ORAL DISEASES

Oral Diseases (2013) 19, 721–727 doi:10.1111/odi.12146

www.wiley.com

ORIGINAL ARTICLE

SNP-based Bayesian networks can predict oral mucositis risk in autologous stem cell transplant recipients

ST Sonis[1,2], JH Antin[3], MW Tedaldi[3], G Alterovitz[1,4]

[1]*Inform Genomics, Boston, MA, USA;* [2]*Brigham and Women's Hospital, Boston, MA, USA;* [3]*Dana-Farber Cancer Institute, Boston, MA, USA;* [4]*Children's Medical Center, Boston, MA, USA*

OBJECTIVE: Approximately 40% of patients receiving conditioning chemotherapy prior to autologous hematopoietic stem cell transplants (aHSCT) develop severe oral mucositis (SOM). Aside from disabling pain, ulcerative lesions associated with SOM predispose to poor health and economic outcomes. Our objective was to develop a probabilistic graphical model in which a cluster of single-nucleotide polymorphisms (SNPs) derived from salivary DNA could be used as a tool to predict SOM risk.

METHODS: Salivary DNA was extracted from 153 HSCT patients and applied to Illumina BeadChips. Using sequential data analysis, we filtered extraneous SNPs, selected loci, and identified a predictive SNP network for OM risk. We then tested the predictive validity of the network using SNP array outputs from an independent HSCT cohort.

RESULTS: We identified an 82-SNP Bayesian network (BN) that was related to SOM risk with a 10-fold cross-validation accuracy of 99.3% and an area under the ROC curve of 99.7%. Using samples from a small independent patient cohort (n = 16), we demonstrated the network's predictive validity with an accuracy of 81.2% in the absence of any false positives.

CONCLUSIONS: Our results suggest that SNP-based BN developed from saliva-sourced DNA can predict SOM risk in patients prior to aHSCT.

Oral Diseases (2013) **19**, 725–731

Keywords: oral mucositis; risk prediction; SNP; stem cell transplant

Introduction

Oral mucositis (OM) is a frequent, painful, and expensive complication of many conditioning regimens for hematopoietic stem cell transplants (HSCTs). The physiologic ramifications of OM are considerable and result in significant morbidity and, in some cases, mortality. Besides its symptomatic toll, OM is linked to a number of negative health and economic outcomes, including increased analgesic and antibiotic use, febrile days, need for parenteral nutrition, length of hospital stay, unplanned office and emergency room visits, and total charges (Sonis *et al*, 2001). In granulocytopenic patients, OM is strongly aligned with an increased risk of bacteremia and sepsis (Reuscher *et al*, 1998).

While OM risk among patients receiving conditioning regimens which include total body irradiation (TBI) exceeds 90%, the risk drops to 30%–50% for individuals being treated with the much more commonly used chemotherapy-based protocols. As the health and economic pressures for personalized medicine increase, the prospective identification of patients at risk of treatment-related toxicities such as OM will become highly desirable as mechanistically based interventions become available.

There is currently no way to accurately predict the risk of OM in an individual patient. Aside from concomitant TBI and aggressive chemotherapy regimens, proposed indicators such as age, sex, body mass, or comorbidities consistently fail to consistently identify patients at risk (Sonis *et al*, 2004).

There is now substantial evidence, however, that genetic factors play a dominant role in determining the likelihood of a patient developing regimen-related toxicities including OM. The original data supporting this supposition were derived from studies concerned with drug metabolism. Not surprisingly, patients genetically incapable of producing enzymes needed for the metabolism of chemotherapeutic drugs such as methotrexate or 5-fluorouracil had disproportionate toxicity outcomes (Schwab *et al*, 2008). From a practical standpoint, however, mutations associated with enzyme deficiencies are rare and therefore are incapable of accounting for the number of patients who develop severe side effects (Ezzeldin and Diasio, 2008). Contrastingly, the complex pathobiology that drives the development of tissue-based side effects such as OM offers the potential for a broad range of genetic risk determinants

Correspondence: Stephen Sonis, DMD, DMSc, Brigham and Women's Hospital, 75 Francis Street, Boston, MA 02115, USA. Tel: 617 525 6864, Fax: 617 525 6899, E-mail: ssonis@partners.org
Received 3 April 2013; accepted 13 May 2013

(Sonis *et al*, 2009). And as recently suggested by the results of the ENCODE Project (Niu and Jiang, 2013), risk is most likely associated with networks of genes functioning together. Thus, any strategy to define genetic risk must account for cooperative interactions among genes.

We hypothesized that a probabilistic graphical model could be identified, using a Bayesian network (BN) of single-nucleotide polymorphisms (SNPs), to determine OM risk. Our objective was to demonstrate proof of concept of this approach in patients receiving common conditioning regimens in preparation for autologous HSCT (aHSCT).

Patients, materials, and methods

Patient identification and sample collection

The study was approved by the Institutional Review Board of the Harvard/Partners Cancer Center. Medical records of patients who had undergone aHSCT for the treatment of Hodgkin, or non-Hodgkin, lymphoma, or multiple myeloma at the Dana-Farber Cancer Institute between January 1, 2006 and June 30, 2010, were reviewed by trained study staff. A total of 144 individuals who did not develop severe OM (mucositis-negative) and 72 participants who did develop severe OM (mucositis-positive) were identified. Severe OM (SOM) was defined as two consecutive days of WHO grade 3 or 4 following the administration of chemotherapy. Demographic information and survival information were collected.

Eligible patients were approached in person at clinic visits or by a letter to request study participation and obtain informed consent. Saliva samples were obtained in standard fashion using DNA Genotek collection tubes (Rogers *et al*, 2007).

DNA isolation

DNA was extracted from coded, de-identified saliva samples by expression analysis, Inc. (Durham, NC), using Qiagen spin column purification technology in which the sample was digested with proteinase K, bound to the spin column, washed, and eluted in a small volume. The resulting DNA was quantified on a Thermo Scientific NanoDrop spectrophotometer and stored at −80°C until use. Samples were qualified for qPCR using standard set primers for glyceraldehyde-3-phosphate dehydrogenase (GAPDH).

Analyses of SNP arrays

Arrays were processed by expression analysis, Inc., using an Infinium DNA analysis BeadChip (Illumina; ~1.1 million SNPs). Genome-wide genotyping was performed using the HumanOmni1-Quad chip (Illumina, San Diego, CA). The array design leverages on the linkage disequilibrium map provided by the HapMap, a directory of human genetic variants, to select markers that maximize the genetic coverage through linkage disequilibrium to the surrounding regions. The chip provided comprehensive genomic coverage of 94% of the white population (CEU) at $r^2 > 0.8$, a standard limit of high linkage disequilibrium. The median spacing between markers was 1.25 kb (mean = 2.63 kb). Additional SNP content covered non-synonymous SNPs, the MHC region, and Y-chromosome SNPs.

Data preprocessing

Before starting data analysis, we used standard data preprocessing operations. SNPs that were monomorphic, did not satisfy Hardy–Weinberg equilibrium (Boccia *et al*, 2010) in the controls, or had ≥10% missing data were excluded because allele frequency differences between cases and controls due to systematic ancestry differences could cause spurious associations. Population stratification was analyzed using Eigenstrat (Price *et al*, 2006), a program that enables explicit detection and correction of population stratification on a genome-wide scale using principal components analysis to explicitly model ancestry differences between cases and controls. The resulting correction was specific to a candidate marker's variation in frequency across ancestral populations, minimizing false associations while maximizing power to detect true associations.

Analytical overview

The analysis was run on an Amazon Web Services Elastic Compute Cloud Quadruple Extra Large Instance running Microsoft Windows Server 2008 on 64-bit architecture with 26 ECU distributed across eight cores and 68.4 GB of RAM (Figures 1 and 2).

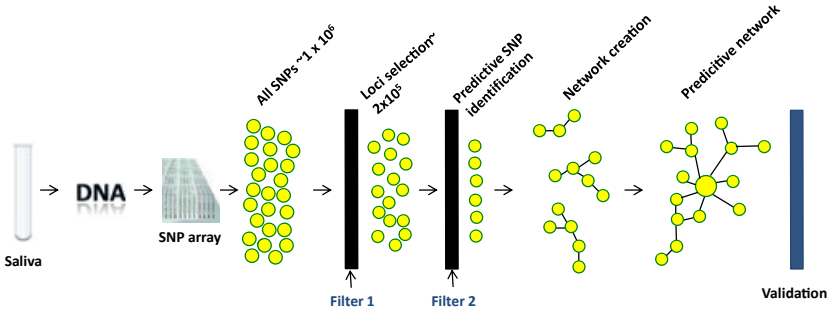

Figure 1 SNP selection and analysis flow. Flow diagram delineating the steps in the generation of the optimal BN for OM risk prediction and model validation. From the SNP array outputs from 153 (102 SOM-negative and 51 SOM-positive), we first applied a heuristic that identified the top 200 000 SNPs as ranked by the chi-square statistic which demonstrated that a difference existed between patients who had developed SOM and those who had not. A second filter, again using the heuristic, further identified 4000 SNPs that served as the basis for building a BN based on gene features. We thereby identified an 82-SNP BN associated with OM risk. Validation was performed in two steps. We technically validated the model by re-running six samples from the training set. We then tested the predictive validity of the BN, by testing SNP array outputs from an additional independent group of 16 patients (8 SOM-negative, 8 SOM-positive)

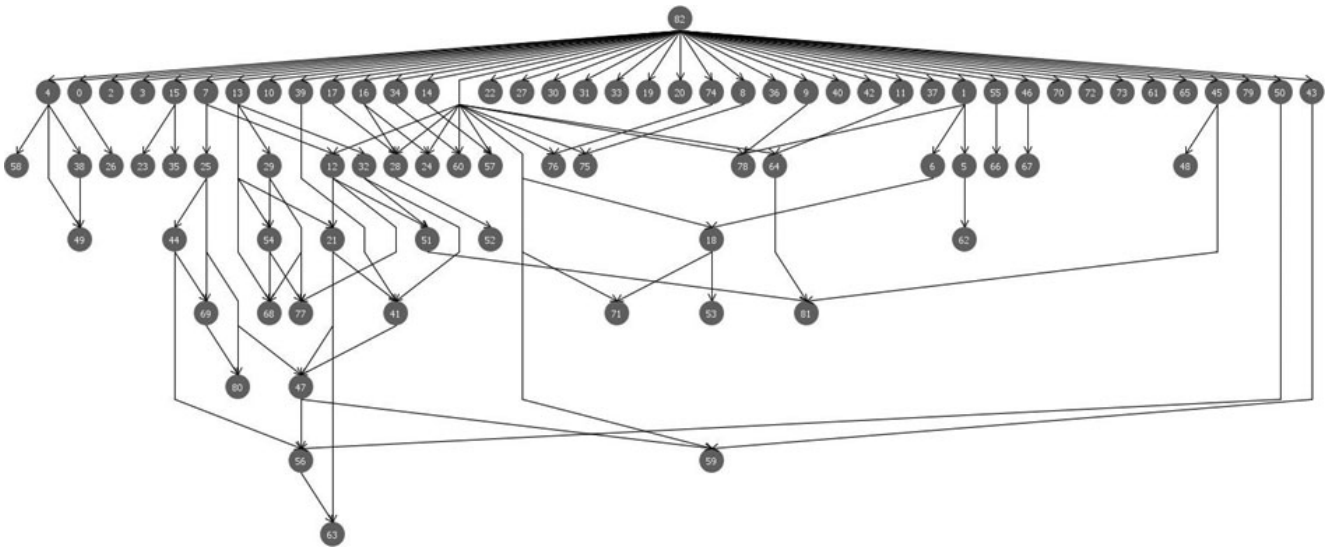

Figure 2 Structure of the Bayesian network as it was applied to the validation set using only training data. The network demonstrates the associations between the 82 individual SNPs. The 'hub' at the top of the figure represents the phenotype (SOM)

To increase the efficiency and accuracy of the discovery of a predictive SNP network, we performed a sequential analysis of the data that allowed us to filter extraneous SNPs and select SNP loci, identify a predictive BN, and then evaluate the networks potential clinical and biological validity.

SNP selection

As an exhaustive search of the potential risk value of every expressed SNP was both impractical and of low yield, we applied a filter in which we limited the number of loci to be considered in building a univariate BN model. From all of the genetic loci (SNPs) generated from the array outputs, those most likely to be the most important were selected using area under the receiver operating characteristic curves (ROC)-based Bayesian optimization. This technique shortened the computation time by filtering SNPs that were highly unlikely to contribute predictive value in a BN. We accomplished this by applying a heuristic before ROC-based Bayesian optimization that limited the number of loci considered in the building of a univariate BN model to the top 200 000 SNPs as ranked by the chi-squared statistic, demonstrating that a difference existed between the SNP distributions between OM-negative and OM-positive patients. We also used the heuristic to limit the number of SNPs (4000) considered in building a BN based on gene features. The number of SNPs considered by this heuristic was conservative, given the amount of computational power available. It was unlikely that limiting the scope of analysis to these candidates would produce a significant change in the final outcome.

We first built a univariate BN model (Alterovitz *et al*, 2007*)* by ranking one SNP at a time by its mucositis risk prediction ROC value. For each SNP in the data set, we built a BN model that included a node for mucositis, a node for the individual locus, and a directed arc from phenotype (mucositis) to the locus. Tenfold cross-validation was used to evaluate the network, with accuracy judged by the area under the ROC. After all SNPs had been evaluated in this way, the SNPs in networks with higher ROC statistics (ROC closer to one) were given better ranks than loci with lower ROC statistics. We then built a BN model using the top N gene features (i.e., SNPs) by ranking. We selected the model with the highest ROC. The predictive accuracy was judged by the ROC associated with 10-fold cross-validation for each network.

Bayesian network analysis: optimal classifier identification

Having determined the particular SNPs to be used in the final network, we identified the BN that best differentiated SOM risk from no SOM risk (Sebastiani *et al*, 2005; Alterovitz *et al*, 2010) by building a large number of networks across a range of plausible network structures and then comparing them via area under the ROC for predictive accuracy and posterior probability. To build the networks, we used the K2 algorithm, which assumed an ordering of the SNPs and an optimal network structure search. To find the optimal structure, different orderings were considered. Tenfold cross-validation was used to evaluate each network structure.

To evaluate the clinical applicability of the BN associated with SOM created from the training sample, we evaluated its ability to 'predict' risk in an independent group of patients (validation set). We reasoned that such information would also be valuable for sample size determination for a larger prospective study. Outputs from SNP arrays derived from 16 patients who had received aHSCT in the same time period as the subjects who were used to develop the exploratory model were used. Clinically, eight patients were SOM-negative. Special attention was paid to formatting the validation set, so that it matched the format mandated by the data set used for learning. The results were recorded for each prediction, and accuracy statistics were generated.

Gene annotation and biological validity

To learn more about the SNPs in the model, we mapped each SNP to a chromosome and the nearest gene using

the Ensembl database. The gene symbols were checked for enrichments of both gene ontology (GO) terms and ontological categories (Ge *et al*, 2008; Pines and Everett, 2008; Da Huang *et al*, 2009).

Results

Patient characteristics

One hundred and fifty-three patients were included in the generation of the training set. Forty-two percent were female. The mean age of study participants was 57 years. All but three were white. Eighty-two subjects received high-dose melphalan for the treatment of multiple myeloma prior to aHSCT. The remaining subjects were treated for lymphoma (Hodgkin or non-Hodgkin) with conditioning regimens consisting of carmustine BCNU, cyclophosphamide, and etoposide. One hundred and two subjects did not develop severe mucositis during their transplant course.

An additional 16 patients were assigned to the validation set. The demographics, diagnosis, and treatments were proportionally the same as in the training set cohort. The cohort consisted of eight subjects who were SOM-negative and eight who were SOM-positive.

Array results and network prediction characteristics

The data set contained two batches that used the 1 140 419 SNPs available on the Illumina HumanOmni1-Quad v1.0 chip. The first batch, used in training, contained 153 samples. The second batch (the validation set) contained 16 completely independent samples. In addition, six technical replicates of samples in the first batch were run as a quality control measure. After the data was converted into an appropriate format, ROC-based Bayesian Optimization was used to significantly pare down the number of loci using the method described above.

The N that maximized the ROC statistic was 82, and accordingly, the top 82 SNP features were chosen for use in the final model. This SNP features list predicted SOM with 99.3% accuracy and 99.7% ROC in 10-fold cross-validation (Table 1).

With an SNP feature list selected, the optimal network to characterize the SNP interactions was identified. The optimal BN was determined to be a complex, multilevel structure (Figure 2).

To assess the predictive efficacy of the final network, it was applied to six technical replicates and the validation set. The technical replicates were derived from samples within the training set. All six of the technical repeats were predicted correctly. The validation set consisted of the 16 independent subjects. The 82-loci BN achieved 76.6% ROC and 81.2% accuracy (Table 2) when applied to this cohort. No false positives were seen: all eight SOM-negative subjects were correctly identified. Five of eight SOM-positive subjects were correctly identified; three of eight SOM-positive subjects were misclassified as SOM-negative. A determination of possible unique clinical features of the misclassified subjects was unrealistic given the small number of patients in this category.

Of the 82 SNPs that defined the predictive network, almost half ($n = 40$) could be mapped to 29 gene symbols. Of these, the BEND2, MAGEA11, DIP2B, PKNOX2, and RALGPS1 were associated with multiple SNPs. Interestingly, the genes we noted were associated with the innate immune response (HMGB3) and the zinc finger family, which has been seen previously to be associated with mucositis.

We further assessed the relationship of the associated genes with established GO terms, although we did not enrich the gene symbol set for GO terms or functional categories. Our findings are seen in Table 3.

Table 1 Cross-validation characteristics and classification accuracies on initial set (training)

	Cross validation
Number of patients	153
Final number of SNPs	82
Accuracy	99.3%
ROC	99.7%

The training set consisted of array outputs from 153 subjects (102 SOM-negative, 51 SOM-positive) and resulted in the generation of an 82-SNP BN that was associated with SOM with an accuracy and area under the ROC curve of greater than 99%.

Discussion

The elimination of TBI as a standard component of the most common aHSCT conditioning regimens has resulted in a reduction in SOM incidence from over 90% (Spielberger *et al*, 2004) to about 40% (McCann *et al*, 2009). Consequently, it can no longer be presumed at the overwhelming number of aHSCT patients will develop SOM associated with their conditioning regimen.

While reduced SOM risk is clearly an asset, it presents potential challenges to caregivers, clinical trialists, and patients. Not being able to differentiate between at-risk patients from those at no risk, typically, results in management strategies that presume that SOM will develop in all individuals or which require symptom onset before the commencement of treatment. This approach is problematic when effective therapy requires the initiation of treatment prior to the appearance of symptoms as it means, in the case of aHSCT, that 60% of patients would be unnecessarily treated. This problem is likely to be compounded as new mechanistically based therapies which require

Table 2 Characteristics and classification accuracies for independent set (validation)

	Validation set
Number of patients	16
Final number of SNPs	82
Accuracy	81.2%
ROC	76.6%

To assess the validity of the SNP BN and to assist with planning the scope of a subsequent prospective study, it was tested against SNP array outputs from a 16 subject independent cohort, half of whom were SOM-negative and half SOM-positive. The small sample size precluded any conclusions regarding clinically 'unique' features of the three subjects who were misclassified. Nonetheless, the predictive results are considered to be 'fair' by conventional definition. The lack of false positives represents an important advantage over other gene-based methods.

Table 3 Genes and GO terms associated with SNP-based network model

GO term	*Associated gene symbols*
Cognition	COL18A1, PDE1C, ATP6V0A4
Chromosome	HIST1H2AC, HMGB3, IKZF1
DNA binding	PKNOX2, HIST1H2AC, HMGB3, IKZF1, ATF1
Ion binding	COL18A1, AGBL3, LIMA1, IKZF1, EGFL6
Cytoskeleton	PKNOX2, LIMA1, ARHGAP26
Cell adhesion	COL18A1, CNTNAP4, EGFL6
Ion transport	GRIK4, HTR3A, ATP6V0A4
Cell junction	LIMA1, GRIK4, HTR3A, ARHGAP26
Cation binding	COL18A1, AGBL3, LIMA1, IKZF1, EGFL6
Plasma membrane	LIMA1, RALGPS1, GRIK4, HTR3A, NRG1, ATP6V0A4, ARHGAP26
Zinc ion binding	COL18A1, AGBL3, LIMA1, IKZF1
Metal ion binding	COL18A1, AGBL3, LIMA1, IKZF1, EGFL6
Sensory perception	COL18A1, PDE1C, ATP6V0A4
Biological adhesion	COL18A1, CNTNAP4, EGFL6
Extracellular space	COL18A1, EGFL6, NRG1
Extracellular region	COL18A1, EGFL6, LYZL4, NRG1
Plasma membrane part	LIMA1, GRIK4, HTR3A, NRG1, ATP6V0A4, ARHGAP26
Integral to membrane	CNTNAP4, GRIK4, HTR3A, NRG1, DPP6, ATP6V0A4
Intrinsic to membrane	CNTNAP4, GRIK4, HTR3A, NRG1, DPP6, ATP6V0A4
Extracellular region part	COL18A1, EGFL6, NRG1
Neurological system process	COL18A1, PDE1C, GRIK4, HTR3A, ATP6V0A4
Neurological system process	COL18A1, PDE1C, GRIK4, HTR3A, ATP6V0A4
Regulation of transcription	PKNOX2, IKZF1, NRG1, ATF1
Transition metal ion binding	COL18A1, AGBL3, LIMA1, IKZF1
Sequence-specific DNA binding	PKNOX2, IKZF1, ATF1
Transcription factor activity	PKNOX2, IKZF1, ATF1
Non-membrane-bounded organelle	PKNOX2, HIST1H2AC, LIMA1, HMGB3, IKZF1, ARHGAP26
Transcription regulator activity	PKNOX2, IKZF1, NRG1, ATF1
Regulation of RNA metabolic process	PKNOX2, IKZF1, ATF1
Regulation of transcription, DNA-dependent	PKNOX2, IKZF1, ATF1
Intracellular non-membrane-bounded organelle	PKNOX2, HIST1H2AC, LIMA1, HMGB3, IKZF1, ARHGAP2

Of the 82 SNPs identified in the BN, 40 could be mapped to 29 gene symbols. Of these, the BEND2, MAGEA11, DIP2B, PKNOX2, and RALGPS1 were associated with multiple SNPs. Interestingly, we noted genes associated with the innate immune response (HMGB3) and the zinc finger family that have been seen previously to be associated with OM.

prophylactic administration become available. Likewise, the efficiency of OM trials is diminished when only 4 of 10 subjects studied are likely to develop SOM. As a result, clinical trials in aHSCT populations require excess accrual to accumulate adequate study numbers of OM-susceptible subjects. There is currently no way to effectively predict SOM risk in the HSCT population, but continuously emerging data suggest that genes associated with OM pathogenesis play a dominant role.

The objective of this proof of concept study was to determine whether a SNP-based BN could form the basis for an accurate and clinically meaningful model to predict SOM risk among patients receiving conditioning regimens in anticipation of aHSCT. We explored the predictive relationship between SNPs and risk in a broader and less restricted way than is typically carried out using the two most common approaches—the candidate gene (SNP) model which relies on a prospective identification of the gene or SNP of interest followed by a search for its presence or the genome-wide association study model (GWAS) in which a sequential analysis is performed between 1 SNP at a time and the phenotype using traditional association tests(Jewell, 2003). We believed that these tactics risked not capturing the multigenic nature of complex traits that characterize SOM. Whereas the single SNP approach suffers from its inability to identify associations because of the simultaneous presence of multiple variants in different DNA regions (Hoh and Ott, 2003), GWAS often identified too many associations (false positives) as a result of dependencies between SNPs in contiguous regions (linkage disequilibrium) and the dependency between SNPs on different chromosomes (Gabriel *et al*, 2002). Multivariate statistical models circumvent these limitations by examining the overall dependency structure between genotypes, phenotype, and environmental/clinical variables.

One of the strengths of the Bayesian approach to model selection is the elimination of the multiple comparisons problem (Hartley *et al*, 2012). The classical method for discovering genotype–phenotype associations involves fixing a threshold to identify statistically significant dependencies. The threshold is known as the significance level and represents the probability of accepting a spurious dependency (Type I error). When several significance tests are performed, the overall probability of the Type I error increases, and adjustments need to be made to ensure that the overall correct significance level is used. The Bayesian approach does not conduct significance tests, but rather ranks alternative models of associations on the basis of their posterior probability. In this way, no threshold is needed to select the most probable BN, and the problem of multiple comparisons is avoided. As this property is not sufficient to remove the potential problem of identifying spurious associations due to unobserved confounders, we assessed the false-positive rate of our model by building networks based on data with the phenotype (SOM) randomly shuffled.

Our results support the potential application of SNP-based BN as a toxicity risk prediction instrument. The 82-SNP BN model we discovered predicted SOM risk in patients receiving standard conditioning regimens prior to aHSCT with an accuracy of 99.3% and an area under the ROC curve of 99.7%. We tested the validity of the network by assessing its ability to discriminate the development of SOM *vs* no SOM in a small, independent sample of 16 patients from the same demographic as those subjects used to develop the model. Despite the small sample size, we were able to achieve an accuracy of 81.2% and an area under the ROC curve of 76.6%. Importantly, application of the BN avoided the false-posi-

tive pitfalls often noted with other analytical approaches. By objective metrics, this result would be classified as a 'fair' predictor (Hoh and Ott, 2003) and is higher than that achieved by widely used tests such as mammography tests (Pisano *et al*, 2005). Of the SNPs that made up the network, 40 mapped to 29 genes. A review of the GO terms for these genes confirms both the independence of the risk predictive aspect of some SNPs and the functional association of others. The latter is important because it reinforces the relationship between risk and pathogenesis. GO terms for cytoskeleton, zinc iron binding, regulation of transcription, and transcription factor activity have been reported in investigational studies reporting gene expression linked to mucositis development. Interestingly, SNPs connected to extracellular space were seen. This observation is consistent with the recent substantiation that changes in cell permeability are a component of mucosal response to injury and are associated with mucositis development (Chen *et al*, 2011). Although mechanistic gene expression studies are independent of SNP-based risk prediction, the consistency of SNP/gene association provides teleological evidence for the hypothetical basis that those SNPs are related to functional gene components.

We elected to use an SNP-based model, rather than a gene-based model for both practical and analytical reasons. The ease with which DNA can be obtained from patients (saliva) and its stability make it easily adaptable to the clinic. In contrast, RNA (for gene arrays) is more friable and requires a blood draw. For patients who suffer from regimen-related xerostomia, buccal swabs also could serve as a DNA source.

Our results suggest that OM risk can be determined using an SNP-based BN that addresses the concept that an interplay between SNPs produces an informative team composite. The application of the BN algorithms developed in this study has been effectively applied to regimen-related risk prediction in patients receiving standard regimens for breast cancer (Schwartzberg *et al*, 2012).

This proof-of-concept study was limited by the number of patients included in both the training and validation sets. As the predictive strength of the learned BN increases with the number of inputs, we anticipate that a larger study will produce a BN of even more utility. Likewise, we were very limited in the size of our validation set. While we were able to show that our method did not result in the high rate of false positives seen with other approaches, the overall predictive accuracy and ROC values were, no doubt, diluted by our limited sample size. We look forward to a large, multicenter validation trial designed and planned to enroll in the next year that will markedly increase the predictive robustness of the BN approach. Ultimately, the clinical application of a saliva-based risk prediction tool should enhance clinical decision making for providers, patients, and their families and afford economic advantages to payers.

Acknowledgements

This work was supported by SOBI Pharmaceuticals, Biomodels LLC, and InformGenomics.

We thank the following individuals for assistance with chart reviews: B. Luong, D. Colantino, J. Hsia, C. Kolstad, Y. Yi, T. Pogal-Sussman, A. Donnell, C. Markova, C. Kim, S. Jin, A. Zhujiang, and H. Mackler.

Author contributions

SS participated in study design, data analysis and manuscript preparation. JA participated in study design and was site principle investigator. MT participated in data collection and was responsible for saliva sample collection. GA participated in study design, data analysis and manuscript preparation.

Conflict of interest

S.S. is a founder, shareholder, and consultant of InformGenomics. G.A. is a shareholder and consultant of InformGenomics. J.A. and M.T. have no competing financial interests to declare.

References

Alterovitz G, Liu J, Afkhami E, Ramoni MF (2007). Bayesian methods for proteomics. *Proteomics* **7**: 2843–2855.

Alterovitz G, Xiang M, Hill DP *et al* (2010). Ontology engineering. *Nat Biotechnol* **28**: 128–130.

Boccia S, De Feo E, Galli P, Gianfagna F, Amore R, Ricciardi G (2010). A systematic review evaluating the methodological aspects of meta-analyses of genetic association studies in cancer research. *Eur J Epidemiol* **25**: 765–775.

Chen P, Lingen M, Sonis ST, Walsh-Reitz MM, Toback FG (2011). Role of AMP-18 in oral mucositis. *Oral Oncol* **47**: 831–839.

Da Huang W, Sherman BT, Lempicki RA (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**: 1–13.

Ezzeldin HH, Diasio RB (2008). Predicting fluorouracil toxicity: can we finally do it? *J Clin Oncol* **26**: 2080–2082.

Gabriel SB, Salomon R, Pelet A *et al* (2002). Segregation at three loci explains familial and population risk in Hirschsprung disease. *Nat Genet* **31**: 89–93.

Ge D, Zhang K, Need AC *et al* (2008). WGAViewer: software for genomic annotation of whole genome association studies. *Genome Res* **18**: 640–643.

Hartley SW, Monti S, Liu C-T, Steinberg MH, Sebastiani P (2012). Bayesian methods for multivariate modeling of pleiotropic SNP associations and genetic risk prediction. *Front Genet* **3**: 176.

Hoh J, Ott J (2003). Mathematical multi-locus approaches to localizing complex human trait genes. *Nat Rev Genet* **4**: 701–709.

Jewell NP (2003). *Statistics for epidemiology*. Chapman and Hall/CRC: Boca Raton, FL.

McCann S, Schwenkglenks M, Bacon P *et al* (2009). The Prospective Oral Mucositis Audit: relationship of severe oral mucositis with clinical and medical resource use outcomes in patients receiving high-dose melphalan or BEAM-conditioning chemotherapy and autologous SCT. *Bone Marrow Transplant* **43**: 141–147.

Niu DK, Jiang L (2013). Can ENCODE tell us how much junk DNA we carry in our genome? *Biochem Biophys Res Commun* **430**: 1340–1343.

Pines JM, Everett WW (2008). *Evidence-based emergency care: diagnostic testing and clinical decision rules (evidence-based medicine)*. Blackwell Publishing: Malden, MA.

Pisano ED, Gatsonis C, Hendrick E *et al* (2005). Digital Mammographic Imaging Screening Trial (DMIST) investigators

group diagnostic performance of digital versus film mammography for breast-cancer screening. *N Engl J Med* **353**: 1773–1783.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904–909.

Reuscher TJ, Sodeifi A, Scrivani SJ, Kaban LB, Sonis ST (1998). The impact of mucositis on alpha-hemolytic streptococcal infection in patients undergoing autologous bone marrow transplantation for hematologic malignancies. *Cancer* **82**: 2275–2281.

Rogers NL, Cole SA, Lan HC, Crossa A, Demerath EW (2007). New saliva DNA collection method compared to buccal cell collection techniques for epidemiological studies. *Am J Hum Biol* **19**: 319–326.

Schwab M, Zanger UM, Marx C *et al* (2008). Role of genetic and non-genetic factors for fluorouracil treatment-related severe toxicity: a prospective clinical trial by the German 5-FU Toxicity Study Group. *J Clin Oncol* **26**: 2131–2138.

Schwartzberg L, Sonis S, Walker M *et al* (2012). Single nucleotide polymorphism (SNP) Bayesian networks (BNs) predict risk of chemotherapy-induced side effects in patients with breast cancer receiving dose dense (DD) Doxorubicin/cyclophosphamide plus paclitaxel (AC+T). Paper presented at San Antonio Breast Cancer Symposium. December 4–8, 2012. San Antonio, TX. Abstract 3032.

Sebastiani P, Ramoni MF, Nolan V, Baldwin CT, Steinberg MH (2005). Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. *Nat Genet* **37**: 435–440.

Sonis ST, Oster G, Fuchs H *et al* (2001). Oral mucositis and the clinical and economic outcomes of hematopoietic stem-cell transplantation. *J Clin Oncol* **19**: 2201–2205.

Sonis ST, Elting LS, Keefe D *et al* (2004). Perspectives on cancer therapy-induced mucosal injury: pathogenesis, measurement, epidemiology, and consequences for patients. *Cancer* **100**(9 Suppl): 1995–2025.

Sonis S, Haddad R, Posner M *et al* (2009). Gene expression changes in peripheral blood cells provide insight into the biological mechanisms associated with regimen-related toxicities in patients being treated for head and neck cancers. *Oral Oncol* **43**: 289–300.

Spielberger R, Stiff P, Bensinger W *et al* (2004). Palifermin for oral mucositis after intensive therapy for hematologic cancers. *N Engl J Med* **351**: 2590–2598.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1. SNP-based risk prediction for oral mucositis.

Figure S1. Supplemental analysis flow diagram.