On the privacy risks of sharing clinical proteomics data

Sujun Li, PhD[1], Nuno Bandeira, PhD[2], Xiaofeng Wang, PhD[1], and Haixu Tang, PhD[1]

[1]School of Informatics and Computing, Indiana University, Bloomington, IN, USA
[2]Department of Computer Science and Engineering, University of California, San Diego, CA, USA

Abstract

Although the privacy issues in human genomic studies are well known, the privacy risks in clinical proteomic data have not been thoroughly studied. As a proof of concept, we reported a comprehensive analysis of the privacy risks in clinical proteomic data. It showed that a small number of peptides carrying the minor alleles (referred to as the minor allelic peptides) at non-synonymous single nucleotide polymorphism (nsSNP) sites can be identified in typical clinical proteomic datasets acquired from the blood/serum samples of individual patient, from which the patient can be identified with high confidence. Our results suggested the presence of significant privacy risks in raw clinical proteomic data. However, these risks can be mitigated by a straightforward pre-processing step of the raw data that removing a very small fraction (0.1%, 7.14 out of 7,504 spectra on average) of MS/MS spectra identified as the minor allelic peptides, which has little or no impact on the subsequent analysis (and re-use) of these datasets.

Introduction

Human genome is considered as an ultimate digital identifier of individuals as many of their phenotypes can be eventually predicted from their genomes by using intelligent computer algorithms. For example, methods are being developed to infer biometric and appearance traits (e.g. the eye and hair colors) of unknown donors from theirs biological materials, a technique called forensic DNA phenotyping [1]. A recent study showed that the surnames of human males can be inferred by profiling Y chromosome short tandem repeats (Y-STRs) from their personal genomes, and, combining them with other publicly available metadata, the identities of these individuals can be revealed [2]. It is known that any human individual can be uniquely identified at high confidence with as few as dozens of single nucleotide polymorphisms (SNPs) in his/her genome[3]. Even aggregate genomic data, such as the allele frequencies at many SNP sites among a group of individuals that are commonly in genome-wide association studies (GWAS) for a case group of a genetic disease, are found to have privacy risks[4]. Homer et. al. proposed a statistical method to infer the presence of an individual in a group of participants from the aggregate genetic data (e.g., the allele frequencies) based on a second sample from the individual [5]. The method was further optimized by follow-up studies [6, 7, 8, 9, 10]. Although it has been shown that the privacy risks within GWAS datasets were often limited [6] due to this inference method, these discoveries received immediate responses from funding agencies and the broad research community. On the one hand, sensitive genomic data are treated as identifiable information and cannot be accessed without signing a user agreement. Up to now, most aggregate DNA data were removed from public databases to protect the privacy of the participants of these studies[11]. On the other hand, novel approaches of informed consent were suggested, such as the open consent approach used by the Personal Genome Project (PGP), in which all participants must "be prepared with sufficient knowledge of genome science", while agree to give up the privacy in their genomic data [12].

Even though privacy issues have been studied intensively in human genomics, surprisingly little research has been done for the other types of high throughput omics techniques, such as mass spectrometry-based proteomics. In fact, mass spectrometry (MS) analysis of human blood/serum proteome has been applied to clinical and biomedical research even before genomic techniques, for discovery of prognostic biomarkers[13], monitoring of disease progression or studies of pharmaceutical effects[14, 15]. In practice, these data were not considered as identifiable when they are shared for a secondary analysis, as long as the associated identifiable information (e.g., patient name or address) is removed. In fact, many of these raw clinical proteomics data were publicly shared through the Proteome Commons[16] service before it was closed recently. The data sharing efforts are invaluable for the secondary analysis and meta-analysis of proteomic data as well as for the development of new algorithms in proteomics, and thus the data sharing effort is continuous with the newly launch of several proteomics data repositories, such as MassIVE Datasets (`http://massive.ucsd.edu/ProteoSAFe/static/massive.jsp`) and ProteomeXchange[17], PRIDE[18], PeptideAtlas[19] and online data journals, such as Scientific Data (`http://www.nature.com/sdata/`). However, it is important to understand the potential privacy risks on sharing raw clinical proteomic data, and to come up simple yet effective approaches to mitigate the risks, if they are indeed present. In particular, the recent advance in mass spectrometry technology has enabled the identification of proteins in a broader dynamic range from a complex proteome sample, and thus subtle identifiable information may be inferred from raw clinical proteomic data.

The most powerful distinct features to distinguish two individuals in their proteomes are their non-synonymous single nucleotide polymorphisms (nsSNPs), i.e., the missense type SNPs causing amino acid changes. And in fact, only a small fraction of SNPs in human genome are non-synonymous. Our statistics on dbSNP database[20, 21] shows that only 1,343,892 (1.2%) are missense class in 111,666,093 known human SNPs. Among the missense type SNPs, 577,747 (43%) records (rsIDs) are reported with minor allele frequency (MAF). The minor allele frequencies of these SNPs are often small, ranging from 4.6^{-4} to 0.5. Nevertheless, because only a small number of SNPs are sufficient to identify a human individual, a single clinical proteomic dataset may be used to infer a small number of minor alleles at nsSNP sites that are sufficient to identify the participant from whom the data is collected.

In this paper, we provided a comprehensive analysis of the identifiability of clinical proteomic data, i.e., the confidence of identifying a human individual based on nsSNPs inferred from mass spectrometry data collected from the same individual. In each of 158 proteomic datasets collected from the serum proteome of 80 breast cancer patients (as cases in the study) and 78 healthy individuals (as controls in the study) by using LTQ-MS (Thermo-Finnigan linear ion-trap mass spectrometer) [22], as well as an additional set of blood/serum samples from 30 individuals, we identified up to 20 minor allelic sites, corresponding up to 25 spectra, at nsSNP sites; according to the minor allele frequencies, we can identify the participants at high confidence (defined as p-value below 10^{-10}), with the lowest as 8.66×10^{-67}. Our results showed that significant privacy risks may be present in raw clinical proteomic data. However, these risks can be mitigated by simple pre-processing of the raw data that removes the MS/MS spectra resulted from the minor allelic peptides before they are shared in public domain. Notably, there are only 0.10% (On average, 7,504 MS/MS spectra are present in each dataset, and 7.14 of them can be identified as minor allelic peptides in our analysis) MS/MS spectra to be removed, indicating the preprocessing has little impact on any secondary analysis on the datasets. We noted that the raw data before removal of the spectra should be under controlled access similar to the clinical genomic data managed by dbGAP (http://www.ncbi.nlm.nih.gov/gap).

Methods

Clinical proteomic datasets

Dataset I: Proteomics dataset I was acquired from a clinical proteomic study consisting of 160 human participants, including 80 samples from women with breast cancer (i.e, the cases) and 80 from healthy volunteer individuals (i.e., the controls)[22]. The cases and controls in the study follow comparable demographic distribution. A total of 158 datasets were downloaded from Tranche (ProteomeCommons, https://proteomecommons.org)[16] initially, 2 of the case datasets were missing at Tranche due to the network instability. As indicated in [22], the proteomic datasets were acquired following a common bottom-up proteomics strategy, in which tryptic peptides were analyzed using Thermo-Finnigan linear ion-trap mass spectrometer (LTQ) coupled with a HPLC system. Peptides were eluted with a gradient from 5% to 45% acetonitrile developed over 120 minutes and LC-MS/MS data were collected. The acquired raw peak list data was generated by using XCalibur (version 2.0) with default parameters. Each of these datasets consists of between 6,734 and 7,797 (on average 7,301, standard deviation 212) MS/MS spectra.

Dataset II: Another clinical plasma/serum dataset was acquired from the study [23] on the complex serum samples from 15 breast cancer patients and 15 healthy control serum samples. Each of these datasets consists of between 6,179 and 10,765 (on average 8,575, standard deviation 847) MS/MS spectra.

Combining Dataset I and Dataset II mentioned above, Our study analysed proteomic data from 188 human individuals

Assembly of nsSNPs dataset in human genome

dbSNP build 142 were downloaded from http://www.ncbi.nlm.nih.gov/projects/SNP/. Specifically, human chromosome data was obtained individually from ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/ASN1_flat/. The variations mapped to multiple chromosomes, or did not map to any human chromosomes, or variations on unplaced chromosomes are excluded from the data. For each RS record in dbSNP, we obtained the global minor allele frequency (MAF) from the database. Per dbSNP[21], the minor allele frequency for each RS included is reported based on a default global population. As in the dbSNP build 142, the current default global population in the database is 5008.

The missense SNP type is then exclusively included in this specific study. To provide a comprehensive database, the missense SNP records are all kept regardless of with or without frequency records in the database. We then extracted the protein ID from the missense annotation, recording the amino acid and its location and mapping to RefSeq protein database[24]. In summary, there are 1,343,892 missense SNPs, mapping to 5,107,618 amino acid variations contained in proteins database, corresponding to 69,737 proteins. Among the ≈ 5 millions mapped amino acid variations, 317,279 variations are recorded at the same positions on the proteins with different amino acids. Moreover, 2,964,582 records have recorded as NA frequency, while 2,143,036 variations (577,747 unique variations) have annotated frequencies.

Construction of nsSNP peptide database for proteomics database searches

The RefSeq protein FASTA file was downloaded from NCBI's FTP site (`ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/`, release 69). Since we obtained $\approx$ 5 millions amino acids changes on 69,737 (total 72,297) RefSeq protein sequences, there is on average 70 amino acid changes per protein. We observed that the average protein length in RefSeq database is 640. It is then expected that on average there will be one amino acid variation every 9 residues on proteins, thus it is possible that multiple amino acid changes occur in one tryptic peptide. Because the identification of multiple amino acids variation is out of our scope in this work, we used a simple approach here. We encoded the SNP peptides as independent records by adding the flanking regions up to 40 amino acid residues (left, right) around the SNP sites[25, 26]. We keep all the variation information in the fasta header, including the rsID from dbSNP, the frequency of the variations and the amino acid changes. For example, the header information ”NP_001193959.1-dbSNP-rs537379392-0.0002-43SR” referred to the rsID ”rs537379392” in dbSNP is on protein ”NP_001193959.1” located in position 43 with minor allelic frequency 0.0002 with the amino acid variation of Serine(S) to Arginine (R). All of the information are provided in fasta file in the supplementary file.

Peptide identification

The total 188 datasets (including both proteomics dataset I and II) were searched against the human RefSEQ protein database[24] in the first round by using MSGF+[27]. Prior to the second round searching, the SNP peptides in the proteins identified in the first searches are constructed from the procedure above. Then the same searching parameters are applied in both searching steps, which contained a combined forward and decoy entries for the target-decoy searching approach[28, 29] to estimate false discovery rate, while the second search employed the additional reversed peptide sequences centered with the minor allelic amino acids. The searching used the following search criteria based on the experimental conditions: (1) tryptic enzyme specificity, (2) with at most one missed cleavage; (3) carbamidomethylation at cysteine residues as a fixed modification; and oxidation at methionine residues as variable modifications; (4) at most two modification on a peptide are allowed; (5) 1.5Da for precursor ion tolerance; (6) maximum precursor charge to consider is 5. Peptide mass calculation was performed using monoisotopic values. The false discovery rate is estimated by using target-decoy searching approach.

Computation of identifiability

Assuming we identified n peptides carrying a minor allele at n nsSNP sites, denoted as $m_1, m_2, ..., m_n$ in a clinical proteomic dataset, we compute the identifiability of the dataset as

$$P = \prod_{i=1}^{n} [f_{m_i}(1 - q_{m_i})] \tag{1}$$

where f_{m_i} is the minor allele frequency at the m_i-th nsSNP, q_{m_i} is the global false discovery rate (FDR) for the peptide containing the nsSNP sites, and P is denoted as the identifiability.

To be noted here, we assume there is no correlation (due to linkage disequilibrium) between the identified minor allele sites in a single sample, and thus the minor allele frequencies are considered to be independent. Because the minor allelic sites identified in the same sample are always from different peptides/proteins encoded at distant genomic loci, this assumption is likely valid.

Results

Identification of peptides containing the nsSNP sites

MS database searches are performed first by using MSGF+[27] (available at `http://proteomics.ucsd.edu/Software/MSGFPlus.html`) against RefSeq Human database[24] (Release 69) composed of 72,297 human protein sequences. The database searching identified 2,100 proteins by using a 1% global false discovery rate cutoff at the spectrum level, among which 2,054 proteins contain one or more nsSNP sites where the major allelic amino acids are present (and are thus referred to as the major allelic proteins).

We then replace the major allelic amino acids in these 2,054 proteins at nsSNP site with the corresponding minor allelic amino acids as indicated in dbSNP, forming the set of minor allelic peptides. In short, A total of 304,820 amino acids were replaced (a full list of these sites can be found in Supplementary Table 1). The flanking regions with up to 80 amino acids are retrieved centered by the minor allelic amino acids and the resulting minor allelic peptides are then appended to the major allelic proteins.

We note that we did not search the MS/MS proteomics data against the complete proteome database composed of proteins with ≈ 5 millions different combinations of major and minor allelic amino acids across all SNP sites to avoid the unnecessary redundancy in the database searching. Instead, we conducted a secondary database searching strategy against a database containing a total of detected 2,100 protein sequences identified in the first searches and their minor allelic peptides by using MSGF+. In order to achieve high confidence on the identified peptides, a threshold corresponding to the estimated FDR 1% is employed to filter the two-round peptide identification results.

On average, 22,512 spectra representing 2,022 unique peptides were identified in the entire 188 proteomic datasets, among which 1,166 peptides have no minor allelic frequencies and 856 peptides have minor allelic frequencies, corresponding to 8,722 spectra in all of the 188 samples. Supplementary table 2 shows the number of spectra and unique minor allelic peptides identified in each clinical proteomic dataset analyzed here.

We also observed that some minor allelic peptides are recurrently identified in a large fraction of clinical proteomics samples. It may be due to the contamination, or to a less degrees of false positive identification or some to relatively high minor allelic frequencies of such sites. In order to avoid the bias in identifiability estimation, we eliminated the peptides that are identified in more than 10 samples. After that, a total of 731 minor allelic peptides and 663 unique minor allelic sites (dbSNP rsID) in 342 proteins were retained, which corresponded to on average 7 unique minor allelic sites (ranging from 1 to 20 minor allelic sites) per sample.

Identifiability in individual samples

The identifiability is defined as the probability of observing a human individual sample with the same set of minor alleles by chance. Therefore, if the identifiability multiplying the current world population is much lower than 1, it implies there is little chance that there is another person in the world with the same set of minor alleles who is not a close relative of the individual.

Figure 1a shows the number of minor alleles sites identified in each proteomic sample. A total of 1-20 minor alleles sites (On average 7 unique minor allelic nsSNPs sites per sample) were identified in each sample as shown in supplementary table 3. Based on this assumption that no correlation (due to linkage disequilibrium) between the minor allelic sites is considered, we computed the identifiability (as defined in the Methods section) of each clinical proteomic sample. Figure 1b shows the distribution of identifiability in 188 individual samples (see supplementary table 3 for the identifiability of each sample) across two datasets. In summary, 27 samples received the identifiability higher than 7.3E-09 (i.e. the multiplication inverse of the world population), while 161 (86%) samples have the identifiability lower than 7.3E-09, indicating that the individuals can be highly reliably identified from most of these samples (considering the world population size is only about 7.3E+09 (i.e., the red line)). Breaking down to each dataset, in Dataset I, 25 out of 158 individuals have lower identifiability and in Dataset II, 2 out of 30 individuals have lower identifiability, showing that the two different datasets have consistent results.

We also tested the identifiability of major allelic peptides, the frequencies of which are defined as (1-MAF). The MAF is extracted from the potential minor allelic sites contained in the major allelic peptides. We retrieved the unique major allelic peptides identified in each dataset without their corresponding minor allelic peptides identified. On average, 27.15 unique major allelic sites (from 55.17 spectra) were identified in each sample. Although this seems to be a large number of nsSNPs containing peptides comparing to 7 unique minor allelic sites, their corresponding major allele frequencies are high (e.g, On average the major allelic frequency of major allelic sites identified in each dataset is 0.992). As a result, the identifiability of these major allelic peptides is low: the lowest identifiability among all 188 datasets is 0.395, while the average identifiability is 0.69. This suggests that in order to mitigate the privacy risks in clinical proteomic datasets, we do not need to remove MS/MS spectra resulted from major allelic peptides, or at most remove only a small number of those spectra corresponding to major allelic peptides with relatively low allelic frequencies.

Discussion and Conclusion

In this study, we intentionally selected human blood/serum samples to investigate the privacy risks in clinical proteomic studies, because in clinical settings, blood/serum samples are commonly used for prognostic biomarker discovery, and thus the privacy risks in these data are of great practical interests.

We showed significant privacy risks are present in raw clinical proteomic data. These risks, however, can be mitigated by a simple pre-processing of the raw data that removes the MS/MS spectra resulted from the peptides carrying a minor allele at an nsSNP site. We recommended this preprocessing step should be carried out before sharing clinical proteomic datasets in public domain.

Here, we took a conservative approach to peptide identification: on one hand, we consider privacy protection as a high priority and thus we intend to remove an MS/MS spectrum from a proteomic dataset even if it is identified as a minor allelic peptide with relatively low confidence; on the other hand, even with the conservative criteria, only a small fraction of MS/MS spectra need to be removed in real-world clinical proteomic data. We note some minor allelic peptides are identified in multiple proteomic datasets, which may be partially due to false positive identifications. Nevertheless, as showed in the supplementary table, 85.5% of the identified minor allelic peptides are identified in less than 10 (out of 188; 5.3%) samples, which is consistent with the allele frequencies. Moreover, when calculating the identifiability, the recurrently identified peptides are excluded with the consideration of avoiding the false positive identification.

We discussed only one identification strategy applied to clinical proteomic datasets, i.e., based on the identified peptide carrying the minor alleles of nsSNPs. Although this is probably the most powerful piece of information for identification, other information exists for the inference of the identity of an individual. For examples, it is known that missense mutations or gene fusions may lead to the expression of complete novel proteins, and the identification of peptides from these proteins may also be used to identify an individual. However, these events are relatively rare [30], and thus, the identification power is expected to be low. On the other hand, we consider the datasets acquired by using relatively old technology (LTQ-MS), which has a limited size (7,000 MS/MS spectra per sample). The new LC-MS instruments (e.g., the OrbiTrap-MS) typically capture over more than 10,000 spectra in a single analysis. Additional separation methods that can be used to further increase the throughput (e.g., multidimensional protein identification technology (MudPIT) [31]) of clinical proteomics. If we assume that a similar rate of minor allelic peptides can be identified from MS/MS spectra, we anticipate at least a few times more minor alleles can be identified in current clinical proteomic datasets, resulting in much stronger identification power than that shown here. Therefore, the identification power presented in this study should be viewed as the lower bound for the estimate of actual privacy risks in unprocessed clinical proteomics data.

Furthermore, the presence/absence of peptides identified in a proteomic experiment is not random: some peptides are highly likely being detected in an experiment if they are present in the sample. The highly detectable peptides[32] can be used to infer the presence of the minor allele at an nsSNP site. Consider a tryptic peptide carrying the major allele of a putative nsSNP that has high detectability. If this peptide is not identified in a specific proteomic experiment but less detectable peptides from the same protein are identified, then one can reasonably infer that it is likely that the individual carries the minor allele at this nsSNP site.

We note that this inference attack cannot be mitigated by simply removing MS/MS spectra resulted from the peptide carrying the minor allele at the same site, because the peptide carrying the major allele is still absent . Even though in this case, we are not fully certain that the individual carries the minor allele at a SNP site, one can design a statistical method to re-identify an individual from a second sample, similar as used for the re-identification from aggregate genomic data[5]. However, as the number of minor allelic peptides identified in a proteomic dataset is relatively small as shown here, we anticipate the risk for such inference attack is low.

Acknowledgement

This work is partially supported by NHGRI/NIH (Grant no: R01 HG007078-01) and NIGMS/NIH (Grant no: R01 GM103725-04).

Footnotes

N.B., X.W., and H.T. designed research; S.L. and H.T. performed research; S.L. analyzed data; S.L. and H.T. wrote the paper; N.B. and X.W. read and revised the paper.

References

[1] Manfred Kayser. “Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes”. In: *Forensic Science International: Genetics* (2015).

[2] Melissa Gymrek et al. “Identifying personal genomes by surname inference”. In: *Science* 339.6117 (2013), pp. 321–324.

[3] Zhen Lin, Art B Owen, and Russ B Altman. “Genomic research and human subject privacy”. In: *Science* (2004), pp. 183–183.

[4] Stephen E Fienberg, Aleksandra Slavkovic, and Caroline Uhler. "Privacy preserving GWAS data sharing". In: (2011), pp. 628–635.

[5] Nils Homer et al. "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays". In: *PLoS genetics* 4.8 (2008), e1000167.

[6] Sriram Sankararaman et al. "Genomic privacy and limits of individual detection in a pool". In: *Nature genetics* 41.9 (2009), pp. 965–967.

[7] Kevin B Jacobs et al. "A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies". In: *Nature genetics* 41.11 (2009), pp. 1253–1257.

[8] Rosemary Braun et al. "Needles in the haystack: identifying individuals present in pooled genomic data". In: *PLoS genetics* 5.10 (2009), e1000668.

[9] Rui Wang et al. "Learning your identity and disease from research papers: information leaks in genome wide association study". In: (2009), pp. 534–544.

[10] David Clayton. "On inferring presence of an individual in a mixture: a Bayesian approach". In: *Biostatistics* (2010), kxq035.

[11] National Institute of Health (US). "Policy for sharing of data obtained in NIH supported or conducted genome-wide association studies (GWAS)". In: (2007).

[12] Misha Angrist. "Eyes wide open: the personal genome project, citizen science and veracity in informed consent". In: *Personalized medicine* 6.6 (2009), pp. 691–699.

[13] Volker Seibert, Matthias PA Ebert, and Thomas Buschmann. "Advances in clinical cancer proteomics: SELDI-ToF-mass spectrometry and biomarker discovery". In: *Briefings in functional genomics & proteomics* 4.1 (2005), pp. 16–26.

[14] Chen Li et al. "Accurate qualitative and quantitative proteomic analysis of clinical hepatocellular carcinoma using laser capture microdissection coupled with isotope-coded affinity tag and two-dimensional liquid chromatography mass spectrometry". In: *Molecular & Cellular Proteomics* 3.4 (2004), pp. 399–409.

[15] Tsung-Heng Tsai et al. "LC-MS/MS based Serum Proteomics for Identification of Candidate Biomarkers for Hepatocellular Carcinoma". In: *Proteomics* (2015).

[16] James A Hill et al. "ProteomeCommons. org collaborative annotation and project management resource integrated with the Tranche repository". In: *Journal of proteome research* 9.6 (2010), pp. 2809–2811.

[17] Juan A Vizcaíno et al. "ProteomeXchange provides globally coordinated proteomics data submission and dissemination". In: *Nature biotechnology* 32.3 (2014), pp. 223–226.

[18] Juan Antonio Vizcaíno et al. "The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013". In: *Nucleic acids research* 41.D1 (2013), pp. D1063–D1069.

[19] Eric W Deutsch, Henry Lam, and Ruedi Aebersold. "PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows". In: *EMBO reports* 9.5 (2008), pp. 429–434.

[20] Stephen T Sherry et al. "dbSNP: the NCBI database of genetic variation". In: *Nucleic acids research* 29.1 (2001), pp. 308–311.

[21] Elizabeth M Smigielski et al. "dbSNP: a database of single nucleotide polymorphisms". In: *Nucleic Acids Research* 28.1 (2000), pp. 352–355.

[22] Fan Zhang and Jake Y Chen. "Discovery of pathway biomarkers from coupled proteomics and systems biology methods". In: *BMC genomics* 11.Suppl 2 (2010), S12.

[23] Anoop Mayampurath et al. "Computational framework for identification of intact glycopeptides in complex samples". In: *Analytical chemistry* 86.1 (2013), pp. 453–463.

[24] Kim D Pruitt et al. "RefSeq: an update on mammalian reference sequences". In: *Nucleic acids research* 42.D1 (2014), pp. D756–D763.

[25] Jing Li et al. "A bioinformatics workflow for variant peptide detection in shotgun proteomics". In: *Molecular & Cellular Proteomics* 10.5 (2011), pp. M110–006536.

[26] Gloria M Sheynkman et al. "Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences". In: *Journal of proteome research* 13.1 (2013), pp. 228–240.

[27] Sangtae Kim and Pavel A Pevzner. "MS-GF+ makes progress towards a universal database search tool for proteomics". In: *Nature communications* 5 (2014).

[28] Joshua E Elias and Steven P Gygi. "Target-decoy search strategy for mass spectrometry-based proteomics." In: *Methods in Molecular Biology (Clifton, NJ)* 604 (2010), pp. 55–71.

[29] Kyowon Jeong, Sangtae Kim, and Nuno Bandeira. "False discovery rates in spectral identification". In: *BMC Bioinformatics* 13.Suppl 16 (2012), S2.

[30] 1000 Genomes Project Consortium et al. "A map of human genome variation from population-scale sequencing". In: *Nature* 467.7319 (2010), pp. 1061–1073.

[31] Michael P Washburn, Dirk Wolters, and John R Yates. "Large-scale analysis of the yeast proteome by multidimensional protein identification technology". In: *Nature biotechnology* 19.3 (2001), pp. 242–247.

[32] Haixu Tang et al. "A computational approach toward label-free protein quantification using predicted peptide detectability". In: *Bioinformatics* 22.14 (2006), e481–e488.

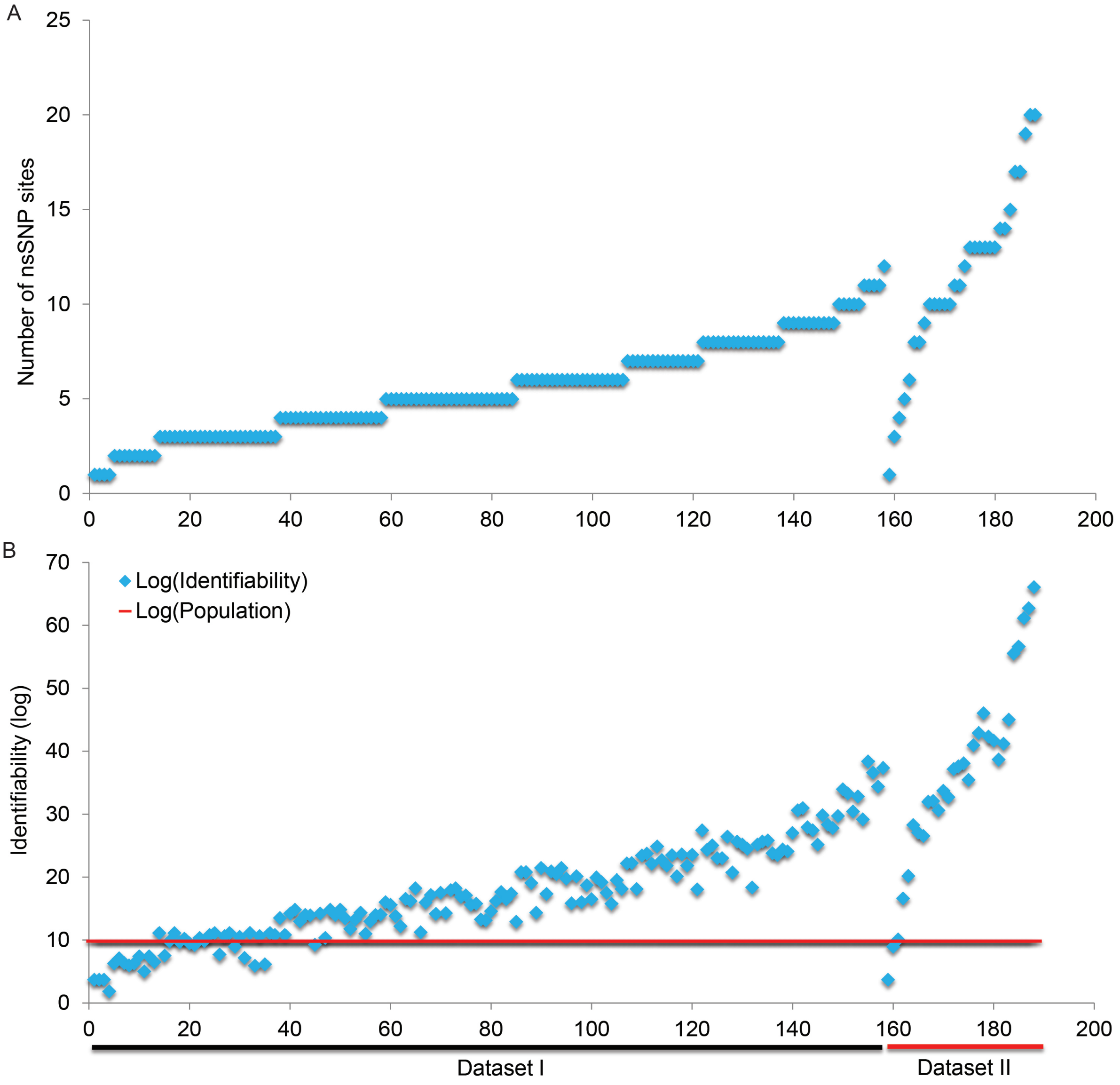

Figure 1: A. Distribution of the number of minor alleles identified in all 188 clinical proteomic samples (x-axis, sorted according to increasing number of identified minor allelic sites in proteomic dataset I and II, respectively). On average, 7 unique minor allelic sites are identified in each sample. B. Distribution of identifiability of all clinical proteomic datasets based on minor allelic sites identification. The red line represents the cutoff at the current world population size (7.3×10^9).