Granularity of *SERPINA1* alleles by DNA sequencing in CanCOLD

Nisha Gupta[1,2], Nathalie Gaudreault[3], Sébastien Thériault[3], Pei Zhi Li[1], Cyndi Henry[3], Miranda Kirby[4], François Maltais[3], Wan Tan[5], Jean Bourbeau[1,2] and Yohan Bossé[3,6] for the Canadian Respiratory Research Network and the CanCOLD study group[7]

Affiliations: [1]Research Institute of the McGill University Health Centre, Montréal, QC, Canada. [2]McGill University, Montréal, QC, Canada. [3]Institut universitaire de cardiologie et de pneumologie de Québec, Québec, QC, Canada. [4]Ryerson University, Toronto, ON, Canada. [5]University of British Columbia, Vancouver, BC, Canada. [6]Dept of Molecular Medicine, Laval University, Québec, QC, Canada. [7]A list of study group members can be found in the acknowledgements section.

Correspondence: Yohan Bossé, Laval University, Department of Molecular Medicine, Institut universitaire de cardiologie et de pneumologie de Québec – Université Laval, Pavillon Marguerite-d'Youville, Y2106, 2725 chemin Sainte-Foy, Québec (Québec), Canada, G1V 4G5. E-mail: yohan.bosse@criucpq.ulaval.ca

ABSTRACT DNA sequencing of the *SERPINA1* gene to detect α_1-antitrypsin (AAT) deficiency (AATD) may provide a better appreciation of the individual and cumulative impact of genetic variants on AAT serum levels and COPD phenotypes.

AAT serum level and DNA sequencing of the coding regions of *SERPINA1* were performed in 1359 participants of the Canadian Cohort Obstructive Lung Disease (CanCOLD) study. Clinical assessment for COPD included questionnaires, pulmonary function testing and computed tomography (CT) imaging. Phenotypes were tested for association with *SERPINA1* genotypes collated into four groups: normal (MM), mild (MS and MI), intermediate (heterozygote MZ, non-S/non-Z/non-I, compound IS, and homozygote SS) and severe (ZZ and SZ) deficiency. Smoking strata and MZ-only analyses were also performed.

34 genetic variants were identified including 25 missense mutations. Overall, 8.1% of alleles in this Canadian cohort were deficient and 15.5% of 1359 individuals were carriers of at least one deficient allele. Four AATD subjects were identified and had statistically lower diffusion capacity and greater CT-based emphysema. No COPD phenotypes were associated with mild and intermediate AATD in the overall cohort or stratified by smoking status. MZ heterozygotes had similar CT-based emphysema, but lowered diffusion capacity compared with normal and mild deficiency.

In this Canadian population-based cohort, comprehensive genetic testing for AATD reveals a variety of deficient alleles affecting 15.5% of subjects. COPD phenotype was demonstrated in severe deficiency and MZ heterozygotes. This study shows the feasibility of implementing a diagnostic test for AATD using DNA sequencing in a large cohort.

Introduction

α_1-Antitrypsin deficiency (AATD) is an inherited disorder associated with accelerated rate of lung function decline and early onset emphysema [1–4]. AATD is caused by genetic mutations in the *SERPINA1* gene located on chromosome 14, which encodes an antiprotease called α_1-antitrypsin (AAT). The prevalence of AATD is estimated at one in 2000 to 5000 in North American populations [5, 6]. Previous studies suggest that severe AATD accounts for 1% to 5% of COPD cases [1, 7, 8], but the exact proportion for the COPD population remains to be described.

Diagnostic methods for AATD vary by country and region, but usually follow a multi-step testing algorithm that includes quantification of AAT in serum or plasma, protease inhibitor phenotyping by isoelectric focussing, and targeted genotyping for the most common mutations (*e.g.* S and Z) [9–11]. The gold standard to detect the deficiency is direct DNA sequencing, which was historically considered laborious, expensive and not available in all centres [2]. In the current genomic era, these arguments are no longer valid. More research and clinical laboratories are transitioning to DNA sequencing as the method of choice [11–15]. In so doing, there is an increasing number of rare variants being identified; however, little is known about their frequencies and clinical impact.

The goals of this study were two-fold: first, to assess the frequencies of AATD alleles in a Canadian population-based cohort of individuals with COPD, at risk of COPD, and never-smokers free of airway obstruction; and, secondly, to evaluate the individual and cumulative impact of AATD alleles on COPD phenotypes including lung function and computed tomography (CT)-based emphysema.

Methods

CanCOLD study

The Canadian Cohort of Obstructive Lung Disease (CanCOLD) is a prospective cohort study built on the Canadian COPD prevalence study "COLD", which evaluated >5000 subjects (male and female subjects aged ⩾40 years) recruited through a random sampling frame from nine urban and suburban areas in Canada (ClinicalTrials.gov identifier: NCT00920348) [16]. Sampling for CanCOLD consists of all COPD subjects as well as age- and sex-matched non-COPD peers (postbronchodilator forced expiratory volume in 1 s (FEV_1)/forced vital capacity (FVC) >0.70) from the COLD study. CanCOLD thus comprises two balanced COPD subpopulations (mild and moderate–severe) and two matched non-COPD subpopulations including ever-smokers (for those at risk) and never-smokers (for the control subjects). Assessments included sociodemographic and clinical status questionnaires (including the COPD Assessment Test, the St. George's Respiratory Questionnaire for COPD (SGRQ-C), mMRC dyspnoea scale and Short Form-36), full pulmonary function tests, CT imaging of the thorax and incremental maximal cardiopulmonary exercise testing. Details about pulmonary function testing and CT imaging are in the supplementary material. Detailed descriptions of the sampling strategy and assessments can be found in the published protocol [16]. A written informed consent was obtained from all subjects and the study was approved by the Institutional Research Ethical Board of each site. For this analysis, only data from the initial visit were used and a final set of 1359 CanCOLD subjects with DNA available were selected for *SERPINA1* sequencing.

DNA sequencing of SERPINA1

The DNA sequences of the coding regions (*i.e.* exons 2 to 5) of the *SERPINA1* gene were obtained by Sanger sequencing in the same laboratory for all subjects. For identified variants with unknown allelic background, allele-specific PCR (AS-PCR) was performed in order to amplify and sequence each allele independently. Details are in the supplementary material.

In silico *functional analysis of genetic variants*

The predicting damaging effects of coding non-synonymous variants were evaluated with PolyPhen-2 [17] and Sorting Intolerant From Tolerant (SIFT) [18]. The deleteriousness of identified genetic variants was also evaluated with the Combined Annotation–Dependent Depletion (CADD) framework [19]. Finally, clinical interpretation was queried in ClinVar [20].

Measurement of AAT serum levels

Blood samples from CanCOLD participants were drawn by venipuncture and processed within 2 h of collection. Serum aliquots were stored at −80°C until analysis. AAT serum levels were measured by immunoturbidimetry on a COBAS INTEGRA 800 analyser (Roche Diagnostics, Laval, QC, Canada). The coefficient of variation for intermediate precision was estimated at 2.8% at a level of 0.76 $g \cdot L^{-1}$ and 2.2% at a level of 1.79 $g \cdot L^{-1}$.

Statistical analysis

The effects of genetic variants on AAT serum levels were evaluated using Wald tests for quantitative phenotype as implemented in PLINK [21]. Clinical data were compared across groups using ANOVA for continuous variables and chi-square for categorical variables. Analyses were also carried out stratified by smoking status (never- and ever-smokers). Analyses were performed with and without adjustment for covariates including age, sex and study sites. *Post hoc* Tukey multiple pairwise-comparison tests were used for statistically significant ANOVA to identify groups that differed.

Results

Identified genetic variants in CanCOLD

Table S1 shows the clinical characteristics of the 1359 CanCOLD participants evaluated in this study. DNA sequencing of exons 2 to 5 in these individuals identified 34 genetic variants. The location of these variants relative to the intron–exon structure of the *SERPINA1* gene is illustrated in figure 1. Table 1 shows the variants based on standardised genetic, nucleotide, protein and protease inhibitor typing nomenclatures. For each variant, the pathogenicity scores and the allele frequency among the 1359 CanCOLD subjects are also provided in table 1. Three new variants were found including two in exon 3 (Lys222Lys and Lys274Asn) and one in exon 4 (Leu318Phe) (figure 1). These new variants are rare with allele frequencies of 0.04% (one heterozygote subject) and predict benign or inconsistent across pathogenicity scores (table 1).

Allele and genotype calling

DNA sequencing provides more granularity compared with the conventional protease inhibitor system. We thus had to use a refined strategy to call alleles and genotypes. Humans are diploid organism, meaning that there are two alleles per individual. For each allele, one or more genetic variants can be present. The absence or presence of a genetic variant is called in relation to an allele (DNA sequence) of reference. For *SERPINA1*, the allele of reference and the most common in human is labelled M1 (Val^{213}) using the protease inhibitor system. In CanCOLD, for example, 1246 out of 2718 alleles (1359 individuals×2 alleles=2718 alleles) are on this background allele. Accordingly, we called the 2718 alleles and the 1359 genotypes of CanCOLD participants based on the absence or presence of the 34 genetic variants identified (figure 1 and table 1). Figure 2 illustrates five representative individuals to better conceptualise the allele and genotype calling process.
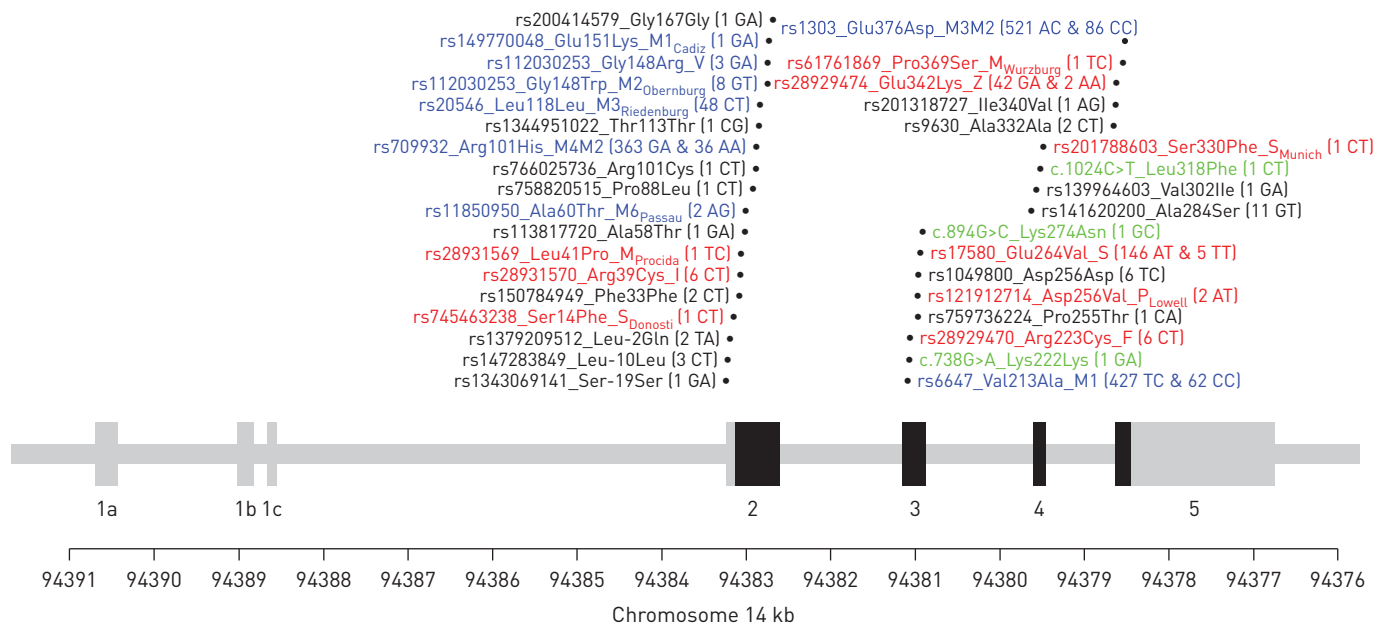

FIGURE 1 The exon–intron structure of the *SERPINA1* gene and the localisation of the identified genetic variants. The coding exons are shown in black and the untranslated regions in grey. Genetic variants are illustrated with their rs numbers (if available), protein nomenclature, and genotyping counts in parentheses for 1359 individuals. Newly identified variants are illustrated in green and named based on standard gene mutation nomenclature [27]. Variants causing a change in conventional protease inhibitor typing are illustrated in blue and red for normal and deficient alleles, respectively. Note that rs112030253 in exon 2 is indicated twice as it is a multi-allelic polymorphism and three alleles are observed in the CanCOLD cohort.

TABLE 1 Nomenclature of the 34 genetic variants identified, allele frequencies in CanCOLD, and pathogenicity scores

Exon	rs number	Nucleotide	Protein	Pi typing	MAF	Effect	AAT p-value[#]	PolyPhen-2[¶]	SIFT[+]	CADD	ClinVar
2	rs1343069141	c.18G>A	Ser-19Ser		0.04%	Unknown	0.789	NA	NA	0.004	NA
2	rs147283849	c.43C>T	Leu-10Leu		0.11%	Unknown	0.068	NA	NA	10.97	Conflicting
2	rs1379209512	c.68T>A	Leu-2Gln		0.07%	Unknown	0.994	0.92 (P)	0.001 (D)	10.56	NA
2	rs745463238[§]	c.113C>T	Ser14Phe	S$_{Donosti}$	0.04%	Deficient	**0.001**	0.708 (P)	0.011 (D)	15.18	NA
2	rs150784949	c.171C>T	Phe33Phe		0.07%	Unknown	0.482	NA	NA	0.991	Conflicting
2	rs28931570[§]	c.187C>T	Arg39Cys	I	0.22%	Deficient	**0.026**	1.0 (D)	0.0 (D)	24.4	Pathogenic
2	rs28931569[§]	c.194T>C	Leu41Pro	M$_{Procida}$	0.04%	Deficient	0.100	0.995 (D)	0.002 (D)	17.59	Pathogenic/ikely pathogenic
2	rs113817720	c.244G>A	Ala58Thr		0.04%	Unknown	0.441	0.276 (B)	0.04 (D)	18.76	Uncertain significance
2	rs111850950	c.250G>A	Ala60Thr	M6$_{Passau}$	0.07%	Normal	0.355	0.999 (D)	0.057 (T)	26.3	Uncertain significance
2	rs758820515	c.335C>T	Pro88Leu		0.04%	Unknown	0.172	0.628 (P)	0.123 (T)	8.747	NA
2	rs766025736[f]	c.373C>T	Arg101Cys		0.04%	Unknown	0.154	0.969 (D)	0.003 (D)	22.3	NA
2	rs709932	c.374G>A	Arg101His	M4/M2	16.00%	Normal	**0.006**	0.0 (B)	1.0 (T)	0.322	Benign/likely benign
2	rs1344951022	c.411C>G	Thr113Thr		0.04%	Unknown	0.892	NA	NA	6.218	NA
2	rs20546	c.424C>T	Leu118Leu	M3$_{Riedenburg}$	1.77%	Normal	0.358	NA	NA	3.036	Benign/likely benign
2	rs112030253	c.514G>T	Gly148Trp	M2$_{Obernburg}$	0.30%	Normal	0.838	0.973 (D)	0.027 (D)	13.59	Conflicting
2	rs112030253	c.514G>A	Gly148Arg	V	0.11%	Normal	0.675	0.0 (B)	0.934 (T)	0.008	Conflicting
2	rs149770048	c.523G>A	Glu151Lys	M1$_{Cadiz}$	0.04%	Normal	NA	0.05 (B)	0.097 (T)	9.991	NA
2	rs200414579	c.573G>A	Gly167Gly		0.04%	Unknown	0.564	NA	NA	0.277	NA
3	rs6647	c.710T>C	Val213Ala	M1	20.27%	Normal	**0.0009**	0.0 (B)	0.597 (T)	0.002	Benign/likely benign
3	New1	c.738G>A	Lys222Lys		0.04%	Unknown	0.803	NA	NA	0.154	NA
3	rs28929470[§]	c.739C>T	Arg223Cys	F	0.22%	Deficient	0.650	0.522 (P)	0.001 (D)	23.1	Pathogenic
3	rs759736224[f]	c.835C>A	Pro255Thr		0.04%	Unknown	0.586	1.0 (D)	0.0 (D)	23.8	NA
3	rs121912714[§]	c.839A>T	Asp256Val	P$_{Lowell}$	0.07%	Deficient	**0.006**	0.847 (P)	0.001 (D)	22.8	Conflicting
3	rs1049800	c.840T>C	Asp256Asp		0.22%	Unknown	0.953	NA	NA	0.085	Benign/likely benign
3	rs17580[§]	c.863A>T	Glu264Val	S	5.74%	Deficient	**1.15×10^{-34}**	0.998 (D)	0.0 (D)	23.7	Pathogenic
3	New2	c.894G>C	Lys274Asn		0.04%	Unknown	0.861	0.003 (B)	0.167 (T)	8.926	NA
4	rs141620200	c.922G>T	Ala284Ser		0.40%	Unknown	0.272	0.009 (B)	0.065 (T)	16.02	Conflicting
4	rs139964603	c.976G>A	Val302Ile		0.04%	Unknown	0.567	0.002 (B)	0.602 (T)	0.004	Conflicting
4	New3	c.1024C>T	Leu318Phe		0.04%	Unknown	0.252	0.776 (P)	0.053 (T)	24.9	NA
4	rs201788603[f]	c.1061C>T	Ser330Phe	S$_{Munich}$	0.04%	Conflicting	0.764	1.0 (D)	0.0 (D)	27.3	Uncertain significance
5	rs9630	c.1068C>T	Ala332Ala		0.07%	Unknown	0.881	NA	NA	0.068	Benign
5	rs201318727	c.1090A>G	Ile340Val		0.04%	Unknown	0.726	0.001 (B)	0.244 (T)	4.739	NA
5	rs28929474[§]	c.1096G>A	Glu342Lys	Z	1.69%	Deficient	**6.23×10^{-51}**	1.0 (D)	0.006 (D)	20.2	Pathogenic
5	rs61761869[§]	c.1177C>T	Pro369Ser	M$_{Wurzburg}$	0.04%	Deficient	**0.033**	1.0 (D)	0.003 (D)	23.4	Conflicting
5	rs1303	c.1200A>C	Glu376Asp	M3/M2	25.50%	Normal	**0.008**	0.0 (B)	1.0 (T)	0.001	Benign/likely benign

MAF: minor allele frequency; AAT: α_1-antitrypsin; SIFT: Sorting Intolerant From Tolerant. [§]: known deficient alleles; [f]:three additional variants considered deleterious by all pathogenicity scores. Note that rs112030253 in exon 2 is indicated twice as it is a multi-allelic polymorphism and three alleles are observed in CanCOLD. [#]: Genetic association with AAT serum levels; [¶]: D, probably damaging; P, possibly damaging; B, benign; [+]: D, damaging; T, tolerated.
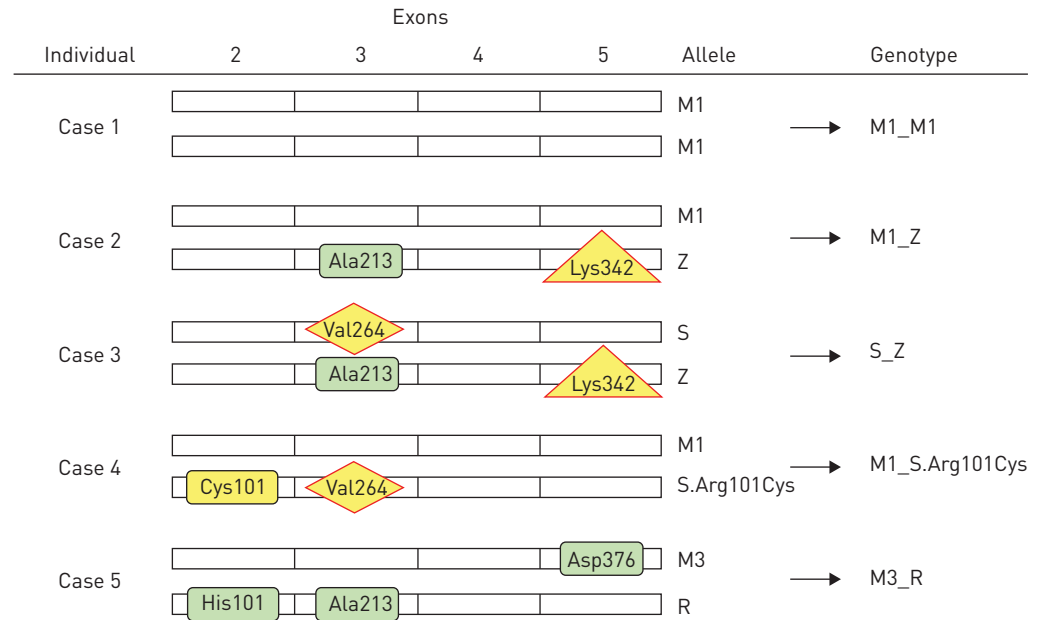
FIGURE 2 Schematic examples of allele and genotype calling derived from DNA sequencing data for five individuals. Case 1 is homozygote for the M1 (Val[213]) allele. This is the most common and normal allele in humans. The DNA nucleotide sequence of M1 is thus considered the reference. Case 2 is a typical MZ individual with one normal M1 allele and one deficient Z allele. The Z allele is the result of a missense mutation (Glu342Lys) in exon 5. Except for very rare cases, the Z allele is found on the M1 (Ala[213]) background. Case 3 is a SZ individual with one Z allele and one S allele characterised by the Glu264Val mutation in exon 3. Case 4 is carrier of the rare mutation Arg101Cys on the S background. Arg101Cys is considered deleterious based on pathogenicity scores (table 1). There are thus two hits (Cys101 and Val264) on the same allele. For this individual, knowing the allelic background was clinically relevant to determine the AATD status; *i.e.* the individual is heterozygote for one deficient allele and not compound heterozygote (two deficient alleles). Case 5 is carrier of a typical M3 allele characterised by Glu376Asp and a second unusual allele with His101 and Ala213, labelled R for rare allele background. These are all considered normal genetic variants, so there is no clinical implication of determining the allelic background of this individual. However, the observation is of great interest for population genetics. Note that this allele and genotype calling process was used to derive all alleles (n=2718) and genotypes (n=1359) of the 1359 CanCOLD individual.

Frequency of alleles

The 34 genetic variants in CanCOLD results in 40 distinct alleles. Table 2 shows the distribution of alleles in the overall CanCOLD cohort and by recruitment sites. The most frequent normal alleles in this population are 45.8% for M1 (Val[213]), 17.5% for M1 (Ala[213]), 14.6% for M2, 9.1% for M3, 1.8% for M3$_{Riedenburg}$, and 1.3% for M4. In total, 91.9% of the 2718 alleles are normal. In contrast, the others are pathogenic, including known deficient or dysfunctional alleles such as S (5.7%), Z (1.7%), F (0.2%), I (0.2%) as well as more rare variants such as P$_{Lowell}$ found in two heterozygote subjects and M$_{Procida}$ and M$_{Wurzburg}$ each found in one heterozygote subject. Four additional rare variants, each found in a single individual, are claimed pathogenic by all scores (table 1) including M3.Pro255Thr, S$_{Donosti}$, S$_{Munich}$ and S. Arg101Cys. Taking together, 220 deficient alleles are identified and indicate that 8.1% of alleles (220 out of 2718) in this Canadian cohort are deficient. The percentage of deficient alleles varies from 5.5% to 11% across recruitment sites (table 2).

Genotype frequencies

The genotype of each individual consists of the combination of both alleles. Figure 3 illustrates the number of individuals in any combinations of normal and deficient alleles. Two subjects carry two copies of the Z allele and are thus confirmed cases of AATD. One other subject was compound heterozygote SZ. A second SZ individual was identified, but carrying also the Ser14Phe mutation. Allele-specific PCR indicated that this mutation occurs on the background of S allele, which is consistent with a previously reported allele labelled S$_{Donosti}$ [14]. As a result, there were a total of four CanCOLD subjects with severe AATD. 40 additional individuals were heterozygotes for the Z allele. Five SS individuals were found, and 143 additional individuals carried the S allele (142 MS and 1 IS). Six subjects were carriers of the F allele. Six carriers of allele I were also found including one IS. Rarer deficient alleles were identified in two

TABLE 2 Allele distribution sorted by allele frequency in the overall CanCOLD cohort and by recruitment sites sorted by geographic location from west to east

Allele	CanCOLD	Vancouver	Calgary	Saskatoon	Toronto	Kingston	Ottawa	Montreal	Quebec	Halifax
M1	1246 (45.84)	391 (48.39)	119 (50)	83 (45.6)	56 (43.75)	84 (42.86)	95 (46.57)	255 (45.05)	63 (38.41)	100 (43.1)
M1(ala)	475 (17.48)	134 (16.58)	43 (18.07)	34 (18.68)	19 (14.84)	31 (15.82)	37 (18.14)	90 (15.9)	30 (18.29)	57 (24.57)
M2	396 (14.57)	114 (14.11)	36 (15.13)	22 (12.09)	26 (20.31)	33 (16.84)	25 (12.25)	88 (15.55)	28 (17.07)	24 (10.34)
M3	246 (9.05)	79 (9.78)	18 (7.56)	14 (7.69)	11 (8.59)	18 (9.18)	19 (9.31)	50 (8.83)	15 (9.15)	22 (9.48)
S[#]	154 (5.67)	29 (3.59)	12 (5.04)	7 (3.85)	5 (3.91)	14 (7.14)	10 (4.9)	53 (9.36)	16 (9.76)	8 (3.45)
M3(Riedenburg)	48 (1.77)	12 (1.49)	2 (0.84)	4 (2.2)	5 (3.91)	4 (2.04)	5 (2.45)	6 (1.06)	5 (3.05)	5 (2.16)
Z[#]	46 (1.69)	20 (2.48)	1 (0.42)	5 (2.75)	1 (0.78)	3 (1.53)	3 (1.47)	6 (1.06)	1 (0.61)	6 (2.59)
M4	36 (1.32)	6 (0.74)	3 (1.26)	7 (3.85)	2 (1.56)	6 (3.06)	2 (0.98)	4 (0.71)	1 (0.61)	5 (2.16)
M1(ala).Ala284Ser	11 (0.4)	3 (0.37)	0 (0)	1 (0.55)	0 (0)	1 (0.51)	4 (1.96)	1 (0.18)	0 (0)	1 (0.43)
M2(Obernburg)	8 (0.29)	1 (0.12)	0 (0)	1 (0.55)	0 (0)	0 (0)	0 (0)	4 (0.71)	2 (1.22)	0 (0)
F[#]	6 (0.22)	3 (0.37)	2 (0.84)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.18)	0 (0)	0 (0)
I[#]	6 (0.22)	2 (0.25)	1 (0.42)	0 (0)	0 (0)	0 (0)	1 (0.49)	1 (0.18)	1 (0.61)	0 (0)
M1.Asp256Asp	5 (0.18)	3 (0.37)	0 (0)	0 (0)	1 (0.78)	0 (0)	1 (0.49)	0 (0)	0 (0)	0 (0)
M1(ala).Leu-10Leu	3 (0.11)	1 (0.12)	0 (0)	0 (0)	1 (0.78)	0 (0)	0 (0)	1 (0.18)	0 (0)	0 (0)
V	3 (0.11)	0 (0)	0 (0)	1 (0.55)	0 (0)	1 (0.51)	0 (0)	1 (0.18)	0 (0)	0 (0)
M1(ala).Leu-2Gln	2 (0.07)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	2 (0.86)
M1.Phe33Phe	2 (0.07)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.49)	0 (0)	1 (0.61)	0 (0)
M6(Passau)	2 (0.07)	1 (0.12)	0 (0)	0 (0)	0 (0)	1 (0.51)	0 (0)	0 (0)	0 (0)	0 (0)
P(Lowell) [#]	2 (0.07)	1 (0.12)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.49)	0 (0)	0 (0)	0 (0)
M(Procida) [#]	1 (0.04)	1 (0.12)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
M(Wurzburg) [#]	1 (0.04)	0 (0)	0 (0)	0 (0)	1 (0.78)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
M1(ala).Ala332Ala	1 (0.04)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.18)	0 (0)	0 (0)
M1(ala).Asp256Asp	1 (0.04)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.18)	0 (0)	0 (0)
M1(ala).Lys274Asn	1 (0.04)	0 (0)	0 (0)	1 (0.55)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
M1(ala).Pro88Leu	1 (0.04)	0 (0)	1 (0.42)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
M1(Cadiz)	1 (0.04)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.43)
M1.Ala332Ala	1 (0.04)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.61)	0 (0)	0 (0)
M1.Ala58Thr	1 (0.04)	0 (0)	0 (0)	1 (0.55)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
M1.Leu318Phe	1 (0.04)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.18)	0 (0)	0 (0)	0 (0)
M1.Lys222Lys	1 (0.04)	1 (0.12)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
M1.Thr113Thr	1 (0.04)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.18)	0 (0)	0 (0)	0 (0)
M1.Val302Ile	1 (0.04)	1 (0.12)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
M3.Gly167Gly	1 (0.04)	1 (0.12)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
M3.Pro255Thr[#]	1 (0.04)	1 (0.12)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
M3.Ser-19Ser	1 (0.04)	1 (0.12)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
R	1 (0.04)	0 (0)	0 (0)	1 (0.55)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
R.Ile340Val	1 (0.04)	1 (0.12)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
S(Donosti) [#]	1 (0.04)	1 (0.12)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
S(Munich) [#]	1 (0.04)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.43)
S.Arg101Cys[#]	1 (0.04)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.18)	0 (0)	0 (0)
Total	2718 (100)	808 (100)	238 (100)	182 (100)	128 (100)	196 (100)	204 (100)	566 (100)	164 (100)	232 (100)
Normal allele	2498 (91.91)	750 (92.82)	222 (93.28)	170 (93.41)	121 (94.53)	179 (91.33)	189 (92.65)	504 (89.05)	146 (89.02)	217 (93.53)
Deficient allele	220 (8.09)	58 (7.18)	16 (6.72)	12 (6.59)	7 (5.47)	17 (8.67)	15 (7.35)	62 (10.95)	18 (10.98)	15 (6.47)

Data are presented as allele count (% in site). [#]: known deficient alleles or considered deleterious by all pathogenicity scores.

P_{Lowell}, one $M_{Procida}$, one $M_{Wurzburg}$, one S_{Munich} and one M3.Pro255Thr heterozygote carriers. Overall, 210 out of 1359 subjects (15.5%) were carriers of at least one deficient allele.

SERPINA1 genotyping grouping

SERPINA1 genotyping groups were then generated from the DNA sequencing data to reflect the resulting AAT serum level or functional activity of AAT (figure 4). At the most detailed level, the 40 distinct alleles found in CanCOLD lead to 77 distinct genotypes with frequencies ranging from 0.07% (one out of 1359 individuals) to 21.3% (289 out of 1359 individuals, *i.e.* genotype M1M1) (see abscissa of figure 4). Although there is no standard for grouping genotypes, we have classified individuals into four groups based on the information of both alleles and the known or predicting effect of each genetic variant on AAT levels or activity. Group 1 (no deficiency) includes all individuals with two normal alleles (n=1149). Group 2 (mild deficiency) are individuals with one normal allele and one S or I allele (n=147, MS and

Allele	R*	M3.Pro255Thr	S.Arg101Cys	V	$M1_{Cadiz}$	S_{Munich}	P_{Lowell}	I	F	$M_{Wurzburg}$	$S_{Donosti}$	Z	S	$M3_{Riedenburg}$	$M2_{Obernburg}$	$M_{Procida}$	$M6_{Passau}$	M4	M3	M2	M1	$M1(Ala^{213})$
$M1(Ala^{213})$	0	0	0	0	0	0	0	1	1	0	0	6	43	12	1	0	0	7	52	59	207	53
M1	0	0	1	3	0	1	2	3	4	1	0	25	61	20	3	0	1	18	116	195	299	
M2	0	1	0	0	0	0	0	1	1	0	0	8	23	7	2	1	0	9	37	26		
M3	2	0	0	0	1	0	0	0	0	0	0	1	9	2	2	0	0	0	13			
M4	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0				
$M6_{Passau}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
$M_{Procida}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0						
$M2_{Obernburg}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0							
$M3_{Riedenburg}$	0	0	0	0	0	0	0	0	0	0	0	0	6	0								
S	0	0	0	0	0	0	0	1	0	0	0	1	5									
Z	0	0	0	0	0	0	0	0	0	0	1	2										
$S_{Donosti}$	0	0	0	0	0	0	0	0	0	0	0											
$M_{Wurzburg}$	0	0	0	0	0	0	0	0	0	0												
F	0	0	0	0	0	0	0	0	0													
I	0	0	0	0	0	0	0	0														
P_{Lowell}	0	0	0	0	0	0	0															
S_{Munich}	0	0	0	0	0	0																
$M1_{Cadiz}$	0	0	0	0	0																	
V	0	0	0	0																		
S.Arg101Cys	0	0	0																			
M3.Pro255Thr	0	0																				
R*	0																					

1359

FIGURE 3 Number of patients identified in any pairwise combinations of normal and deficient alleles. Alleles are annotated based on the legacy nomenclature, *i.e.* protease inhibitor system, if available, or on protein nomenclature. Normal alleles without protease inhibitor naming were combined with their respective background to ease visualisation (for example, allele M1.Val302Ile was merged with M1, this allele does not have a protease inhibitor naming and Val302Ile is benign). Each individual is represented in only one cell and the combinations of alleles for the 1359 CanCOLD participants are indicated. The colour of the squares illustrates the expected level of deficiency on a black and white scale where white indicates no deficiency and black indicates severe deficiency. *: Rare allelic background (1. His101 and Ala213 on the same allele, and 2. His101, Ala213, and Val340 on the same allele).

MI). Group 3 (intermediate deficiency) consists of individuals that are heterozygotes (MZ, MF, $MM_{Procida}$, $MM_{Wurzburg}$, MP_{Lowell}, MS_{Munich}, MS.Arg101Cys, MM3.Pro255Thr), homozygotes for the S allele (SS) and compound heterozygotes for deficient alleles (IS) (n=59). Group 4 (severe deficiency) are AATD individuals (n=4, ZZ, SZ, and $S_{Donosti}Z$).

Association of SERPINA1 *genotyping groups with AAT serum levels and COPD phenotypes*
Table 3 shows the clinical characteristics of CanCOLD participants according to the SERPINA1 genotyping groups. Serum AAT levels were strongly associated with genotyping groups (ANOVA p<0.001; figure 5a). The 6-heterozygote individuals for the F allele have serum AAT levels ranging with those with two normal alleles. Diffusion capacity to carbon monoxide (D_{LCO}) and the percentage of lung voxels below −950 Hounsfield units ($LAA_{−950}$) were also statistically different across these groups (table 3 and figure 5b and c). These associations were driven by AATD individuals (group 4: severe deficiency) as no significant differences were observed between group 1 (no deficiency) and group 3 (intermediate deficiency). No statistically significant differences were observed across these genotyping groups for smoking habits, lung function, symptoms and quality of life, self-reported comorbidities and respiratory medications (table 3).

Intermediate AAT deficiency, mostly defined by MZ heterozygotes, was previously associated with an increased risk of COPD, especially in the presence of smoking exposure [22–25]. We thus repeated the analyses by smoking status (never- and ever-smokers). As observed for the entire cohort, the SERPINA1 genotyping groups were significantly different for $LAA_{−950}$ in both never- and ever-smokers, and D_{LCO} in
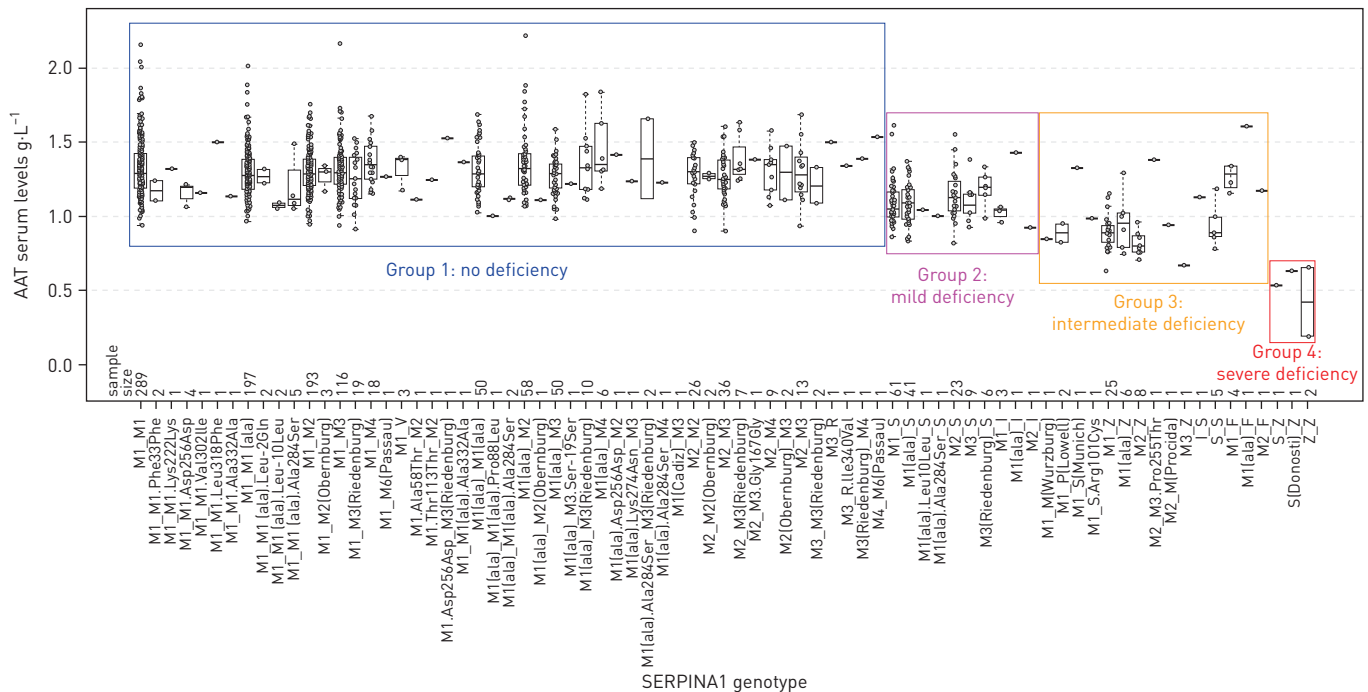
FIGURE 4 Serum levels of α₁-antitrypsin (AAT) by *SERPINA1* genotypes. *SERPINA1* genotyping groups are depicted by rectangles. Groups were generated to reflect the resulting impact of the genetic variants on AAT serum level or functional activity of AAT. For rare mutations of unknown biological and clinical significance, the grouping was guided based on the actual serum level of AAT among carriers and pathogenicity scores. This classification system results in four groups: Group 1 (no deficiency) includes all individuals with two normal alleles (n=1149). Group 2 (mild deficiency) are individuals with one normal allele and one S or I allele (n=147, MS and MI). Group 3 (intermediate deficiency) consists of individuals that are heterozygotes (MZ, MF, MM$_{Procida}$, MM$_{Wurzburg}$, MP$_{Lowell}$, MS$_{Munich}$, MS.Arg101Cys, MM3.Pro255Thr), homozygotes for the S allele (SS) and compound heterozygotes for deficient alleles (IS) (n=59). Group 4 (severe deficiency) are AATD individuals (n=4, ZZ, SZ and S$_{Donosti}$Z). AAT serum levels are not available for 15 out of 1359 individuals including the single individual carrying the M1$_{Cadiz}$ allele, which explains the absence of boxplot for the M1(Cadiz)_M3 genotype. M3_R, "R" indicates rare allelic background.

never-smokers (table S2). Again, *post hoc* tests revealed that group 4 (severe deficiency), but not group 3 (intermediate deficiency), is significantly different from group 1 (no deficiency). We have also evaluated the interaction between *SERPINA1* genotyping groups and smoking (ever *versus* never smokers) on COPD phenotypes, but no significant interactions were identified (table 3).

Analysis of MZ heterozygotes

Finally, COPD phenotypes were compared by *SERPINA1* genotyping groups, but this time, limiting intermediate deficiency to MZ heterozygotes (n=40). Compared to normal and mild deficiency, MZ had similar lung function and CT-based emphysema (Table S3). However, D_{LCO} was statistically lower (p<0.001) in MZ (95.5%±21.2%) and severe deficiency (69.4%±19.1%) compared with normal (106.4%±24.3%) and mild deficiency (109.9%±26.5%).

Discussion

In this Canadian cohort, 34 genetic variants were identified by sequencing the coding regions of the *SERPINA1* gene. This includes genetic variants accounting for common normal alleles, such as M1, M1 (Ala[213]), M2, M3 and M4, but also more rare normal alleles, such as M3$_{Riedenburg}$, M2$_{Obernburg}$, V, M6$_{Passau}$ and M1$_{Cadiz}$. Collectively, 91.9% of all the alleles evaluated were considered normal. In contrast, a variety of 220 deficient alleles was found in these 1359 individuals. Not surprisingly, the most common were S and Z, but mutations causing alleles I, F, P$_{Lowell}$, M$_{Procida}$, M$_{Wurzburg}$, M3.Pro255Thr, S$_{Donosti}$, S$_{Munich}$ and S.Arg101Cys were also identified. Deficient alleles were observed in 210 individuals, indicating that 15.5% of CanCOLD subjects carried at least one deficient allele. Grouping subjects by genotypes reflecting the serum level or functional activity of AAT indicated that severe deficiency (ZZ, SZ and S$_{Donosti}$Z) had lower diffusion capacity and greater CT-based emphysema compared with subjects with no deficiency (MM). These differences were not observed in mild (MS and MI) and intermediate (heterozygote MZ, heterozygote non-S/non-Z/non-I, homozygote SS and compound heterozygote IS) deficiency. However, in MZ-only analysis, MZ heterozygotes had lower diffusion capacity compared with normal and mild deficiency.

TABLE 3 Clinical characteristics by *SERPINA1* genotyping groups

Characteristics	Group 1: no deficiency, n=1149	Group 2: mild deficiency, n=147	Group 3: intermediate deficiency, n=59	Group 4: severe deficiency, n=4	p-value[#]	p-value adjusted for age, sex and study sites	p-value for interaction with smoking[¶]
Age years	66.8±9.7	64.9±9.6	68.8±8.6	68.3±10.6	0.057	–	–
AAT levels g·L^{-1}	1.31±0.17; 12[+a]	1.10±0.14; 1[+b]	0.96±0.20; 2[+c]	0.50±0.21[d]	**<0.001**	**<0.001**	0.445
Sex					0.553		–
Male	647 (56.3)	85 (57.8)	36 (61.0)	1 (25.0)			
Female	502 (43.7)	62 (42.2)	23 (39.0)	3 (75.0)			
BMI kg·m^{-2}	27.6±5.0; 1[+]	27.9±5.3	27.0±4.3	25.7±4.8	0.650	0.881	0.876
Smoking status					0.692	0.801	–
Never	401 (34.9)	55 (37.4)	15 (25.4)	2 (50.0)			
Former	561 (48.8)	70 (47.6)	35 (59.3)	2 (50.0)			
Current	187 (16.3)	22 (15.0)	9 (15.3)	0 (0.0)			
Pack-years (ever-smokers)	16.7±22.5; 19[+]	16.9±21.9; 1[+]	16.2±19.8; 1[+]	11.7±22.2	0.954	0.982	–
Post-bronchodilator FEV$_1$ L	2.55±0.81	2.62±0.88	2.58±0.82	2.18±0.62	0.693	0.921	0.424
Post-bronchodilator FEV$_1$ % predicted	91.5±20.4	90.8±19.7	91.9±21.5	90.1±27.8	0.883	0.992	0.33
Post-FEV$_1$/FVC	69.5±10.4	69.4±9.8	68.9±12.4	59.7±13.0	0.406	0.271	0.464
FEV$_1$ reversibility, %	5.4±7.9	6.7±8.9	5.1±10.9	7.9±12.9	0.284	0.346	0.553
FEF$_{25-75}$	1.77±1.04	1.79±1.04	1.79±1.00	1.05±0.68	0.492	0.519	0.662
D$_{LCO}$ % predicted	106.4±24.3; 56[+a]	109.9±26.5; 11[+a]	102.5±25.3; 6[+a]	69.4±19.1[b]	**0.005**	**0.008**	0.289
RV % predicted	120.1±34.8; 61[+]	125.1±34.9; 10[+]	117.1±32.7; 3[+]	145.9±33.1	0.099	0.12	0.636
FRC % predicted	110.3±25.6; 54[+]	113.6±25.2; 10[+]	108.6±25.6; 3[+]	136.9±19.7	0.052	0.067	0.892
TLC % predicted	118.1±17.4; 52[+]	120.2±18.3; 11[+]	116.2±15.7; 3[+]	133.8±11.9	0.127	0.097	0.942
LAA$_{-950}$ %	4.30±4.52; 135[+a]	4.11±4.23; 25[+a]	4.22±4.92; 2[+a]	19.89±12.18[b]	**0.014**	**<0.001**	0.169
PD15	−917.2±21.5; 135[+a]	−916.5±19.8; 25[+a]	−918.7±18.6; 2[+a]	−954.5±15.0[b]	**0.015**	**0.003**	0.939
Pi10_all	3.96±0.16; 135[+]	3.97±0.14; 25[+]	3.99±0.17; 2[+]	4.00±0.14	0.34	0.283	0.817
Pi10_leq20	3.87±0.11; 135[+]	3.88±0.11; 25[+]	3.90±0.12; 2[+]	3.90±0.08	0.478	0.371	0.714
MRC scales (level 1–5)	1.44±0.65; 51[+]	1.40±0.67; 11[+]	1.49±0.81; 4[+]	1.75±0.50	0.343	0.761	0.634
MRC scales ⩾3 n (%)	58 (5.3); 51[+]	8 (5.9); 11[+]	6 (10.9); 4[+]	0 (0.0)	0.335	0.421	0.996
CAT score	6.8±5.9; 25[+]	7.9±6.7; 3[+]	6.5±8.1	5.8±2.9	0.078	0.131	0.42
SGRQ total score	13.3±14.4; 255[+]	15.5±17.0; 26[+]	15.2±19.0; 15[+]	20.0±4.4; 1[+]	0.332	0.299	0.302
Self-reported comorbidities n (%)							
Hypertension	402 (35.0)	51 (34.7)	19 (32.2)	2 (50.0)	0.88	0.699	0.624
CVD (any CVD excluding hypertension)	325 (28.3)	40 (27.2)	21 (35.6)	1 (25.0)	0.618	0.854	0.345
Diabetes	123 (10.7)	13 (8.8)	8 (13.6)	0 (0.0)	0.724	0.903	**0.021**
Asthma	267 (23.2)	32 (21.8)	16 (27.1)	1 (25.0)	0.813	0.725	0.57

Continued

TABLE 3 Continued

Characteristics	Group 1: no deficiency, n=1149	Group 2: mild deficiency, n=147	Group 3: intermediate deficiency, n=59	Group 4: severe deficiency, n=4	p-value[#]	p-value adjusted for age, sex and study sites	p-value for interaction with smoking[¶]
Respiratory medications n (%)							
SAMA/SABA	35 (3.0)	6 (4.1)	5 (8.5)	0 (0.0)	0.128	0.163	1
LABA±SAMA/SABA	3 (0.3)	0 (0.0)	0 (0.0)	0 (0.0)	1.000	1.000	1
LAMA±SAMA/SABA	10 (0.9)	3 (2.0)	1 (1.7)	0 (0.0)	0.248	0.567	1
LAMA+LABA±SAMA/SABA	71 (6.2)	11 (7.5)	2 (3.4)	0 (0.0)	0.690	0.826	0.485
ICS±SAMA/SABA	2 (0.2)	0 (0.0)	0 (0.0)	0 (0.0)	1.000	1.000	1
LABA+ICS±SAMA/SABA	99 (8.6)	10 (6.8)	1 (1.7)	1 (25.0)	0.098	0.232	0.442
LAMA+ICS±SAMA/SABA	3 (0.3)	1 (0.7)	0 (0.0)	0 (0.0)	0.490	0.819	1
LAMA+LABA+ICS±SAMA/SABA	32 (2.8)	5 (3.4)	3 (5.1)	0 (0.0)	0.469	0.799	1
Any above medications	255 (22.2)	36 (24.5)	12 (20.3)	1 (25.0)	0.855	0.936	0.995

Data are presented as mean±SD or n (%), unless otherwise stated. [#]: *Post hoc* Tukey multiple pairwise-comparison tests were used for statistically significant ANOVA. Means with different letters indicate groups that significantly differed. [¶]: p-value for the interaction term between *SERPINA1* genotyping groups and smoking (ever *versus* never smokers), adjusted for age, sex, and study sites. [+]: missing values. For lung function measurements, the maximum values among all the trials are indicated. Group 1 includes all individuals with two normal alleles (n=1149). Group 2 are individuals with one normal allele and one S or I allele (n=147, MS and MI). Group 3 consists of individuals that are heterozygotes (MZ, MF, MM$_{Procida}$, MM$_{Wurzburg}$, MP$_{Lowell}$, MS$_{Munich}$, MS.Arg101Cys, MM3.Pro255Thr), homozygotes for the S allele (SS) and compound heterozygotes for deficient alleles (IS) (n=59). Group 4 are AATD individuals (n=4, ZZ, SZ, and S$_{Donosti}$Z). AAT: α_1-Antitrypsin; BMI: body mass index; FEV$_1$: forced expiratory volume in 1 s; FVC: forced vital capacity; FEF$_{25-75}$: forced expiratory flow at 25–75% of FVC; D_{LCO}: diffusion capacity to carbon monoxide; RV: residual volume; FRC: functional residual capacity; TLC: total lung capacity; LAA$_{-950}$: percentage of lung voxels below −950 Hounsfield units; PD15: 15th percentile of the computed tomography attenuation histogram; Pi10_all: airway wall thickness of an airway with internal perimeter of 10 mm generated using all computed tomography airways; Pi10_leq20: airway wall thickness of a theoretical airway with internal perimeter of 10 mm generated using all computed tomography airways with an internal perimeter ⩽20 mm; MRC: modified British Medical Research Council Questionnaire; CAT: COPD Assessment Test; SGRQ: St. George's Respiratory Questionnaire for COPD; CVD: cardiovascular diseases; SAMA: short-acting muscarinic antagonist; SABA: short-acting β-agonist; LABA: long-acting β-agonist; LAMA: long-acting muscarinic antagonist; ICS: inhaled corticosteroid.
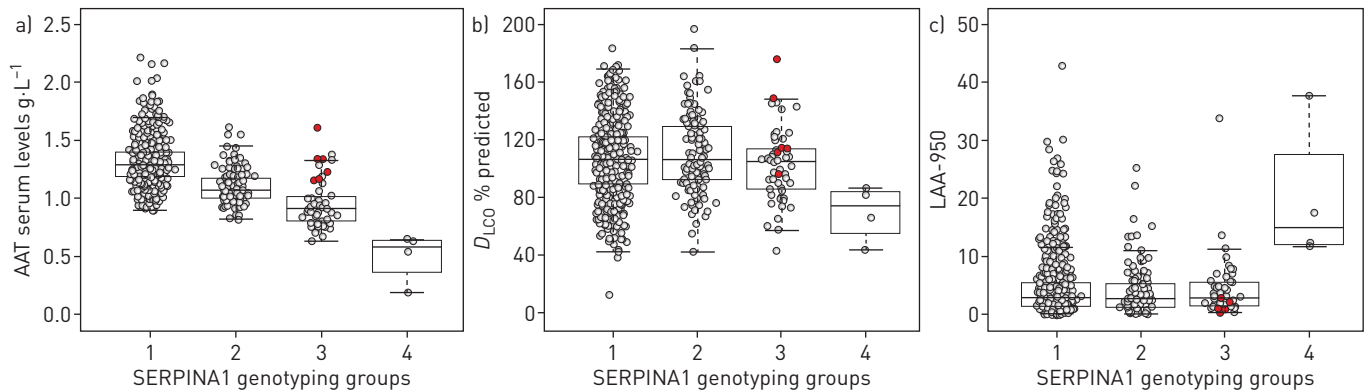
FIGURE 5 Serum levels of α_1-antitrypsin (AAT) (a), diffusing capacity of the lung for carbon monoxide (D_{LCO}) (b), and percentage of lung voxels below −950 Hounsfield units (LAA$_{-950}$) (c) by *SERPINA1* genotyping groups derived from DNA sequencing data. Genotyping groups are on the abscissa. Group 1 includes all individuals with two normal alleles (n=1149). Group 2 are individuals with one normal allele and one S or I allele (n=147, MS and MI). Group 3 consists of individuals that are heterozygotes (MZ, MF, MM$_{Procida}$, MM$_{Wurzburg}$, MP$_{Lowell}$, MS$_{Munich}$, MS.Arg101Cys, MM3.Pro255Thr), homozygotes for the S allele (SS) and compound heterozygotes for deficient alleles (IS) (n=59). Group 4 are AAT deficient individuals (n=4, ZZ, SZ and S$_{Donosti}$Z). Boxplot boundaries represent the first and third quartiles, whiskers are the most extreme data points which are no more than 1.5-times the interquartile range, and the centre mark represents the median. Red dots illustrate the six individuals carrying one F allele. F allele have dysfunctional proteins but normal AAT blood levels [28], which exemplifies the intrinsic limitation of relying solely on serum measurement of AAT and commercially available genotyping tests to establish the diagnosis of AAT deficiency.

Four subjects (0.29%) had AATD in CanCOLD, *i.e.* ZZ, SZ or S$_{Donosti}$Z. This is lower than the 1.9% of ZZ found in 965 severe COPD patients [7], but higher than the frequency of severe AATD commonly reported at one out of 2000–5000 individuals [5, 6]. Random sampling in this study [16] which includes subgroups without airway obstruction, those at-risk of COPD, and individuals with COPD with a range of severity may explain this difference. Nevertheless, deficient alleles together including Z, S, I, F, P$_{Lowell}$, M$_{Procida}$, M$_{Wurzburg}$, M3.Pro255Thr, S$_{Donosti}$, S$_{Munich}$ and S.Arg101Cys account for 8.1% of all alleles tested (220 out of 2718). Overall, these alleles affect 15.5% of individuals, which may have a genetic predisposition to develop COPD [22, 23]. Our results highlight that less frequent deficient alleles, beyond S and Z, affects 20 individuals (1.5%) in CanCOLD.

In this study, we identified 59 individuals with intermediate AATD. The novelty of our study is that we have included genotypes associated with intermediate deficiency beyond MZ including heterozygote non-Z (MF, MM$_{Procida}$, MM$_{Wurzburg}$, MP$_{Lowell}$, MS$_{Munich}$, MS.Arg101Cys, MM3.Pro255Thr), homozygote SS, and compound heterozygote IS. However, the COPD phenotypes of this group was not statistically different compared to individuals with no deficiency (MM), which was consistent across smoking status strata. Sample size may have limited our ability to detect the effect of intermediate deficiency on COPD phenotypes. Alternatively, the ascertainment method based on random sampling in CanCOLD may also explain this lack of association. For many COPD cohorts, ascertainment is based on patients referred specifically for obstructive airway disease. In the natural history of AATD, we know that index cases, those seen by medical care facilities, are more ill compared with nonindex cases (family members) [26]. In a population-based cohort like CanCOLD, fewer individuals have developed pulmonary symptoms and associated inflammatory milieu that would favour disease progression in the intermediate AAT deficiency state. Finally, lower smoking exposure may also have limited expressivity.

SERPINA1 genotyping grouping used in this study cannot be directly compared with previous studies. To make a fair comparison, we have performed MZ-only analysis and showed lower diffusing capacity in MZ heterozygotes. This result seems consistent with previous studies that have showed that MZ individuals are characterised by a lower FEV$_1$/FVC ratio and more radiographic emphysema [24], accelerated forced expiratory force at 25–75% of FVC (FEF25–75%) decline in the presence of smoking or obesity [25], lower lung function and CT-based emphysema [22], and lower lung function and greater COPD risk, especially in ever-smokers [23].

This study demonstrates that DNA sequencing is feasible in a large cohort. DNA sequencing provides a complete assessment of the gene causing AATD, namely *SERPINA1*. It can detect conventional protease inhibitor-deficient alleles (*e.g.* Z and S), but also rare and novel genetic variants. Diagnosis at the DNA level is the modern gold standard information needed for hereditary disorders. Importantly, DNA sequencing can be implemented in any laboratory with a standard DNA sequencing system, has a fast turnaround time, and may eventually replace the historical multistep testing algorithms to detect AATD used in most countries [9–11]. With established consistency in diagnosing AATD at the DNA level, there

is potential to expedite the clinical decision making process recommended by many national and international lung societies for targeted testing and augmentation therapy [1–4].

In this study, we have not performed cost-effectiveness analysis of *SERPINA1* DNA sequencing relative to current multistep-testing algorithms to detect AATD. For practical considerations, we provide costs for DNA sequencing. In Canadian currency, the cost of materials and supplies to sequence the 1359 CanCOLD participants is estimated at CAD 31 500 ((CAD 1.85 per PCR reaction+CAD 3.75 per sequencing reaction)×4 amplicons per sample×1359 samples). About CAD 10 per sample must be added for DNA extraction. An effective workforce and the number of samples processed at the same time are key to keep the cost low. It takes about a day and a half to complete the protocol. The salary of a laboratory technician for that time is thus the main factor fixing the cost per sample. In this research project, we were able to process samples in 96-well plate format and save costs. Lower throughput is expected in a realistic clinical healthcare system. For example, in our laboratory, the overall cost of *SERPINA1* DNA sequencing per sample is CAD 650, which drops to CAD 150 when eight samples are processed at the same time. This implies that all equipment are in place, including a sequencer and about CAD 20 000 of ancillary equipment (thermocycler, centrifuges, pipettes, water bath, vortex and electrophoresis system). DNA sequencers come in different throughput formats and prices, but one that can process four samples at the time is available at approximately CAD 50 000. Sequencing reactions can also be outsourced if a sequencer is not available. All these costs must be balanced with the resulting benefits. Here, we tested 1359 individuals and obtained a conclusive AATD diagnostic for all of them, *i.e.* success rate=100%. The single step test can be performed in 2 days (more realistically offered within a week) and thus has the potential to end the diagnostic odyssey of patients with AATD.

This study has many strengths, including being the first Canadian COPD cohort that has recruited its participants from the general population rather than convenient sampling often utilised in clinical research. Furthermore, the research participants have been extensively phenotyped for COPD (clinical, physiology, CT scan with qualitative and quantitative evaluation). The limitations include the relatively limited sample size. As with other cohort studies, participation bias may affect the representativeness of the cohort relative to the underlying population. The current study has been limited to cross-sectional data. Longitudinal data collection is ongoing, and future analyses are needed with respect to COPD development and progression. We attempted to classify all 34 genetic variants identified in CanCOLD participants as normal or deficient (table 1). This was based on all possible sources of information including previous literature, location of the variants in the protein structure, pathogenicity scores and AAT serum levels among carriers. However, some uncertainties remain about the classification of rare variants and further functional studies will be needed. Finally, the four-group classification scheme used in this study (no, mild, intermediate and severe deficiency) that reflects the combination of the two alleles observed in each individual is likely to be challenged and refined with time. However, grouping genotypes has always been inherent (consciously or not) in the field of AATD. For example, how many individuals would be MM in CanCOLD based on the conventional protease inhibitor system? With DNA sequencing, we demonstrated that the no deficiency group (individuals with two normal alleles) is in fact a collection of 49 different genotypes (see abscissa of figure 4). Groups were thus formed in this study using an unprecedented level of granularity. The strong association and progressive decline in AAT levels with *SERPINA1* genotyping groups (figure 5a) support our grouping scheme.

In conclusion, this study identified 34 genetic variants in the *SERPINA1* gene resulting in 220 deficient alleles from 1359 individuals. Four AATD cases were identified and the remaining deficient alleles were carried by 206 individuals, indicating that deficient alleles affected 15.5% of subjects in this Canadian population-based cohort. Patients with genotypes resulting in severe AATD (AATD cases) are more susceptible to develop airway obstruction as demonstrated by lower diffusion capacity and greater CT-based emphysema. In contrast, mild and intermediate deficiency genotypes were not associated with COPD phenotypes in this random sampling cohort except a lower diffusion capacity in MZ heterozygotes. Overall, this study demonstrates the frequency of deficient alleles beyond the most common S and Z alleles in a population-based cohort, and the feasibility of DNA sequencing on a large scale to provide an accurate and definitive diagnosis for AATD. The longitudinal evaluation will be important to determine the effects of *SERPINA1* genotypes on COPD risk and progression.

References

1   American Thoracic Society/European Respiratory Society statement: standards for the diagnosis and management of individuals with alpha-1 antitrypsin deficiency. *Am J Respir Crit Care Med* 2003; 168: 818–900.

2   Marciniuk DD, Hernandez P, Balter M, *et al.* Alpha-1 antitrypsin deficiency targeted testing and augmentation therapy: a Canadian Thoracic Society clinical practice guideline. *Can Respir J* 2012; 19: 109–116.

3   Miravitlles M, Dirksen A, Ferrarotti I, *et al.* European Respiratory Society statement: diagnosis and treatment of pulmonary disease in alpha1-antitrypsin deficiency. *Eur Respir J* 2017; 50: 1700610.

4   Alpha 1-antitrypsin deficiency: memorandum from a WHO meeting. *Bull World Health Organ* 1997; 75: 397–415.

5   de Serres FJ. Worldwide racial and ethnic distribution of alpha1-antitrypsin deficiency: summary of an analysis of published genetic epidemiologic surveys. *Chest* 2002; 122: 1818–1829.

6   Luisetti M, Seersholm N. Alpha1-antitrypsin deficiency. 1: epidemiology of alpha1-antitrypsin deficiency. *Thorax* 2004; 59: 164–169.

7   Lieberman J, Winter B, Sastre A. Alpha 1-antitrypsin Pi-types in 965 COPD patients. *Chest* 1986; 89: 370–373.

8   Abboud RT, Ford GT, Chapman KR. Alpha1-antitrypsin deficiency: a position statement of the Canadian Thoracic Society. *Can Respir J* 2001; 8: 81–88.

9   Snyder MR, Katzmann JA, Butz ML, *et al.* Diagnosis of alpha-1-antitrypsin deficiency: An algorithm of quantification, genotyping, and phenotyping. *Clin Chem* 2006; 52: 2236–2242.

10   Ferrarotti I, Scabini R, Campo I, *et al.* Laboratory diagnosis of alpha1-antitrypsin deficiency. *Transl Res* 2007; 150: 267–274.

11   Graham RP, Dina MA, Howe SC, *et al.* SERPINA1 full-gene sequencing identifies rare mutations not detected in targeted mutation analysis. *J Mol Diagn* 2015; 17: 689–694.

12   Maltais F, Gaudreault N, Racine C, *et al.* Clinical experience with SERPINA1 DNA sequencing to detect alpha-1 antitrypsin deficiency. *Ann Am Thorac Soc* 2018; 15: 266–268.

13   Silva D, Oliveira MJ, Guimaraes M, *et al.* Alpha-1-antitrypsin (SERPINA1) mutation spectrum: three novel variants and haplotype characterization of rare deficiency alleles identified in Portugal. *Respir Med* 2016; 116: 8–18.

14   Matamala N, Lara B, Gomez-Mariano G, *et al.* Characterization of novel missense variants of SERPINA1 gene causing alpha-1 antitrypsin deficiency. *Am J Respir Cell Mol Biol* 2018; 58: 706–716.

15   Ortega VE, Li X, O'Neal WK, *et al.* The effects of rare SERPINA1 variants on lung function and emphysema in SPIROMICS. *Am J Respir Crit Care Med* 2020; 201: 540–554.

16   Bourbeau J, Tan WC, Benedetti A, *et al.* Canadian Cohort Obstructive Lung Disease (CanCOLD): fulfilling the need for longitudinal observational studies in COPD. *COPD* 2014; 11: 125–132.

17   Adzhubei IA, Schmidt S, Peshkin L, *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* 2010; 7: 248–249.

18   Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009; 4: 1073–1081.

19   Kircher M, Witten DM, Jain P, *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014; 46: 310–315.

20   Landrum MJ, Lee JM, Benson M, *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 2018; 46: D1062–D1067.

21   Purcell S, Neale B, Todd-Brown K, *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; 81: 559–575.

22   Foreman MG, Wilson C, DeMeo DL, *et al.* Alpha-1 antitrypsin PiMZ genotype is associated with chronic obstructive pulmonary disease in two racial groups. *Ann Am Thorac Soc* 2017; 14: 1280–1287.

23   Molloy K, Hersh CP, Morris VB, *et al.* Clarification of the risk of chronic obstructive pulmonary disease in alpha1-antitrypsin deficiency PiMZ heterozygotes. *Am J Respir Crit Care Med* 2014; 189: 419–427.

24   Sorheim IC, Bakke P, Gulsvik A, *et al.* Alpha(1)-antitrypsin protease inhibitor MZ heterozygosity is associated with airflow obstruction in two large cohorts. *Chest* 2010; 138: 1125–1132.

25   Thun GA, Ferrarotti I, Imboden M, *et al.* SERPINA1 PiZ and PiS heterozygotes and lung function decline in the SAPALDIA cohort. *PLoS One* 2012; 7: e42728.

26   Stoller JK. The National Institutes of Health/National Heart, Lung, and Blood Institute Registry for patients with severe deficiency of alpha 1-antitrypsin. *In*: Crystal RG, ed. Alpha 1-antitrypsin deficiency: biology, pathogenesis, clinical manifestations, therapy. New York, Taylor & Francis Group, 1996.

27   den Dunnen JT, Antonarakis SE. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum Mutat* 2000; 15: 7–12.

28   Ringenbach MR, Banta E, Snyder MR, *et al.* A challenging diagnosis of alpha-1-antitrypsin deficiency: identification of a patient with a novel F/Null phenotype. *Allergy Asthma Clin Immunol* 2011; 7: 18.