A new mathematical model for the interpretation of translational research evaluating six CTLA-4 polymorphisms in high-risk melanoma patients receiving adjuvant interferon compared with healthy controls

Petr Pancoska[1], John M Kirkwood[2], Spyros Bouros[3] Maria Spyropoulou-Vlachou[4], Eirini Pectasides[3], Dimosthenis Tsoutsos[5], Aristidis Polyzos[3], Christos Markopoulos[3], Petros Panagiotou[5], Urania Kastana[6], Dimitrios Bafaloukos[7], George Fountzilas[7], Helen Gogas [3]

1. University of Pittsburgh, Department of Medicine, Center for Clinical Pharmacology, Center for Craniofacial and Dental Genetics, Pittsburgh, Pennsylvania, USA

2. University of Pittsburgh Cancer Institute, Hillman Cancer Center, Pittsburgh, Pennsylvania, USA

3. University of Athens, 1st Department of Medicine, Medical School, Athens, Greece

4. Department of Immunology, National Tissue Typing Center, General Hospital of Athens, Greece

5. Department of Plastic Surgery and Microsurgery, G. Gennimatas General Hospital of Athens, Greece

6. Department of Plastic Surgery, Evangelismos General Hospital of Athens, Greece

7. Hellenic Cooperative Oncology Group, Data Office, Athens, Greece

Address for correspondence

Helen Gogas, M.D.
Associate. Professor in Medical Oncology
1st Department of Medicine
University of Athens
P.O. Box 14120
Athens 11510
GREECE
Tel. +30 6944 6811 59
Fax. +30 210 7781517
e-mail: hgogas@hol.gr

Acknowledgement: P. Pancoska was supported in part by ERC-CZ LL1201 program CORES.

INTRODUCTION

Adjuvant therapy of patients with stage IIB/III melanoma (high-risk) with interferon was approved by FDA (United States Food and Drug administration) and subsequently by regulatory authorities worldwide (1). Despite the ability of this regimen to reduce relapse and mortality by up to 33% (2) acceptance has been limited due to toxicity of this regimen. Attempts to identify the subset of patients destined to benefit from adjuvant treatment with IFNα-2b have failed to discover clinical or demographic features of the patient population that are capable of predicting the benefit from high dose interferon (HDI) therapy. Correlative studies have been undertaken over the years, demonstrating a variety of immunological responses subsequent to therapy (3,4).

We recently published a paper in which six CTLA-4 polymorphisms were evaluated in a cohort of patients treated with adjuvant interferon (5). The human CTLA-4 gene is located on chromosome 2q33, in a region that is associated with susceptibility for autoimmune disease (6) and multiple polymorphisms of the CTLA-4 gene have been found to be associated with susceptibility to autoimmune diseases (e.g. the GG allele of the +49 AG polymorphism is associated with decreased expression of CTLA-4 upon T-cell activation and thus a higher proliferation of T-cells) (7-10).

We genotyped DNA isolated from the peripheral blood of a total of 286 patients with high-risk melanoma who participated in a prospective multicenter randomized phase III trial of adjuvant interferon and a panel of 288 randomly selected healthy unrelated Greek individuals from the Donor Marrow Registry of the National Tissue Typing Center, Athens, Greece that served as a control population for 6 CTLA4-SNPs of potential interest – namely CT 60, AG 49, CT 318, JO 27, JO 30 and JO 31. CT 318 is located within the promoter region of the CTLA-4 gene A/G49 is located at exon 1, while the rest of the SNPs tested are located at the 3' untranslated region of CTLA-4.

High levels of association among the different polymorphisms were found (Fisher's exact p value< 0.001 for all associations). Genotypes corresponding to the six CTLA-4 polymorphisms did not significantly deviate from the Hardy-Weinberg equilibrium. This indicates significant linkage disequilibrium among the six polymorphisms. We analyzed the segregation pattern of CT 318, AG 49, CT 60, JO 27, JO 30, JO 31 SNPs on 572 chromosomes and identified 5 major haplotypes. No statistically significant differences for RFS or OS were found for the presence of each of the 3 most common haplotypes. When the respective polymorphisms were considered separately for outcome analysis by the allele status, or when the three most significant haplotypes were considered, two results emerged:

1. No significant differences were found between the distributions of CTLA-4 polymorphisms in the melanoma population compared with healthy controls.

2. Relapse free survival (RFS) and overall survival (OS) did not differ significantly among patients with the alleles represented by these polymorphisms. No correlation between autoimmunity and specific alleles was evident.

These results on CTLA4 genotype profile as risk factor are the basis for the analysis designed and undertaken in this study. A novel method of pattern analysis, referred below as network phenotyping strategy (NPS), was introduced for integrative, relationship-based analysis of general clinical data (11-13). In this particular application, NPS replaces analysis of individual alleles and allele frequencies by the analysis of relationships between CTLA-4 alleles for every individual in the study. NPS solves the "power" problem of methods that approach such complete-relationship based analysis by using large number of interaction terms, which requires large number of subjects for informative statistical analyses. NPS captures instead the actual polymorphism relationship patterns cumulatively into special mathematical graphs. In our CTLA-4 genotyping data, we thus do not analyze independent interrelationships among the 153 possible combinations of AA, AB and BB alleles of the six studied CTLA-4 polymorphism. Instead, we take advantage of the fact that all those 153 relationships can be captured in a single relationship pattern graph. A path in this graph then encodes the actual complete experimental CTLA-4 genotyping results for every studied subject. In this way, the complete information about all allele relationships for an individual is captured by a single mathematical object. An important property of the NPS analysis is that, from the collection of all individual SNP relationship patterns, we can additionally compute (in a deterministic, non-statistical way) a framework of directly clinically and functionally interpretable reference relationship profiles (RRP). These RRP's represent ”landmarks” in the (multidimensional) clinical/genotypic relationship data space. The clinical significance of the RRP landmarks is then measurable in terms of how many patients have close (but not necessarily identical) personal relationship patterns to those “landmarks”. For the concrete example of CTLA-4 polymorphisms studied in this paper, RRP's represent limiting characterization of the CTLA-4 SNP co-occurrence patterns. The main advantage of the NPS approach is its identification of any significant heterogeneity that might be captured in the data from the clinical, or in this case the CTLA-4 based immune regulation mechanism that we focused upon in this study of subjects with and without melanoma. These results can be then used in designing follow-up clinical studies.

MATERIALS AND METHODS

MATERIALS

Genotyping of DNA isolated from the peripheral blood of a total of 286 patients with melanoma and a panel of 288 randomly selected healthy unrelated Greek individuals that served as a control population was described in detail previously. Six CTLA-4 SNPs were studied, namely CT 60 (rs3087243), AG 49 (rs231775), CT 318 (rs5742909), JO 27 (rs11571297), JO 30 (rs7565213) and JO 31 (rs11571302). CT 318 is located within the promoter region of the CTLA-4 gene, A/G49 is located at exon 1, while the rest of the SNPs tested are located at the 3' untranslated region of CTLA-4.

METHODS:

Characterization of personal CTLA-4 genotype relationship pattern by 6-partite graphs: Identifying the part of the study data in which we have maximal information to extract additional components of information. We present two levels of CTLA-4 genotype analysis. In the first one, we do not distinguish between homozygous or heterozygous status of the six alleles. In the second one, we will expand the genotype characterization using the known zygosity of the six SNP's. Fig. 1a shows how an observed CTLA-4 genotype for one patient may be represented by a 6-partite graph that will be called a personal relationship profile **prp**, which we use for the purpose of the first analysis type, considering the major/minor allele relationships only (Fig.1a). In Fig. 1b we define the type of personal relationship profile, for which symbol **PRP** is used to emphasize that allele relationships include observed allele zygosity. In both these representations, each assayed SNP is represented by one of six partitions in the **prp** or **PRP**. Each partition contains two or three vertices, representing the allele for a given polymorphism (a = major allele, b = minor allele in **prp**, a = major homozygous, ab=heterozygous, b = minor homozygous allele in **PRP**). Edges in both graphs connect only those vertices in different partitions that represent observed (genotyped) alleles in the two different polymorphic sites. The complete CTLA-4 genotype profile for an individual is then a collection of edge-connected vertices in **prp/PRP**, forming a cycle in **prp/PRP**. Because the edges in **prp/PRP** represent relationships between the allelic states of the studied SNP's, there is clear meaning for each segment of the CTLA-4 genotype illustrated in the hexagonal cycle. We can understand these lines in as conditional relationships of type "if AG49 contains minor allele then CT60 contains also minor allele and J031 contains minor allele and then …. ". Note that the experimentally defined cycle in e.g. **prp** represents not only the pair wise conditional relationships shown by lines such as (AG49 = b when CT60 = b), but also all other co-occurrences such as (AG49 = b when JO30 = b) etc. The **prp/PRP** cycle representation of the CTLA-4 SNP allele status co-occurrences is the simplest one capturing all co-occurrence relationships while maintaining convenient mathematical simplicity.

CT318	AG49	CT60	JO31	JO30	JO27
C/C	A/G	A/G	G/T	G/A	T/C
aa	ab	ab	ab	ab	ab

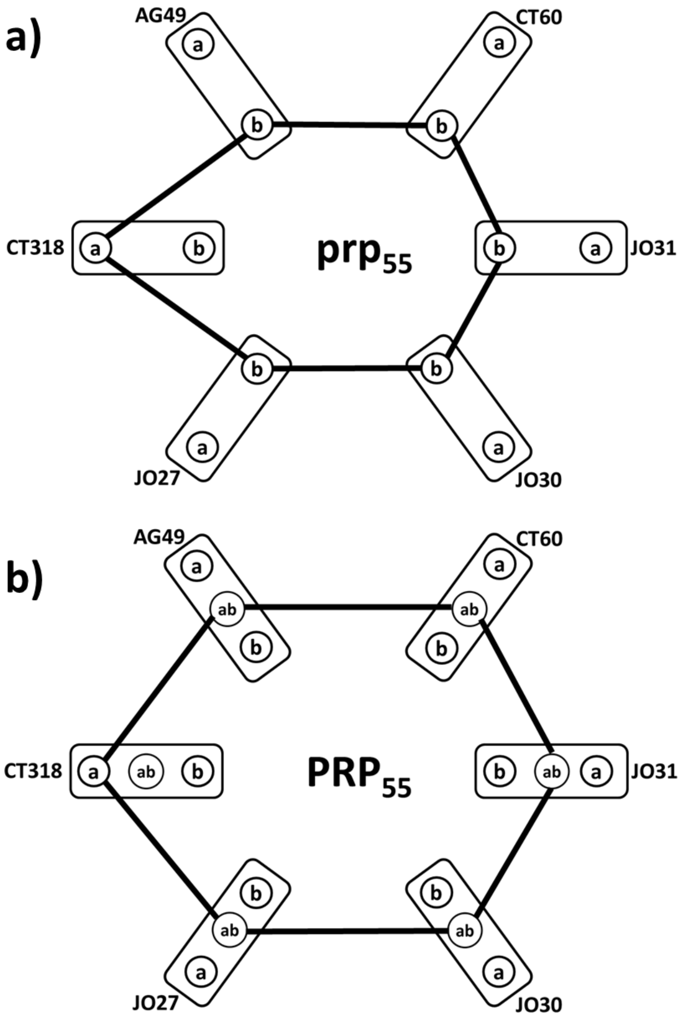

Figure 1. example how experimentally determined CTLA-4 genotype (top panel) for a patient (id=55) is transformed into a) ***prp*** graph and b) ***PRP*** graph. **a**-major allele, **b**-minor allele, **ab**-heterozygous allele status vertices. Each SNP is represented by a graph partition (rectangles), identified by the SNP code. Lines - graph edges, representing the co-occurrences of all alleles in the patient's CTLA-4 genotype .

Collective characterization of CTLA-4 genotype profile distribution in a cohort by cumulative weighted 6-partite graph G. While **PRP**'s are exact "qualitative" representation of the studied polymorphism relationship patterns in CTLA-4, we need to convert this qualitative information into quantitative characterization of these individual relationship patterns. It has been shown by exact mathematical theorem (14) that the maximal quantitative information captured by graphs is obtained when **PRP**'s are compared to one another in graphs of the same type, which we call reference relationship patterns (***RRP***). Therefore, the next step of NPS transformation of the CTLA-4 polymorphism relationship patterns into quantitative descriptors is to use the actual data to derive the 6-partite graphs, representing the ***RRP***'s we need.

For this purpose, the individual **prp** or **PRP** graphs, describing the SNP co-occurrences for all subjects were assembled into cumulative 6-partite "study graphs" **g** and **G**. By adding every individual patient CTLA-4 genotype profile representation **prp** to the cumulative **g** graph, the weightings of every edge in **g** is increased by one, and similarly but independently for PRP's and **G.** As a consequence of this construction, these **g** and **G** graphs will have weighted edges defined by the co-occurrence frequencies of all SNP pairs. The distribution of all individual CTLA-4 genotype profiles in case cohort is now represented by graph **g**.

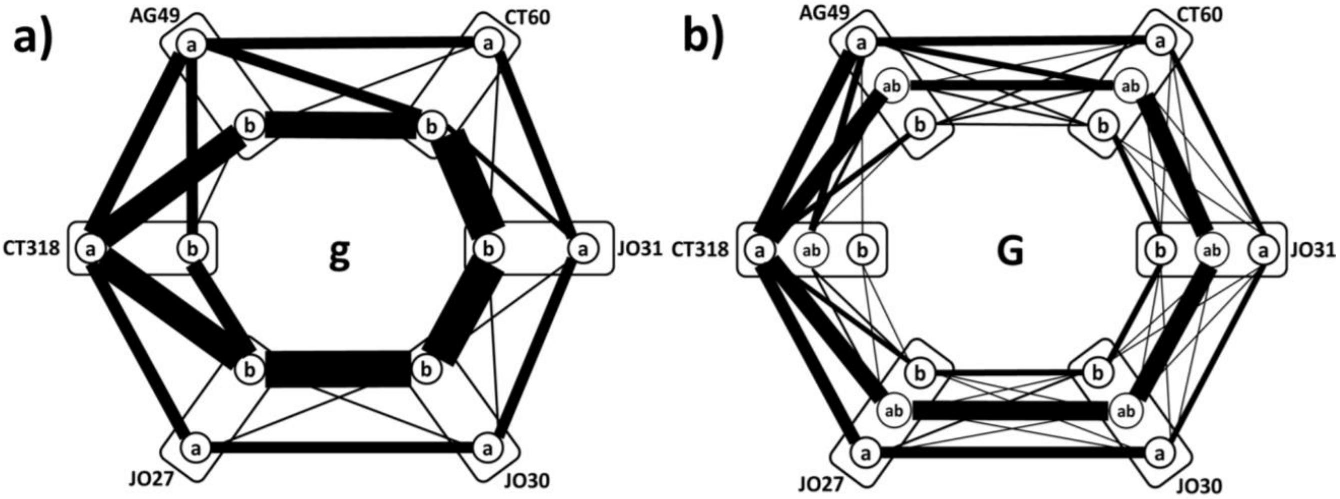

Figure 2. Study graphs **g** (a) and **G** (b) constructed as union of all **prp**'s (g) or **PRP**'s (G). Symbols as in Fig.1, thickness of edges in **g** and **G** are proportional to co-occurrence frequencies of respective SNP pairs, connected by the edge.

In Fig. 2, the relative edge weights, resulting from adding all individual case graphs **prp** and **PRP** to **g** and **G**, respectively, are graphically represented by the variable relative thickness of the edge lines. By converting these edge counts to frequencies, statistical interpretation of the basic vertex-weighted edge-vertex (a-b), (a-a), (b-a) and (b-b) motifs in study graphs is obtained**.** The weights of study graph edges connecting, for example, the major and minor allele vertices in the AG49 and CT60 partitions define the estimates of the following conditional probabilities:

$$a - b \sim P(AG49\ is\ major|CT60\ is\ minor)$$

$$a - a \sim P(AG49\ is\ major|CT60\ is\ major)$$

$$b - a \sim P(AG49\ is\ minor|CT60\ is\ major)$$

$$b - b \sim P(AG49\ is\ minor|CT60\ is\ minor)$$

In the next step, the complete sets of reference relationship patterns for CTLA-4 genotypes in both study graphs g and G are identified and in case of g identified as haplotypes. Haplotype is defined as a series of polymorphisms in CTLA-4 genotype profile that are co-occurring with identical probabilities, P(1)~P(2)~ … ~ P(6). Using the conditional probability interpretation of edges in the study graphs shown in above example, we can derive from the Bayes' theorem, that if sub-graphs of the study graph with equal weights (co-occurrence frequency components) are found, the condition of P(1)~P(2)~ … ~ P(6) is automatically fulfilled. Thus, in our representation, a complete set of haplotypes is represented by all **RRP** cycle subgraphs with equal weights of all edges, which can be found in **g** or **G** by "greedy" algorithm (see Supplement).

For validation of this study graph-based approach to haplotype identification, established procedures were additionally used where the maximum likelihood estimates of haplotype frequencies given a multi-locus sample of genetic marker genotypes [3 different genotypes of the 6 polymorphisms] were generated using the expectation-maximization (EM) algorithm under the assumption of Hardy-Weinberg equilibrium (HWE). Linkage disequilibrium was explored for each pair of the 6 polymorphisms (PROC HAPLOTYPE). SAS 9.1 (SAS Institute Inc., Cary, NC, USA), was used for the statistical analysis (reported in (5)).

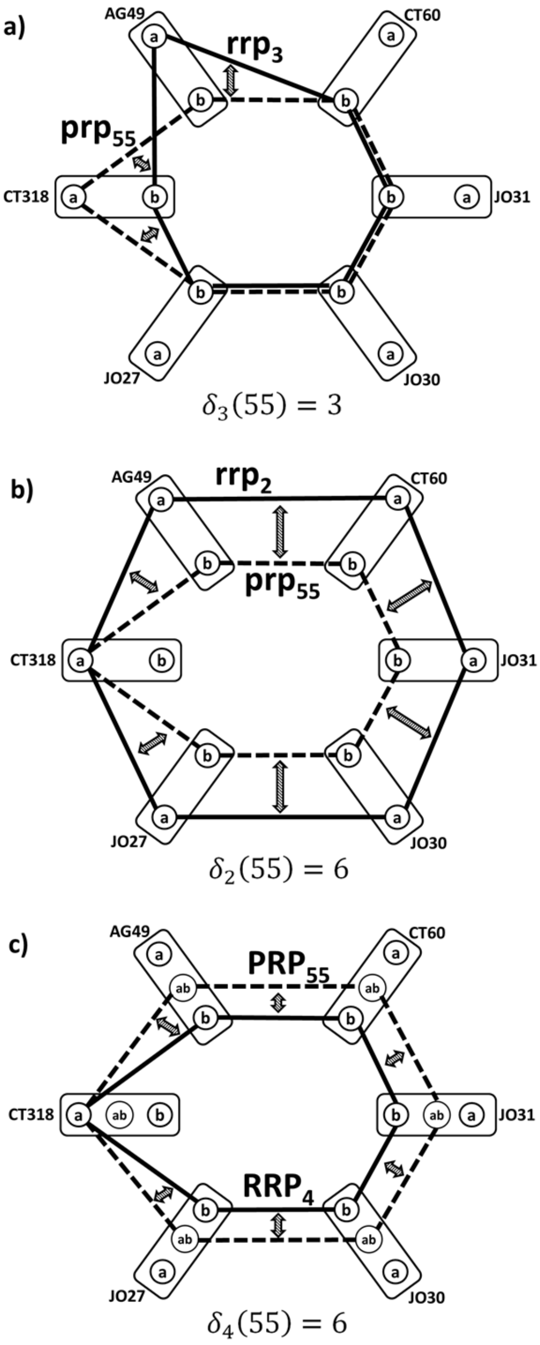

Figure 3. Three examples showing how elements of distance vectors $\vec{\delta}_j$ are computed for the same patient #55. In all figures, **prp (RRP** in c)) for this patient = dashed lines, **rrp**'s (or **RRP** in c)) = solid lines. Double arrows indicate mismatch in SNP co-occurrences. Elements of $\vec{\delta}_j$ are sums of these mismatches (in computations, we add negative sign to make identity (zero mismatches) mathematically largest). **a,b)** comparison of patient's genotype to the second and third reference SNP relationship patterns **rrp c)** Comparison of patient's genotype to the 4th reference SNP relationship pattern **RRP$_4$.**

Quantitative characterization of differences of personal CTLA-4 genotype profiles prp and PRP from haplotypes, represented by *rrp*'s and *RRP*'s.

For the quantification of the graph-graph distances between individual patient relationship patterns and haplotype-reference relationship patterns, we use the mathematical results of (14,15), showing that one of the possible definitions of graph-graph distances with all necessary mathematical properties is obtained simply by counting the number of edge mismatches between the two graphs, as is shown by example in Fig.3. As the result, with haplotype decomposition of study graph **g** resulting in **8** haplotype components, each subject **(j)** is characterized by an **8**-element vector $\vec{\delta}_j = [\delta_j(1), \delta_j(2), \dots, \delta_j(8)]$ of eight distances of the personal CTLA-4 genotype profile from compositions of all **8** respective haplotypes identified. Difference vectors $\vec{\delta}_j$

were computed for all patients and controls using a) the control cohort-defined haplotypes and b) the case cohort-defined haplotypes.

Developing the hierarchical model for differentiating between healthy controls and melanoma cases using the CTLA-4 based personal genotype profiles from haplotypes. Weka package (v. 3-6-6) implementation of J48 pruned tree algorithm was used to construct optimal model recognizing the controls from cases using personal difference vectors $\vec{\delta}_j$. Tenfold cross-validation was used and characterized the model quality by confusion matrices and ROC parameters.

RESULTS:

Fig. 4 shows decomposition of the **g** graphs for healthy controls (Fig. 2a) and melanoma cases (Fig. 2b) into component cycles ***rrp***$_i$, representing the haplotypes derived from individual genotyped profiles, containing CT60 (rs3087243), AG49 (rs231775), CT318 (rs5742909), JO27 (rs11571297), JO30 (rs7565213) and JO31 (rs11571302) SNPs.

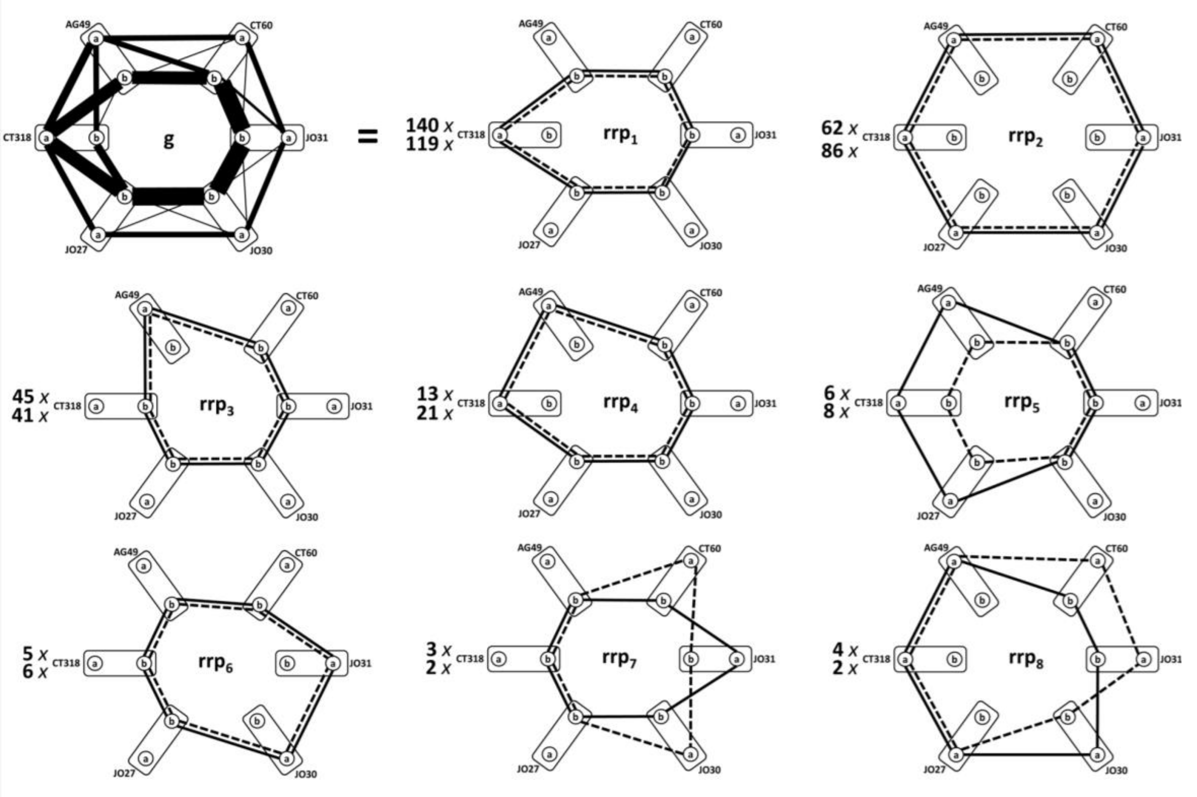

Figure 4: Decomposition of study graphs **g** (picture represents both cases and control subcohorts) into ***rrp***'s 1-8. Case rrp's are shown by solid, control by dashed edges. Coefficients show the multiplicities of respective ***rrp***'s in the **g**-decompositions (top=case graph, bottom=control graph). Symbols as in Fig.1.

Decomposing the 6-partite graph G constructed with explicit 3 allele states resulted in 20 ***RRP***. We then computed a 20-component vector of distances $\vec{\boldsymbol{\delta}}_i$ for every personal CTLA-4 genotype relationship pattern from all 20 RRPs.

Results for study graph g

In both cohorts, the respective **g** graphs were decomposed into 8 cycles ***rrp****ᵢ* (**i**=1…8). Interestingly (and importantly) the three haplotype graphs with the largest frequency were identical for control and case cohorts. Table 1 shows that our **g**-based graph algorithm also identified the same dominant haplotypes and comparable frequencies of occurrence as the statistical algorithm in (PROC HAPLOTYPE). SAS 9.1 (SAS Institute Inc., Cary, NC, USA).

Table 1 CTLA-4 most frequent haplotypes

AG49	CT60	CT318	JO27	JO30	JO31	Chromosomes (ref JOTM)		
						Frequency	Standard Error	Frequency this work
A	A	C	C	A	T	46.99	2.089	45.7 %
G	G	C	T	G	G	29.34	1.91	23.0
A	G	T	T	G	G	9.77	1.24	10.2
A	G	C	T	G	G	6.49	1.031	6.0
A	G	C	C	A	T	2.81	0.69	2.5

A unique feature of this approach in comparison to the analysis of differences in haplotype frequencies that were tested in our previous paper is that we can quantitatively characterize the difference of the individual genotype profile from "averaged" CTLA-4 haplotype profiles. Fig. 3 demonstrates the meaning of the differences. In this example, patient's **P55** CTLA-4 genotype profile captured into **ppr(55)** matches the composition of the graph representation of haplotype ***rrp₃*** in just three edges, thus the $\boldsymbol{\delta_3(55)}$ is 3. In the second example, CTLA-4 genotype profile of the same patient is compared to ***C₂*** haplotype. Here no edges in **ppr(55)** coincide with those of ***rrp₂***, thus the $\boldsymbol{\delta_2(55)}$ is 6. This is the example of maximal difference between any haplotype subgraph ***rrp**$_i$* and individual CTLA-4 genotype profile **prp**(***j***) that can be found in **g**.

Fig. 5 explains the main finding of this paper. Top level of CTLA-4 genotype profile-based differentiation

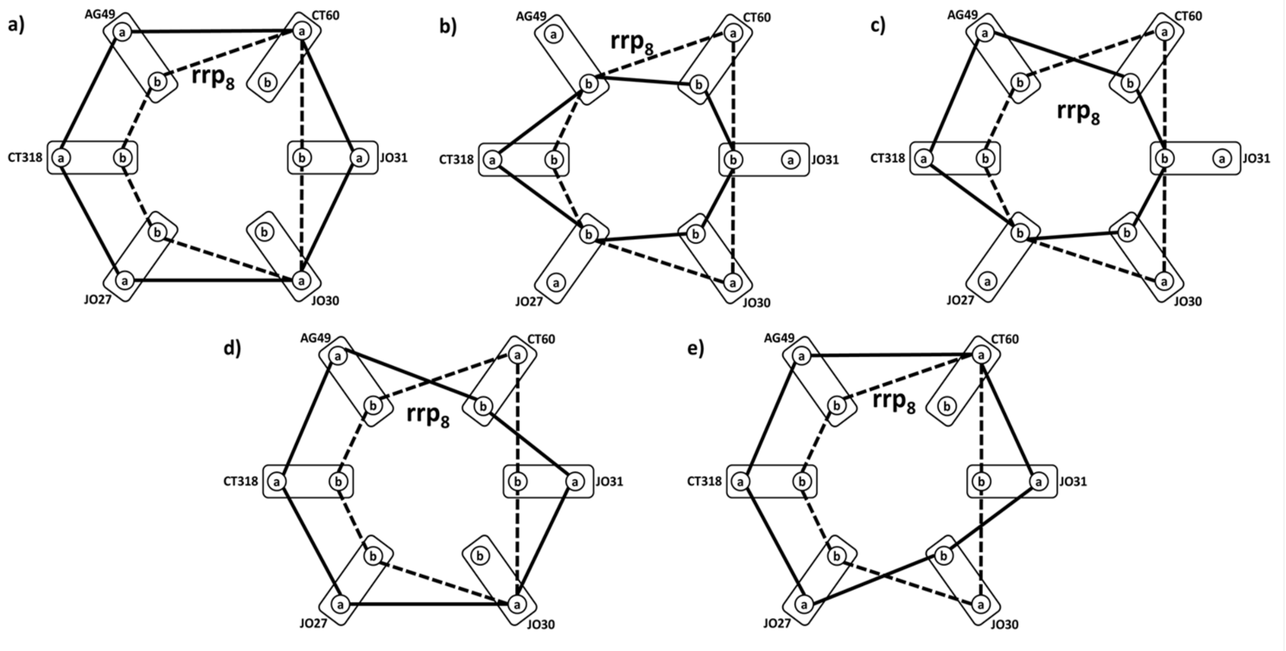

Figure 5. Case-control discrimination by "missing" CTLA-4 genotype reference profile***rrp**$_8$* (dashed lines in all figures). In schemes **a**) – **e**) are shown by solid lines five **prp** CTLA-4 genotype profiles, found exclusively for 219 (77%) patients identified from the complete case cohort by condition that their prp have maximal possible distance from the ***rrp**$_8$*. Symbols as in Fig.1.

between cases and controls is related to SNP pattern ***rrp**$_8$* = (bbabab) for (CT318-AG49-CT60-JO30-JO27) cycle (see Fig. 4 and 5). 77% of melanoma cases (219 patients) are recognized from healthy controls by the ABSENCE of the ***rrp**$_8$* = (bbabab) allele pattern for (CT318-AG49-CT60-JO30-JO27) SNP cycle. By surveying all 219 CTLA-4 individual genotype profiles for patients with $\boldsymbol{\delta_8(i) = 6}$ it was found that all have one of the five co-occurring patterns, shown by solid line cycles in Fig. 5a-e. By overlaying the ***rrp**$_8$* = (bbabab) case-control differentiating pattern (dashed line cycles) over these actual case-specific genotype profiles it is shown that the ***rrp**$_8$* pattern does not share any relationship with these 5 melanoma-characteristic CTLA-4 SNP co-occurrence patterns, indicating the possibility of disease risk identification not by presence, but actually absence of specific genotype profile.

Graph mathematics opens the previously overlooked half of the marker identification "Universe" - allowing us to study invariants (such as our personalized differences of CTLA-4 genotype profiles from the haplotype reference) and identifying multiple SNP relationship patterns that share certain properties (simultaneous presence or absence of a specific combination of parameters).

Results for study graph G

Because it was known that there are no characteristic simple CTLA-4 genotype patterns that would differentiate healthy controls from melanoma cases, we instead looked for differences in distances from the all possible ***RRP*$_1$-*RRP*$_{20}$** pairs that would maximize the separation of the two sub-cohorts. The motivation for this approach is as follows: 1. The pattern-based genotype data transformation captures more details of inter-subject differences in genetic status of CTLA-4 than can be captured by any conventional analytical approach. This information enhancement can be further increased by explicitly considering the actual allele statuses. as discussed above, the identification of the clinically relevant context relationship CTLA-4 genotype pattern is

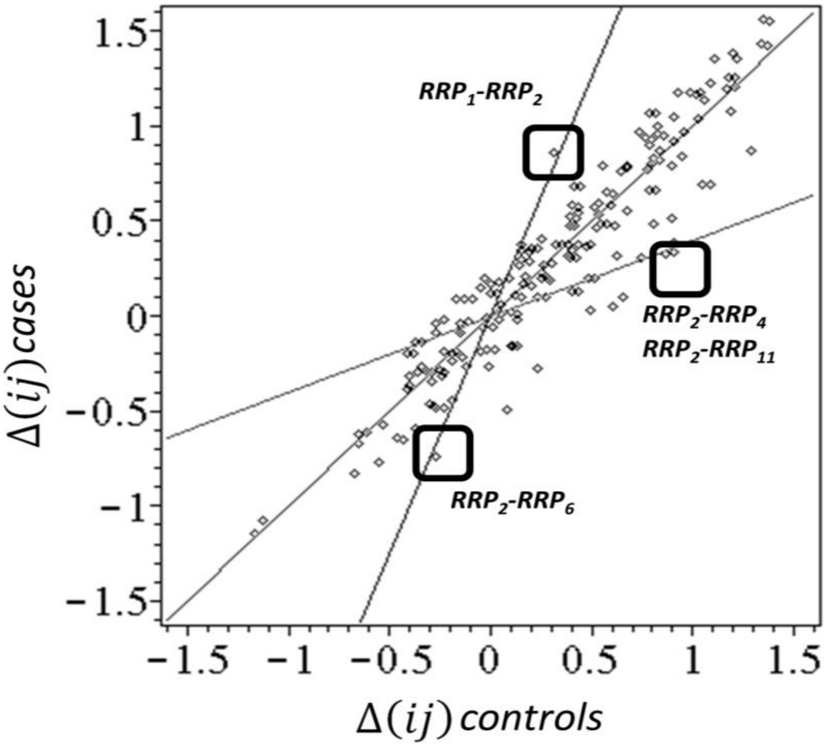

Figure 6. Selection of maximally discriminating combination of distances from all ***RRP***'s. Points are defined by the $[\Delta(ij_c), \Delta(ij_p)\,]$ coordinates (see text) computed by averaging the distance differences separately in case and control sub-cohorts for all 190 possible ***RRP*** pairs. In the neighborhood of diagonal line $\Delta(ij_c) = \Delta(ij_p)$ are non-discriminatory combinations. The two lines are used to identify the combinations, with maximal case – control and control-case bias in ***PRP*-*RRP*** distances. The optimal selection is shown by boxes.

obtained by looking for a higher frequency of patients or controls with smaller distances from selected ***RRP's***, relative to others.

An element of the $\vec{\boldsymbol{\delta}}_{\boldsymbol{i}}$ vector characterizes the distance of the personal CTLA-4 genotype pattern from reference, but does not include directionality and distances of the personal CTLA-4 genotype pattern from other reference patterns. To include that information into processed data, we therefore computed a complete set of 190 pairwise distance differences $\vec{\boldsymbol{\delta}}_{\boldsymbol{i}} - \vec{\boldsymbol{\delta}}_{\boldsymbol{j}}$, with ***i*** and ***j*** going through all 20 elements of the CTLA-4 differences from the four maximally case-control biased reference patterns ***RRP$_i$-RRP$_j$*** identified in Fig.6. These differences include directionality of the closeness of the personal genotype to one of the reference genotype patterns: $\Delta(ij) = \delta_{RRPi}(k) - \delta_{RRPj}(k)$ can be positive or negative. Assume that $\delta_{RRPi}(k) = -7, \delta_{RRPj}(k) = -3$. Then $\Delta(ij) = -7 - (-3) = -7 + 3 = -4 < 0$. Thus, $\Delta(ij) < 0$ indicates that a personal CTLA-4 genotype profile is closer to $\boldsymbol{RRP_j}$, while $\Delta(ij) > 0$ indicates that personal CTLA-4 genotype profile is closer to $\boldsymbol{RRP_i}$ and $\Delta(ij) = 0$ means that the personal CTLA-4 genotype profile has the same number of differences when compared either to reference profile $\boldsymbol{RRP_i}$ or $\boldsymbol{RRP_j}$. We computed the $\Delta(ij)$ using distances from all 190 possible ***RRP***'s pairs, separately for cases and controls and averaged them for each sub-cohort, obtaining case mean ($\Delta(ij_p)$) and control mean $\Delta(ij_c)$ for each ***RRP***'s pair. Plotting these case and cohort averages against each other in the two-dimensional scheme allows direct identification of the reference CTLA-4 genotype pattern combinations that separate maximally the two sub-cohorts. For uniformly or randomly distributed CTLA-4 genotype pattern positions we obtain $\Delta(ij_c) = \Delta(ij_p)$ seen in the 2D plot as the diagonal y=x line. The combinations with maximal $\Delta(ij_c) > \Delta(ij_p)$ or $\Delta(ij_c) < \Delta(ij_p)$, which are the desired clinically characteristic contexts will be in the 2D plot maximally distant from the diagonal. Fig. 6 shows the resulting 2D plot with the extreme combinations of the references indicated. The region of $\Delta(ij)$smaller than 0.5 is not considered, as there the subject’s CTLA-4 genotype patterns are on average equally distant from both reference pairs.

Fig. 7 shows histogram of patients with observed valued of $\Delta(ij)$. The patient or control distribution in the CTLA-4 genotype pattern space is not uniform or normal. We see clear heterogeneity: In both groups, there are three main patient subgroups. One, common for cases and controls has CTLA-4 genotypes equally different from all reference CTLA-4 allele relationships (central peak). Then there are two groups with their individual CTLA-4 genotype patterns significantly closer to one than to the other reference genotype relationship network.

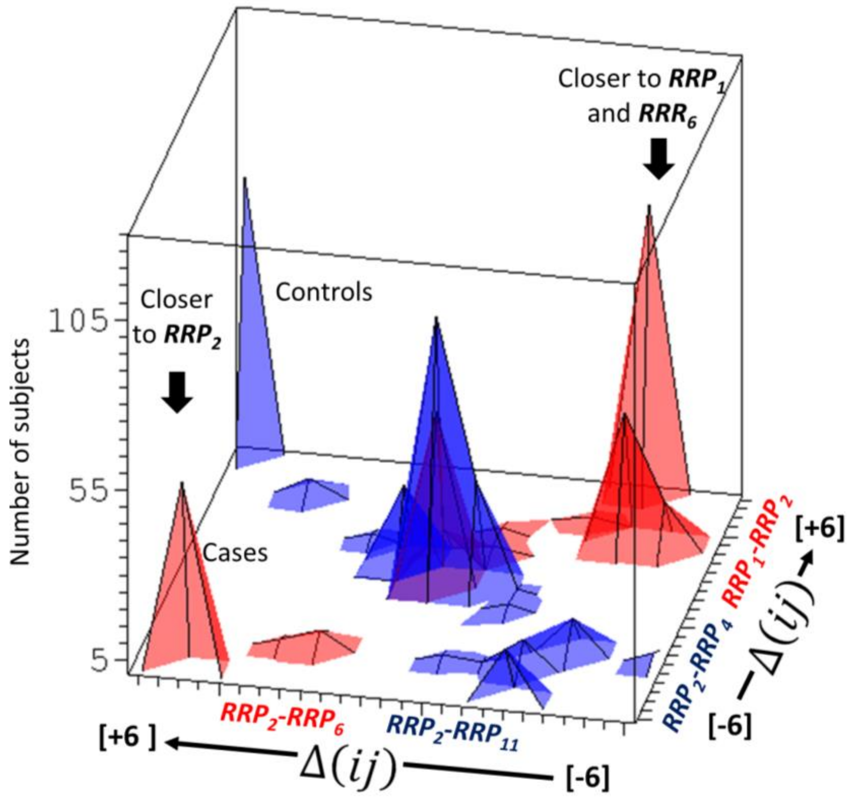

Figure 7. Histograms showing heterogeneity of distributions of individuals shown in the CTLA-4 genotype landscape, defined by the inter-personal differences in prp’s for the five most discriminating RRP combinations. Two selected combination of $\boldsymbol{\delta_j(i) - \delta_k(i)}$ distance differences are plotted on x and y axes, on the z axis are numbers of subjects having a given combination of the distance differences. Blue-controls. red-cases.

Fig. 8 shows the actual composition of these reference CTLA-4 genotype patterns for cases and controls. For controls, the dominant reference CTLA-4 genotype pattern is all major allele combination (***RRP₂***) while for cases, ***RRP₁*** dominates, where majority of studied CTLA-4 polymorphisms are in the heterozygous state. This heterogeneity might be utilized in focused prospective study of patients within the three subgroups identified: One being characterized by the minimally genetically affected CTLA-4, another having majority of CTLA-4 polymorphisms with heterozygous state and the third with mixed CTLA-4 genotype relationship patterns, equally different from the two extremes. It is clear that, contrary to melanoma patients, the healthy biosystem of controls can accommodate the CTLA-4 genetic variation where a majority of studied polymorphisms relate to the minor allele states that are identified as reference contexts for two groups with CTLA-4 genotype patterns different from "normal" ***RRP*$_2$**.

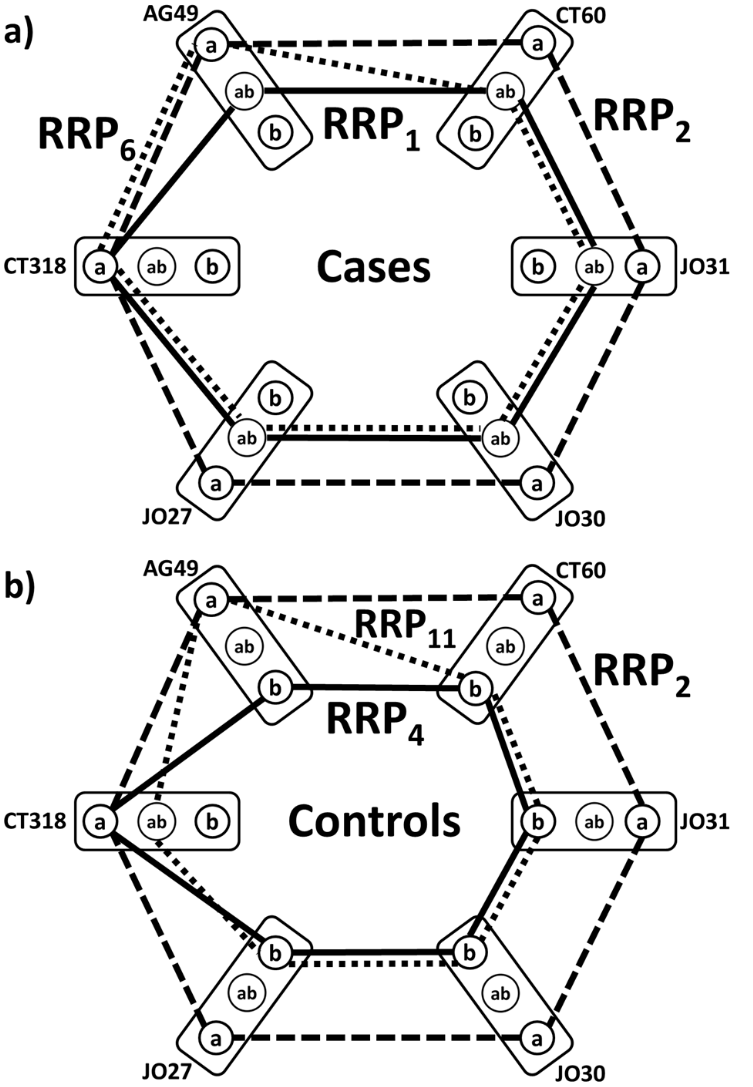

Figure 8. Comparison of CTLA-4 genotype relationship profiles of five most case-control discriminating RRP's. RRP2 (dashed edges) is shown in both panels for reference. Symbols as in Fig.1

Differentiation of the CTLA-4 genotype contexts between the long and short surviving sub cohorts of melanoma patients.

Out of the 386 melanoma cases, we had 282 with survival data. Characterization of the possible differences between the long- and short-surviving patients now requires a different analysis strategy. First, we tested the choice of CTLA-4 genotype reference relationship patterns. After separate testing of results from NPS analysis of melanoma case CTLA-4 genotype relationship profiles, we found the simplest and statistically most significant results were obtained when the ***RRP*$_1$*-RRP*$_{20}$** resulting from the analysis of combined case/control cohort were used. That makes sense in light of previous standard statistical analysis indicating no significant differences in the actual CTLA-4 genotype patterns. A larger cohort combined from cases and controls provided better coverage of the possible reference CTLA-4 genotype relationship patterns. Moreover, the results were significant when the case sub-cohort was analyzed separately, and overlapped with the patterns identified using differences of distances from the combined analysis.

For the analysis of CTLA-4 genotype relationship pattern differences between the survival categories, we used a different strategy to make sure that what was found was indeed significant. We defined an overall survival threshold and separated the cohort into patients who lived longer or shorter than the selected threshold. We then ran the complete analysis described below and compared the statistical significance and performed logistic regression models to recognize the survival categories from the $\Delta(ij)$. We systematically iterated through a threshold of 800 days to a threshold 1900 days, and found the optimal threshold at 1820 days (5 years). This threshold separated the cohort into balanced sub cohorts of 145 shorter and 137 longer surviving patients.

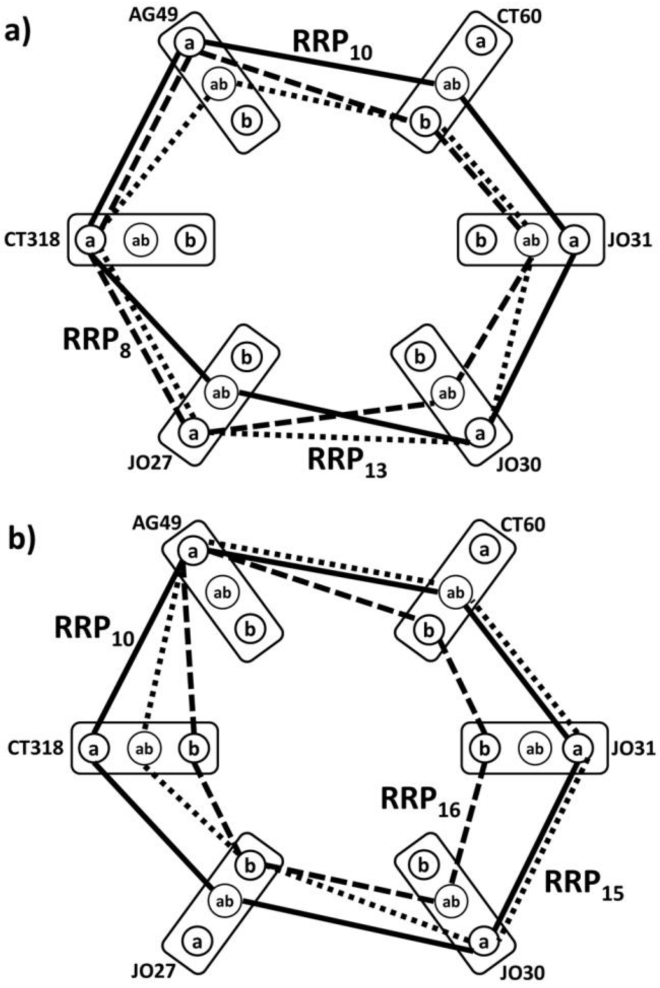

Figure 9. Comparison of ***RRP***'s, differentiating two survival groups ($\lesseqgtr 5\,years$) – see text. ***RRP$_{10}$*** is shown (solid edges) in both panels for reference.

We then computed the $\Delta(ij)$ separately for both these survival-defined sub cohorts and tested the distributions of the results for all 190 CTLA-4 reference relationship pattern pairs. Out of the 190, only 4 combinations resulted in the statistically significantly different means of these distributions (see p-value Table II). Here, ***RRP$_{10}$*** reference pattern is the common context in all these CTLA-4 genotype relationship patterns, which are significantly biased between the longer and shorter surviving melanoma patients. Similar interpretation is now possible for the localization of the typical CTLA-4 genotype relationship patterns for these outcome different patients: For example, shorter surviving patients have typically positive $\Delta(ij)$ for ***RRP$_8$*-*RRP$_{10}$***, so they are closer to ***RRP$_8$***, meaning that their CTLA-4 genotype tend to converge to 4 minor, one heterozygous and one major allele (see Figure 9). Similar interpretation is possible for remaining significantly different genotype pattern pairs: RRP$_{10}$-RRP$_{13}$ pairing have typically zero $\Delta(ij)$ for shorter surviving patients, and positive for longer survivals, indicating that ***RRP$_{10}$*** pattern with 4 major and 2 heterozygous alleles provides better functioning CTLA-4.

Table II. *p*-values for difference in mean difference distributions for distances of ***prp***'s from ***RRP***'s pairs, differentiating two survival groups ($\lesseqgtr 5\ years$).

RRP combination	*p*-values
RRP$_8$*-*RRP$_{10}$	0.022
RRP$_{10}$*-*RRP$_{13}$	0.024
RRP$_{10}$*-*RRP$_{16}$	0.025
RRP$_{10}$*-*RRP$_{15}$	0.043

DISCUSSION

Using a novel approach to the analysis of SNP results for the CTLA4 gene, we have hypothesized that recognition of melanoma risk genotype profile requires an added dimension of analysis. This second step in the analysis progression moves from analyzing the means and variance of independent SNPs to analyzing the distributions of differences of individual CTLA-4 genotype profiles in the studied cohorts, in reference to normative reference profiles.

We argue that the observed haplotypes are the proper reference for this purpose, and that we need to use them to account for interpersonal variability in CTLA-4 genotype profiles. The approach generates 6-partite graphical depictions which are based upon algorithms that identified the same haplotypes and their frequencies in established statistical procedures. Importantly, this algorithm has shown that the haplotypes are not markers by themselves, but rather that their averaged constructs, identifying common co-occurrences of CTLA-4 SNPs in case and control cohorts are useful. Having both personal CTLA-4 genotype profiles and the normative reference co-occurring CTLA-4 SNP haplotype patterns represented by the K-partite graphs has two main advantages:

A. It determines from the data used to construct the **g** and through the decomposition algorithm we developed from the statistical conditions used in general characterization of haplotype the actual TOTAL number of haplotypes in the cohort (8 in both our cohorts). Considering that the theoretical number of haplotypes for **g** is 64, this is an important data reduction outcome of this approach. We know from other applications that in cases where deconstructed 6-partite graphs are close to random distributions of the conditional probabilities, the number of components needed to fully deconstruct the model increases significantly. Thus, small number of components in the **g** deconstruction implies the commonality/regularity in the CTLA-4 genotype profile composition and frequency in our study population. This is in agreement with the previous study results.

B. The component graphs $\boldsymbol{rrp}_i$ are data-driven, information-rich references for exact quantitative computation of the $\boldsymbol{\delta(j,i)}$ descriptors, which are tools enabling to change the focus of the analysis from means and averages to where we need it (i.e. towards differentiating features). Importantly, the $\boldsymbol{rrp}_i$'s are NOT just mathematical constructs, but have well-defined genomic meaning, being haplotypes. This facilitates clinically relevant interpretation of the results in general and the individual (personalized) disease related markers in particular. Results validate the hypothesis.

Another important aspect of this work is its “translation” of the main molecular result of this paper to design of tools and algorithms that use the relationship-patterns between genotyped CTLA-4 variants to enable differential outcome analysis. Our approach allows to show, that in the relationship patterns picture of the individual CTLA-4 genotype, differential outcome can be caused by a “majority rule”, understood as a larger than critical deviation from an ideal, reference haplotype relationship pattern. Thus, same impact can be observed for different combinations of the personal CTLA-4 variants, which is clearly quantitatively captured in our NPS (relationship) based analysis, but causes problems in conventional approaches. This sharing of a certain level of differences from a reference normative pattern is very specific in relation to the kinds of patterns that share a particular property. This linkage of several heterogeneous patterns to one "functional" patient's individual difference is that other side of clinical data understanding, which can be brought to the plate using this approach. Without the pattern-based approach, we would never recognize the relationship between those patterns and could not ask what is unique about them. More importantly, this common distance of personal CTLA-4 genotype profiles from reference genotype patterns may group patients that would conventionally not have been thought to be potentially grouped for interpretation. By definition, they have different patterns of CTLA-4 parameters, the conventional approach will tell you that THESE ARE DIFFERENT, so that you would never ask whether they have something in common. Our approach – by contrast – has brought together patients with five different CTLA-4 genotypes so that we are forced to ask what these patterns have in common. We can now clearly identify that the ABSENCE of ONE common pattern from these five different, is what distinguishes cases and controls.

The combination of SNP’s, shared by all individual patients’ profiles that satisfy the condition of having the largest distance from one specific haplotype allows then discussing the mechanistic details (why it is just this combination of major and minor allele in the 6 genotyped loci, which separates cases from healthy controls).

Key issue is that detailed characterization of the genotype by explicit consideration of the actual state of each SNP provides the significant clustering (for one survival group) or difference/distance (for the other survival group) of the prp's relatively to perhaps interesting and interpretable CTLA-4 genotype relationship patterns.

We therefore argue that for disease outcome differentiation, the analysis tools need to evaluate networks of relations among the CTLA-4 polymorphisms and hetero- and homozygosity of each SNP.

This prepares the stage for the categorization of disease outcome via analysis of thermodynamic changes in the in CTLA-4 SNPs discovered by entromics, and quantified by the differences in matrices that quantify the energy weights associated with the various genotype profiles in individual patient entromic coherence networks.

Pattern based polymorphism relationship analysis revealed that in healthy controls, the context in which the CTLA-4 and its genetic variants operates is compatible with the genotype with relationship pattern with "consensus" alleles in all six sites. While we see some relationship pattern differences between long and short overall survival groups, these are not independently recognized, we need to know who is long and who is short surviving. To obtain really independent, statistically significant, prediction of the long or short survival we thus need to go one additional step: consider that there is coherence pattern between assayed regions of CTLA-4 gene and that this coherence pattern is affected by the polymorphisms in the personal genotype in exactly computable way. This is provided by the entromic characterization of the individual polymorphism patterns.

We can characterize from entromics physical principles, describing the origin and functional significance of these coherence changes, that there will be two extreme cases of CTLA-4 genotypes – one (A) with actual individual polymorphisms, that would in sum of their contributions to the CTLA-4 coherence patter change STRENGHTEN these function-related CTAL-4 relationships and the other case (B), where these coherence changes would sum to weakening the function-related CTLA-4 relationships. We therefore can plausibly hypothesize, that for patients with personal CTLA-4 genotype (A) we can expect longer overall survival, as their modifications are actually somehow improving the CTLA-4 performance, while in the case (B) we have deterioration of the CTLA-4 function because of negative alteration of the communication of the different gene parts with each other or with the rest of the genome, so we expect the shorter survival. These hypotheses are formulated solely a priori; known survival is not used to state them in any way. Also the finding patients, who are (A) or (B) is derived solely from their experimental genotype, just processed in a way that computes from the known alleles those summary changes of the coherence in incorporation entropies for CTLA-4. The main contributions of entromic characterization of the individual CTLA-4 polymorphic relationship patterns to understanding of their clinical relevance are: We see that there is clustering of patients in three areas of the changes – one group is close to (A), the other is close to (B) and the third group is in between those two. Thus, differences in CTLA-4 genotype related changes in the incorporation entropy coherence = communication between different parts of the CTLA-4 gene revealed the heterogeneity of the CTLA-4 genotypes. Importantly, some of the allele combinations are “compensatory”. Final finding is, that when we do independent Kaplan-Meier analysis of survivals in the three groups, we get what was hypothesized: (A) has significantly longer overall survival compared to group with entromic characteristics of their CTLA-4 SNP-pattern (B) and patients with the intermediate entromic CTLA-4 variation profiles (C) have survival probability curve in between (A) and (B).

Supplement:

The algorithm for identifying the presence and relative frequency of <u>all haplotypes</u> in a genotyped cohort (see Fig. S1):

1. In the complete **g**, identify the 6 SNP alleles represented by the cycle **C$_{\mathbf{max}}$** with the highest sum of all 6 edge weights.
2. In **C$_{\mathbf{max}}$,** find the edge with the minimal weight **W$_{\mathbf{min}}$** out of all 6.
3. From **g**, remove the cycle ***rrp$_i$,*** which has all edges with weight **W$_{\mathbf{min}}$**. Vertices, connected by edge in ***rrp$_i$*** define the molecular composition of this haplotype, **W$_{\mathbf{min}}$** is the frequency of this haplotype.

$$g^{i-1} = g^{i} - \sum_{x=1}^{i} w_{min,i} \times rrp_i$$

4. The reminder is again subjected to steps 1-3 above, until all edges from **g** are removed. This generates the series of all haplotypes ***rrp$_i$*** (***i***=1..***k***) in the cohort, together with their frequencies (**W$_{\mathbf{min}}$(i)**, ***i***=1..***k***).

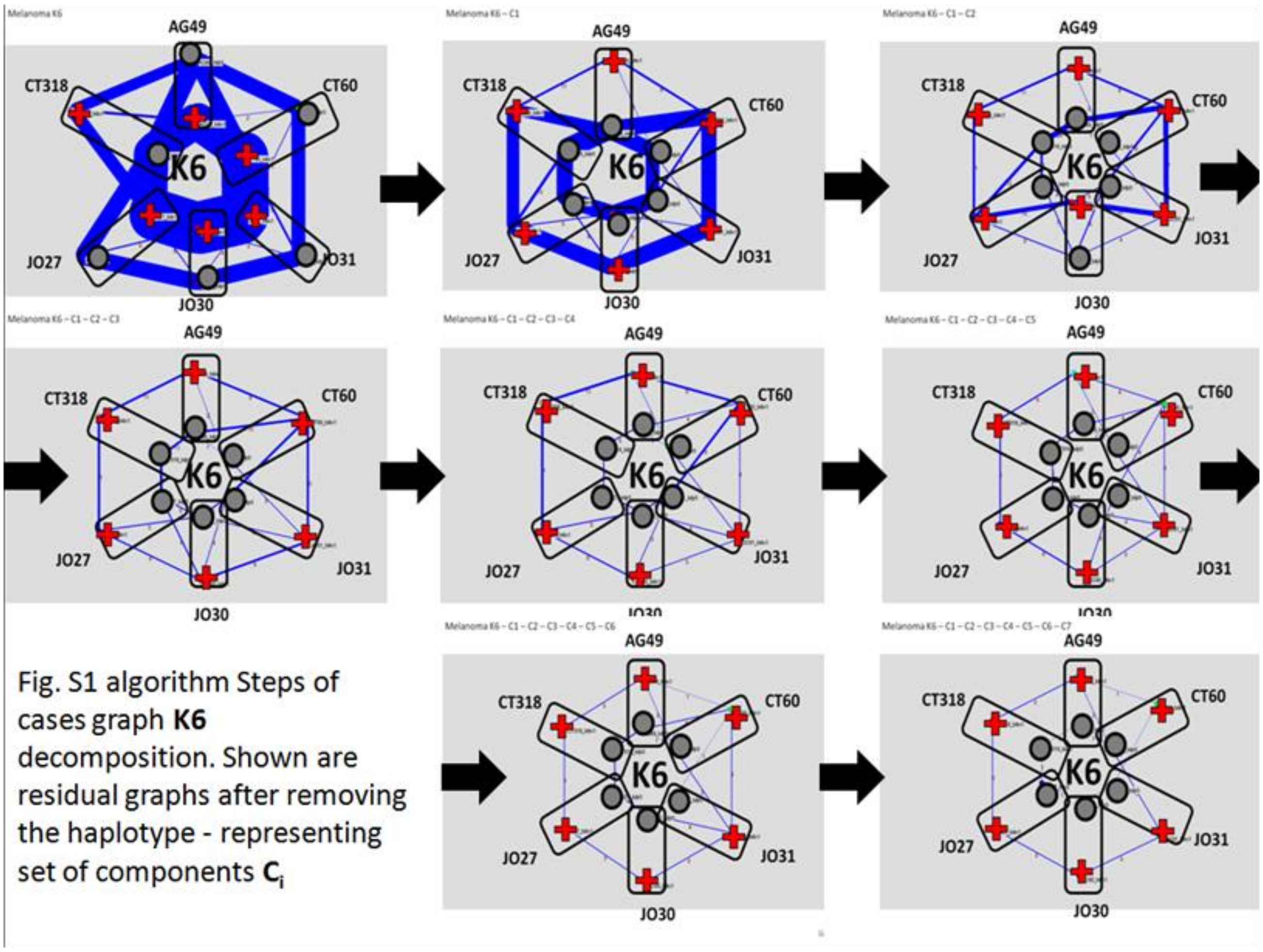
Melanoma K6
Melanoma K6 – C1
Melanoma K6 – C1 – C2
Melanoma K6 – C1 – C2 – C3
Melanoma K6 – C1 – C2 – C3 – C4
Melanoma K6 – C1 – C2 – C3 – C4 – C5
Melanoma K6 – C1 – C2 – C3 – C4 – C5 – C6
Melanoma K6 – C1 – C2 – C3 – C4 – C5 – C6 – C7
AG49
CT318
CT60
K6
JO27
JO31
JO30

Fig. S1 algorithm Steps of cases graph **K6** decomposition. Shown are residual graphs after removing the haplotype - representing set of components **C$_i$**

REFERENCES

1. Kirkwood, J.M., Strawderman, M.H., Ernstoff, M.S., Smith, T.J., Borden, E.C. and Blum, R.H. (1996) Interferon alfa-2b adjuvant therapy of high-risk resected cutaneous melanoma: the Eastern Cooperative Oncology Group Trial EST 1684. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, **14**, 7-17.
2. Kirkwood, J.M., Ibrahim, J.G., Sosman, J.A., Sondak, V.K., Agarwala, S.S., Ernstoff, M.S. and Rao, U. (2001) High-dose interferon alfa-2b significantly prolongs relapse-free and overall survival compared with the GM2-KLH/QS-21 vaccine in patients with resected stage IIB-III melanoma: results of intergroup trial E1694/S9512/C509801. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, **19**, 2370-2380.
3. Kirkwood, J.M., Richards, T., Zarour, H.M., Sosman, J., Ernstoff, M., Whiteside, T.L., Ibrahim, J., Blum, R., Wieand, S. and Mascari, R. (2002) Immunomodulatory effects of high-dose and low-dose interferon alpha2b in patients with high-risk resected melanoma: the E2690 laboratory corollary of intergroup adjuvant trial E1690. *Cancer*, **95**, 1101-1112.
4. Yurkovetsky, Z.R., Kirkwood, J.M., Edington, H.D., Marrangoni, A.M., Velikokhatnaya, L., Winans, M.T., Gorelik, E. and Lokshin, A.E. (2007) Multiplex analysis of serum cytokines in melanoma patients treated with interferon-alpha2b. *Clinical cancer research : an official journal of the American Association for Cancer Research*, **13**, 2422-2428.
5. Gogas, H., Dafni, U., Koon, H., Spyropoulou-Vlachou, M., Metaxas, Y., Buchbinder, E., Pectasides, E., Tsoutsos, D., Polyzos, A., Stratigos, A. *et al.* (2010) Evaluation of six CTLA-4 polymorphisms in high-risk melanoma patients receiving adjuvant interferon therapy in the He13A/98 multicenter trial. *Journal of translational medicine*, **8**, 108.
6. Dariavach, P., Mattei, M.G., Golstein, P. and Lefranc, M.P. (1988) Human Ig superfamily CTLA-4 gene: chromosomal localization and identity of protein sequence between murine and human CTLA-4 cytoplasmic domains. *European journal of immunology*, **18**, 1901-1905.
7. Gough, S.C., Walker, L.S. and Sansom, D.M. (2005) CTLA4 gene polymorphism and autoimmunity. *Immunological reviews*, **204**, 102-115.
8. Kristiansen, O.P., Larsen, Z.M. and Pociot, F. (2000) CTLA-4 in autoimmune diseases--a general susceptibility gene to autoimmunity? *Genes and immunity*, **1**, 170-184.
9. Thompson, C.B. and Allison, J.P. (1997) The emerging role of CTLA-4 as an immune attenuator. *Immunity*, **7**, 445-450.
10. Ueda, H., Howson, J.M., Esposito, L., Heward, J., Snook, H., Chamberlain, G., Rainbow, D.B., Hunter, K.M., Smith, A.N., Di Genova, G. *et al.* (2003) Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature*, **423**, 506-511.
11. Carr, B.I., Lu, S.N. and Pancoska, P. (2011) Small hepatocellular carcinoma in Chinese patients. *Hepato-gastroenterology*, **58**, 1334-1342.
12. Pancoska, P., Carr, B.I. and Branch, R.A. (2010) Network-based analysis of survival for unresectable hepatocellular carcinoma. *Seminars in oncology*, **37**, 170-181.
13. Pancoska, P., De Giorgio, M., Fagiuoli, S. and Carr, B.I. (2011) Small HCCs identified by screening. *Digestive diseases and sciences*, **56**, 3078-3085.
14. Banks, D. and Constatntine, G.M. (1998) Metric models for randomg graphs. *Journal of Classification*, **15**, 199-223.
15. Hamming, R. (1950) Error detecting and error correcting codes. *Bell System Technical Journal*, **29**, 147-160.