Using whole exome sequencing and bacterial pathogen sequencing to investigate the genetic basis of pulmonary non-tuberculous mycobacterial infections

Luke J. Tallon[1], Lisa D. Sadzewicz[1], Xinyue Liu[1], Sushma Nagaraj[1], Sandra Ott[1], Xuechu Zhao[1], Scott Devine[1], Steven M. Holland[2], Kenneth Olivier[2], Hervé Tettelin[1]

[1]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD
[2]National Institutes of Health, Bethesda, MD

Abstract

Pulmonary non-tuberculous mycobacterial (PNTM) infections occur in patients with chronic lung disease, but also in a distinct group of elderly women without lung defects who share a common body morphology: tall and lean with scoliosis, pectus excavatum, and mitral valve prolapse. In order to characterize the human host susceptibility to PNTM, we performed whole exome sequencing (WES) of 44 individuals in extended families of patients with active PNTM as well as 55 additional unrelated individuals with PNTM. This unique collection of familial cohorts in PNTM represents an important opportunity for a high yield search for genes that regulate mucosal immunity. An average of 58 million 100bp paired-end Illumina reads per exome were generated and mapped to the hg19 reference genome. Following variant detection and classification, we identified 58,422 potentially high-impact SNPs, 97.3% of which were missense mutations. Segregating variants using the family pedigrees as well as comparisons to the unrelated individuals identified multiple potential variants associated with PNTM. Validations of these candidate variants in a larger PNTM cohort are underway.

In addition to WES, we sequenced the genomes of 52 mycobacterial isolates, including 9 from these PNTM patients, to integrate host PNTM susceptibility with mycobacterial genotypes and gain insights into the key factors involved in this devastating disease. These genomes were sequenced using a combination of 454, Illumina, and PacBio platforms and assembled using multiple genome assemblers. The resulting genome sequences were used to identify mycobacterial genotypes associated with virulence, invasion, and drug resistance.

Methods

Exome Sequencing and Analysis

Exome target enrichment was performed using Agilent SureSelect All Exon V4 kits according to protocol. Illumina HiSeq 2000 100 bp paired-end reads from all individuals were first mapped onto the NCBI human reference genome NCBI GRCh37 using BWA[4]. Duplicate mapping artifacts were then marked using PICARD[10]. After that, the sorted, duplicate-marked BAMs were input to the GATK (v2.6) pipeline for INDEL realignment, base quality recalibration, and read reduction[2]. The resulting BAMs were merged together to call variants using GATK's UnifiedGenotyper and filtered using VQSR. The resulting VCF containing the filtered SNPs and INDELs was further annotated using SNPEff and ANNOVAR[1,9]. Finally, a summary table was generated to include for each variant call: allele frequency (AF) in PNTM cases vs. controls, 1000 Genomes Project, dbSNP, ESP6500, and various gene annotations containing gene impact (nonsense, missense, frameshift, etc.), amino acid change, SIFT and Polyphen scores. This approach identified 58,422 high impact SNPs, 97.3% of which were missense mutations.

We studied the segregation of variants within the family pedigrees because analysis of related individuals helps reduce the background noise of variants that are not associated with susceptibility to PNTM infections. We created a profile of presence/absence of traits (e.g. PNTM infection, bronchiectasis, scoliosis, gender, *etc.*) and variants (e.g. SNPs, indels) for each family.

Variants were further separated as homozygous or heterozygous, and models for dominant, recessive, and dominant compound heterozygote segregation were applied to the data. Candidate mutated genes identified by these methods were then checked against public variant databases (*e.g.* dbSNP and the 1,000 genomes project[3]) to determine their frequency and whether they have already been associated with unrelated diseases. Novel candidate genes were then compared across families as well as within the 55 unrelated PNTM patients.

Mycobacterial Genome Sequencing & Analysis

Two sets of bacterial isolates were sequenced as part of this project. The first set contained 38 serial clinical isolates from 6 patients. These were sequenced using a combination of Roche 454 FLX 3 Kbp paired-end libraries and 300 bp libraries on Illumina HiSeq2000 100bp paired-end runs, generating an average of >20x coverage of 454 reads and >100x of Illumina reads. These data were assembled together using Celera Assembler 7.0[5]. The resulting genome sequences were aligned to the type strain sequence and variants were called using an in-house pipeline called Skirret.

The second set of isolates from PNTM patients were sequenced using a combination of PacBio RS long-fragment reads and Illumina MiSeq 250 bp paired-end reads, generating >100x coverage of RS I reads and >200x coverage of MiSeq reads for each genome. The genomes were assembled using both PacBio HGAP and Celera Assembler 7.0, and the best assembly was selected based on contiguity, contig N50, and manual inspection of the resulting contigs[5,6].

References

1. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6: 80-92.

2. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics 43: 491-498.

3. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.

4. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.

5. Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, et al. (2008) Aggressive assembly of pyrosequencing reads with mates. Bioinformatics 24: 2818-2824.

6. Shin SC, Ahn do H, Kim SJ, Lee H, Oh TJ, et al. (2013) Advantages of Single-Molecule Real-Time Sequencing in High-GC Content Genomes. PLoS ONE 8: e68824.

7. Tettelin H, Sampaio EP, Daugherty SC, Hine E, Riley DR, et al. (2012) Genomic Insights into the Emerging Human Pathogen Mycobacterium massiliense. Journal of Bacteriology 194: 5450.

8. Tettelin H, Davidson RM, Agrawal S, Aitken M, Shallom S, et al. (2014) High-level relatedness among Mycobacterium abscessus subsp. massiliense strains from widely separated outbreaks. Emerging Infectious Diseases 20: 370-377.

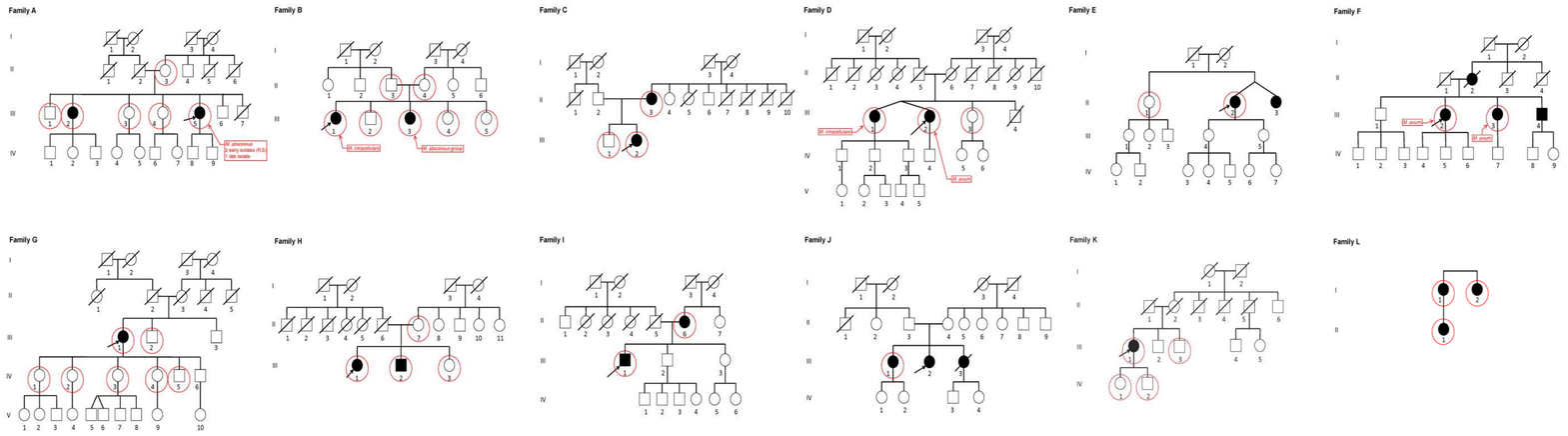
9. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research 38: e164.

10. http://picard.sourceforge.net/

Acknowledgements

This project has been funded with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under contract number HHSN272200900009C.

Pedigrees



Human Exomes

Exome sequencing of 99 individuals was conducted in two phases. The first phase sequenced 44 individuals from extended families shown in the pedigrees above (individuals sequenced denoted by red circles). The second phase added 55 unrelated individuals with PNTM. Validations of candidate variants associated with PNTM are underway.

Exome Sequencing Stats

	Phase I	Phase II	Total
Total Exomes	44	55	99
Avg Read Count	57,760,173	57,949,866	58,299,715
Avg % mapped	98.33%	99.13%	98.77%
Avg % mapped on target (+/-200bp)	86.58%	85.51%	85.98%
Avg proper pairs	97.88%	98.63%	98.27%
Avg on target cvg	78	83	80.66
Avg p10 on target cvg	22	36	30.01

Variant Stats Per Exome (Phase I & II)

Avg Total SNPs per exome	60,142
Avg PF SNPs per exome	58,669
Avg Novel PF SNPs per exome	**716**
Avg Total Indels per exome	8,720
Avg PF Indels per exome	7,523
Avg Novel PF Indels per exome	**660**

Variant Stats – Merged

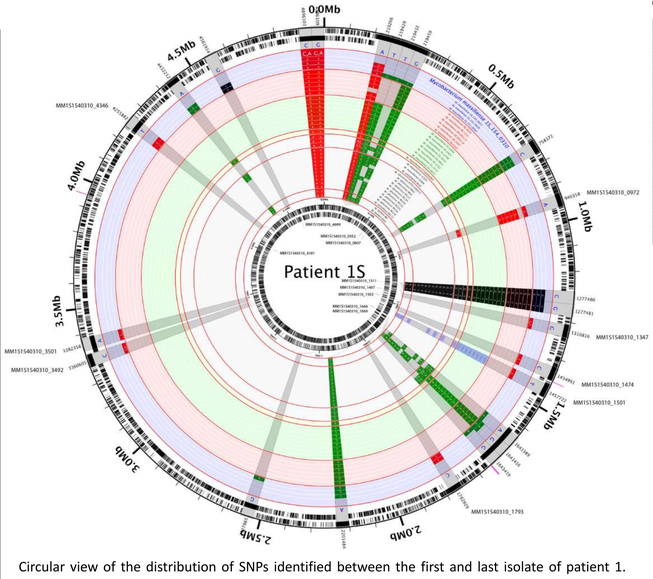
Total Merged SNP positions	273,497
Total Merged PF SNP positions	245,309
Total Merged PF Novel SNP positions	34,718
Total Merged High Impact SNPs	**58,422**
Total Merged High Impact PF SNPs	54,125
Total Merged High Impact Novel SNPs	**7,903**

Mycobacterial Genomes

Phase I of bacterial genome sequencing included 38 longitudinally gathered and curated specimens obtained from patients predominantly with lung infection followed over long periods (up to 11 years). We are using the genome sequences for comparisons at various levels to explore changes associated with virulence, invasion, and drug resistance. These isolates were sequenced using a combination of 454 and Illumina data, resulting in genome assemblies with an average of <10 contigs and a contig N50 of >1.8 Mbp.

Isolate	Scaffolds	Contigs	N50 Contig Bp	Genome Size
1S_151_0930	2	6	2,517,669	4,895,763
1S_152_0914	1	2	4,138,301	4,897,666
1S_153_0915	1	13	708,027	4,894,488
1S_154_0310	2	5	1,439,704	4,895,947
2B_0107	1	7	929,360	4,811,369
2B_0307	2	9	2,520,989	4,809,972
2B_0626	1	6	966,400	4,810,868
2B_0912_R	1	9	804,333	4,809,994
2B_0912_S	1	12	832,198	4,809,311
2B_1231	1	9	720,581	4,811,264
3A_0119_R	4	13	701,772	5,281,259
3A_0122_R	5	33	267,876	5,231,383
3A_0122_S	3	18	683,876	5,233,665
3A_0731	6	12	863,785	5,386,958
3A_0810_R	2	20	878,257	5,287,644
3A_0930_R	4	13	702,908	5,273,535
3A_0930_S	4	12	1,294,116	5,249,987
4S_0116_R	2	6	1,566,263	4,840,738
4S_0116_S	1	7	2,879,709	4,841,923
4S_0206	3	11	3,408,769	4,855,358
4S_0303	2	8	1,039,762	4,864,284
4S_0726_RA	1	6	3,091,077	4,842,208
4S_0726_RB	2	6	2,446,836	4,862,293
5S_0304	1	7	3,678,170	5,253,081
5S_0421	2	6	2,977,626	5,242,408
5S_0422	4	13	1,028,114	5,323,965
5S_0708	1	6	2,955,020	5,253,252
5S_0817	3	12	3,340,348	5,248,885
5S_0921	1	12	682,040	5,254,648
5S_1212	3	7	3,514,598	5,239,840
5S_1215	3	5	4,075,344	5,206,300
6G_0125_R	1	5	3,629,393	5,139,423
6G_0125_S	2	8	1,271,054	5,328,817
6G_0212	2	13	678,741	5,144,692
6G_0728_R	3	8	948,650	5,341,825
6G_0728_S	2	6	1,990,736	5,315,024
6G_1108	2	16	918,577	5,337,035
CCUG 48898*	1	5	2,849,228	5,195,205
AVERAGE	2.2	9.8	1,840,532	5,087,428
MEDIAN	2	8	1,282,585	5,200,753

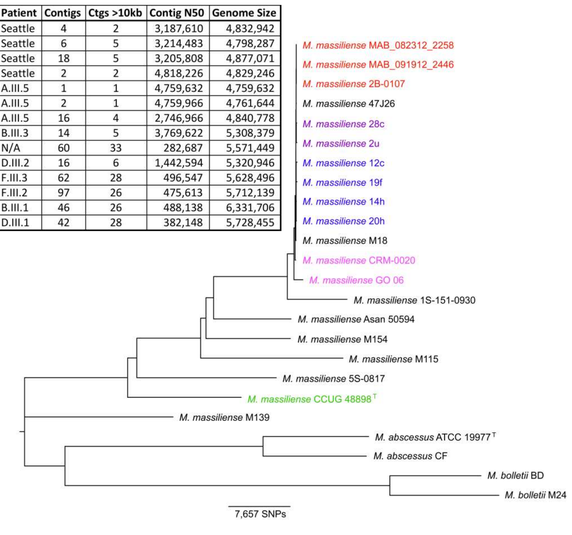
* Type strain[7]



Circular view of the distribution of SNPs identified between the first and last isolate of patient 1. Green: synonymous SNPs, red: non-synonymous, black: intergenic, light purple "NH": SNP region not found (no hit) in that isolate. Rim shading – blue: patient 1, red: patient 2, green: patient 5, white (none): public genomes. The 23s rRNA mutation at nt 940358 correlated with the development of clarithromycin resistance, and appears only in the last isolate of patient 1, while it appears in the second isolate of patient 2 and remains stable until the last isolate.

Phase II included 9 samples from patients whose exomes were sequenced as part of the family pedigrees in Phase I, together with 5 samples from unrelated individuals. These isolates were sequenced using a combination of PacBio RS I and Illumina data. As shown in the table below, assembly results followed a species-specific trend with *M. abscessus* samples producing the highest quality genome sequences.

Isolate	Species	Patient	Contigs	Ctgs >10kb	Contig N50	Genome Size
MAB_091912_2446	*M. abscessus*	Seattle	4	2	3,187,610	4,832,942
MAB_082312_2258	*M. abscessus*	Seattle	6	5	3,214,483	4,798,287
MAB_082312_2272	*M. abscessus*	Seattle	18	5	3,205,808	4,877,071
MAB_091912_2455	*M. abscessus*	Seattle	2	2	4,818,226	4,829,246
MAB_030201_1061	*M. abscessus*	A.III.5	1	1	4,759,632	4,759,632
MAB_030201_1075	*M. abscessus*	A.III.5	2	1	4,759,966	4,761,644
MAB_110811_1470	*M. abscessus*	A.III.5	16	4	2,746,966	4,840,778
MAB_110811_2726	*M. abscessus*	B.III.3	14	5	3,769,622	5,308,379
MAC_080597_8934	*M. avium* comp.	N/A	60	33	282,687	5,571,449
MAV_061107_1842	*M. avium*	D.III.2	16	6	1,442,594	5,320,946
MAV_120709_2344	*M. avium*	F.III.3	62	28	496,547	5,628,496
MAV_120809_2495	*M. avium*	F.III.2	97	26	475,613	5,712,139
MIN_052511_1280	*M. intracellulare*	B.III.1	46	26	488,138	6,331,706
MIN_061107_1834	*M. intracellulare*	D.III.1	42	28	382,148	5,728,455



Neighbor-joining phylogenetic tree based on whole-genome multiple alignment of 24 *M. abscessus* group genomes[8]. Genomes were aligned using Mugsy, core segments of the alignment were identified using Phylomark, and resulting concatenated nucleotide sequences were used for construction of the midpoint-rooted NJ tree using MEGA. Strains from an outbreak at a cystic fibrosis center in Seattle (sequenced as part of this study) are in red; strains from an outbreak at a cystic fibrosis center in the UK are in blue and purple; strains from Brazil are in magenta; and the type strain is in green[7].

Discussion

This study represents a unique opportunity to investigate the association between host and pathogen genotypes in patients with PNTM. As demonstrated above, ongoing analysis of the mycobacterial genome sequences will provide insights into the appearance and stabilization of SNPs associated with antibiotic resistance, bacterial phenotypes *in vitro* and *in vivo*, transmissibility (outbreaks), and adaptation to the lung environment, including cystic fibrosis microbiota. These findings will lead to the development of novel PCR-based diagnostics for early detection of adapted, drug resistant, and/or transmissible mycobacteria in the clinical microbiology laboratory. The data generated here will be used to form recommendations for the modification of official Infection Control Guidelines, especially for handling cystic fibrosis patients.

The human exome sequences generated by this project have led to the discovery of a number of potentially novel and high-impact variants associated with PNTM susceptibility. These variants are currently being validated across the entire set of 99 subjects sequenced here, as well as a larger cohort of individuals with CF and PNTM. Although host susceptibility to PNTM is likely to be multi-factorial, the discovery of associated variants could lead to earlier diagnosis and modified treatment for PNTM. Combining the exome data with the mycobacterial genome sequences is further enabling investigation into associations between segregating PNTM host mutations and mycobacterial surface determinants potentially involved in host-pathogen interactions.