



## OPEN

# Microbiomic Signatures of Psoriasis: Feasibility and Methodology Comparison

SUBJECT AREAS:

MACHINE LEARNING

MICROBIOLOGY

COMPUTATIONAL MODELS

GENOMICS

Alexander Statnikov<sup>1,2</sup>, Alexander V. Alekseyenko<sup>1,2</sup>, Zhiguo Li<sup>1</sup>, Mikael Henaff<sup>1</sup>, Guillermo I. Perez-Perez<sup>2,3</sup>, Martin J. Blaser<sup>2,3,5</sup> & Constantin F. Aliferis<sup>1,4</sup>Received  
1 July 2013Accepted  
22 August 2013Published  
10 September 2013

Correspondence and requests for materials should be addressed to A.S. (Alexander.Statnikov@med.nyu.edu) or C.F.A. (Constantin.Aliferis@nyumc.org)

<sup>1</sup>Center for Health Informatics and Bioinformatics (CHIBI), New York University Langone Medical Center, New York, New York, <sup>2</sup>Department of Medicine, New York University School of Medicine, New York, New York, <sup>3</sup>Department of Microbiology, New York University School of Medicine, New York, New York, <sup>4</sup>Department of Pathology, New York University School of Medicine, New York, New York, <sup>5</sup>Medical Service, Department of Veterans Affairs New York Harbor Healthcare System, New York, New York.

Psoriasis is a common chronic inflammatory disease of the skin. We sought to use bacterial community abundance data to assess the feasibility of developing multivariate molecular signatures for differentiation of cutaneous psoriatic lesions, clinically unaffected contralateral skin from psoriatic patients, and similar cutaneous loci in matched healthy control subjects. Using 16S rRNA high-throughput DNA sequencing, we assayed the cutaneous microbiome for 51 such matched specimen triplets including subjects of both genders, different age groups, ethnicities and multiple body sites. None of the subjects had recently received relevant treatments or antibiotics. We found that molecular signatures for the diagnosis of psoriasis result in significant accuracy ranging from 0.75 to 0.89 AUC, depending on the classification task. We also found a significant effect of DNA sequencing and downstream analysis protocols on the accuracy of molecular signatures. Our results demonstrate that it is feasible to develop accurate molecular signatures for the diagnosis of psoriasis from microbiomic data.

Psoriasis is a chronic inflammatory disease of unknown cause. Few studies of the microbiota in psoriatic patients have used molecular methods for the detection of bacterial and fungal taxa<sup>1–3</sup>. Such studies have involved relatively small numbers of subjects, and relatively low depths of coverage. Similarly, no studies have developed multivariate molecular signatures of psoriasis from high-throughput molecular data from the skin and assessed their accuracy. By “multivariate molecular signature” or “molecular signature” of psoriasis, we mean a multivariate computational or mathematical model that can either classify or diagnose psoriasis from molecular data. Studies that have generated high-throughput molecular data from the psoriatic skin lesions focus primarily on differential gene expression<sup>4–6</sup>. Outside of the domain of psoriasis, there have been few efforts to develop and assess multivariate molecular signatures using microbiomic data for clinical tasks. Most related efforts involved classification of body sites or subject/population identification<sup>7–9</sup>.

As part of the Human Microbiome Project of the National Institutes of Health, we sought to assess feasibility and compare methodologies for developing molecular signatures of psoriasis using microbiomic data from the skin. Because the cutaneous microbiota is complex<sup>10,11</sup>, and its composition is site-specific<sup>12,13</sup>, we matched lesional skin with unaffected contralateral skin from the same subject. Likewise, the samples from demographically matched healthy controls were collected at standardized sites where psoriasis lesions commonly appear. The experimental specimen matching and sample collection protocols were designed to measure predictive microbial markers that are common across body-sites regardless of the inter-site variability. Each site may have additional specific markers which may not be detected by the current design. We used this set of specimens to obtain high-throughput sequencing of the 16S rRNA gene for both the V1–V3 and the V3–V5 loci<sup>14</sup>. Using resulting microbiomic data, we then developed and evaluated multivariate molecular signatures of psoriasis aimed at the separation of cutaneous psoriatic lesions, the clinically uninvolved contralateral skin, and the similar skin loci in matched healthy control subjects. The availability of two different datasets, based on the V1–V3 and the V3–V5 16S rRNA gene loci, enabled us to study the effect of the DNA sequencing protocol on the classification accuracy of molecular signatures. Similarly, we experimented with 37 protocols for downstream analysis by using various methods for selecting features/taxa (taxonomic features) for inclusion in the molecular signatures.



This approach allowed us to study the effects of downstream analysis protocols on the classification accuracy of molecular signatures.

## Results

**Microbiomic signatures can diagnose psoriasis with high accuracy.** Using the high-throughput 16S rRNA gene survey data from the V3–V5 locus, we developed and evaluated molecular signatures of psoriasis for four binary classification tasks aimed at separating psoriasis lesions (PL: psoriatic plaque), psoriasis normal (PN: contralateral area of clinically uninvolved skin), and healthy control (CC) samples: (i) PN vs. CC, (ii) PL vs. CC, (iii) PL vs. PN, and (iv) CC vs. PL and PN.

The resulting classification accuracy (AUC) and the number of taxa participating in molecular signatures derived from the data from the V3–V5 locus are shown in **Table 1, Panel A**. Using very few taxa (2–4) it is possible to separate classes of samples with high and statistically significant classification accuracy. Of the four classification tasks, separation of CC vs. PL and PN yields the highest classification accuracy (AUC = 0.894), while separation of PN vs. PL has lower (AUC = 0.754) but still statistically significant classification accuracy (**Table 1, Panel A, Columns 1 and 2**). The number of taxa in the constructed molecular signatures is provided in **Table 1, Panel A, Columns 3 and 4**. Detailed information on (a) the features/taxa selected by GLL in all data as well as (b) the 20 most frequently selected features/taxa by GLL in the training sets during cross-validation is provided in **Supplementary File 1**. The file reports individual AUC's of the features as well as p-values from Fisher's Z-test; both AUC and p-values were computed on all data samples. In addition, the file reports frequency of selection for the 20 most frequently selected features/taxa (the maximum frequency is 500 because this is the number of training sets in which GLL was applied). **Supplementary File 3** provides a list of the 50 most univariately discriminative taxonomic features for each classification task along with their individual AUC's and mean relative abundances per group. Features in this list that have also been selected by GLL in all data are highlighted with yellow.

**DNA sequencing protocol affects the accuracy of microbiomic signatures.** We also investigated the effect of using a survey of microbiota from a different region (V1–V3) of the 16S rRNA gene on classification accuracy of molecular signatures by repeating the same experiment described above. We focused on the same four binary classification tasks (PN vs. CC, PL vs. CC, PL vs. PN, and CC vs. PL and PN) and followed a nested cross-validation design to select features/taxa and build molecular signatures.

The resulting classification accuracy (AUC) and the number of taxa participating in the molecular signatures are shown in **Table 1, Panel B**. Detailed information on features/taxa selected by GLL is provided in **Supplementary File 2**. Information about the 50 most univariately discriminative taxonomic features for each classification task is provided in **Supplementary File 4**. Only the separation of PL from CC has statistically significant accuracy. The observation that other classification tasks do not yield statistically significant accuracies indicates that the V1–V3 locus, unlike V3–V5, captures less information about the different skin types. Although the V1–V3 based survey can still be useful for diagnostic purposes, the ability to distinguish between the unaffected skin types (normal vs. control) or the skin within the affected individuals (normal vs. lesion) has been lost compared to V3–V5. The general reason for differences in performance is because amplification of V1–V3 and V3–V5 regions provides different surveys of the variability of 16S rRNA gene in the communities.

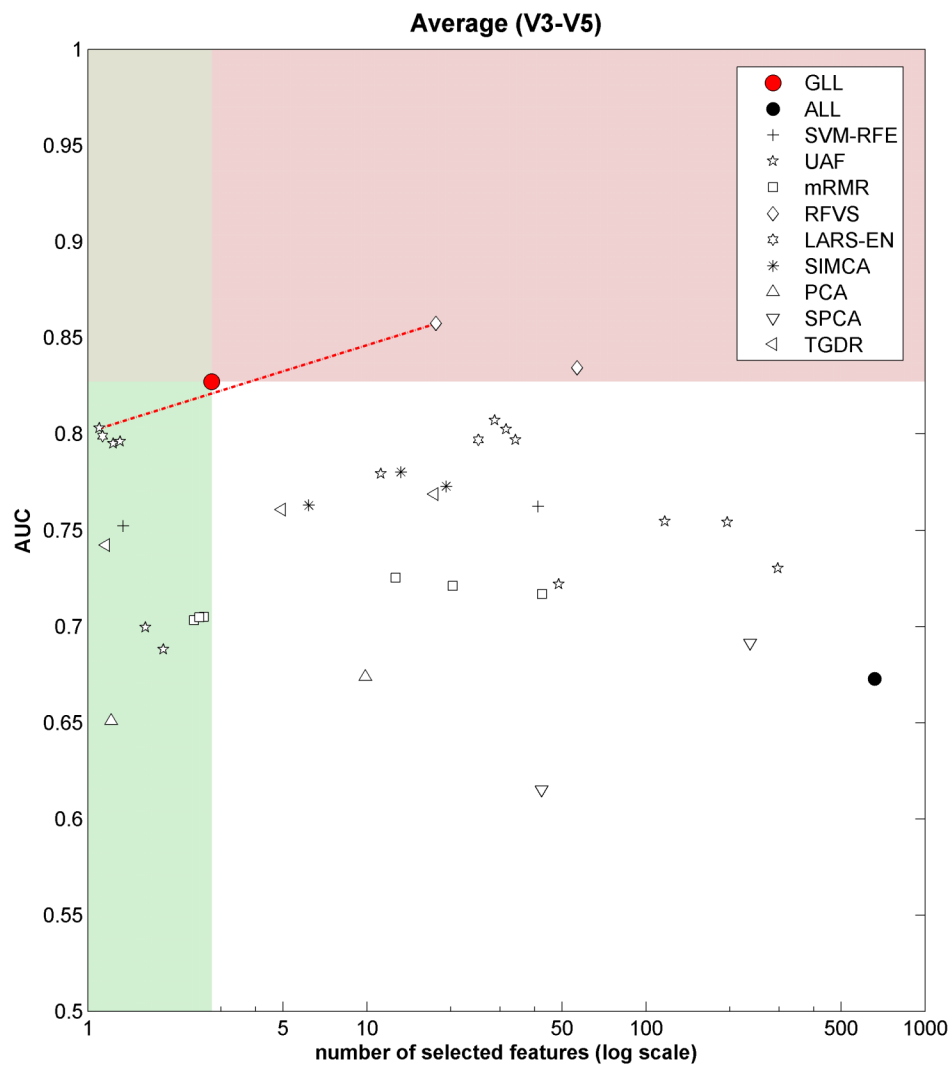
**Downstream analysis protocol affects the accuracy of microbiomic signatures.** **Table S1 in Supplementary Information** describes all downstream analysis protocols evaluated using the 16S rRNA data both from the V1–V3 and the V3–V5 loci for all four binary classification tasks (PN vs. CC, PL vs. CC, PL vs. PN, and CC vs. PL and PN). Each protocol differs from the others by using a different feature/taxa selection method.

The resulting classification accuracy (AUC) and the number of features/taxa participating in molecular signatures is shown for results averaged over all four classification tasks (**Figure 1**) and for individual results for each classification task (**Figure 2**) for data from

**Table 1 | Classification accuracy (AUC) and the number of taxa in the molecular signatures of psoriasis constructed by the GLL and SVM algorithms using the V3–V5 16S rRNA locus data (Panel A) and using the V1–V3 16S rRNA locus data (Panel B)\***

<b>Panel A: V3–V5 16S rRNA locus</b>				
	1	2	3	4
	Classification accuracy (AUC)		Number of selected taxa	
Classification task	Cross-validation estimate	Statistical significance (p-value)	From cross-validation (mean)	From the entire dataset
PN vs. CC	<b>0.854</b>	< <b>0.001</b>	<b>2.8</b>	<b>2</b>
PL vs. CC	<b>0.806</b>	<b>0.002</b>	<b>2.5</b>	<b>2</b>
PL vs. PN	<b>0.754</b>	<b>0.004</b>	<b>2.1</b>	<b>3</b>
CC vs. PL and PN	<b>0.894</b>	< <b>0.001</b>	<b>3.7</b>	<b>4</b>
<b>Panel B: V1–V3 16S rRNA locus</b>				
	1	2	3	4
	Classification accuracy (AUC)		Number of selected taxa	
Classification task	Cross-validation estimate	Statistical significance (p-value)	From cross-validation (mean)	From the entire dataset
PN vs. CC	0.405	0.985	2	1
PL vs. CC	<b>0.751</b>	< <b>0.001</b>	<b>3.8</b>	<b>4</b>
PL vs. PN	0.576	0.080	3.1	3
CC vs. PL and PN	0.482	0.618	4.2	3

\*Results for molecular signatures with statistically significant classification accuracy (at 5% alpha-level adjusted for multiple comparisons) are shown with bold font. **Column 4** reports the number of selected taxa when the GLL method was applied to the entire dataset (all samples). **Column 3** reports the number of taxa selected on average in all training sets of five-fold cross-validation (repeated 100 times with different splits into five folds). These represent results of the application of GLL to (5 × 100 = 500) training sets of the dataset with 80% of samples.



**Figure 1** | Classification accuracy (AUC) versus number of selected features/taxa for 37 feature selection methods *averaged over 4 classification tasks* (PN vs. CC, PL vs. CC, PL vs. PN, and CC vs. PL and PN) in data based on the V3–V5 16S rRNA locus. Methods from the same algorithmic family are shown with the same markers in the figure. The pink area contains methods that have nominally higher classification accuracy (AUC) than GLL. The green area contains methods that have selected fewer taxa than GLL. The red dash-dotted line indicates a Pareto frontier constructed over non-GLL methods. Methods on the Pareto frontier are such that no other non-GLL method has both higher AUC and a smaller number of selected features averaged over the four classification tasks.

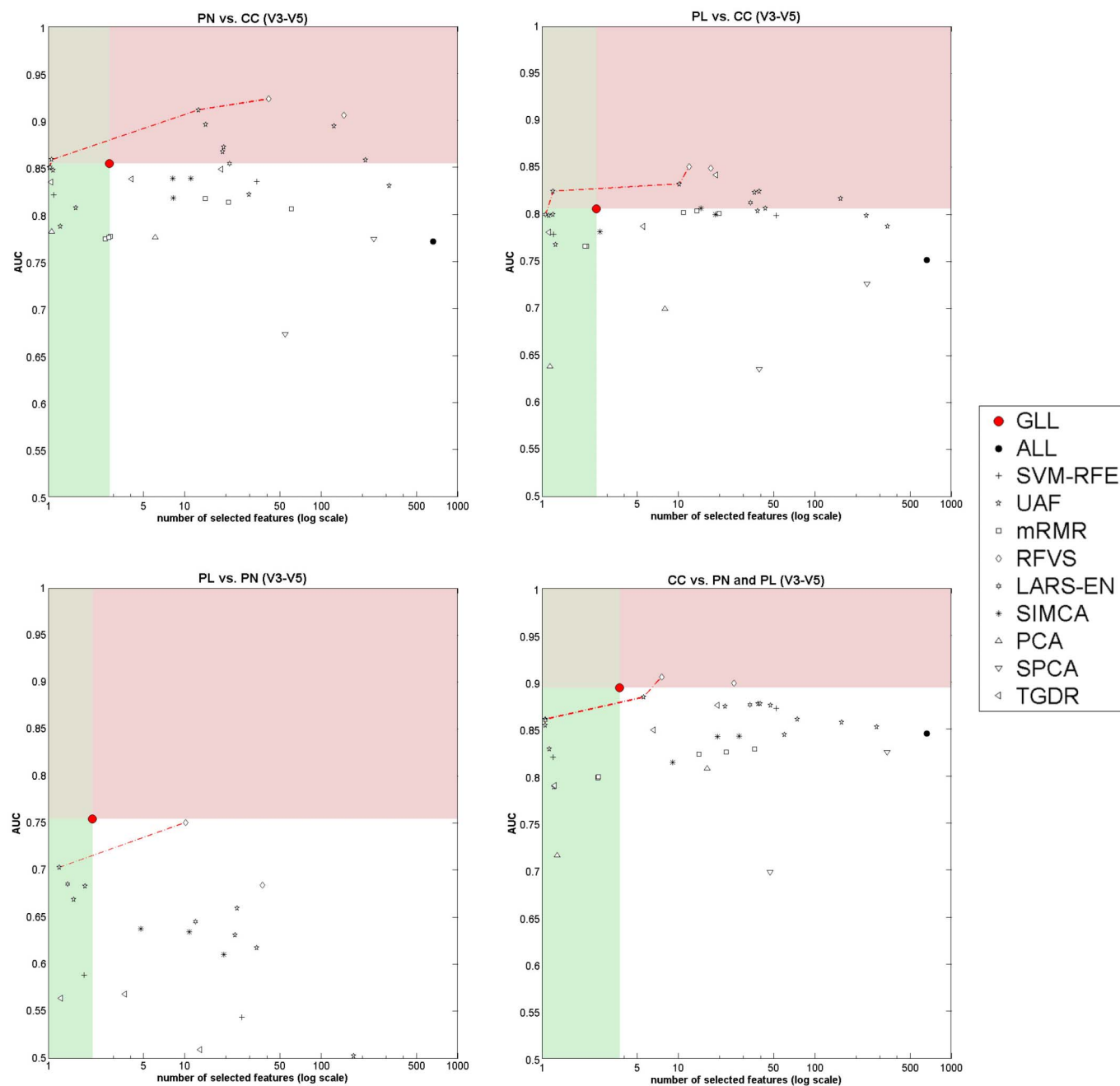
the V3–V5 locus. Detailed results and statistical comparison are given in **Table S2** (for classification accuracy) and **Table S3** (for number of selected features/taxa) in **Supplementary Information**. With respect to classification accuracy, there are 8 methods (including GLL) that yield molecular signatures with statistically significant classification accuracies for all 4 tasks (GLL, UAF-KW1, UAF-SN1, UAF-T1, UAF-T2, RFVS1, RFVS2, LARS-EN1). The GLL feature/taxa selection method outperforms 34 of the 36 comparator methods on average over four classification tasks. While there are 2 methods (RFVS1 and RFVS2) that yield slightly higher classification accuracy than GLL, the differences in accuracy are not statistically significant. With respect to the number of selected features/taxa, GLL selects fewer features/taxa than 24 of the 36 comparator methods (including RFVS1 and RFVS2) on average over the four classification tasks. There are 12 methods that select fewer features than GLL (6 of these 12 methods select significantly fewer features), however there is no single method that both selects fewer features and yields higher classification accuracy than GLL. Combined with the previous observation that there is no method that yields *significantly* higher classification accuracy than GLL regardless of the number of

features, we conclude that GLL is the optimal method for the development of molecular signatures from this data.

Detailed results and statistical comparison for data using the V1–V3 locus are given in **Table S4** (for classification accuracy) and **Table S5** (for number of selected features/taxa) in **Supplementary Information**. None of the other comparator methods yield molecular signatures with statistically significant classification accuracies for all 4 tasks. Therefore, we did not perform comparison between methods using the data from the V1–V3 locus.

## Discussion

The original analysis of the V1–V3 dataset (Alekseyenko AV, Perez-Perez GI, D’Souza A, Strober B, Gao Z, Bihan M, Li K, Methé BA, Blaser MJ): Population differentiation of the cutaneous microbiota in psoriasis, forthcoming) consisted of ecological analysis, which identified increasing intra-group (Unifrac) beta-diversity in the psoriasis specimens (diversity CC < PN < PL). The PL and PN specimens also showed decreased taxonomic richness and distribution evenness (Shannon index) compared with the CC (Control samples). These differences could be explained by the combined increased abundance



**Figure 2** | Classification accuracy (AUC) versus number of selected features/taxa for 37 feature selection methods for each of the four classification tasks (PN vs. CC, PL vs. CC, PL vs. PN, and CC vs. PL and PN) in data based on the V3–V5 16S rRNA gene locus. Methods from the same algorithmic family are shown with the same markers in the figure. The pink area contains methods that have nominally higher classification accuracy (AUC) than GLL. The green area contains methods that have selected fewer taxa than GLL. The red dash-dotted line indicates a Pareto frontier constructed over non-GLL methods. Methods on the Pareto frontier are such that no other non-GLL method has both higher AUC and a smaller number of selected features for each classification task.

of the four major skin-associated genera (*Corynebacterium*, *Propionibacterium*, *Staphylococcus*, and *Streptococcus*). The univariate analysis using the AUC determined that the three genera *Cupriavidus*, *Methylobacterium*, and *Schlegelella* statistically significantly (at 5% alpha-level) classify the groups of specimens. Interestingly, these genera also often participate in the multivariate molecular signatures of psoriasis constructed in the present study (detailed comparison is provided in Table S6 in Supplementary Information). Finally, based on partitioning around medoids clustering, the ecological study identified two distinct cutaneotypes: (1) *Proteobacteria*-associated, and (2) *Firmicutes*- and *Actinobacteria*-associated. Cutaneotype 2

was enriched in PL specimens compared to CC (OR 3.52 [1.44–8.98],  $p < 0.01$ ).

A useful theoretical property of the GLL method is that under broad distributional assumptions, it provably discovers the local pathway membership around the response variable of interest<sup>15</sup>. Specifically, this means that under sufficient assumptions of the method<sup>15</sup>, which are commonly accepted and used in causal discovery research<sup>16</sup>, GLL will output from observational data the set of direct causes, direct effects, and (optionally, depending on the version of GLL) direct causes of direct effects of the response variable (in this study, the response variable is a binary indicator of the



phenotype). This theory has been tested in both biological and non-biological fields; see for example<sup>15,17,18</sup>. The output of non-causal feature selection methods, which are comparator methods used in this study, cannot in general be interpreted causally and often results in associated variables that do not have a causal relation with the phenotype; they can merely be passengers<sup>15,17,19</sup>. Nevertheless, we consider the current work to be an initial step in the discovery process. To causally validate and interpret the findings, several considerations must be addressed: the genetic composition and immune status of the host, prior exposure and entrance routes of potential pathogens, pathogenetic mechanisms of implicated taxa, and functional assessment of the impact of microbial community composition on the observed phenotype.

In conclusion, this work demonstrates that it is possible to develop accurate molecular signatures for the diagnosis of psoriasis from microbiomic data. We have shown that the accuracy of molecular signatures depends on the DNA sequencing and downstream analysis protocols. Using the 16S rRNA data from the V3–V5 locus leads to accurate and statistically significant molecular signatures, whereas data from the V1–V3 locus carries limited diagnostic signal. With respect to the downstream analysis protocol, our results suggest that the Markov boundary-based Generalized Local Learning (GLL) method works well for this type of data, returning molecular signatures based on very few (2–4) features/taxa with statistically optimal classification accuracy. Furthermore, the features/taxa selected by GLL have putative causal interpretation. This warrants their further study and quantifying their effect on phenotype. The approach presented in this study was designed as a general-purpose downstream analysis methodology and as such may facilitate the design and development of microbiomic molecular signatures of other diseases in addition to psoriasis.

## Methods

**Study population.** We obtained informed written consent (using the model consent forms for the HMP demonstration projects) and enrolled a total of 199 subjects (75 patients with psoriasis and 124 healthy controls) under New York University School of Medicine IRB #08-709 between June 2008 and September 2011. Psoriatic patients were diagnosed at an academic dermatology practice, and psoriasis was clinically classified based on characteristic morphologic features of the individual skin lesions and their cutaneous distribution. Among the patients with psoriasis, 57 (76%) had not been exposed to antibiotics or received a relevant clinical treatment for at least one month before skin samples were obtained, and of these, a total of 54 patients were studied. Among the healthy controls, 112 (90.3%) had not been exposed to antibiotics or received any relevant treatment for at least one month before skin samples were obtained.

For the 54 subjects with psoriasis, we sought control subjects of the same gender and ethnicity, and of similar age ( $\pm 5$  years), for whom a cutaneous specimen was obtained in a region proximate to the psoriasis lesion. In total, we obtained matching control specimens from 37 (29.8%) subjects. One or more sites from each of these controls were matched to the 54 subjects with psoriasis. A control subject could be matched to more than one patient, since we also matched for cutaneous site. However, each control cutaneous site was uniquely mapped to only one triplet, thus there was no duplication of specimens in the analysis.

**Specimen collection and processing.** In the patients with psoriasis, we sampled a typical psoriatic plaque (designated as Psoriasis, lesion), and as a reference site, the contralateral area of clinically uninvolved skin (designated as Psoriasis, normal). We also sampled skin from matched healthy (control) individuals at locations that approximated the location of the psoriatic lesion. We accomplished this by obtaining 4 skin swabs from each control person, from scalp (posterior-temporal, above the ear crease), elbow (inner aspect), lower lateral abdomen, and knee. The selected sampling sites reflect common general locations on the head, torso, and extremities where psoriatic lesions arise in patients<sup>2</sup>. Both the lesions and the control specimens were predominantly obtained from dry and oily cutaneous sites<sup>12</sup>, reflecting the distribution of psoriatic lesions in general<sup>1</sup>.

Methods for specimen processing are described elsewhere<sup>20</sup>. In brief, a 2 by 2-cm area of the cutaneous surface at each of the locations was sampled by swabbing the skin with a cotton pledget that had been soaked in sterile 0.15 M NaCl with 0.1% Tween 20 (Fisher Scientific, Fair Lawn NJ). DNA was extracted from the swab suspensions in a PCR-free clean room by using the DNeasy blood and tissue kit (Qiagen, Chatsworth CA); glass beads (0.5 to 1 mm) were added to the specimens and vortex-mixed at maximum speed for 40 s, followed by DNA extraction, using the manufacturer's protocol for genomic DNA isolation from Gram-positive bacteria, and samples were eluted in 100  $\mu$ l AE buffer (DNeasy Blood and Tissue kit; Qiagen).

To eliminate bacterial or DNA contamination, lysozyme (Sigma-Aldrich, St. Louis MO) was passed through a microcentrifuge filter (molecular mass threshold, 30,000 Da; Amicon, Bedford MA) at  $18,514 \times g$  for 20 min before adding to the enzymatic lysis buffer.

**DNA sequencing and upstream processing.** Samples were prepared for amplification and sequencing at the J. Craig Venter Institute (JCVI) Joint Technology Center (JTC)<sup>14</sup>. Genomic DNA sample concentrations were normalized. The V1–V3 region of the 16S rRNA gene was amplified using forward primer attached to the Roche B adapter for 454-library construction and reverse primer attached to the Roche A adapter and a 10-nt barcode [5'-A-adapter-N (10) + 16S primer-3]. A barcoded primer design was completed using a set of algorithms developed at the JCVI. PCR reactions were completed as follows (per reaction): 2  $\mu$ l of gDNA,  $1 \times$  final concentration of Accuprime PCR Buffer II (Invitrogen, Carlsbad CA, USA), 200 nM forward and reverse primers, 0.75 U of Accuprime TaqDNA polymerase high fidelity (Invitrogen), and nuclease-free water to bring the final volume to 20  $\mu$ l. PCR cycling conditions consisted of an initial denaturation of 2 min at 95°C, 30 cycles of 20 s at 95°C, 30 s at 50°C, and 5 min at 72°C. A negative control (water blank) reaction was examined after 35 cycles. Samples then were quantified, cleaned, and sequenced on the Roche 454-FLX (454 Life Sciences, Branford CT, USA) as described<sup>21</sup>, and a read processing pipeline consisting of a set of modular scripts designed at the JCVI were employed for deconvolution, trimming, and quality filtering, as described previously<sup>21</sup>. Finally, we included in our analysis only samples with more than 2,000 reads (the median number of reads per sample was 8,621). The sequencing effort was statistically similar across clinical skin types. The above protocol was repeated to obtain sequencing data from the V3–V5 locus of the 16S rRNA gene.

As a result, we obtained two 16S rRNA datasets (one for the V1–V3 locus and another one for the V3–V5 locus) with the number of evaluable samples for each class given in Table 2. The number of samples in the dataset for the V3–V5 locus was smaller than for the V1–V3 locus primarily because of the lower sequencing quality and yield at the former locus, however as shown in the Results section the number of samples was sufficient to obtain molecular signatures and evaluate their statistical significance. The passing sequences were analyzed using the RDP classifier executed at a bootstrap confidence cut-off of 80%<sup>22</sup> and were normalized. This criterion resulted in processed datasets with 791 and 660 taxonomic features (relative taxonomic abundance at phylum, class, order, family, and genus levels) for the V1–V3 and the V3–V5 loci, respectively. These datasets were used for development of molecular signatures and assessment of their classification accuracy with the methods described in the following subsection.

**Downstream analyses for development of microbiomic signatures.** The features/taxa for inclusion in molecular signatures were selected by several feature selection methods. We used Generalized Local Learning (GLL), a supervised multivariate method, as a key technique for feature selection in this study<sup>15,19</sup>. Under broad distributional assumptions, the taxa selected by GLL are theoretically guaranteed to achieve both maximal classification accuracy and maximal parsimony for the dataset at hand<sup>15,19</sup>. In addition, because the GLL method is based on Markov Boundary theory (a branch of formal causal graph induction), under broad distributional assumptions, it returns the local pathway members around the response variable of interest. GLL was executed with a default setup: Fisher's Z-test was used for assessing conditional independence at a 5% alpha level, and the max-k parameter, which represents the maximum size of the conditioning sets, was set to 3. In addition to GLL, we applied 36 comparator methods/variants that have been successful in other types of genomic data, as well as recent applications to microbiomics<sup>7,23</sup>. The reasons for such an extensive range of methods is to shed light on relative strengths and weaknesses of state-of-the-art methods for microbiomic molecular signature development, and to ensure that we obtain as highly predictive and as compact signatures as possible from our data. In brief, these methods include: support vector machine-based recursive feature elimination (SVM-RFE)<sup>24</sup>; univariate methods based on various statistical tests (Kruskal-Wallis non-parametric one-way ANOVA test, signal-to-noise ratio, ratio of between-group to within-group sum of squares, two-sample t-test, and  $\chi^2$ -test<sup>23,25–27</sup>, denoted as UAF-KW, UAF-S2N, UAF-BW, UAF-T, and UAF-X2, respectively); minimum redundancy and maximum relevancy methods (MRMR)<sup>28,29</sup>; random forest-based variable selection (RFVS)<sup>30</sup>; least angle regression elastic net (LARS-EN)<sup>31</sup>; soft independent modeling of class analogy (SIMCA)<sup>32</sup>; principal component and sparse principal component analysis-based methods (PCA and SPCA)<sup>33</sup>; threshold gradient descent regularization (TGDR)<sup>34</sup>; and using all features/taxa in the dataset without feature selection (ALL). All feature/taxa selection methods and details about their variants are listed in Table S1.

Once the features/taxa were selected, molecular signatures were developed using the support vector machine (SVM) supervised classification algorithm<sup>35</sup>. We chose to use SVM classifiers because they have high utility in many genomic datasets<sup>23</sup>, and have outstanding empirical performance in microbiomic data (see<sup>36</sup> and chapter 6 of<sup>23</sup>). Furthermore, several theoretical considerations warrant application of SVMs to microbiomic data: SVMs perform well in data with limited sample size, are relatively insensitive to high dimensionality of the data, prevent overfitting by using regularization techniques, and can learn both simple and complex decision functions<sup>35,37,38</sup>. SVMs were applied with a polynomial kernel; the degree of the kernel ( $d$ ) and penalty parameter ( $C$ ) were optimized by nested cross-validation over values  $d = (1, 2, 3)$  and  $C = (0.01, 0.1, 1, 10, 100)$  as detailed below. We used the libSVM version 3.12 implementation of SVMs<sup>39</sup>.



Table 2 | Number of samples for each class in 16S rRNA gene data from the V1–V3 and the V3–V5 loci

	Psoriasis, lesion (PL)	Psoriasis, normal (PN)	Healthy controls (CC)
V1–V3 locus	51	51	49
V3–V5 locus	22	22	22

For estimation of classification accuracy and optimization of classifier hyper-parameters, we used nested five-fold cross-validation repeated 100 times with different splits of the data into five folds<sup>40,41</sup>. This protocol proceeded as follows. First, all samples in the data were randomly split into five non-overlapping and approximately equal parts (folds) such that each fold retained a similar proportion of samples from two classes. Next, four of five folds (80% of the data; “original training set”) were used for feature/taxa selection and development of the molecular signature, and the remaining fold (20% of the data; “original testing set”) was used for estimation of classification accuracy of the molecular signature. The molecular signature was developed and evaluated five times so that each of the five folds could play the role of the original testing set, and its complement could play the role of the original training set. Then the entire process was repeated 100 times with different splits of the data into five folds. The resulting classification accuracy estimate was the mean over five different testing sets from each of 100 different data splits into folds. Since the methodologies for feature/taxa selection and development of molecular signature have parameters that require tuning, each of the original training sets (consisting of four folds) was further split into a smaller training set with three folds and a validation set with one fold, and a four-fold cross-validation protocol was used to estimate accuracy of each parameter configuration and select the one with the best average accuracy for application to the original training set. This protocol avoids overfitting and yields unbiased estimates since the testing set is not used for development of the molecular signature (i.e., no “information leak” occurs from the response variable back to the model selection, feature selection, and classifier fitting procedures).

As a metric of classification accuracy, we used the area under the ROC curve (AUC)<sup>42,43</sup>. This metric has greater statistical power to detect signal than the commonly used proportion of misclassifications and is insensitive to the ratio of cases to controls<sup>44</sup>. The ROC curve is the plot of sensitivity versus one minus the specificity for a range of classification threshold values. AUC ranges from 0 to 1, with an AUC equal to 0 indicating the classifier opposite to the perfect, 0.5 indicating a random (i.e., uninformative) classifier, ~0.70–0.75 indicating a mediocre classifier, ~0.80–0.85 indicating a very good classifier, ~0.90 or better indicating an excellent classifier, and 1 indicating perfect classification<sup>42,43</sup>.

Statistical comparison/testing of results was conducted using two types of permutation tests following the theory developed by Good<sup>45</sup>. The first test assessed whether a molecular signature had statistically significant classification accuracy and tested the null hypothesis of AUC = 0.5. This test was implemented with 1,000 permutations of the response variable, as described<sup>46</sup>. The second test assessed whether a molecular signature obtained with the GLL methodology was significantly better or worse than a molecular signature obtained with a different feature/taxa selection method. This test was implemented with 10,000 permutations, as described<sup>47</sup>. Since both statistical tests assessed a large number of comparisons, we performed adjustment for multiple comparisons using methods to control false-discovery<sup>48,49</sup>. We used a 5% alpha-level adjusted for multiple comparisons to establish statistical significance.

All experiments presented in this manuscript were run on the Asclepius Compute Cluster at the Center for Health Informatics and Bioinformatics (CHIBI) at New York University Langone Medical Center (<http://www.nyuinformatics.org>).

- Fahlen, A., Engstrand, L., Baker, B. S., Powles, A. & Fry, L. Comparison of bacterial microbiota in skin biopsies from normal and psoriatic skin. *Archives of dermatological research* **304**, 15–22 (2012).
- Gao, Z., Tseng, C. H., Strober, B. E., Pei, Z. & Blaser, M. J. Substantial alterations of the cutaneous bacterial biota in psoriatic lesions. *PLoS ONE* **3**, e2719 (2008).
- Paulino, L. C., Tseng, C. H., Strober, B. E. & Blaser, M. J. Molecular analysis of fungal microbiota in samples from healthy human skin and psoriatic lesions. *Journal of clinical microbiology* **44**, 2933–2941 (2006).
- Reischl, J. et al. Increased expression of Wnt5a in psoriatic plaques. *The Journal of investigative dermatology* **127**, 163–169 (2007).
- Yao, Y. et al. Type I interferon: potential therapeutic target for psoriasis? *PLoS ONE* **3**, e2737 (2008).
- Nair, R. P. et al. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nat Genet* **41**, 199–204 (2009).
- Knights, D., Costello, E. K. & Knight, R. Supervised classification of human microbiota. *FEMS Microbiol. Rev.* **35**, 343–359 (2011).
- Arumugam, M. et al. Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).
- Yatsunencko, T. et al. Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
- Fierer, N. et al. Forensic identification using skin bacterial communities. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 6477–6481 (2010).

- Gao, Z., Tseng, C. H., Pei, Z. & Blaser, M. J. Molecular analysis of human forearm superficial skin bacterial biota. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 2927–2932 (2007).
- Grice, E. A. & Segre, J. A. The skin microbiome. *Nature reviews. Microbiology* **9**, 244–253 (2011).
- Costello, E. K. et al. Bacterial community variation in human body habitats across space and time. *Science* **326**, 1694–1697 (2009).
- Human Microbiome Project, C. A framework for human microbiome research. *Nature* **486**, 215–221 (2012).
- Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S. & Koutsoukos, X. D. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification. Part I: Algorithms and Empirical Evaluation. *Journal of Machine Learning Research* **11**, 171–234 (2010).
- Spirtes, P., Glymour, C. N. & Scheines, R. *Causation, prediction, and search*. Vol. 2nd (MIT Press, 2000).
- Narendra, V., Lytkin, N. I., Aliferis, C. F. & Statnikov, A. A comprehensive assessment of methods for de-novo reverse-engineering of genome-scale regulatory networks. *Genomics* **97**, 7–18 (2011).
- Alekseyenko, A. V. et al. Causal graph-based analysis of genome-wide association data in rheumatoid arthritis. *Biology Direct* **6**, 25 (2011).
- Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S. & Koutsoukos, X. D. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification. Part II: Analysis and Extensions. *Journal of Machine Learning Research* **11**, 235–284 (2010).
- Gao, Z., Perez-Perez, G. I., Chen, Y. & Blaser, M. J. Quantitation of major human cutaneous bacterial and fungal populations. *Journal of clinical microbiology* **48**, 3575–3581 (2010).
- Cho, I. et al. Antibiotics in early life alter the murine colonic microbiome and adiposity. *Nature* **488**, 621–626 (2012).
- Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology* **73**, 5261–5267 (2007).
- Statnikov, A., Aliferis, C. F., Hardin, D. P. & Guyon, I. *A Gentle Introduction to Support Vector Machines in Biomedicine, Volume 2: Case Studies & Benchmarks*. (World Scientific Publishing, 2013).
- Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Machine Learning* **46**, 389–422 (2002).
- Hollander, M. & Wolfe, D. *Nonparametric statistical methods*. Vol. 2nd (Wiley, 1999).
- Golub, T. R. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
- Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D. & Levy, S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics* **21**, 631–643 (2005).
- Ding, C. & Peng, H. Minimum redundancy feature selection from microarray gene expression data. *J Bioinform. Comput. Biol.* **3**, 185–205 (2005).
- Peng, H., Long, F. & Ding, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE transactions on pattern analysis and machine intelligence* **27**, 1226–1238 (2005).
- Diaz-Uriarte, R. & Alvarez de Andres, S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**, 3 (2006).
- Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B(Statistical Methodology)* **67**, 301–320 (2005).
- Bicciato, S., Luchini, A. & Di, B. C. PCA disjoint models for multiclass cancer analysis using gene expression data. *Bioinformatics* **19**, 571–578 (2003).
- Zou, H., Hastie, T. & Tibshirani, R. Sparse principal component analysis. *Journal of Computational and Graphical Statistics* **15**, 265–286 (2006).
- Ma, S. & Huang, J. Clustering threshold gradient descent regularization: with applications to microarray studies. *Bioinformatics* **23**, 466–472 (2007).
- Vapnik, V. N. *Statistical learning theory*. (Wiley, 1998).
- Statnikov, A. et al. A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome* **1**, 11 (2013).
- Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* **2**, 121–167 (1998).
- Cristianini, N. & Shawe-Taylor, J. *An introduction to support vector machines and other kernel-based learning methods*. (Cambridge University Press, 2000).
- Fan, R. E., Chen, P. H. & Lin, C. J. Working set selection using second order information for training support vector machines. *Journal of Machine Learning Research* **6**, 1918 (2005).
- Braga-Neto, U. M. & Dougherty, E. R. Is cross-validation valid for small-sample microarray classification? *Bioinformatics* **20**, 374–380 (2004).



41. Statnikov, A., Tsamardinos, I., Dosbayev, Y. & Aliferis, C. F. GEMS: a system for automated cancer diagnosis and biomarker discovery from microarray gene expression data. *Int.J.Med.Inform.* **74**, 491–503 (2005).
42. Fawcett, T. ROC Graphs: Notes and Practical Considerations for Researchers. *Technical Report, HPL-2003-4, HP Laboratories* (2003).
43. Harrell, F. E., Jr., Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat.Med* **15**, 361–387 (1996).
44. Ling, C. X., Huang, J. & Zhang, H. AUC: a statistically consistent and more discriminating measure than accuracy. *Proceedings of the Eighteenth International Joint Conference of Artificial Intelligence (IJCAI)* (2003).
45. Good, P. I. *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Vol. 2nd (Springer, 2000).
46. Aliferis, C. F. *et al.* Factors Influencing the Statistical Power of Complex Data Analysis Protocols for Molecular Signature Development from Microarray Data. *PLoS ONE* **4**, e4922 (2009).
47. Menke, J. & Martinez, T. R. Using permutations instead of student's t distribution for p-values in paired-difference algorithm comparisons. *Proceedings of 2004 IEEE International Joint Conference on Neural Networks* **2**, 1331–1335 (2004).
48. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society.Series B (Methodological)* **57**, 289–300 (1995).
49. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann.Statist* **29**, 1165–1188 (2001).

## Acknowledgements

This research was supported in part by the grants UH2 AR057506-01S1 from the Human Microbiome Project, 1UL1 RR029893 from the National Center for Research Resources, and R01 LM011179-01A1 from the National Library of Medicine, National Institutes of Health; by the Diane Belfer Program in Human Microbial Ecology; and by the Department of Veterans Affairs Office of Research and Development. The authors thank Dr. Efstratios Efstathiadis and Dr. Eric Peskin for providing high performance computing support.

## Author contributions

C.F.A., M.J.B., A.V.A., A.S. conceived the research study and designed the methods and experiments. M.H., Z.L., A.V.A., A.S. prepared the data, implemented the methods and conducted experiments. All authors participated in the interpretation of the results. All authors have contributed to, read, and approved the final manuscript.

## Additional information

**Supplementary information** accompanies this paper at <http://www.nature.com/scientificreports>

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Statnikov, A. *et al.* Microbiomic Signatures of Psoriasis: Feasibility and Methodology Comparison. *Sci. Rep.* **3**, 2620; DOI:10.1038/srep02620 (2013).



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported license. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/3.0>