New Microbiologica, 40, 1, 62-69, 2017, ISSN 1121-7138

SHORT COMMUNICATION

Genomic polymorphisms in a laboratory isolate of *Mycobacterium tuberculosis* reference strain H37Rv (ATCC27294)

Francesco Santoro[1], Valentina Guerrini[1,2], Elisa Lazzeri[1], Francesco Iannelli[1], Gianni Pozzi[1]

[1]*Laboratory of Molecular Microbiology and Biotechnology (L.A.M.M.B.), Department of Medical Biotechnologies, University of Siena, Siena, Italy;*
[2]*Present address: Public Health Research Institute, New Jersey Medical School, Rutgers, The State University of New Jersey, Newark, NJ, 07103, USA*

SUMMARY

The complete genome sequence of *Mycobacterium tuberculosis* reference strain H37Rv (ATCC27294) was determined on an isolate carried in our laboratory collection for almost 20 years and named H37RvSiena. DNA sequence analysis showed that the genome of H37RvSiena was 4,410,911 bp in size and contained 101 genetic polymorphisms compared to H37Rv: 83 single nucleotide polymorphisms, 10 insertions, and 8 deletions of which one was 617-bp long and seven ranged from 1 to 7 bp. Comparison with the genomes of two other H37Rv derivatives allowed identification of 28 polymorphisms specific for H37RvSiena.

Received May 16, 2016 *Accepted* October 14, 2016

Mycobacterium tuberculosis H37Rv is a derivative of clinical strain H37, isolated from a patient with pulmonary tuberculosis in 1905. Two variants of strain H37 were obtained in 1934 by rounds of dissociation and subculturing on media with controlled pH. The less virulent variant was selected by serial passages in both liquid and solid media at pH 6.1, produced crater-like colonies, became lysis-resistant and was initially named "R". The more virulent variant was propagated on media at pH 7.2, produced stippled, raised colonies and was initially named "S" (Steenken *et al.,* 1934). As both variants produced rough colonies, the names of the strains were later modified in H37Ra, for the avirulent "R" variant, and H37Rv, for the virulent "S" variant (Steenken, 1935). H37Rv was maintained for many years in the strain collection of Trudeau Institute of New York under the name TMC102 (for Trudeau Mycobacterial Collection) (*Figure 1*). H37Rv was deposited twice at the American Type Culture Collection (ATCC). The isolate deposited by A. G. Karlson (Mayo Clinic, Rochester, NY, USA) in 1970 was named ATCC25618, whereas the TMC102 strain deposited by G. P. Kubica (Trudeau Institute, Saranac Lake, NY, USA) in 1972 was named ATCC27294 (Bifani *et al.,* 2000; Kubica *et al.,* 1972) (*Figure 1*). H37Rv was designated the neotype of *M. tuberculosis* in 1972 (Kubica *et al.,* 1972) and is used worldwide as a reference strain for *M. tuberculosis* since it is virulent in animal models, susceptible to drugs, and amenable to genetic manipulation. The first genome sequence of H37Rv was published in 1998 (Cole *et al.,* 1998), and in 2008 Zheng *et al.* partially resequenced a H37Rv derivative of TMC102 (Zheng *et al.,* 2008). The polymorphisms identified in the "Zheng strain" were described in the paper (Zheng *et al.,* 2008) but not deposited in GenBank. Further sequencing of H37Rv included work by Ioerger *et al.* (Ioerger *et al.,* 2010) on strains H37RvAE, H37RvHA, H37RvMA, and H37RvCO (*Figure 1*) which were maintained in different laboratories.

Genome sequencing was performed on an isolate of ATCC27294 carried in our strain collection at the University of Siena since early 1997, which we named H37RvSiena. The strain was a kind gift of Lanfranco Fattorini (Istituto Superiore di Sanità, Roma, Italy), who, in turn, obtained it from ATCC in September 1996 (Fattorini *et al.,* 2003) (*Figure 1*). Genomic DNA was extracted from bacterial colonies plated on 7H11 following a previously described protocol (Larsen *et al.,* 2007) with modifications. Briefly, cells were scraped from plate, suspended in 10 ml of 7H9 containing 20 glass beads (3 mm in diameter) and vortexed for 1 minute to disrupt clumps. Cells were then inactivated for 3 hours at 85°C and centrifuged for 20 minutes at 4,000 x *g*. Pellets were suspended in 5 ml of GTE (50 mM glucose, 25 mM Tris-Cl, pH 8.0, 10 mM EDTA) with 1 mg/ml lysozyme and 0.5 ml of this suspension were aliquoted in 10 Eppendorf tubes containing each 0.4 g of acid washed glass beads (150-212 µm, Sigma). The tubes were vortexed 3 times for 1 minute on a Vortex-Genie 2 (Scientific Industries, NY) at maximum speed and incubated 1 hour at 37°C. This treatment was repeated two more times, then SDS and proteinase K were added to the pooled supernatants at a final concentration of 2% and 1 mg/ml, respectively. The suspension was incubated at 55°C for 1 hour and 2 ml of 5 M NaCl were added. After mixing by inversion, 1.6 ml of preheated CTAB was added,

Key words:
Mycobacterium tuberculosis, Genome, H37Rv, Genetic polymorphism, Illumina sequencing.

Corresponding author:
Francesco Santoro
E-mail: santorof@unisi.it

and the solution was incubated at 65°C for 10 minutes. DNA was extracted twice with chloroform/isoamyl alcohol (24:1, v/v) and precipitated in two volumes of ice cold ethanol. Genome sequencing of H37RvSiena was performed at IGA Technology Services S.r.l. (Udine, Italy) generating 100 bp paired-end reads with Illumina HiSeq 2000 system. Reads were trimmed, quality filtered and mapped against H37Rv as a reference sequence (GenBank NC_000962.3) using the Mosaik Assembler software (Lee *et al.,* 2014). Single nucleotide polymorphisms (SNPs), insertions and deletions were retrieved with VarScan software (Koboldt *et al.,* 2009) setting a minimum read depth of fifteen and a minimum variant frequency of 50% required to call variants.

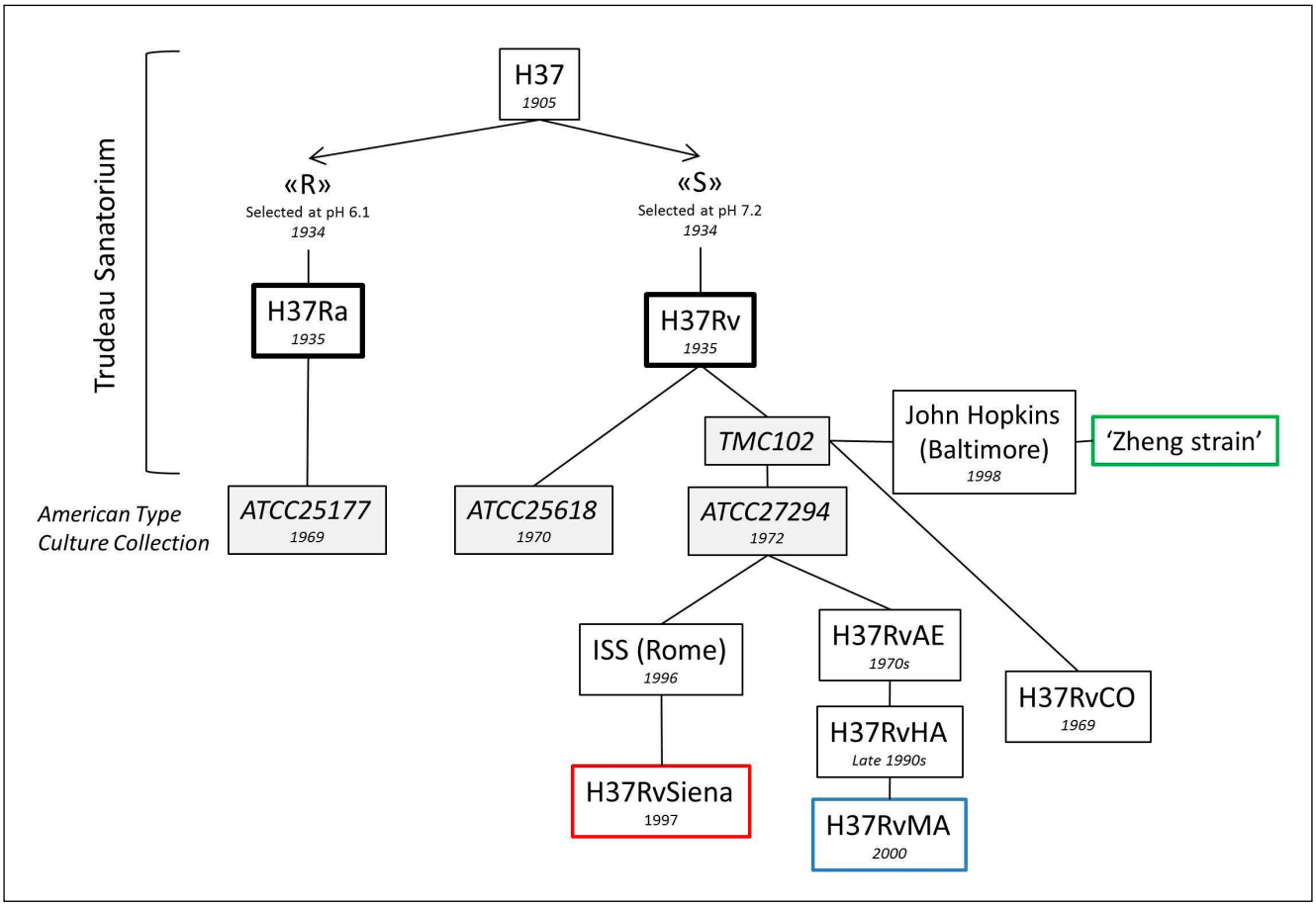

Figure 1 - ***Mycobacterium tuberculosis H37 and derivatives.*** *The clinical strain H37 was isolated at Saranac Laboratory in 1905 by Dr E. R. Baldwin from the sputum of a 19-year-old patient with chronic pulmonary tuberculosis. The "R" and "S" variants were selected in 1934 by a forced dissociation process on culture media with controlled pH. The "R" variant produced crater-like colonies, became lysis-resistant and was less virulent in the animal models, while the "S" variant produced stippled, raised colonies and retained full virulence in animal models. As both variants produced rough colonies, in 1935, the names of the strains were modified in H37Ra, for the avirulent "R" variant, and H37Rv, for the virulent "S" variant (Steenken, 1935). H37Rv was maintained for many years in the strain collection of Trudeau Institute of New York under the name of TMC102 (for Trudeau Mycobacterial Collection). H37Ra was deposited at the American Type Culture Collection (ATCC) as ATCC25177 in 1969 by C. L. Larson (University of Montana, Missoula, MT, USA) while H37Rv was deposited twice at the ATCC. The isolate deposited by A. G. Karlson (Mayo Clinic, Rochester, NY, USA) in 1970 was named ATCC25618, whereas the TMC102 strain deposited by G. P. Kubica (Trudeau Institute, Saranac Lake, NY, USA) in 1972 was named ATCC27294 (Bifani et al., 2000; Kubica et al., 1972). The "Zheng strain" was acquired from John Hopkins Center for TB Research, which obtained it from the Trudeau Institute in 1998 (Zheng et al., 2008). H37RvCO was obtained directly from the Trudeau Institute by the Colorado State University in 1969. H37RvAE, H37RvHA and H37RvMA all derive from the same ATCC27294 stock. ATCC27294 strain was acquired by: 1) Albert Einstein College of Medicine (New York, NY, USA) and renamed H37RvAE, 2) from there it was transferred to Harvard University (Boston, MA, USA) in 1990s and stored as H37RvHA, 3) transferred again to the University of Massachusetts Medical School (Worchester, MA, USA) in 2000 and called H37RvMA. Istituto Superiore di Sanità (ISS, Rome, Italy) acquired strain ATCC27294 in September 1996 from ATCC, at the beginning of 1997 it was sent to our laboratory and renamed H37RvSiena. Arrows indicate derivation based on a selection process, while lines indicate transfer between laboratories. Strain names are indicated in black boxes, year of isolation/acquisition (when available) is reported in italics under the strain name. Red, blue and green boxes are used for H37RvSiena, H37RvMA and "Zheng strain", respectively, reference strains ATCC25177, ATCC25618, ATCC27294, and TMC102 are in shaded boxes. Whole genome sequences are available for H37RvCO, H37RvAE, H37RvMA, and H37RvSiena (GenBank acc. no. CM001515, CM002882, CM002884, and CP007027, respectively).*

Table 1 - *Genetic polymorphisms of H37RvSiena.*

Locus_tag H37Rv	*H37Rv genome position*	*H37RvSiena genome position*	*Genetic polymorphism*[1]	*H37RvSiena*[2]	*H37RvMA*[2]	*H37RvZheng*[2]
Rv0012	14785	14785	T > C	●	●	●
Rv0050	55536	55533	C > T	●	●	●
Rv0064	69989	69986	G > A	●	●	●
Rv0101	116000	115997	T > G	●	●	●
intergenic Rv0108c-Rv0109	131176-131177	131173-131175	G–T > GGT	●	●	●
Rv0197	234477	234475	T > G	●	●	●
Rv0197	234496-234497	234494-234497	C––C > CGTC	●	●	●
Rv0323c	390828	390828	T > C	●	●	●
Rv0354c	424322-424323	424322-424324	C–T > CCT	●	●	●
Rv0388c	467516	467519	G > C	●	●	●
Rv0388c	467526	467529	C > G	●	●	●
Rv0388c	467546	467549	G > C	●	●	●
Rv0388c	467557	467560	A > C	●	●	●
Rv0388c	467564	467567	A > C	●	●	●
Rv0388c	467585	467588	G > C	●	●	●
Rv0388c	467590	467593	T > C	●	●	●
Rv0388c	467621	467624	T > G	●	●	●
Rv0388c	467638	467641	G > T	●	●	●
Rv0890c	990001	990004	G > C	●	●	●
Rv0907	1010206-1010207	1010209-1010211	G–T > GGT	●	●	●
Rv0919	1025106	1025110	T > C	●	●	●
Rv0930	1037911	1037915	C > T	●	●	●
Rv1181	1315884	1315889	G > A	●	●	●
Rv1266c	1414021	1414026	C > T	●	●	●
nc RNA	1471659	1471664	C > T	●	●	●
Rv1809	2051746	2051134	T > C	●	●	●
Rv1815	2057774	2057162	A > T	●	●	●
intergenic Rv1963c-Rv1964	2207591-2207592	2206979-2206981	T–T > TCT	●	●	●
intergenic Rv2005c-Rv2006	2251999	2251388	A > G	●	●	●
Rv2037c	2282787	2282176	C > T	●	●	●
Rv2250A-Rv2251	2525726-2525728	2525115-2525116	GGA > G-A	●	●	●
intergenic Rv2421c-Rv2422	2718852	2718240	T > G	●	●	●
Rv2450c	2751804	2751192	C > T	●	●	●
Rv2495c	2809621	2809009	T > C	●	●	●
Rv3021c	3379708	3379089	G > C	●	●	●
Rv3021c	3379712	3379093	G > C	●	●	●
Rv3021c	3379718	3379099	T > C	●	●	●
Rv3021c	3379726	3379107	C > A	●	●	●
Rv3021c	3379730	3379111	G > C	●	●	●
Rv3021c	3379732	3379113	C > T	●	●	●
Rv3021c	3379735	3379116	A > C	●	●	●
Rv3021c	3379736	3379117	C > A	●	●	●
Rv3021c	3379742	3379123	T > C	●	●	●
Rv3021c	3379751	3379132	A > C	●	●	●
Rv3021c	3379757	3379138	A > C	●	●	●
Rv3021c	3379763	3379144	G > A	●	●	●

→

Locus_tag H37Rv	*H37Rv genome position*	*H37RvSiena genome position*	*Genetic polymorphism*[1]	*H37RvSiena*[2]	*H37RvMA*[2]	*H37RvZheng*[2]
Rv3021c	3379784	3379165	C > A	●	●	●
Rv3021c	3379788	3379169	C > G	●	●	●
intergenic Rv3202-Rv3203	3580638-3580640	3580020-3580021	TTG > T–G	●	●	●
intergenic Rv3212-Rv3213c	3590686-3590687	3590067-3590069	G-T > GCT	●	●	●
Rv3479	3896340	3895721	T > G	●	●	●
Rv3911	4400662-4400664	4400042-4400043	CCG > C–G	●	●	●
Rv0109	132417	132415	C > G	●	○	●
Rv0388c	467500-4675501	467501-4675503	G–T > GGT	●	○	●
Rv0388c	467509-4675510	467511-4675513	G–C > GGC	●	○	●
Rv0050	55532-55536	55532-55533	GCCGC > G – – – T	●	●	○
intergenic Rv0383c-Rv0384c	459399	459400	A > C	●	●	○
Rv0532	623508	623511	C > G	●	●	○
Rv0543c	635633	635636	C > T	●	●	○
Rv0861c	958922	958925	C > A	●	●	○
Rv0978c	1093406	1093410	A > G	●	●	○
Rv1046c	1168717-1168718	1168721-1168723	T–C > TTC	●	●	○
Rv1180	1315191	1315196	A > C	●	●	○
Rv1520	1711627	1711632	C > T	●	●	○
Rv1755c[3]	1987085-1987703	1987090-1987091	G(...)T > G(617 bp deletion)T	●	●	○
Rv1771	2006032	2005420	A > G	●	●	○
Rv1907c	2153410	2152798	A > G	●	●	○
Rv2126c	2387733	2387122	T > C	●	●	○
Rv2627c	2954439	2953827	T > C	●	●	○
Rv3021c	3380439-3380440	3379821-3379823	C-T > CCT	●	●	○
Rv3331	3718357	3717739	C > T	●	●	○
intergenic Rv3443c-Rv3444c	3862473-3862475	3861855-3861856	AAC > A–C	●	●	○
Rv3655c	4095001-4095003	4094382-4094383	CGA > C–A	●	●	○
Rv0278c	333892	333892	G > C	●	○	○
Rv0279c	337959	337959	A > C	●	○	○
Rv0279c	338020	338020	A > C	●	○	○
Rv0279c	338100	338100	T > C	●	○	○
Rv0279c	338453	338453	A > G	●	○	○
Rv0516c	608331	608334	T > G	●	○	○
Rv0532	623472	623475	A > G	●	○	○
Rv0578c	672491	672494	C > G	●	○	○
Rv0746	836272	836275	A > G	●	○	○
Rv0746	836291	836294	A > G	●	○	○
Rv0746	836426	836429	A > C	●	○	○
Rv0746	836454	836457	A > G	●	○	○
Rv0746	836538	836541	A > G	●	○	○
Rv0746	836658	836661	A > G	●	○	○
Rv0746	837033	837036	A > G	●	○	○
Rv0747	838990	838993	C > G	●	○	○
Rv0747	839334	839337	A > G	●	○	○
Rv0747	839348	839351	A > G	●	○	○
Rv0747	840496	840499	C > G	●	○	○
Rv0833	927110	927113	A > G	●	○	○

→

Locus_tag H37Rv	*H37Rv genome position*	*H37RvSiena genome position*	*Genetic polymorphism*[1]	*H37RvSiena*[2]	*H37RvMA*[2]	*H37RvZheng*[2]
Rv2209	2474823	2474212	G > A	●	○	○
Rv2741	3054724	3054112	A > G	●	○	○
Rv2931	3246316-3246324	3245704-3245705	ACTGTCACC > A−−−−−−−C	●	○	○
Rv3508	3935335	3934716	T > C	●	○	○
Rv3511	3940802	3940183	A > G	●	○	○
Rv3514	3948404	3947785	T > G	●	○	○
Rv3514	3948414	3947795	T > G	●	○	○
Rv3514	3948417	3947798	T > G	●	○	○

[1] The '>' symbol indicates the nucleotide changes, hyphens indicate deletions.
[2] A solid dot indicates the presence of the described polymorphism, while an empty dot indicates absence.
[3] A 617-bp deletion occurred in the *plcD* pseudogene (Rv1755c) in both H37RvSiena and H37RvMA.

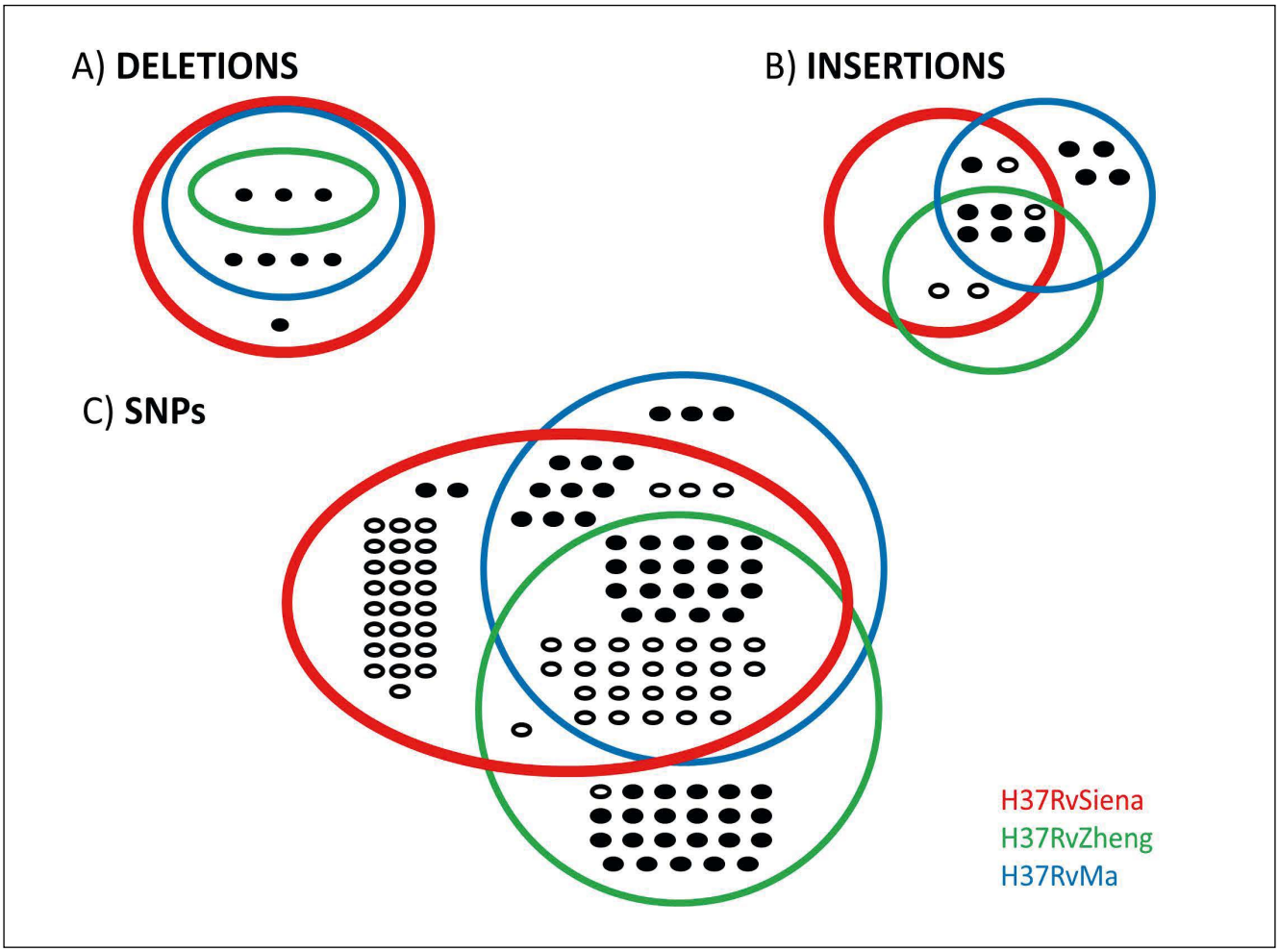

Figure 2 - ***Venn diagram of the genetic polymorphisms in different H37Rv genomes.*** *Red, blue and green circles contain the polymorphisms of H37RvSiena, H37RvMA and "Zheng strain" (Zheng et al., 2008), respectively, compared to the H37Rv reference sequence (GenBank NC_000962). Polymorphisms are indicated by black circles, while open black circles indicate polymorphisms falling in highly repetitive genes (pe_pgrs and ppe genes). A) 8 deletion events are present in H37RvSiena compared to the reference sequence. Of those, 7 are shared with H37RvMA and 3 with both H37RvMA and "Zheng strain". One 7-bp deletion in Rv2931 is unique to H37RvSiena, one 3-bp deletion in Rv0050 and one 617-bp deletion in plcD pseudogene are shared by H37RvSiena and H37RvMA, the remaining 5 deletions are 1-bp long. B) Nine 1-bp insertions are present in H37RvSiena, of those 5 are shared among all genomes, 2 are common to H37RvSiena and H37RvMA and 2 are common to H37RvSiena and "Zheng strain". One 2-bp insertion in Rv0197 is shared among all genomes. C) 83 single nucleotide polymorphisms are present in H37RvSiena, of those 43 are shared among all genomes, 12 are common to H37RvSiena and H37RvMA and 1 is common to H37RvSiena and "Zheng strain", while the remaining 27 are unique to H37RvSiena.*

Table 2 - *Regions of H37RvSiena with no read coverage.*

Locus_tag H37Rv	*Gene/gene family*	*H37Rv genome position*	*H37RvSiena genome position*
Rv0279c	*pe_pgrs4*	336680-336684 336690-336691	336680-336684 336690-336691
Rv0336	13E12 repeat family	400722-400752 400774-401021 401046-401119	400722-400752 400774-401021 401046-401119
Rv0746	*pe_pgrs9*	837291-837301	837294-837304
Rv0796	IS*6110*	889604-889610 889632-890125 890560-890640	889607-889613 889635-890128 890563-890643
Rv1047	IS*1081*	1169877-1169896 1169928-1170162	1169882-1169901 1169933-1170167
Rv1199c	IS*1081*	1341882-1341884 1341906-1342063 1342083-1342128 1342148-1342159	1341887-1341889 1341911-1342068 1342088-1342133 1342153-1342164
Rv1313c	IS*1557*	1469063-1469069	1469068-1469074
Rv1369c	IS*6110*	1542523-1542577 1542607-1542694 1542716-1542737	1542528-1542582 1542612-1542699 1542721-1542742
Rv1450c	*pe_pgrs27*	1633562-1633608	1633567-1633613
Rv1756c	IS*6110*	1988284-1988328 1988362-1988473	1987672-1987716 1987750-1987861
Rv1759c	*wag22*	1991190-1991193	1990578-1990581
Rv1763	IS*6110*	1996692-1996693 1996716-1996770 1996794-1996805 1996824-1996872	1996080-1996081 1996104-1996158 1996182-1996193 1996212-1996260
Rv2106	IS*6110*	2365995-2365997 2366032-2366042 2366060-2366118 2366140-2366171 2366209-2366225	2365384-2365386 2365421-2365431 2365449-2365507 2365529-2365560 2365598-2365614
Rv2167c	IS*6110*	2430714-2430715 2430740-2430861 2430882-2430883	2430103-2430104 2430129-2430250 2430271-2430272
Rv2279	IS*6110*	2550627-2550786	2550015-2550174
Rv2355	IS*6110*	2636171-2636310 2636328-2636345	2635559-2635698 2635716-2635733
Rv2479c	IS*6110*	2785205-2785264 2785308-2785317 2785375-2785381	2784593-2784652 2784696-2784705 2784763-2784769
Rv2490c	*pe_pgrs43*	2804982-2804985 2805071-2805076	2804370-2804373 2804459-2804464
Rv2649	IS*6110*	2972724-2972887	2972112-2972275
Rv2814c	IS*6110*	3121069-3121100 3121119-3121173 3121191-3121294	3120457-3120488 3120507-3120561 3120579-3120682
Rv2933	*ppsC*	3258354-3258374	3257735-3257755
Rv3185	IS*6110*	3551837-3551899 3551920-3552007	3551219-3551281 3551302-3551389
Rv3187	IS*6110*	3553300-3553450 3553468-3553480 3553519-3553524	3552682-3552832 3552850-3552862 3552901-3552906
Rv3326	IS*6110*	3711044-3711070 3711091-3711111 3711158-3711163	3710426-3710452 3710473-3710493 3710540-3710545
Rv3508	*pe_pgrs54*	3931692-3931706 3931775-3931792 3932016-3932030	3931073-3931087 3931156-3931173 3931397-3931411
Rv3514	*pe_pgrs57*	3947752-3947754	3947133-3947135

Analysis of genetic polymorphisms in the genome of H37RvSiena with respect to H37Rv reference sequence (GenBank NC_000962.3) revealed 101 polymorphic sites (83 SNPs, 10 insertions, and 8 deletions of which one was 617-bp long and seven ranged from 1 to 7 bp) which are reported in *Table 1*. The H37RvSiena genome sequence was assembled using the genetic variants reported in *Table 1* to modify the H37Rv reference sequence, obtaining a genome length of 4,410,911 bp, deposited in GenBank with accession number CP007027. Polymorphisms of H37RvSiena were then compared with polymorphisms of H37RvMA (GenBank CM002884) (Ioerger *et al.,* 2010) and of the "Zheng strain" (Zheng *et al.,* 2008) (*Figure 2*). 52 polymorphisms (3 deletions, 6 insertions and 43 SNPs) were conserved among the 3 sequences: 1 insertion and 24 SNPs fell into highly repetitive sequences of *ppe7*/Rv0354c, *ppe9*/Rv0388c, *ppe33*/Rv1809, and *ppe47*/Rv3021c genes. 28 polymorphisms are unique to H37RvSiena:

1. a 7-bp deletion in the *ppsA* gene;
2. a T→G transversion in Rv0516c;
3. a G→A transition in Rv2209;
4. 25 SNPs located into highly repetitive sequences of the following genes *pe_pgrs3*/Rv0278c, *pe_pgrs4*/Rv0279c, *pe_pgrs6*/Rv0532, *pe_pgrs7*/Rv0578c, *pe_pgrs9*/Rv0746, *pe_pgrs10*/Rv0747, *pe_pgrs47*/Rv2741, *pe_pgrs53*/Rv3508, *pe_pgrs55*/Rv3511, *pe_pgrs57*/Rv3514.

Finally, 3 polymorphisms (2 insertions in *ppe9*/Rv0388c and 1 SNP in *pe_pgrs1*/Rv0109) are shared only between H37RvSiena and the "Zheng strain", and another 18 (4 deletions, 2 insertions and 12 SNPs) between H37RvSiena and H37RvMA. Six out of these 18 polymorphisms (a 3-bp deletion in Rv0050/*ponA1*, and SNPs in Rv0543c, Rv0861c/*ercc3*, Rv1520, Rv1771, Rv1907c) are present in ATCC27294 and TMC102 derivatives (*Figure 1*). H37RvSiena had a 617 bp deletion within the *plcD* pseudogene which was also present in 3 ATCC27294 derivatives (H37RvAE, H37RvMA and H37RvHA) (Ioerger *et al.,* 2010). The latter 3 strains shared 4 polymorphisms, while H37RvCO had an IS*6110* insertion in *plcD*. It is likely that the deletion in *plcD* occurred in ATCC27294 before H37RvSiena was separated from the other 3 strains and after TMC102 was given to the Colorado State University and renamed H37RvCO (*Figure 1*).

The present work obtained a good coverage (>20×) in most of repetitive sequences of *pe_pgrs*, *ppe* and 13E12 repeat family genes, prophage elements and insertion sequences (ISs), encompassing 99.92% of the total genome length. This allowed the identification of 25 polymorphic sites in regions where re-sequencing had previously been unsuccessful (Ioerger *et al.,* 2010; Zheng *et al.,* 2008). *Table 2* reports the regions of H37RvSiena which obtained zero coverage depth and for which we assumed that the sequence was identical to H37Rv reference genome. Highly repetitive sequences are

1. difficult to sequence;
2. subject to mutations driven by recombination of the repeats, in fact highly polymorphic loci are often used for intraspecies molecular typing of bacteria [e.g. MIRU-VNTR typing in *M. tuberculosis* (Supply *et al.,* 2006) and *spa* typing in *Staphylococcus aureus* (Shopsin *et al.,* 1999)].

In H37Rv, *ppsA* is a 5,631-bp gene which codes for a 1,876-aa type I polyketide synthase involved in phthiocerol dimycocerosate (PDIM) synthesis. The 7-bp deletion in *ppsA* of H37RvSiena corresponded to nucleotides 873-879 at the 5' end of H37Rv-*ppsA* and caused a frameshift. After the deletion event in H37RvSiena, a start codon in a different reading frame restored the frame of the last 4,752 nucleotides of *ppsA*. We have no evidence that the resulting 4,809-bp open reading frame is transcribed and that its gene product is functional. Mutations in the PDIM synthesis pathway have been reported to impair virulence in mice and occurred in strains which had only been maintained in acidic culture media, as was the case for H37RvSiena (Domenech *et al.,* 2009).

The point mutation in the H37RvSiena homolog of Rv0516c gene, caused an amino acid substitution (M69L) in the STAS (Sulphate Transporter and Anti Sigma Factor Antagonist) domain of predicted protein sequence. The point mutation in the Rv2209 homolog caused an amino acid substitution (R485H) in the C-terminal portion of the deduced amino acid sequence. This protein, reported to be upregulated following antitubercular drug treatment (Gupta *et al.,* 2010), is predicted to be a transmembrane protein, the mutation occurred in a hypothetical cytoplasmic region and caused a substitution of an arginine with an histidine, which are both positively charged amino acids.

In conclusion, the 4,410,911 bp-long genome of H37RvSiena harbors 101 genetic polymorphisms compared to the H37Rv reference sequence, 73 were described in previous resequencing projects, but here we showed 3 polymorphic sites unique to our strain and were able to confirm the presence of 25 polymorphisms in highly repetitive genes.

Nucleotide sequence accession number

H37RvSiena genome was deposited at GenBank under accession number CP007027.1.

Acknowledgments

We thank Lanfranco Fattorini of Istituto Superiore di Sanità (Rome, Italy) for providing the H37Rv strain. This study was funded by EU TB PAN NET project FP7-HEALTH 223681 and by Italian Ministry of University and Research, project PRIN 2012 (2012WJSX8K) "Host-microbe interaction models in mucosal infections: development of novel therapeutic strategies".

References

Bifani P., Moghazeh S., Shopsin B., Driscoll J., Ravikovitch A. et al. (2000). Molecular characterization of *Mycobacterium tuberculosis* H37Rv/Ra variants: distinguishing the mycobacterial laboratory strain. *J Clin. Microbiol.* **38**, 3200-3204.

Cole S.T., Brosch R., Parkhill J., Garnier T., Churcher C. et al. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature.* **393**, 537.

Domenech P., Reed M.B. (2009). Rapid and spontaneous loss of phthiocerol dimycocerosate (PDIM) from *Mycobacterium tuberculosis* grown *in vitro*: implications for virulence studies. *Microbiology.* **155**, 3532-3543.

Fattorini L., Tan D., Iona E., Mattei,M., Giannoni F. et al. (2003). Activities of moxifloxacin alone and in combination with other antimicrobial agents against multidrug-resistant *Mycobacterium tuberculosis* infection in BALB/c mice. *Antimicrob. Agents Chemother.* **47**, 360-362.

Gupta A.K., Katoch V.M., Chauhan D.S., Sharma R., Singh M. et al. (2010). Microarray analysis of efflux pump genes in multidrug-resistant *Mycobacterium tuberculosis* during stress induced by common anti-tuberculous drugs. *Microb. Drug Resist.* **16**, 21-28.

Ioerger T.R., Feng Y.C., Ganesula K., Chen X.H., Dobos K.M. et al. (2010). Variation among Genome Sequences of H37Rv Strains of *Mycobacterium tuberculosis* from Multiple Laboratories. *J. Bacteriol.* **192**, 3645-3653.

Koboldt D.C., Chen K., Wylie T., Larson DE., McLellan M.D. et al. (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics.* **25**, 2283-2285.

Kubica G.P., Kim T.H., Dunbar F.P. (1972). Designation of Strain H37Rv as the Neotype of *Mycobacterium tuberculosis*. *Int J Syst Evol Microbiol.* **22**, 99-106.

Larsen M.H., Biermann K., Tandberg S., Hsu T., Jacobs W. R. (2007). Genetic Manipulation of *Mycobacterium tuberculosis*. *Curr Protoc Microbiol.* **6**, 10A.2.1-10A.2.21.

Lee W.P., Stromberg M.P., Ward A., Stewart C., Garrison E.P. et al. (2014). MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS One.* **9**, e90581.

Shopsin B., Gomez M., Montgomery S.O., Smith T.H., Waddington M. et al. (1999). Evaluation of protein A gene polymorphic region DNA sequencing for typing of Staphylococcus aureus strains. *J Clin Microbiol.* **37**, 3556-3563.

Steenken W. (1935). Lysis of Tubercle Bacilli *in vitro*. *Exp Biol Med*. **33**, 253-255.

Steenken W., Oatway W.H., Petroff S.A. (1934). Biological Studies of the Tubercle Bacillus: III. Dissociation and Pathogenicity of the R and S Variants af The Human Tubercle Bacillus (H(37)). *J Exp. Med.* **60**, 515-540.

Supply P., Allix C., Lesjean S., Cardoso-Oelemann M., Rusch-Gerdes S. et al. (2006). Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. *J. Clin. Microbiol.* **44**, 4498-4510.

Zheng H.J., Lu L.D., Wang B.F., Pu S.Y., Zhang X.L. et al. (2008). Genetic Basis of Virulence Attenuation Revealed by Comparative Genomic Analysis of *Mycobacterium tuberculosis* Strain H37Ra versus H37Rv. *PLoS One.* **3**.