	272	173	248	213	213	119	139	172		141	123	Yugoslavian
115	157	223	94	166	82	150	94	119	159	248		29	Finnish
70	172	152	77	124	99	88	61	63	118	136	115		Hungarian

ment of $> 50\%$), the consensus applies only to the outlier populations and not to the core structure of the tree. This lack of robustness, confirmed by the test of treeness [10], has a simple interpretation: the European populations have not evolved according to a tree of descent. A basic assumption for giving the tree a phylogenetic meaning is that each of its branches evolves independently from the others. This could be true for very distant or isolated populations, but it is a very unrealistic model in the case of Europe where migrations in prehistorical and historical times are known to have played a major role.

The Genetically Most Deviant Populations

For the seven outlier populations mentioned above, whose special genetic structure is confirmed by the statistical technique of bootstrapping, it is possible to formulate hypotheses explaining their genetic isolation.

The most deviant population is *Lapp* (Saame), limited numbers of whom live in the extreme north of Scandinavia, in four countries (Norway, Sweden, Finland and Russia). There are seven or more Lapp groups distinct by territory, dialect, and preferential occupation (fishing, reindeer breeding, and others). Although they are heavily mixed with Scandi-

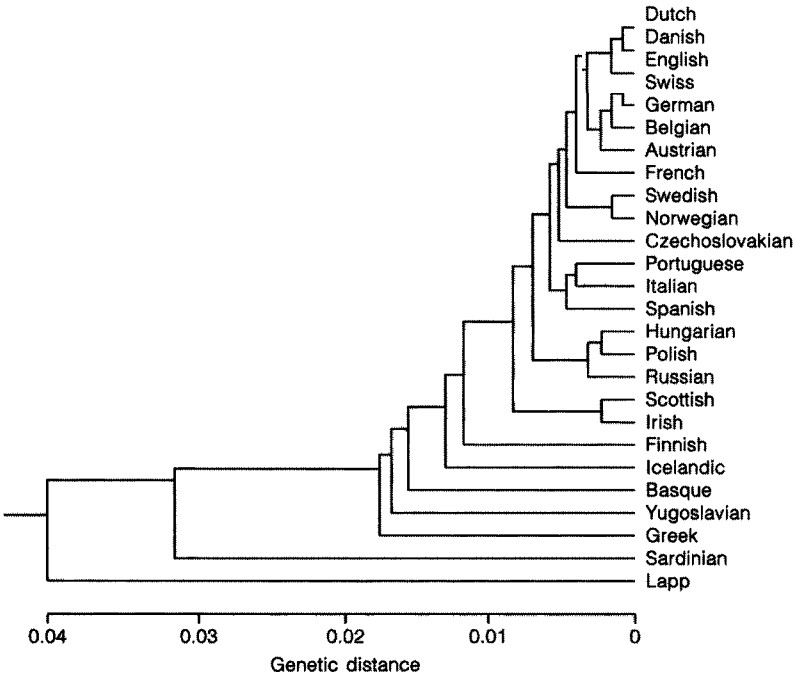

Fig. 1. Genetic tree of 26 European populations. Genetic distances are based on the allele frequencies shown in table 1. Details on the data and methods can be found in Cavalli-Sforza et al. [5].

navians and are by now, on average, less pigmented than southern Europeans [11] they have variable external phenotypes, and a fraction of them retain a phenotype characteristic of northern Siberian people, in particular Samoyed, who speak a language of the same family (Uralic). Classical polymorphisms show Lapps to be an admixture, in which European genes predominate, but genes in common with people from the Uralic region may reach between 20 and 50% [12]. This could be a reasonable hypothesis for their outlier nature in the genetic tree.

Sardinians are the next outliers. The island was settled at least 10,000 years ago [13] and the local population had reached substantial numbers (200,000 and more) 3,000 years ago [14], before other foreign colonizers, Carthaginians, arrived, especially in the south of the island. There were no Greek settlers at the time of Greek colonization of the western Mediterranean. The Roman occupation had little genetic consequence; however, it did affect the language, which is of Latin origin, with substrata of earlier, probably non-Indo-European languages [15]. Later settlers from Italy and Spain had a local and limited influence on certain coastal regions. The earliest founders may have been pre-Neolithic; later contacts and migrations were sufficiently few so that the population may have undergone considerable genetic drift and is therefore

rather different from other populations. Some genetic similarity with more direct descendants of Paleolithic people from Europe, such as Basques or Caucasians [16], indicate that the first immigrants may have been Paleolithic, and the first archaeological date available for the island is in agreement with this notion. The arrival of Neolithic farmers (who originally came from the Middle East) at some unknown but probably early date, and genetic contributions from both Phoenicians and Carthaginians may help to explain why Sardinians show a genetic resemblance to the Lebanese, second but only by a small amount to that of Italians, who are their closest geographic neighbors, and who have contributed to the island's colonization since late Roman times.

Among the group of the less extreme outliers, *Basques* are probably the most direct descendants of the earliest post-Neanderthal settlers of Europe. They are easily distinguishable from neighbors because of their unique language, which has no known relative in Europe except for languages of the North Caucasian family [17]. Caucasian is also believed to have some ancient relationship with the American Na-Dene and the East Asian Sino-Tibetan families [18]. If so, this group of languages may be a very ancient superfamily that spread widely east and west in northern Asia and Europe during the Paleolithic, perhaps at the time of the replacement of Neanderthals in Europe around 40,000 years ago. It is very difficult to find traces of a common origin in languages after such a long time, therefore the matter is inevitably highly speculative. The fact remains that Basque is still spoken in southwestern France and northeastern Spain, and toponymy shows it was spoken over a wider area at an earlier time [19]. Our analysis shows that there is a marked genetic similarity among the people living today in the regions corresponding to the French and Spanish areas with Basque toponymy, and that these areas show the greatest difference from the rest of western Europe, which was probably settled later. It is presumably not a coincidence that the same area shows the greatest concentration of cave art in the upper Paleolithic. Basques probably maintained an endogamy that was not very rigid but still sufficiently so to conserve a distinctive genotype that was probably diluted to some extent by later admixture with new neighbors, beginning especially in Neolithic times. Conservation of a distinct language must have been an important factor in maintaining social and genetic identity. It is very likely that the genetic uniqueness of Basques is a product of the remarkable isolation of western from eastern Europe at the time of the last glaciation, which peaked around 18,000 years ago, and there may have been very limited genetic and cultural exchanges between the two halves of Europe during the early Paleolithic [20].

Another interesting outlier population shown by the tree analysis is *Iceland.* According to tradition, this island was settled in the 9th century by an estimated 20,000 Norwegians from the middle of Norway [21]. There were very few later immigrants. The Norwegian origin has been confirmed [22] in spite of earlier indications of greater similarity of Icelanders to the Scots and Irish. This misunderstanding was partly caused by earlier analyses which relied on genetic systems like ABO which are more sensitive than most others to environmental changes, and by the considerable similarity of people from northern Ireland and Scotland to Scandinavian Vikings, who settled the coasts of these countries before the colonization of Iceland. Moreover, Icelandic cattle have their genetic origin in mid-Norway. Iceland had a small population for a long time and there are still only 200,000 people today. It is the smallest European

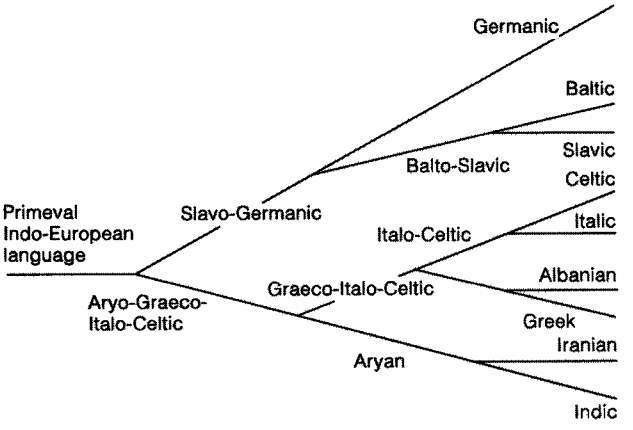

Fig. 2. The family tree model for the Indo-European languages proposed by Schleicher [23].

country and it probably owes its outlier position to having been subjected to much more drift and less immigration than most other European countries.

Finns, Greeks and Yugoslavs are also outliers in the tree for reasons which may correlate with their linguistic isolation. *Finns* speak a Uralic language [see also ref. 12] of the Ugro-Finnic family which also includes Hungarians; the latter, however, do not seem to segregate clearly from other East Europeans even if their genetic relationships with Uralic populations from West Siberia have been shown [12]. *Greeks* speak a language which is an almost entirely separate branch of Indo-European languages (as in the classical tree by Schleicher [23] represented in fig. 2; see also [24]). *Yugoslavs* (Southern Slavs) are also linguistically separated and genetically rather heterogeneous. Each of these populations has a complex history which may help in understanding their genetic patterns. Historical explanations are untestable by experiment, but future research may add more evidence in favor or against using them in genetic interpretations even if totally unambiguous conclusions can never be reached.

Genetic and Linguistic Trees of Descent

The remaining populations of figure 1 tend to associate in clusters in which one might discover frequent affinities of a linguistic nature. The first cluster from below includes the two *Celtic*-language-speaking samples (*Scots* and *Irish*). The second includes two of the four *Slavic*-speaking populations (*Russians* and *Poles*); *Hungarians* have been mentioned above; the third *(Czechoslovakians)* is isolated in the tree probably because of its intermediate geographic position (it is located between *Slavic-* and *German*-speaking people and was part of the Austrian empire for some time). *Spaniards, Portuguese* and *Italians* are grouped together as southwestern Europeans speaking a *Romance* language. *Swedes* and *Norwegians* are associated both geographically and linguistically. The Romance-speaking *French,* genetically rather heterogeneous, seem intermediate between this last group, the southwestern Europeans and the Saxons. The *Saxon* samples are grouped in two subclusters: the 'northern' subcluster is made up of *Dutch, Danish,* and *English* people, the 'central' subcluster of *Austrian, Swiss, Germans,* and *Belgians.* All of them speak Ger-

manic languages but Belgium and Switzerland are linguistically divided.

It is quite interesting to compare our genetic tree with the historical classical tree of Indo-European languages proposed by Schleicher [23] and shown in figure 2. In addition to some similarity in the two clustering structures, note the analogy between the isolation of Greek- and Germanic-speaking people in both trees.

On the world scale, most Europeans show relatively few differences. Averaging over all genes or taking a more robust statistics, the median, Europe has the lowest F_{ST} of all continents (see the last two rows of table 1). European 'races' descriptions between 1850 and 1950 were based on a few anthroposcopic features, such as skin, hair and eye color, stature, and cephalic index, and reflected the strong correlation between pigmentation and climate, most probably because skin pigmentation is an adaptive response to the intensity of solar radiation. Other genes also tend to show some differences between northern and southern Europeans, but they offer no confirmation of the clustering of Europeans according to any of the old anthropological classifications of European races [summarized e.g. in ref. 25, 26].

Trees of Descent and 'Synthetic Maps'

Trees have relatively little use in areas which, like Europe, have had a very active genetic exchange generating a network more than a tree-like structure, as shown by the low average F_{ST} value and the results of bootstrap tests. The use of other methods for reconstructing trees also seems rather unrewarding in the case of Europe for the same reason. The average linkage tree we showed, however, has the advantage of pointing to populations which have some degree of uniqueness, i.e. are sufficiently different from neighbors, most probably because of isolation and drift, so that they stand out as outliers. Not surprisingly, they are almost all geographically peripheral.

Other methods can be more informative than trees for describing the relationships among European populations, e.g. principal component analysis. This is a linear transformation of the observed gene frequencies (in geometrical terms a rotation of the coordinate axes) whose coefficients are chosen so as to maximize the variation of the transformed data (having as coordinates new principal component values or principal coordinates) measured along each new coordinate axis (principal component). Principal components are ranked according to the fraction of total variation each of them can independently explain: for instance the 'first' principal component (explaining say 30% of the total gene frequency variation) is by definition more informative than the 'second' principal component (explaining $< 30\%$ of the total gene frequency variation), and all principal components are independent of one another [for the method see ref. 27, and for applications to genetic data see ref. 5]. A widely employed graphical display is that of the first two highest ranking component values of the populations: in a Cartesian diagram abscissa and ordinate represent the first and second principal component axes, respectively, and each point places a population in this transformed space. An interesting property of this representation is that the geometric distance between any pair of populations represents the 'true' multidimensional genetic distance with the least possible error. Figure 3 shows this kind of presentation calculated from the genetic distance matrix of table 2.

In 1978 we showed how a quite different application of the principal component analysis, namely geographic contour mapping of the highest component values by smoothing the original gene frequency data over the

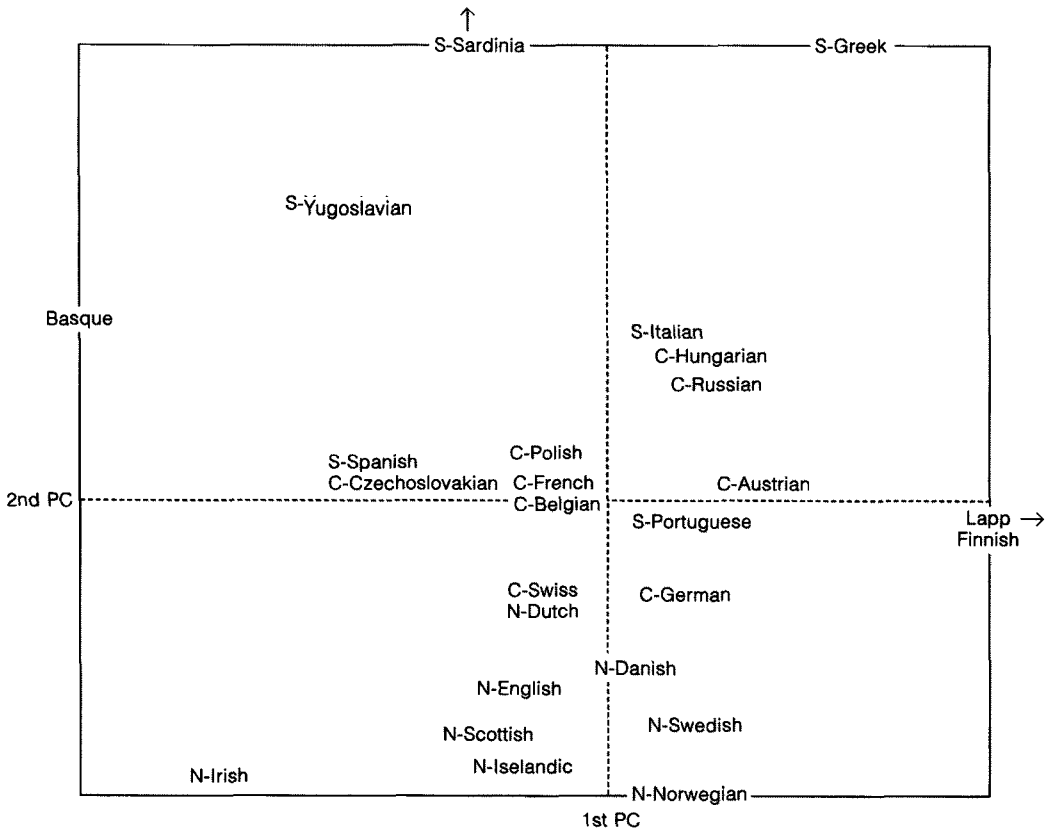

Fig. 3. Principal component (PC) map of European populations. N = Northern, C = central, S = southern.

whole geographical surface, could resolve the genetic picture of Europe into a number of 'genetic landscapes' each of which most probably summarizes a particular set of events and thus highlighted a particular historical scenario [28]. Of the first three principal component representations shown in that research, each of which corresponded to a geographical contour map of principal component values, or, as we called them, 'synthetic maps', the map representing the first principal component agreed very closely with that expected from the demographic expansion of Neolithic farmers. This had been previously hypothesized [29, 30] on the basis of archaeological information. A simulation of the process of expansion of Neolithic farmers in Europe with compatible growth rates confirmed this interpretation of the genetic data [31, 32].

The Major Genetic Landscapes of Europe

The leading component of the European genetic landscape is a gradient originating in the Middle East and directed to the north-

west. It has been confirmed [33, 34] that this gradient was generated by a migration of Neolithic farmers from Anatolia, directed west along the coast of the Mediterranean and northwest via the Balkans and central Europe to France, England, and Scandinavia. The genetic gradient is the result of continuous, partial admixture of the expanding farmers with local hunter-gatherers. There is a very high correlation ($r = 0.792 \pm 0.051$) between the value of the first principal component shown in the synthetic map, and the archaeological map of the first arrival of farming in over 100 archaeological sites radiocarbon-dated in Europe. The first PC explains 28.1% of the genetic variation of Europe.

The second principal component explains 20% of the total genetic variation and shows a clear north-south gradient, with a peak where Lapps are living today. This component is highly correlated with latitude, and hence probably with temperature. It is worth noting, however, that it is also correlated with a partition of Europe into two linguistic families, Indo-European and Uralic. We are currently testing the statistical significance of the two findings, but it is quite possible that both correlations are correct. Lapps have up to 48% of genetic admixture with Eastern Uralic-speaking people [12]. People speaking Uralic languages may have spread westwards along the Arctic coast from an unknown area of origin: today Samoyeds who are perhaps the most representative speakers of Uralic languages live not far from the Arctic Ocean east of the Urals.

The third principal component explains 10.6% of the total genetic variation and shows a correlation with a possible expansion of pastoral nomads from a region north of the Caucasus and Black Sea, which, according to Gimbutas [35], is the area of origin of Indo-European speakers [Piazza A, et al., unpubl. results]. Barbujani and Sokal [36] found a correlation between linguistic and genetic boundaries in Europe. In the majority of cases (22 out of 33) there were also physical barriers that may be the cause of both genetic and linguistic boundaries. In 9 cases there were only linguistic and genetic boundaries but not physical ones. It remains to be established if in these cases or in some of them linguistic boundaries have generated or enhanced the genetic ones, or if both are the consequence of political, cultural and social boundaries (as in the case of Lapps and non-Lapps) that have played a role similar to that of physical barriers.

Much more of the demographic history and prehistory of Europe can be understood by an accurate study of its human population genetics. This knowledge may contribute greatly to archaeological, historical, and linguistic information, and the Gimbutas study from all these perspectives will be especially illuminating. Modern genetic techniques have brought analysis to an unprecedented degree of sophistication, and the knowledge from nongenetic disciplines that can enhance the understanding of the history of human evolution is more developed in Europe than on any other continent. This is the time to join forces and take full advantage of the current trend towards cooperation among Europeans.

Testing Human Genome Diversity in Europe

Europe is in a special situation with respect to other continents. Most aboriginal groups outside Europe are difficult to reach, and many are likely to become extinct soon. This does not hold for most European populations, and it is questionable whether one should establish transformed immortalized cell lines from individuals of most European countries. Taking a sample of 50 individuals from each of, say, 25 countries, immortalization would

require a basic expense in the order of a million dollars, but this would not provide a sufficiently detailed data base. As most Europeans show few genetic differences, samples of 50 individuals may rarely give statistically significant results in intra-European comparisons unless a very large number of genes is tested: the high genetic homogeneity of Europe puts a high cost on the study of local genetic variation.

In Europe there are, however, much cheaper ways of obtaining DNA from large numbers of individuals other than by immortalization, e.g. via a good network of well-equipped blood banks. In the countries where this holds, DNA can be obtained from blood donors. This would be a nonrenewable resource, but samples from the same blood donors might be obtained again, and in most cases there would be no scientific loss using DNA from different individuals of the same population, after a first set of DNA samples from a given population is exhausted.

One may nevertheless want to immortalize some populations, especially the outliers mentioned above, which are of special interest and some of which may not be so easily accessible through blood banks, for example the most important, least acculturated groups of Lapps.

Blood donors have already been used to extract DNA from a population of Celtic origin in Trino Vercellese, Italy [37], as well as from many European samples collected during the 11th Histocompatibility Workshop [38]. The transformed cultures collected in Bergamo, Italy, by Ferrara et al. [unpubl. data] are also produced by blood donations. They have the advantage that white cells in toto can also be used for analysis of immunoglobulin regions [39, 40] whereas transformed B lymphocytes may be unsatisfactory.

It would be important to explore the feasibility of testing human genome diversity in each European country. Local committees should be formed, including geneticists who have experience and interest in testing DNA samples by molecular techniques, blood bank experts, and a group of specialists in history, archaeology, cultural anthropology, ethnography and linguistics. This would also be an interesting experiment of collaboration between the 'two cultures' where the 'other' culture group should suggest geographical areas and populations worthy of special attention. Major criteria in this choice are historical, archaeological and linguistic information on the origin of the people, and the need to avoid cities as sources of samples, concentrating instead on rural and mountain areas, i.e. on regions where recent immigration is exceptional. Ideally, one would like to select individuals whose four grandparents are from the area of interest. Where this information is impossible to obtain, an appropriate analysis of surnames could supply useful selection criteria.

Sample Sizes and Selection of DNA Markers

It is not necessary to sample all individuals in a single village but just a few from each village belonging to the same region worth examining for that particular population. The sample size should be around 70–100 individuals who should be unrelated or distantly related (less than first cousins).

A large, standard sample of markers must be tested in all populations, both in Europe and the rest of the world, otherwise comparisons will not be possible. One of the major difficulties in using published data for classical markers is the diversity of markers tested by different research workers. Fortunately, a few very informative markers have been analyzed fairly systematically and by a homogeneous

set of reagents in a number of populations: the HLA system is the best known example [41], but some earlier identified genetic systems like GM have also been tested extensively [42]. The need to examine as many markers as possible, and to use standard testing conditions can only be achieved by widespread agreement among researchers. It is probably superfluous to add that the burgeoning 'molecular paleontology' cannot give meaningful results unless it is accompanied by the accumulation of control data from modern Europeans and other populations.

There is at the moment much innovative activity in the development of testing and sequencing methods, as well as increasing opportunities for automated research. Only a few sets of DNA markers are sufficiently well established, but information is beginning to accumulate on many populations. It is possible that for some populations, like many E-Sokal ropean ones that are extremely similar and have therefore only a short differentiation history limited to the last few millennia, markers with high mutation rates (like CA repeats) may be especially valuable [43], but it is necessary to test if these can be used for comparisons between genetically more unrelated populations.

Acknowledgements

This work was supported by National Institute of Health GMS20467, by MURST (Italy), and by CNR Projects 'Biological Archive' and 'Biotechnology and Bioinstrumentation'.

References

1 Cavalli-Sforza LL: How can one study individual variation for three billion nucleotides of the human genome? Am J Hum Genet 1990;46: 649–651.
2 Cavalli-Sforza LL, Wilson AC, Cantor CR, Cook-Degan RM, King MC: Call for a worldwide survey of human genetic diversity: A vanishing opportunity for the Human Genome Project. Genomics 1991;11: 490–491.
3 Pääbo S: Ancient DNA: Extraction, characterization, molecular cloning, and enzymatic amplification. Proc Natl Acad Sci USA 1989;86:1939–1943.
4 McEvedy C, Jones R: Atlas of World Population History. New York, Penguin, 1978.
5 Cavalli-Sforza LL, Menozzi P, Piazza A: History and Geography of Human Genes. Princeton, Princeton University Press, in press.
6 Wright S: The genetical structure of populations. Ann Eugen 1951;15: 323–354.
7 Reynolds J, Weir BS, Cockerham CC: Estimation of the coancestry coefficient: Basis for a short-term genetic distance. Genetics 1983; 105:767–779.
8 Sokal RR, Michener CD: A statistical method for evaluating systematic relationship. Univ Kansas Sci Bull 1958;38:1409–1438.
9 Efron B: The Jacknife, Bootstrap, and Other Resampling Plans. Philadelphia, Society for Industrial and Applied Mathematics, 1982.
10 Cavalli-Sforza LL, Piazza A: Analysis of evolution: Evolutionary rates, independence and treeness. Theor Popul Biol 1975;8:127–165.
11 Eriksson AW, Lehmann W, Simpson NE: Genetic studies on circumpolar populations; in Milan FA (ed): The Human Biology of Circumpolar Populations. Cambridge, Cambridge University Press, 1929.
12 Guglielmino-Matessi CR, Piazza A, Menozzi P, Cavalli-Sforza LL: Uralic genes in Europe. Am J Phys Anthropol 1990;83:57–68.
13 Spoor CF, Sondaar PY: Human fossils from the endemic island fauna of Sardinia. J Hum Evol 1986;15: 399–408.
14 Lilliu G: La civiltà dei Sardi dal neolitico all'età dei Nuraghi. Torino, Radio Italiana, 1983.
15 Contini M, Capello N, Griffo R, Rendine S, Piazza A: Géolinguistique et géogénétique: Une démarche interdisciplinaire. Géolinguistique 1989;4:129–197.
16 Piazza A, Cappello N, Olivetti E, Rendine S: The Basques in Europe: A genetic analysis. Munibe Antropologia-Arqueologia 1988;6:168–176.
17 Trombetti A: Le origini della lingua basca. Bologna, Forni, 1925.
18 Ruhlen M: A Guide to the World's Languages. Stanford, Stanford University Press, 1987.
19 Bernard J, Ruffié J: Hématologie et culture: Le peuplement de l'Europe de l'ouest. Ann Ecol Sup Prat Hautes Etudes 1976;661–676.

20 Soffer O, Gamble C (ed): The World at 18,000 BP. London, Unwin Hyman, 1990.
21 Bjarnason O, Biarnason V, Edwards JH, Fridriksson S, Magnusson M, Mourant AE, Tills D: The blood groups of Icelanders. Ann Hum Genet 1973;36:425–455.
22 Wijsman EM: Techniques for estimating genetic admixture and applications to the problem of the origin of the Icelanders and the Ashkenazi Jews. Hum Genet 1984;67:441–448.
23 Schleicher A: Die Darwinsche Theorie und die Sprachwissenschaft. Weimar, 1863.
24 Mallory JP: In Search of the Indo-Europeans: Language, Archaeology and Myth. London, Thames and Hudson, 1989.
25 Coon CS: The Origin of Races. New York, Knopf, 1963.
26 Garn SM: Human Races. Springfield, Thomas, 1971.
27 Hotelling H: Analysis of a complex of statistical variables into principal components. J Educ Psychol 1933; 24:417–441, 498–520.
28 Menozzi P, Piazza A, Cavalli-Sforza LL: Synthetic maps of human gene frequencies in Europe. Science 1978;201:786–792.
29 Ammerman AJ, Cavalli-Sforza LL: A population model for the diffusion of early farming in Europe; in Renfrew C (ed): The Explanation of Culture Change. London, Duckworth, 1973.
30 Ammerman AJ, Cavalli-Sforza LL: Neolithic Transition and the Genetics of Populations in Europe. Princeton, Princeton University Press, 1984.
31 Sgaramella-Zonta L, Cavalli-Sforza LL: A method for the detection of a demic cline; in Morton NE (ed): Genetic Structure of Populations. Honolulu, University of Hawaii Press, 1973, pp 128–135.
32 Rendine S, Piazza A, Cavalli-Sforza LL: Simulation and separation by principal components of multiple demic expansions. Am Naturalist 1986;128:681–706.
33 Sokal RR, Oden NL, Wilson C: Genetic evidence for the spread of agriculture in Europe by demic diffusion. Nature 1991;351:143–145.
34 Sokal RR, Oden NL, Wilson C: Patterns of population spread. Nature 1992;355:214.
35 Gimbutas M: The civilization of the Goddess: The world of old Europe. San Francisco, Harper, 1991.
36 Barbujani G, Sokal RR: Zones of sharp genetic change in Europe are also linguistic boundaries. Proc Natl Acad Sci USA 1990;87:1816–1819.
37 Griffo RM, Cappello N, Danese P, Mangione AM, Matullo G, Turco E, Trucco M, Piazza A: The village of Rigomagus (Trino Vercellese), Italy: A settlement of Celtic origin?; in Tsuji K, Sasazuki T, Juii T, Aizawa M (ed): Histocompatibility Testing 1991. Oxford, Oxford University Press, 1992.
38 Tsuji K, Sasazuki T, Juii T, Aizawa M (ed): Histocompatibility Testing 1991. Oxford, Oxford University Press, 1992.
39 Migone N, Feder J, Cann H, Van West B, Hwang J, Takahashi N, Honjo T, Piazza A, Cavalli-Sforza LL: Multiple DNA fragment polymorphisms associated with immunoglobulin μ-switch-like regions in man. Proc Natl Acad Sci USA 1983; 80:467–471.
40 Johnson MJ, Natali AM, Cann HM, Honjo T, Cavalli-Sforza LL: Polymorphisms of a human variable heavy chain gene show linkage with constant heavy chain genes. Proc Natl Acad Sci USA 1984;81:7840–7844.
41 Dausset J, Colombani J (ed): Histocompatibility Testing 1972. Copenhagen, Munksgaard, 1973.
42 Steinberg AG, Cook CE: The Distribution of the Human Immunoglobulin Allotypes. Oxford, Oxford University Press, 1981.
43 Jeffreys AJ, Neumann R, Wilson V: Repeat unit sequence variation in minisatellites. A novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. Cell Mol Neurobiol 1990;60:473–485.