Genomic Data and Disease Forecasting: Application to Type 2 Diabetes (T2D)

Lawrence Sirovich*

Center for Studies in Physics & Biology, Rockefeller University, New York, New York, United States of America

Abstract

A general approach is presented for the extraction of a classifier of disease risk that is latent in large scale disease/control databases. Novel features are the following: (1) a data reorganization into a regularized *standard* form that emphasizes individual alleles instead of the single nucleotide polymorphism (Snp) allele pair to which they belong; (2) from this a procedure that significantly enhances the discovery of high value genomic loci; (3) an investigative analysis based on the hypothesis that disease represents a very small signal (small signal-to-noise) that is latent in the data. The resulting analyses applied to the FUSION T2D database leads to the polling of thousands of genomic loci to classify disease. This large genomic kernel of loci is shared by non-diabetics at nearly the same high level; but a small well defined separation exists and it is speculated that this might be due to unconventional disease mechanisms. Another analysis demonstrates that the FUSION database size limits its disease predictability, and only one third of the resulting classifier loci are estimated to relate to T2D. The remainder is associated with hidden features that might contrast the disease and control populations and that more data would eliminate.

Citation: Sirovich L (2014) Genomic Data and Disease Forecasting: Application to Type 2 Diabetes (T2D). PLoS ONE 9(1): e85684. doi:10.1371/journal.pone.0085684

Editor: Gustavo Stolovitzky, Thomas J. Watson Research Center, United States of America

Received June 28, 2013; **Accepted** December 6, 2013; **Published** January 17, 2014

Funding: There were no external funding sources for this study. The high quality Fusion database was acquired from NIH at the request of the Rockefeller University Committee for Clinical and Translational Science, grant # UL1RR024143, National Center for Research Resources, National Institutes of Health.

Competing Interests: The author has declared that no competing interests exist.

* E-mail: lsirovich@rockefeller.edu

Introduction

In rough approximation about 99.5% of the 3 billion DNA base pairs of the human genome are shared by all homosapiens. Somatic cells contain two copies, paternal and maternal, and so about 99.5% of the pairs are homozygous. The remaining pairs appear as two alleles and are termed single nucleotide polymorphisms (Snps): based on the criterion that the rarer of the two alleles is present in greater than 1% of the *population* [1]. This heterozygous content of the Snps is generally regarded as the determinant of human diversity including the potential for acquiring diseases [2].

Genome wide association studies (GWAS) refer to extensive investigations, past and present, that endeavor to find correlations between clinically diagnosed diseases (phenotype) with their Snp counterparts (genotype), and have as a goal the determination of genomic classifiers of disease risk. Location of the two Snps associated with age related macular degeneration [3] represents a triumph of this approach. Although other disease associations have been found [4] disappointment has been expressed on the paucity of DNA linkages that have been found for human diseases, particularly in the case of complex disorders [5]. The "predictive power of DNA" has also been questioned [6] (also see [7]). Additionally doubt has been raised in regard to the relationship of identified Snps with actual disease mechanisms [8].

The methods presented here for associating Snp loci with disease lie outside traditional statistical approaches [9–11]. Instead the present framework originates in methods that have their origin in extracting extremely weak signals (small "signal-to-noise" ratio) such as appear in optical imaging [12–14]. The view of the present study is that GWAS data show a weak disease signal. In a genomic setting similar methods have been applied to taxonomic studies [15,16]. For *complex* diseases such as diabetes it may be an error to focus on the role of individual genomic loci. Instead the disease/control contrast appears to be better sought in a large genomic framework of risk loci common to the disease cohort.

A principal goal of this effort will be to isolate out of the vast collection of genomic loci, potentially in the millions, a smaller set of loci along with an appropriate nucleotide symbol at each locus which is associated with the disease. If the number of loci (and symbols) is N then the $2\times N$ matrix of loci over genomic symbols will be referred to as the *indicator (vector)*, and the properly ordered symbols which form a word in the general sense will be termed the disease *classifier*, in the N loci subspace.

The next section presents a verbal description of the Methods (technical summaries appear in the Appendix) and also provides a graphical illustration of the superior informational content of the present data reorganization; and a second illustration convincingly shows the structural difference of case/control sets in the classifier subspace. After a section describing the FUSION data, and more preliminary results, a section entitled Validation and Prediction follows. This deals with the self-consistency of the classifier that was determined in the previous section, but also points out limitations on the predictability of the classifier when applied to other data. The final Discussion section presents additional results, implications and speculations.

Methods

The Standard Organization

A typical database is composed of disease and control genomic records. Each such record is a sequence of symbols at Snp locations common to the database. Information on the chromosome number, chromosome location, and Snp alleles is included in a typical database, as encoded in the rs (ref Snp) number, which can contain as many as eight digits (NCBI Resources, 2013). The two alleles of a Snp are chosen from the nucleotide symbols [A,C,G,T] or equivalently as aliased by [1,2,3,4]. Generally there is no particular order in the acquisition of allele pairs. Without loss of generality we adopt a convention that places the higher number first in each Snp. Since the allele content of a Snp is known from its rs number the two symbols therefore can be further aliased by 2 and 1, with the higher number going first, see Table 1 in Appendix.

On this basis any representative sequence of a population appears as a sequence of Snps each of which contains an odd numbered and then an even numbered locus. Each allele pair then appears as: 22, 21, or 11 which gives the essential content of the Snp of a sequence. Henceforth this description, which is general and free of bias, will be referred to as the standard organization. The standard organization divides a SNP into the two allele compartments that will be referred to as odd, o, and even, e. The informational content of the two compartments always exceeds the Snps form, see Appendix.

High Value Loci

Computing challenges and rational considerations dictate an à priori search for those loci that are likely to be associated with a specific disease. These will be referred to as *high value* loci. Customary treatment of GWAS data dwells on the Snp disease/control odds ratio, Ω, and relatively large values suggest a locus of interest. Adoption of the standard organization now permits calculation of the odds ratio for each allele, ω. Figure 1 displays Snp and allele odds-ratios, as defined in the Appendix, for the Fusion database described in the next section. Histograms are based on 2,000 bins and viewed as densities ρ_s and ρ_B of Snps and alleles, respectively. Allele locations are far more effective locators of risk, as suggested above by their higher informational content.

The Indicator Vector

In contrasting the disease and control populations, one must be mindful that many other variabilities are at work. For example for type 2 diabetes, T2D, specifically studied here, see next Section, the control and disease populations can be extremely diverse, since there are manifold ways of having and not having T2D, for example by possessing any number of additional diseases as well as to ethnic and other (irrelevant for us) phenotype factors. As a second step in the procedure we restrict attention to the subspace of high value loci determined from the full database. Within this sub-space a classifier is sought which is optimally correlated with the disease population, and minimally correlated with the control population, see Appendix for specific details.

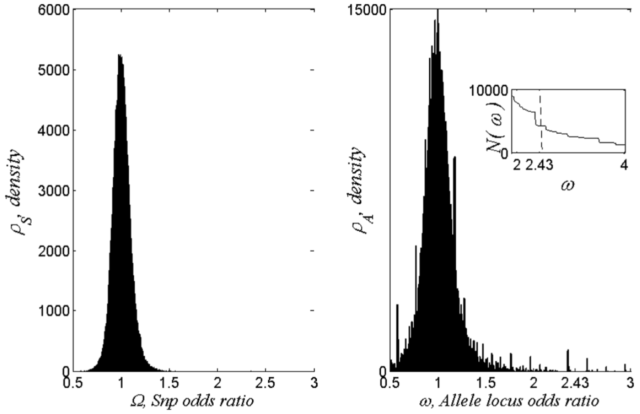

Figure 1. Histograms of odds-ratios for the Fusion database. At the left the customary Snps odds-ratio, Ω, is shown, while at the right individual base pair odds-ratios, ω, are shown. Note the ordinate range at the right is twice that at the left.
doi:10.1371/journal.pone.0085684.g001

It follows from Figure 1 that there are about 8,000 loci in the subspace defined by odds ratios, ω, greater than 2. Figure 2 shows the histograms of all intra-population distances in this space. Distance between sequences of a population is defined as the number of letter substitutions to obtain agreement, the Hamming distance, d_H. The central limit theorem suggests that both histograms are well fit by gaussians with the indicated parameters. The more widely distributed quality of the control population underlines the above remark that "there are manifold ways of not having T2D". It is the present contention that the much more closely grouped form of the disease set, as well as the other structural differences of the two populations seen in Figure 2 provides convincing evidence that this collection of loci provides a framework that separates disease and control.

We will see that for T2D, the resulting number of loci, i.e., the dimension of the indicator is reduced but remains in the thousands. Thus the classifier is a *word* of an equal number of characters, viewed (projected) in this sub-space of allele loci.

Application to Data: Type 2 Diabetes

Figures 1 and 2 are based on "The Finland-United States Investigation of NIDDM Genetics (FUSION) Study", which was obtained from NIH-dbGap. This study which focuses on type 2 diabetes (T2D) has been well described in the literature [17–21].

This database contains 919 T2D cases and 787 normal glucose tolerant (NGT) controls. Each genomic record contained 315,693 common Snps, or 631,386 allele pairs. Although the mean level of

Table 1. Nominal Snps and their transformations as described in the text.

1	Locus	…	N	N+1	N+2	N+3	N+4	…
2	Snps	…	(A,T)	(C,G)	(C,T)	(A,T)	(A,G)	…
3	Acquired Sequence	...	AT	CC	TT	TA	AG	…
4	Standard Form	…	TA	CC	TT	TA	GA	…
5	Alias	…	2 1	1 1	2 2	2 1	2 1	…

doi:10.1371/journal.pone.0085684.t001

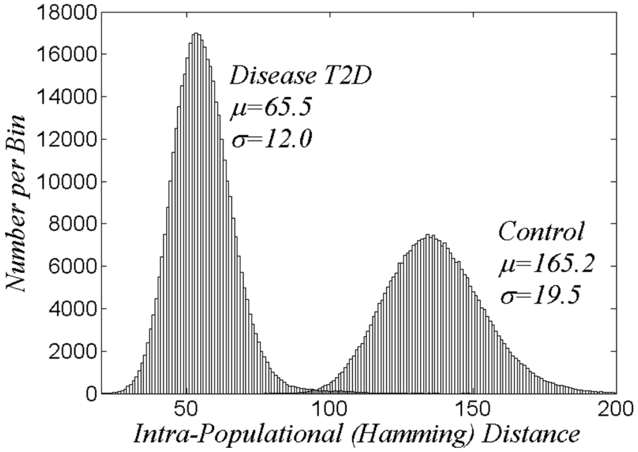

Figure 2. Histograms of intra-populational sequences in a sub-space of roughly 8,000 base pair loci, as obtained for odds-ratios $\omega > 2$ from Figure 1.
doi:10.1371/journal.pone.0085684.g002

missing data was low, .014%, individual loci had as many as 10% missing symbols. Rational procedures exist for dealing with missing data but this was not deemed to be a priority. Instead all loci which in totality had more than 2 missing symbols over all sequences were dropped. This left 272,423 Snps or 544,846 pairs. The few remaining missing symbols were then replaced by the appropriate modal symbol at the allele locus.

Analysis Criteria

For Figure 2 the criterion for choosing high value loci was taken to be that the odds ratio, ω, be larger than 2. This produced a set of $\approx 8{,}000$ loci. Clearly Figure 2 shows a structural case/control difference. To further specify this we denote by $l(\omega)$ the loci set such that $\omega' > \omega$ and by $N(\omega)$ the number of these loci, shown in the inset of Figure 1. Further define $\mathcal{C}(\omega)$ to be mode word (or mode classifier) gotten by choosing the mode symbol at each locus of $l(\omega)$. The $2 \times N$ matrix

$$\mathbf{v}_m(\omega) = \begin{bmatrix} l(\omega) & \rightarrow \\ \mathcal{C}(\omega) & \rightarrow \end{bmatrix}; \qquad (1)$$

will be referred to as the modal indicator (vector).

Any sequence, S, case or control, when projected on to the set $l(\omega)$, denoted by $S(\omega)$, has agreement with the disease classifier given by

$$S_c(\omega) = N(\omega) - d_H(\mathcal{C}(\omega), S(\omega)). \qquad (2)$$

This will be referred to as the score, and clearly $S_c \leq N(\omega)$.

Figure 3(A) shows the scores for all 1706 sequences of T2D for $\omega = 2$; $N(\omega = 2) = 7962$, the highest possible score and as indicated the disease and control scores have mean and standard deviation (μ_d, σ_d) and (μ_c, σ_c) as given in the Figure legend. Since scores are sums of random variables, the central limit theorem might be regarded as applicable and the gaussian fit is also plotted on the ordinate scale. Two other cases, $\omega = 3$ and $\omega = 4$, are shown in Figures 3(B) and 3(C). Across this range $2 \leq \omega \leq 4$ there is an accurate scaling of parameters given by

$$(\mu_d - \mu_c)/N \approx .006, \mu_d/\mu_c \approx 1.006, \sigma_d/N \approx .0011. \qquad (3)$$

but not for σ_c.

An ROC analysis, given in the Appendix, shows that total error, false positives plus false negatives, for $\omega = 2,3,4$ is given by 2, 18, and 66, respectively each being a relatively small fraction of the scores. The error is based on the gaussian fits; the actual data show a slightly larger error, which is due to the usual poorer fit at the tail of a distribution.

This results in nested ranges of loci which are based on the odds ratio and lead to different degrees of success in distinguishing case/control. Calculation shows that when $\omega = 1.9$ $N(1.9) \approx 9000$, while $N(2) \approx 8000$, $N(2.1) \approx 7000$ and thus there is high sensitivity in the neighborhood of $\omega = 2$, see inset of Figure 1. A detailed investigation of Figure 1 shows that $N(2.38) - N(2.48) = 25$ so that the interval (2.38, 2.48) is a relatively insensitive range, a sweet spot. We therefore focus on the results obtained when the odds ratio has the threshold $\omega = 2.43$, and this will be our reference case. As shown in the inset of Figure 1, $N(\omega)$, is virtually flat in this interval.

A simple argument shows that 1906 loci can precisely fit the Fusion disease/control outcomes, which raises the issue of overfitting. Figure 3(C) refutes this and larger values of ω further reduce the estimate of needed loci. A later discussion will imply that two thirds of the above loci are irrelevant for T2D prediction.

Indicator Vector

The modal symbol which appears in the mode indicator, (1) does not reflect the degree of probability of the symbol, only that it is greater than 1/2. This is improved on by the indicator vector. To obtain this we embed allele space into a Euclidean space, which has the advantage of having a distance which is an inner product, see *Indicator Vectors* in the Appendix. This choice of distance transforms the disease and control matrices, $\mathbf{M}_d$ and $\mathbf{M}_c$, to a numerical form. Next an optimal disease classifier, $\mathbf{v}$, is obtained, from a reasonable criterion, see (A.13). In plain terms $\mathbf{v}$ is a *word* that is highly correlated with the disease set and minimally correlated with the control set. The procedure outlined in the Appendix leads to the refinement that only loci having the most highly probable symbols are selected (see Appendix).

For the reference case with threshold $\omega = 2.43$, $N(\omega) = 4315$ along with an estimated 6 errors. The plots for this case resemble those depicted in Figure 3. Use of he indicator vector reduces the number of loci to 4300 and an error estimate of 4. This modest reduction is largely due to the insensitivity of the odds ratio threshold in the neighborhood of $\omega = 2.43$. For the odds ratios shown in Figure 3, indicator reductions in the number of loci is roughly 25%.

Validation and Prediction

Consistency of the T2D classifier is next explored by comparing it with results from data generated by randomizing the phenotypical Fusion sequences in a manner consistent with the data. The resulting classifiers computed at the same $\omega = 2.43$ criterion level are then compared with the T2D classifier. This randomization produces three possible alterations of the 4300 distinguished loci: (1) symbol change $1 \rightleftarrows 2$; (2) high → low value at a major locus; (3) low → high value at a minor locus. Non-parametric statistics thus implies a 1/8 overlap of the randomized classifier with the T2D classifier, i.e., an intersection $I = 538$, which is confirmed by the results obtained for 50 randomized trials:

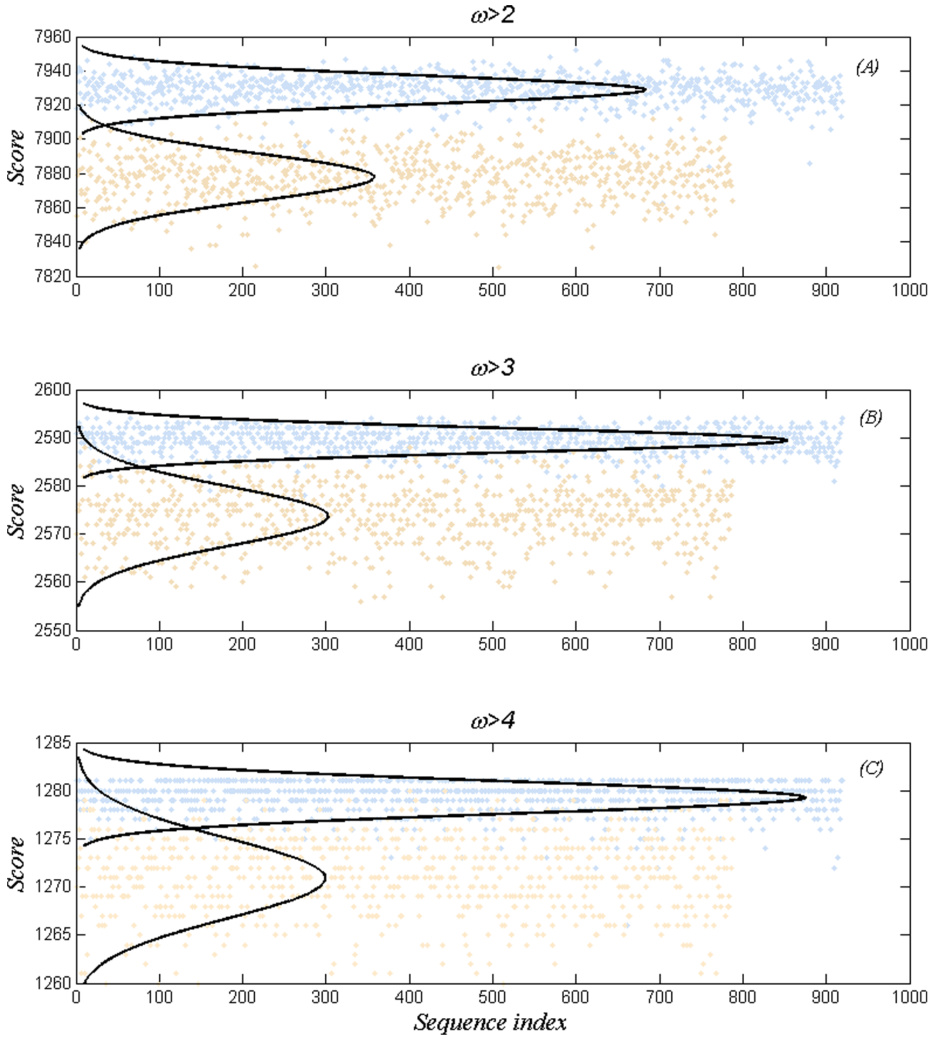

Figure 3. Agreement scores and gaussian fits at 3 odds ratios: (**A**) $\omega > 2$, $\mu_d = 7929$, $\sigma_d = 8.58$, $\mu_c = 7878$, $\sigma_c = 14.12$ (**B**) $\omega > 3$, $\mu_d = 2.590$, $\sigma_d = 2.58$, $\mu_c = 2574$, $\sigma_c = 6.26$ (**C**) $\omega > 4$, $\mu_d = 1279$, $\sigma_d = 1.68$, $\mu_c = 1270$, $\sigma_c = 4.21$. Light blue dots represent the 919 case sequences, light pink dots the 787 control sequences.

doi:10.1371/journal.pone.0085684.g003

$$I = 552 \pm 46 \ \& \ M = 4{,}361 \pm 134, \tag{4}$$

where M is the set size at the $\omega = 2.43$ threshold. It is an underlying hypothesis of the Fusion data acquisition that the (roughly 540,000) non T2D loci should be statistically the same for disease and control. The above 3 *alterations* show that there are only T2D loci shifts. Under the randomization the overall distribution of odds ratios, ω, only shows changes in nearby T2D allele loci.

The results of (4) should be compared with the estimate of I if M is chosen at random without replacement; in this case classifier overlaps would be given by

$$I = M^2/N, \tag{5}$$

where $N = 544{,}846$.

If $M = 4300$, this gives $I = 34$, and if the symbols are also random this gives $I = 17$. To emphasize this point 4300 loci were chosen at random and for 50 trials if the symbol is chosen as the mode we obtain $I = 31.06 \pm 4.82$; and if the symbol is chosen at random $I = 14.62 \pm 4.82$, is obtained, confirming (5).

These deliberations show that the T2D classifier is certainly not only not an artifact, but also clearly emphasizes the special role of the T2D loci for type 2 diabetes.

The best evidence for the T2D indicator would be an objective test of prediction. Internal consistency was shown by splitting the Fusion data into a training set, 760 cases and 650 controls; and a test set of 159 cases and 137 controls. The indicator vector for the training set was then determined, and applied to the test set for disease/control designation. The error rate was about 1%, about the same as for the training set. Since all sequences figured in the determination of high value loci, this is irrelevant for purposes of prediction.

A successful prediction was achieved in one limiting case. At random, one disease and one control sequence was removed from the data and reserved as the 'test set'. The classifier for the remaining data was then determined. At each locus of the classifier, the difference of the 'correct' classifier frequency for the disease and control populations was calculated. *For no apparent reason* if the bottom 3,000 so ranked loci replaced the original classifier a statistical advantage resulted. Over the course of 1,000 such trials a 52% correct prediction rate was obtained, yielding a p-value of.03. No specific risk loci were obtained, only the certainty that such loci exist within the reduced classifier.

It is the present contention that statistical fluctuations in the data possibly allow for hidden contrasts between the case/control sets besides T2D, and that this confounds prediction. To investigate this we randomly chose sets of 60%, 70%, 80%, 90% of the Fusion populations of disease and controls. Many repeats of this showed a high level of consistency in the size of the corresponding indicator spaces. (Note this procedure always leads to a nested set of loci that at 100% is the T2D indicator that has been obtained here).

If we denote the average indicator loci size at the five values (60%, 70%,..., 100%) by L_i and the corresponding population sizes by P_i then a simple regression shows that

$$L = L_o + \frac{a}{P} \tag{6}$$

with

$$a = 5.1367 \times 10^6, \ \ L_o = 1{,}310, \tag{7}$$

fits the data to within a fraction of one percent. Thus in the limit of unbounded data the estimated number of T2D loci for the classification is $L_o = 1{,}310$. Since the degree of diversity in the Fusion data is small this is likely an over estimate, which also implies that over two thirds of the allele loci are irrelevant for prediction.

Further Results and Discussion

For the representative case, $\omega = 2.43$, about 4300 loci were found for the classifier, however the above analysis suggests that with sufficient data this number might be reduced to about 1300. About two thirds of the 4300 are due to possible hidden contrasts, which in turn are due to data fluctuations and the limited amount of data. Unfortunately, the present analysis cannot suggest which 1300 loci are *correct*, and our further remarks can only be stated in general terms.

The literature contains suggestions that many loci figure in the genomics of complex diseases, however that thousands of genomic markers are involved might not have been anticipated. It is tempting when confronted with such large collections of loci to suggest that a network is involved, but the analysis does not have the capacity to reveal potential interactions as implied by this terminology.

To further the issue of possible patterns recall that Figure 2 was constructed from the histograms of $\mathbf{S}_D$ and $\mathbf{S}_C$ the matrices of Hamming distances of intra-disease and intra-control sequences, sometimes referred to as structure matrices. Next we adopt a re-ordering of sequences based on increasing average Hamming distance of a sequence to all others of the set; which in spirit is similar to the ordering generated by dendrograms in taxonomy. The result is shown in Figure 4 where the image of the reordered structure matrix $\mathbf{S}_D$ is shown above, and below that for $\mathbf{S}_C$. The difference in appearance of the two sets is striking; note the different color bars ranges. The upper panel of Figure 4 depicting disease shows a strong granularity and in analogy with taxonomy might indicate possible sub-types of T2D and also a lack of structure for $\mathbf{S}_C$.

Next we consider the score calculations of Figure 3 for the case of $\omega > 2.43$, but make use of the reordering of Figure 4. This is shown in Figure 5. A small number of *false negatives* and *false positives* have been removed for display purposes. The result appears as a structured generally decreasing trace. The granularity found in Figure 4 has a counterpart in Figure 5. The steps and their vertical columns suggests nearby sets of sequences; an issue for future investigation.

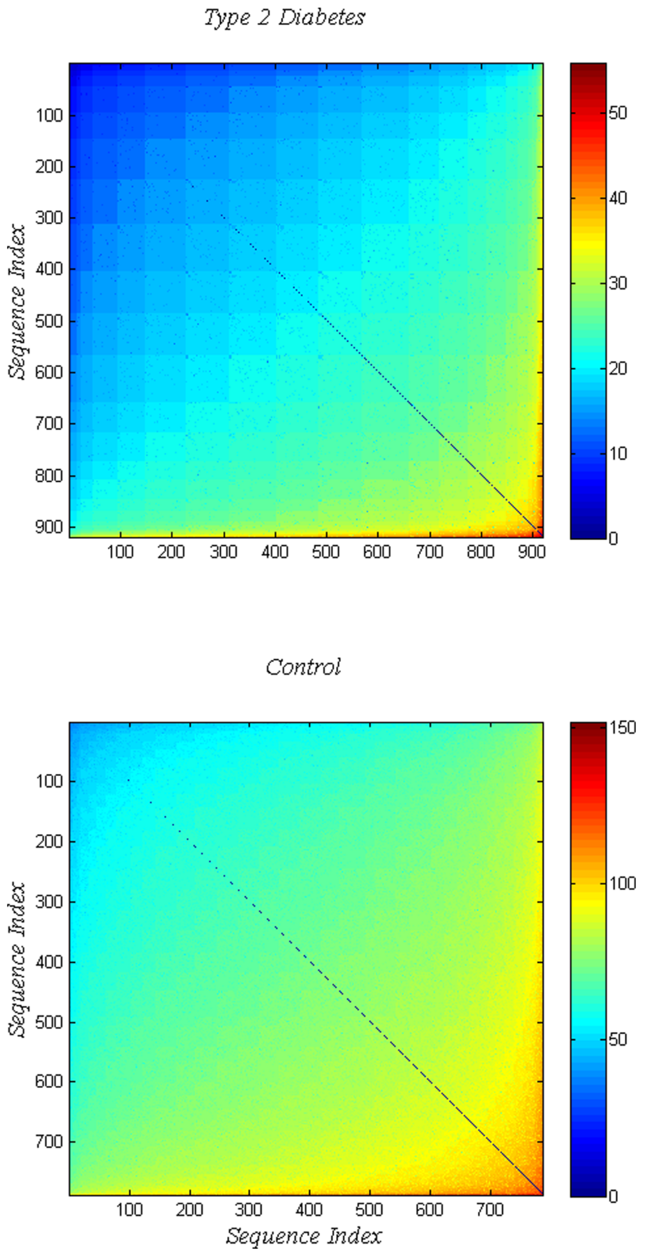

Figure 4. Reordering the reference case structure matrices for T2D on top, and for controls bottom. Reordering is based on ascending values of the mean Hamming distance of a sequence from all others.
doi:10.1371/journal.pone.0085684.g004

The present results suggest that if the presence of Snps is due to mutations, then single mutations are irrelevant since they are shared by both the disease and control sets, an indication that attention to individual loci may be futile. The situation is consistent with the view that a threshold number of such mutations

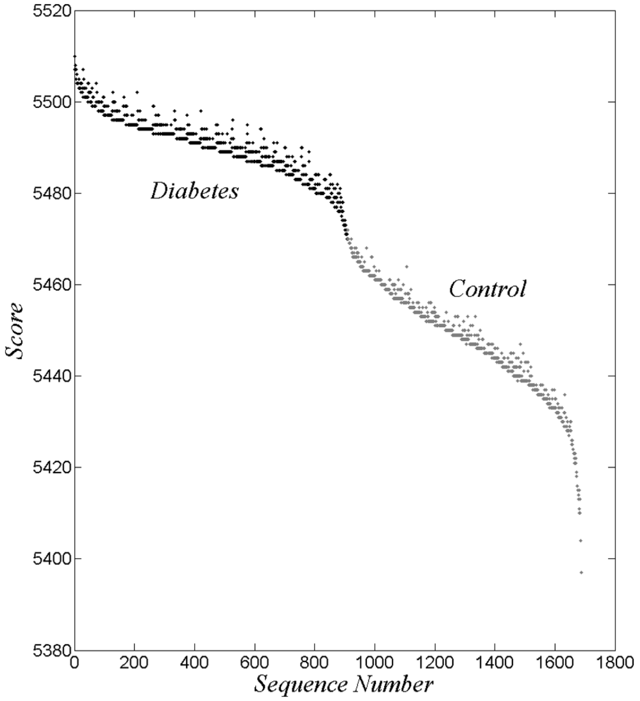

Figure 5. Scores reordered according to Figure 4.
doi:10.1371/journal.pone.0085684.g005

is required for disease to manifest itself; an implication of Figures 3, and 5.

As a metaphor for this situation consider the case of a neuron which gathers small incremental membrane potentials, and only when all inputs sum to a threshold does the neuron perform its function of firing an action potential. Perhaps a more relevant attempt at modeling, is to note that the distinguished collection of 4300 loci is of unknown organization, but that for a score above 4286 the collection manifests itself as diabetes, and below that as the absence of disease. In such terms the score can be compared to a morphogen, which above some threshold produces a form of tissue different than below the threshold [22,23], with 'function' or 'organization' replacing 'tissue'.

Recent studies of T2D data [24,25] produced a compilation of *variants* that might be associated with T2D. In total that list contains 121 *risk loci*, but only 32 are shared with the full Fusion dataset and none of these belong to the reference set of 4300 loci. Both cited investigations made use of the observation that Snps might be acquired as segments of DNA and therefore of associated genomic material, referred to as haplotypes. This concept played no role in the present development, which regards all loci as independent. Such locational correlations might be present in the present results, and a Hapmap of the 4300 loci might prove to be revealing. Along similar lines since roughly 98% of DNA is non-coding [26] it may be of interest to determine how much of this kernel of 4300 genomic loci is coding.

The observation that a disease mechanism might not be revealed by analysis of data from genome wide associative studies [8] is not refuted by what is presented here. As already observed there have been many allusions in the literature to the modest role that individual Snps play in a disease as complex as T2D, but the notion that large numbers of loci might play a role, is probably a surprise. On the other hand since this number of loci appears to successfully distinguish disease from control it would be remarkable if the mechanism of the disease is not to be found in the set. It is also an implication of the present analysis that more case/control records would significantly shrink the classification set.

Since the distinguished kernel of genomic loci is heavily present in the *healthy* control set, a more subtle question is: 'What is its role?' in the control set.

Appendix: Analytic Summary

Standard Order

The n^{th} Snp of a sequence is recorded as two allele loci, say $(2n-1,2n)$ referred to as the odd and even members of the n^{th} Snp pair. According to the *standard order* the higher number in the symbol alias is recorded first and is then re-aliased as 2 and the lower number is recorded second and re-aliased as 1. No information is lost since the accompanying rs number, part of the database, of a Snp furnishes chromosome location and alleles themselves. The essential Snp information is thus determined by whether 22, 21, or 11 is recorded. This is illustrated in the following table.

Probabilities

For a population of sequences, at any Snp, one can calculate the probability $p^o(2)$ as the frequency of symbol 2 at the odd locus and similarly $p^e(1)$ at the even locus.

It is clear from this formulation that the probability of the 22 pair, $P(22)$ is

$$P(22)=1-p^e(1) \tag{A.1}$$

and similarly for

$$P(11)=1-p^o(2), \tag{A.2}$$

and finally

$$P(21)=p^o(2)+p^e(1)-1. \tag{A.3}$$

If $P(2)$ denotes the probability of symbol 2 for a Snp then

$$P(1)=\frac{p^o(1)+p^e(1)}{2} \tag{A.4}$$

and

$$P(2)=\frac{p^o(2)+p^e(2)}{2} \tag{A.5}$$

Information

For the Snps case the information (entropy), from (A.5) is

$$S_{Snps}=-\frac{p^o(1)+p^e(1)}{2}\ln(\frac{p^o(1)+p^c(1)}{2}) \tag{A.6}$$

$$-\frac{p^o(2)+p^e(2)}{2}\ln(\frac{p^o(2)+p^e(2)}{2})$$

and the allele version has two compartments

$$S_{Al}=S^{o}_{Al}+S^{e}_{Al}=-[(p^{o}(1)\ln p^{o}(1)+p^{o}(2)\ln p^{o}(2)] \\ -[p^{e}(1)\ln p^{e}(1)+p^{e}(2)\ln p^{e}(2)]. \tag{A.7}$$

Odds Ratios

For a disease and counterpart control population the allele odds ratios are given by

$$\omega^{s}=\frac{p^{s}_{d}}{1-p^{s}_{d}}/\frac{p^{s}_{c}}{1-p^{s}_{c}};\ s=o,e. \tag{A.8}$$

where the subscripts d and c indicate the disease and control set, respectively.

The Snp odds ratio is

$$\Omega^{a}=\frac{P_d(a)}{1-P_d(a)},\frac{1-P_c(a)}{P_c(a)} \tag{A.9}$$

where a is the major allele, and in which (A.4) and (A.5) can be substituted.

Table 2 shows a string of three Snps of risk loci. The $p^o(2)$ and $p^e((1)$ probabilities, for disease and control sets are shown on the first two lines. The last two lines show odds ratios based on allele loci and Snp loci, respectively. See Figure 1.

If the probabilities of risk loci lie close to unity for both disease and controls so we can write $p_d=1-\epsilon_d$ and $p_c=1-\epsilon_c$ then (A.8) shows $OR\approx\epsilon_c/\epsilon_d$, which accounts for the bold face values of line 3 of the table. On the other hand from (A.4) and (A.5) $P_d\approx P_c$ and leads to $\mathbf{\Omega}$ values near unity.

Indicator Vectors

To pass from a symbolic sequence to a numerical vector we set

$$\begin{aligned}1&\rightarrow[1,0]\\2&\rightarrow[0,1]\end{aligned} \tag{A.10}$$

which is a reduced form of the more general case [15,16]. Thus

$$\begin{aligned}S&=(A,G,A)\rightarrow\\ \mathbf{S}&=[1,0,0,1,1,0].\end{aligned} \tag{A.11}$$

The transformation (A.10) of a sequence of N alleles becomes a vector in a Euclidean space of dimension $2N$. The Euclidean distances, d_E, between sequences of the same length is related their Hamming distances d_H by the relation

$$d_H=d_E^2/2. \tag{A.12}$$

The vectorized matrix of disease and control sequences will be denoted by $\mathbf{M}_d$ and $\mathbf{M}_e$ respectively, i.e., the rows of each are the vectorized genomic sequences of the corresponding data. To generate the indicator vector, $\mathbf{v}$, of the disease class we seek the maximum of the criterion functional

$$\langle\|\mathbf{M}_d\mathbf{v}\|^2\rangle-\langle\|\mathbf{M}_e\mathbf{v}\|^2\rangle \tag{A.13}$$

under the condition that

$$\|\mathbf{v}\|^2=1 \tag{A.14}$$

and where $\langle\ \rangle$ indicates the average over rows. This leads to

$$\frac{1}{N_d}\mathbf{M}^{\dagger}_{d}\mathbf{M}_d\mathbf{v}-\frac{1}{N_c}\mathbf{M}^{\dagger}_{c}\mathbf{M}_c\mathbf{v}=\lambda\mathbf{v}, \tag{A.15}$$

and simple arguments show there must be at least one positive λ.

The dimensionality of the problem can be substantially reduced by recognizing that $\mathbf{v}$ must be an admixture of the rows of $\mathbf{M}_d$ and $\mathbf{M}_c$, thus

$$\mathbf{v}=\mathbf{M}^{\dagger}_{d}\alpha+\mathbf{M}^{\dagger}_{c}\beta, \tag{A.16}$$

known as the method of snapshots [27]. From (A.16) it then follows that

$$\begin{bmatrix}\mathbf{M}_d\mathbf{M}^{\dagger}_{d}/N_d & \mathbf{M}_d\mathbf{M}^{\dagger}_{c}/N_d\\ -\mathbf{M}_c\mathbf{M}^{\dagger}_{d}/N_c & -\mathbf{M}_c\mathbf{M}^{\dagger}_{c}/N_c\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix}=\lambda\begin{bmatrix}\alpha\\ \beta\end{bmatrix} \tag{A.17}$$

Solution of (A.17) leads to a set of loci and the distinguished genomic symbol. As mentioned in the main text, the combination is the indicator vector and the ordered symbols, or word, is the classifier in the corresponding locus space.

ROC Analysis

The proper odds ratio threshold might sensibly be formulated in terms of true and false positives and negatives as customarily treated by *receiver operation characteristic*, ROC, curves. For simplicity it will be assumed that the distribution of scores $G_d(s)$ and $G_c(s)$, being sums of large numbers of loci, can by the central limit theorem be fit by normal distributions,

$$G_d(s)=N(\mu_d,\sigma_d,s) \tag{A.18}$$

and

$$G_c(s)=N(\mu_c,\sigma_c,s) \tag{A.19}$$

of mean μ and standard deviation σ.

For N_d and N_c the number of disease and control sequences, and a discrimination value T

Table 2. Odds-ratios, ω and Ω, at three Snp locations.

	Snp (1)		**Snp (2)**		**Snp (3)**	
	Odd	**Even**	**Odd**	**Even**	**Odd**	**Even**
Disease	.2302	.9934	.2329	.9947	.9987	.1105
Control	.2415	.9754	.2354	.9785	.9923	.1000
OR ω	.94	**3.80**	.9862	**4.13**	**5.96**	1.118
OR Ω	1.12		1.090		.9608	

doi:10.1371/journal.pone.0085684.t002

$$r_d(T)=\frac{N_d}{N_d+N_c}\int_T^{\infty}G_d(s)ds \qquad (A.20)$$

is the fraction of true positives. The plot of this versus false positives, for the range of T furnishes a ROC curve. The error fraction is

$$\epsilon_r(T)=\frac{N_d}{N_d+N_c}\int_{-\infty}^{T}G_d(s)ds+\frac{N_c}{N_d+N_c}\int_T^{\infty}G_c(s)ds. \qquad (A.21)$$

the minimum T, is easily calculated and occurs at the point of the ROC curve of slope -1. For any odds-ratio threshold this is considered the ideal value.

Pseudo-probabilities

An issue is the fact that choosing the modal symbol at a locus only requires a probability $>1/2$. All probabilities as defined above must satisfy

$$p(1)+p(2)=1 \qquad (A.22)$$

whether defined for alleles or Snps. Under the vectorization (A.10) and the optimization (A.10) it can be shown that (A.22) is preserved. However, in the space of all probabilities, probabilities may leave the first orthant. In simple terms under (A.13) individual probabilities can be greater than unity and so also can be negative. This circumstance has a history in theoretical physics [28–30] and has been given a sound mathematical basis [31], and these are sometimes termed pseudo-probabilities. A part of the optimization is to only retain those loci which are overprobable, by choice the level is $p>1$, which is not critical.

Acknowledgments

The author takes pleasure in expressing gratitude to Yunguo Yu for the help afforded by his considerable skill and expertise in dealing with computer issues and data management. Once again it is a pleasure to thank Ellen Paley for the careful and patient preparation of a manuscript. The author also thanks his long standing colleague, Bruce Knight, for his sustained encouragement, and for his critical reading of earlier versions of the manuscript. Special thanks to my family Carole, Brenda, and Matthew for helpful comments and support. An essential element of this effort was the high quality Fusion database, which was acquired from NIH at the request of the Rockefeller University Committee for Clinical and Translational Science, grant # UL1RR024143, National Center for Research Resources, National Institutes of Health. Finally grateful thanks to Mitchell Feigenbaum for affording me the hospitality of Rockefeller University.

Author Contributions

Conceived and designed the experiments: LS. Performed the experiments: LS. Analyzed the data: LS. Contributed reagents/materials/analysis tools: LS. Wrote the paper: LS.

References

1. Brookes A, Lehvaslaiho H, Siegfried M, Boehm J, Yuan Y, et al. (2000) HGBASE: a database of SNPs and other variations in and around human genes. Nuclei Acids Res 28: 356–360.
2. Stenson P, Mort M, Ball E, Howells K, Phillips A, et al. (2009) The human gene mutation database: 2008 update. Genome Medicine 1: 13.
3. Klein R, Zeiss C, Chew E, Tsai JY, Sackler R, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308: 385–389.
4. Johnson A, O'Donnell C (2009) An open access database of genome-wide association results. BMC Genet 10.
5. Chakravarti A (2011) Genomics is not enough. Science 334: 15.
6. Roberts N, Vogelstein J, Parmigiani G, Kinzler K, Vogelstein B, et al. (2012) The predictive capacity of personal genome sequencing. Sci Transl Med 4: 133ra58.
7. Kolata G (2012) Study says DNA's power to predict illness is limited. NYTimes, April 2.
8. Manolio T (2010) Genome-wide association studies and assessment of the risk of disease. N Engl J Med 363: 166–176.
9. Gonzalez J, Armengol L, Sole X, Guino E, Mercader J, et al. (2007) SNPassoc: an r package to perform whole genome association studies. Bioinformatics 23: 654–655.
10. Aulchenko Y, Ripke S, Isaacs A, van Duijn C (2007) GenABEL: an r library for genome-wide association analysis. Bioinformatics 23: 1246–1296.
11. Purcell S, Neal B, Todd-Brown K, Thomas L, Ferreira M, et al. (2006) PLINK: A tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81: 559–575.
12. Sirovich L, Everson R, Kaplan E, Knight B, O'Brien E, et al. (1996) Modeling the functional organization of the visual cortex. Physica D 96: 355–366.
13. Everson R, Prashanth A, Gabbay M, Knight B, Sirovich L, et al. (1998) Representation of spatial frequency and orientation in the visual cortex. PNAS 95: 8334–8338.
14. Sirovich L, Kaplan E (2002) Chapter 3: Analysis methods for optical imaging. In: Methods for In Vivo Optical Imaging of the Central Nervous System (R Frostig, ed.). CRC Press, Boca Raton, 43–76.
15. Sirovich L, Stoeckle M, Zhang Y (2009) A scalable method for analysis and display of DNA sequences. PLoS ONE 4: e7051.
16. Sirovich L, Stoeckle M, Zhang Y (2010) Structural analysis of biodiversity. PLoS ONE 5: e9266.
17. Valle T, Tuomilehto J, Bergman R, Ghosh S, Hauser E, et al. (1998) Mapping genes for niddm. design of the finland-united states investigation of niddm genetics (fusion) study. Diabetes Care 21: 949–958.
18. Ghosh S, Watanabe R, Valle T, Hauser E, Magnuson V, et al. (2000) The finland-united states investigation of non-insulin-dependent diabetes mellitus genetics (FUSION) study. I. an autosomal genome scan for genes that predispose to type 2 diabetes. Am J Hum Genet 67: 1174–1185.
19. Watanabe R, Ghosh S, Langefeld C, Valle T, Hauser E, et al. (2000) The finland-united states investigation of non-insulin-dependent diabetes mellitus genetics (FUSION) study, II. an autosomal genome scan for diabetes-related quantitative-trait loci. Am J Hum Genet 67: 1186–1200.
20. Silander K, Scott L, Valle T, Mohlke K, HM S, et al. (2004) A large set of finnish affected sibling pair families with type 2 diabetes suggests susceptibility loci on chromosomes 6, 11, and 14. Diabetes 53: 821–829.
21. Scott L, Mohlke K, Bonnycastle L, Willer C, Li Y, et al. (2007) A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. Science 316: 1341–1345.
22. Wolpert L (1969) Positional information and the spatial pattern of cellular differentiation. Theor Biol 25: 147.
23. Wolpert L, Jessell T, Lawrence P, Meyerowitz E, Robertson E, et al. (2007) Principles of Development (Third Edition). Oxford: Oxford University Press.
24. Morris A, Voight B, Teslovich T, Ferreira T, Segre A, et al. (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature Genetics 44: 981–990.
25. Zeggini E, Scott L, Saxena R, Voight B (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nature Genetics 40: 638–645.
26. Elgar G, Vavouri T (2008) Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. Trends in Genetics 24: 344–352.
27. Sirovich L (1987) Turbulence and the dynamics of coherent structures, parts i, ii, and iii. Quarterly of Applied Mathematics XLV: 561–590.
28. Dirac P (1942) The physical interpretation of quantum mechanics. Proc Roy Soc London (A 180): 1–39.
29. Wigner EP (1932) On the quantum correction for thermodynamic equilibrium. Phys Rev 40: 749–759.
30. Feynman RP (2000) Techniques in molecular systematics and evolution. Basel: Birkaauser Verlag. 31. Bartlett MS (1945) Negative probability. Math Proc Camb Phil Soc 41: 71–73.