Supplementary data:

Table S1 - List of 80 SNPs used for SNP typing the dataset in study.

Position[a]	Locus	Posit (bp)	Posit (aa)	Alleles (bp)	Alleles (aa)	S or NS	Reference
2532	Rv0002	481	161	C/T	Leu/Leu	S	(Dos Vultos et al., 2008)
5671	Rv0005	549	183	C/T	Tyr/Tyr	S	(Kasai and Ezaki, 2000)
5752	Rv0005	630	210	A/G	Val/Val	S	(Kasai and Ezaki, 2000)
6307	Rv0005	1185	395	G/T	Ala/Ala	S	(Niemann et al., 2000)
6406	Rv0005	1284	428	C/T	Asn/Asn	S	(Kasai and Ezaki, 2000)
7539	Rv0006	238	80	A/G	Thr/Ala	NS	(Hershberg et al., 2008)
9304	Rv0006	2003	668	A/G	Gly/Asp	NS	(Hershberg et al., 2008)
37031	Rv0034	165	55	C/G	Ala/Ala	S	(Filliol et al., 2006)
43945	Rv0041	384	128	A/G	Val/Val	S	(Filliol et al., 2006)
92199	Rv0083	1800	600	G/T	Thr/Thr	S	(Filliol et al., 2006)
157292	Rv0129c	309	103	A/G	Glu/Glu	S	(Hershberg et al., 2008)
220050	Rv0189c	1674	558	A/G	Leu/Leu	S	(Filliol et al., 2006)
311613	Rv0260c	1047	349	A/C	Val/Val	S	(Filliol et al., 2006)
519806[b]	Rv0432	207	69	A/G	Ala/Ala	S	(Filliol et al., 2006)
720863	Rv0629c	870	290	A/C	Ala/Ala	S	(Dos Vultos et al., 2008)
726703	Rv0631c	1604	535	G/T	Arg/Met	NS	(Dos Vultos et al., 2008)
797736	Rv0697	804	268	C/T	Leu/Leu	S	(Filliol et al., 2006)
909166	Rv0815c	153	51	A/G	Leu/Leu	S	(Filliol et al., 2006)
918316	Rv0824c	435	145	A/G	Gln/Gln	S	(Filliol et al., 2006)
923065	Rv0831c	645	215	A/T	Gly/Gly	S	(Filliol et al., 2006)

949221	Rv0852	663	221	C/T	Asp/Asp	S	(Filliol et al., 2006)
1047683	Rv0938	1548	516	G/T	Leu/Phe	NS	(Dos Vultos et al., 2008)
1068151	Rv0956	591	197	C/T	His/His	S	(Filliol et al., 2006)
1139222	Rv1020	256	86	A/G	Ala/Thr	NS	(Dos Vultos et al., 2008)
1163134	Rv1040c	243	81	A/G	Gly/Gly	S	(Filliol et al., 2006)
1178116	Rv1056	489	163	C/T	Thr/Thr	S	(Filliol et al., 2006)
1294398	Rv1165	231	77	C/T	Asp/Asp	S	(Filliol et al., 2006)
1477588	Rv1316c	44	15	C/G	Thr/Ser	NS	(Dos Vultos et al., 2008)
1479085	Rv1317c	34	12	A/G	Ile/Val	NS	(Dos Vultos et al., 2008)
1548149	Rv1375	318	106	A/G	Pro/Pro	S	(Filliol et al., 2006)
1588456	Rv1411c	27	9	C/T	Arg/Arg	S	(Hershberg et al., 2008)
1595342	Rv1420	1301	434	C/T	Val/Ala	NS	(Dos Vultos et al., 2008)
1692069	Rv1501	180	60	A/G	Ala/Ala	S	(Filliol et al., 2006)
1692685	Rv1501	796	266	A/C	Arg/Arg	S	(Filliol et al., 2006)
1830295	Rv1628c	267	89	C/T	Phe/Phe	S	(Filliol et al., 2006)
1884697	Rv1662	2994	998	A/G	Gly/Gly	S	(Filliol et al., 2006)
1892017	Rv1665	792	264	C/T	His/His	S	(Filliol et al., 2006)
1920120	Rv1696	438	146	G/T	Pro/Pro	S	(Dos Vultos et al., 2008)
1960391	Rv1733c	97	33	C/T	Pro/Ser	NS	(Hershberg et al., 2008)
2134215	Rv1884c	47	16	A/G	His/Arg	NS	(Hershberg et al., 2008)
2156025	Rv1908c	87	29	A/C	Pro/Pro	S	(Filliol et al., 2006)
2223682	Rv1980c	348	116	C/G	Val/Val	S	(Filliol et al., 2006)
2239349	Rv1996	346	116	A/G	Ala/Thr	NS	(Hershberg et al., 2008)
2278376	Rv2030c	111	37	C/T	Asp/Asp	S	(Hershberg et al., 2008)

2376135	Rv2115c	156	52	C/T	Ser/Ser	S	(Filliol et al., 2006)
2603797	Rv2330c	426	142	C/T	Ile/Ile	S	(Hershberg et al., 2008)
2627946	Rv2349c	753	251	C/T	Arg/Arg	S	(Filliol et al., 2006)
2643653	Rv2362c	606	202	C/T	Gly/Gly	S	(Dos Vultos et al., 2008)
2825581	Rv2510c	1509	503	A/C	Ile/Ile	S	(Filliol et al., 2006)
2880702	Rv2560	628	210	C/G	Val/Leu	NS	(Filliol et al., 2006)
2891267	Rv2567	1473	491	C/T	Gly/Gly	S	(Filliol et al., 2006)
2990040	Rv2673	750	250	C/T	Phe/Phe	S	(Filliol et al., 2006)
3207250	Rv2897c	693	231	A/T	Leu/Leu	S	(Filliol et al., 2006)
3300104	Rv2949c	467	156	A/G	Arg/His	NS	(Hershberg et al., 2008)
3300196	Rv2949c	375	125	C/T	Phe/Phe	S	(Hershberg et al., 2008)
3312632	Rv2959c	207	69	A/G	Trp/stop	NS	(Hershberg et al., 2008)
3332254	Rv2976c	501	167	A/G	Leu/Leu	S	(Dos Vultos et al., 2008)
3335708	Rv2979c	41	14	C/G	Pro/Arg	NS	(Dos Vultos et al., 2008)
3426795	Rv3062	1212	404	C/G	Ser/Ser	S	(Dos Vultos et al., 2008)
3438386	Rv3075c	588	196	C/T	Ile/Ile	S	(Filliol et al., 2006)
3440464	Rv3077	924	308	G/T	Arg/Arg	S	(Filliol et al., 2006)
3440542	Rv3077	1002	334	A/G	Gly/Gly	S	(Filliol et al., 2006)
3450725	Rv3084	729	243	C/T	Val/Val	S	(Filliol et al., 2006)
3455686	Rv3088	1347	449	C/G	Leu/Leu	S	(Filliol et al., 2006)
3544710	Rv3176c	591	197	A/G	Pro/Pro	S	(Filliol et al., 2006)
3572834	Rv3200c	836	279	A/C	Thr/Asn	NS	(Dos Vultos et al., 2008)
3597737	Rv3221c	30	10	A/G	Val/Val	S	(Hershberg et al., 2008)
3641447	Rv3261	905	302	C/T	Thr/Met	NS	(Hershberg et al., 2008)

3681548	Rv3297	229	77	A/C	Arg/Arg	S	(Dos Vultos et al., 2008)
3783058	Rv3370c	1719	573	C/T	Ser/Ser	S	(Filliol et al., 2006)
4024273	Rv3581c	75	25	A/G	Val/Val	S	(Filliol et al., 2006)
4081987	Rv3644c	735	245	C/G	Ala/Ala	S	(Dos Vultos et al., 2008)
4081996	Rv3644c	726	242	C/G	Pro/Pro	S	(Dos Vultos et al., 2008)
4119246	Rv3679	471	157	C/T	Asp/Asp	S	(Filliol et al., 2006)
4137829	Rv3695	624	208	C/T	Ala/Ala	S	(Filliol et al., 2006)
4156239	Rv3711c	491	164	C/T	Val/Ala	NS	(Dos Vultos et al., 2008)
4156503	Rv3711c	227	76	A/G	Gly/Asp	NS	(Dos Vultos et al., 2008)
4182695	Rv3731	938	313	A/G	Arg/His	NS	(Dos Vultos et al., 2008)
4254006	Rv3798	1014	338	G/T	Leu/Leu	S	(Filliol et al., 2006)
4255922	Rv3799c	27	9	C/T	His/His	S	(Filliol et al., 2006)

[a] - SNP position in the reference strain H37Rv.

[b] - SNP discarded from the analyzes since it showed more than 5% of no-calls.

Table S2 - List of strains from Heshberg et al. 2008 used for identification of strain groups.

Name	SNP lineage	Strain group	TBDB database[a]	NCBI database[b]
canettii	Outgroup	-	Present	Present
95_0545	Lineage 1	EA1	Present	-
K21	Lineage 1	EA1	Present	-
K67	Lineage 1	EA1	Present	-
K93	Lineage 1	EA1	Present	-
T46	Lineage 1	EA1	-	Present
EAS054	Lineage 1	EA1	-	Present
T17	Lineage 1	EA1	Present	Present
94_M4241A	Lineage 2	-	-	Present
02_1987	Lineage 2	Beijing	-	Present
00_1695	Lineage 2	Beijing	Present	-
210	Lineage 2	Beijing	-	Present
T85	Lineage 2	Beijing	Present	Present
91_0079	Lineage 3	CAS	Present	-
SG1	Lineage 3	CAS	Present	-
K49	Lineage 3	CAS	Present	-
H37Rv	Lineage 4	T	Present	Present
4783_04	Lineage 4	Cameroon	Present	-
Haarlem	Lineage 4	Haarlem	Present	Present
F11	Lineage 4	LAM	Present	Present
GM_1503	Lineage 4	LAM	Present	Present
K37	Lineage 4	Uganda	Present	-
C-strain	Lineage 4	X	Present	Present
CDC1551	Lineage 4	X	Present	Present
5444_04	Lineage 5	AFRI2	Present	-
11821_03	Lineage 5	AFRI2	Present	-
CPHL_A	Lineage 5	AFRI2	-	Present
4141_04	Lineage 6	AFRI1	Present	-
GM_0981	Lineage 6	AFRI1	Present	-
K85	Lineage 6	AFRI1	-	Present
bovis	Animal	-	-	Present
BCG	Animal	-	-	Present

[a] - sequence information obtained from http://genome.tbdb.org/.

[b] - sequence information obtained from http://www.ncbi.nlm.nih.gov/.

Table S3 - Characterization of the SNPs used in the study in regards to presence in one or several strain groups. Dataset analyzed consisted of strains from Hershberg et al. (2008).

SNP ID[a]	Group-specific[b]	Intra-group[c]	Supra-group[d]	Reference
Rv0002_0481			Lineage 2, 3 and 4	
Rv0005_0549		Animal[e]		(Bouakaze et al., 2010)
Rv0005_0630	Animal			(Bouakaze et al., 2010)
Rv0005_1185		Animal[e]		(Bouakaze et al., 2010)
Rv0005_1284	Animal			(Bouakaze et al., 2010)
Rv0006_0238	Uganda			(Comas et al., 2009)
Rv0006_2003	T[f]			(Comas et al., 2009)
Rv0034_0165	T[f]			
Rv0041_0384			Lineage 2, 3 and 4	
Rv0083_1800	T[f]			
Rv0129_0309	LAM			(Comas et al., 2009)
Rv0189_1674			Haarlem and X	
Rv0260_1047	T[f]			
Rv0629_0870	EAI			
Rv0631_1604	LAM			
Rv0697_0804		Beijing		
Rv0815_0153	Beijing[g]			
Rv0824_0435	X			
Rv0831_0645			Haarlem and X	
Rv0852_0663		X		
Rv0938_1548		LAM		
Rv0956_0591	T[f]			
RV1020_0256	EAI			
Rv1040_0243	T[f]			
Rv1056_0489	T[f]			
Rv1165_0231		Animal		(Filliol et al., 2006)
Rv1316_0044	Haarlem			(Comas et al., 2009)
Rv1317_0034			Lineage 4	
Rv1375_0318			Animal and Lineage 6	(Filliol et al., 2006)
Rv1411_0027		LAM		
Rv1420_1301			Lineage 2 and 3	
Rv1501_0180		Beijing		
Rv1501_0796		X		
Rv1628_0267	Animal			(Filliol et al., 2006)
Rv1662_2994	X[g]			
Rv1665_0792		Beijing		
Rv1696_0438			Lineage 2, 3 and 4	
Rv1733_0097	X			
Rv1884_0047		LAM		
Rv1908_0087	Animal			(Filliol et al., 2006)
Rv1980_0348		X		
Rv1996_0346		X		
Rv2030_0111		X[h]		

Rv2115_0156		Beijing[g]		
Rv2330_0426	X			(Comas et al., 2009)
Rv2349_0753			Lineage 2 and 3	
Rv2362_0606	EAI			(Abadia et al., 2010)
Rv2510_1509		Beijing		
Rv2560_0628			Cameroon, T and Uganda	
Rv2567_1473	T[f]			
Rv2673_0750		X		
Rv2897_0693	Animal			(Filliol et al., 2006)
Rv2949_0467	Cameroon			(Comas et al., 2009)
Rv2949_0375	Uganda			(Comas et al., 2009)
Rv2959_0207		LAM		
Rv2976_0501	Haarlem			
Rv2979_0041			Lineage 4	
Rv3062_1212	LAM			(Abadia et al., 2010)
Rv3075_0588			Animal and Lineage 6	(Filliol et al., 2006)
Rv3077_0924	T[f]			
Rv3077_1002			Lineage 2, 3 and 4	
Rv3084_0729			Cameroon, LAM, T and Uganda[g]	
Rv3088_1347			Cameroon, LAM, T and Uganda	
Rv3176_0591			Haarlem and X	
Rv3200_0836		EAI [e]		
Rv3221_0030	X			(Comas et al., 2009)
Rv3261_0905		X[h]		
Rv3297_0229			Lineage 4	
Rv3370_1719			Haarlem and X	
Rv3581_0075	T[f]			
Rv3644_0735	EAI			
Rv3644_0726	EAI			
Rv3679_0471			Lineage 2, 3 and 4	
Rv3695_0624		Beijing		
Rv3711_0491			Lineage 2, 3 and 4	
Rv3711_0227			Lineage 4	
Rv3731_0938	T[f]			
Rv3798_1014			Lineage 2 and 3[g]	
Rv3799_0027	T[f]			

[a] - locus ID and SNP position within the locus.

[b] - unique to a family and present in all strains.

[c] - unique to a family but not present in all strains.

[d] - unique to more than one family.

[e] - monomorphic in the reference set but polymorphic in the cited study.

[f] - dataset contains only one strain of T group, it is not possible to distinguish between group-specific

and intra-group SNPs in this group.

[g] - reference set with missing data for this SNP.

[h] - monomorphic in the reference set but polymorphic in the cited study and in the considered dataset.

Table S4 - Identification of genotype geno_7 using SNPs relevant to Lineage 4 (except for Uganda and Cameroon).

SNP ID[a]	Observed bp[b]	Group-specific[c]	Supra-group[d]	Strain group[e]
Rv0129_0309	G	LAM (A)		not LAM
Rv0189_1674	G		Haarlem and X(A)	not Haarlem nor X
Rv0631_1604	G	LAM(T)		not LAM
Rv0824_0435	A	X(G)		not X
Rv0831_0645	A		Haarlem and X(T)	not Haarlem nor X
Rv1316_0044	C	Haarlem (G)		not Haarlem
Rv1733_0097	C	X(T)		not X
Rv2330_0426	C	X (T)		not X
Rv2560_0628	C		T (G)	not T
Rv2976_0501	G	Haarlem (A)		not Haarlem
Rv3062_1212	C	LAM (G)		not LAM
Rv3084_0729	C		LAM and T (T)	not LAM nor T
Rv3088_1347	C		LAM and T (G)	not LAM nor T
Rv3176_0591	A		Haarlem and X (G)	not Haarlem nor X
Rv3221_0030	G	X (A)		not X
Rv3370_1719	C		Haarlem and X (T)	not Haarlem nor X

[a] - locus ID and SNP position within the locus.

[b] – base pair of the genotype considered.

[c] - unique to a family and present in all strains.

[d] - unique to more than one family.

[e] - identification of the strain group of the genotype based on the considered SNP.

Table S5 - Identification of genotype geno_17 using SNPs relevant to Lineage 4 (except for Uganda and Cameroon).

SNP ID[a]	Observed bp[b]	Group-specific[c]	Supra-group[d]	Strain group[e]
Rv0129_0309	G	LAM (A)		not LAM
Rv0189_1674	G		Haarlem and X(A)	not Haarlem nor X
Rv0631_1604	G	LAM(T)		not LAM
Rv0824_0435	A	X(G)		not X
Rv0831_0645	A		Haarlem and X(T)	not Haarlem nor X
Rv1316_0044	C	Haarlem (G)		not Haarlem
Rv1733_0097	C	X(T)		not X
Rv2330_0426	C	X (T)		not X
Rv2560_0628	C		T (G)	not T
Rv2976_0501	G	Haarlem (A)		not Haarlem
Rv3062_1212	C	LAM (G)		not LAM
Rv3084_0729	T		LAM and T (T)	LAM or T
Rv3088_1347	G		LAM and T (G)	LAM or T
Rv3176_0591	A		Haarlem and X (G)	not Haarlem nor X
Rv3221_0030	G	X (A)		not X
Rv3370_1719	C		Haarlem and X (T)	not Haarlem nor X

[a] - locus ID and SNP position within the locus.

[b] – base pair of the genotype considered.

[c] - unique to a family and present in all strains.

[d] - unique to more than one family.

[e] - identification of the strain group of the genotype based on the considered SNP.

Table S6 - Identification of genotype geno_31 using SNPs relevant to Haarlem and X.

SNP ID[a]	Observed bp[b]	Group-specific[c]	Supra-group[d]	Strain group[e]
Rv0189_1674	A		Haarlem and X(A)	Haarlem or X
Rv0824_0435	G	X(G)		not X
Rv0831_0645	T		Haarlem and X(T)	Haarlem or X
Rv1316_0044	C	Haarlem (G)		not Haarlem
Rv1733_0097	C	X(T)		not X
Rv2330_0426	C	X (T)		not X
Rv2976_0501	G	Haarlem (A)		not Haarlem
Rv3176_0591	G		Haarlem and X (G)	Haarlem or X
Rv3221_0030	G	X (A)		not X
Rv3370_1719	T		Haarlem and X (T)	Haarlem or X

[a] - locus ID and SNP position within the locus.

[b] – base pair of the genotype considered.

[c] - unique to a family and present in all strains.

[d] - unique to more than one family.

[e] - identification of the strain group of the genotype based on the considered SNP.

Table S7- Identification of genotype geno_37 using SNPs relevant to Lineage 4 (except for Uganda and Cameroon).

SNP ID[a]	Observed bp[b]	Group-specific[c]	Supra-group[d]	Strain group[e]
Rv0129_0309	G	LAM (A)		not LAM
Rv0189_1674	G		Haarlem and X(A)	not Haarlem nor X
Rv0631_1604	G	LAM(T)		not LAM
Rv0824_0435	A	X(G)		not X
Rv0831_0645	A		Haarlem and X(T)	not Haarlem nor X
Rv1316_0044	C	Haarlem (G)		not Haarlem
Rv1733_0097	C	X(T)		not X
Rv2330_0426	C	X (T)		not X
Rv2560_0628	G		T (G)	T
Rv2976_0501	G	Haarlem (A)		not Haarlem
Rv3062_1212	C	LAM (G)		not LAM
Rv3084_0729	T		LAM and T (T)	LAM or T
Rv3088_1347	G		LAM and T (G)	LAM or T
Rv3176_0591	A		Haarlem and X (G)	not Haarlem nor X
Rv3221_0030	G	X (A)		not X
Rv3370_1719	C		Haarlem and X (T)	not Haarlem nor X

[a] - locus ID and SNP position within the locus.

[b] – base pair of the genotype considered.

[c] - unique to a family and present in all strains.

[d] - unique to more than one family.

[e] - identification of the strain group of the genotype based on the considered SNP.

References

Abadia, E., Zhang, J., dos Vultos, T., Ritacco, V., Kremer, K., Aktas, E., Matsumoto, T., Refregier, G., van Soolingen, D., Gicquel, B., Sola, C., 2010. Resolving lineage assignation on Mycobacterium tuberculosis clinical isolates classified by spoligotyping with a new high-throughput 3R SNPs based method. Infection, genetics and evolution 10, 1066–1074.

Bouakaze, C., Keyser, C., de Martino, S.J., Sougakoff, W., Veziris, N., Dabernat, H., Ludes, B., 2010. Identification and genotyping of Mycobacterium tuberculosis complex species by use of a SNaPshot Minisequencing-based assay. Journal of clinical microbiology 48, 1758–66.

Comas, I., Homolka, S., Niemann, S., Gagneux, S., 2009. Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies. PLoS one 4, e7815.

Dos Vultos, T., Mestre, O., Rauzier, J., Golec, M., Rastogi, N., Rasolofo, V., Tonjum, T., Sola, C., Matic, I., Gicquel, B., 2008. Evolution and diversity of clonal bacteria: the paradigm of Mycobacterium tuberculosis. PLoS one 3, e1538.

Filliol, I., Motiwala, A., Cavatore, M., Qi, W., 2006. Global phylogeny of Mycobacterium tuberculosis based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy. Journal of bacteriology 188, 759–772.

Hershberg, R., Lipatov, M., Small, P.M., Sheffer, H., Niemann, S., Homolka, S., Roach, J.C., Kremer, K., Petrov, D. a, Feldman, M.W., Gagneux, S., 2008. High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography. PLoS biology 6, e311.

Kasai, H., Ezaki, T., 2000. Differentiation of phylogenetically related slowly growing mycobacteria by their gyrB sequences. Journal of clinical microbiology 38, 301–308.

Niemann, S., Harmsen, D., Rusch-Gerdes, S., Richter, E., 2000. Differentiation of clinical Mycobacterium tuberculosis complex isolates by gyrB DNA sequence polymorphism analysis. Journal of clinical microbiology 38, 3231–3234.