Functional coding haplotypes and machine-learning feature elimination identifies predictors of Methotrexate Response in Rheumatoid Arthritis patients

Ashley J.W. Lim,a,1 Lee Jin Lim,a,1 Brandon N.S. Ooi,a Ee Tzun Koh,b Justina Wei Lynn Tan,b TTSH RA Study Group b

Samuel S. Chong,c Chiea Chuen Khor,d Lisa Tucker-Kellogg,e Khai Pang Leong,b,f,#** and Caroline G. Lee,a,g,h,i*,#

aDept of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
bDepartment of Rheumatology, Allergy and Immunology, Tan Tock Seng Hospital, Singapore
cDept of Pediatrics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
dDivision of Human Genetics, Genome Institute of Singapore, Singapore
eCentre for Computational Biology, and Cancer and Stem Cell Biology, Duke-NUS Medical School, Singapore
fClinical Research & Innovation Office, Tan Tock Seng Hospital, Singapore
gDiv of Cellular & Molecular Research, Humphrey Oei Institute of Cancer Research, National Cancer Centre Singapore, Singapore
hDuke-NUS Medical School, Singapore
iNUS Graduate School, National University of Singapore, Singapore

Summary

Background Major challenges in large scale genetic association studies include not only the identification of causative single nucleotide polymorphisms (SNPs), but also accounting for SNP-SNP interactions. This study thus proposes a novel feature engineering approach integrating potentially functional coding haplotypes (pfcHap) with machine-learning (ML) feature selection to identify biologically meaningful, possibly causative genetic factors, that take into consideration potential SNP-SNP interactions within the pfcHap, to best predict for methotrexate (MTX) response in rheumatoid arthritis (RA) patients.

Methods Exome sequencing from 349 RA patients were analysed, of which they were split into training and unseen test set. Inferred pfcHaps were combined with 30 non-genetic features to undergo ML recursive feature elimination with cross-validation using the training set. Predictive capacity and robustness of the selected features were assessed using six popular machine learning models through a train set cross-validation and evaluated in an unseen test set.

Findings Significantly, 100 features (95 pfcHaps, 5 non-genetic factors) were identified to have good predictive performance (AUC: 0.776-0.828; Sensitivity: 0.656-0.813; Specificity: 0.684-0.868) across all six ML models in an unseen test dataset for the prediction of MTX response in RA patients.

Interpretation Majority of the predictive pfcHap SNPs were predicted to be potentially functional and some of the genes in which the pfcHap resides in were identified to be associated with previously reported MTX/RA pathways.

Funding Singapore Ministry of Health’s National Medical Research Council (NMRC) [NMRC/CBRG/0095/2015; CG12Aug17; CGAug16M012; NMRC/CG/017/2013]; National Cancer Center Research Fund and block funding Duke-NUS Medical School.; Singapore Ministry of Education Academic Research Fund Tier 2 grant MOE2019-T2-1-138.

#*Address correspondence to: Caroline GL. Lee, Ph.D. Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore c/o MD7, Level 2, 8 Medical Drive, Singapore 117597. Tel: (+65) 64368353. **Leong Khai Pang, MBBS, Department of Rheumatology, Allergy and Immunology Tan Tock Seng Hospital, 11 Jln Tan Tock Seng, Singapore 308433. Tel: (+65) 63577821.

E-mail addresses: khai_pang_leong@ttsh.com.sg (K.P. Leong), bchlee@nus.edu.sg (C.G. Lee).

† Lee Jin Lim and Ashley J.W. Lim contributed equally to this work

# Equal Contribution to this work
**Research in Context**

**Evidence before this study**

Methotrexate (MTX) is a first-line medication for rheumatoid arthritis (RA) patients despite low monotherapy response of only 25 to 45 percent. Identification of non-responders is necessary to help mitigate disease progression. However, evaluation of treatment response often takes three to six months following MTX administration. Current research on MTX response in RA are primarily focused on genetic variation in specific genes involved in MTX-related pathways and RA susceptibility with few interrogating the entire genome using genome-wide association studies (GWAS). Nonetheless, several challenges are associated with current GWAS including high data dimensionality and the identification of causative variations.

**Added value of this study**

In this study, we employed a cost-effective exome sequencing approach to interrogate haplotype of single nucleotide polymorphisms (SNPs) in the coding region of genes which is deemed as one of the most informative functional regions of the genome. To reduce the high dimensionality data, and capitalize on the property that SNPs within a functional unit (e.g. coding region) interact to modulate structure/function of the target protein, we inferred SNP haplotypes in the coding region and employed machine-learning (ML) to identify potentially functional coding haplotypes (pfcHaps) that best predicts MTX response in RA patients. Notably, the predictive pfcHap SNPs and genes were predicted to be functional and associated with previously reported MTX/RA pathways, respectively, highlighting the promise of this approach.

**Implications of all the available evidence**

Taken together, we envision that the best predictors identified will be effective in aiding decision-making for the treatment of RA patients after further validation in larger multi-institutional studies. Furthermore, we believe that our analysis pipeline for handling and interpretation of genetic data will also be applicable in other contexts beyond MTX response in RA patients.

**Introduction**

Of the diverse factors influencing drug response, elucidating the genetic basis that underlie differences in drug response may facilitate the development of tools to predict drug response even before the drug is administered. Although there is growing interest in pharmacogenomics, the focus has thus far been primarily on the identification of common gene variants with large effect size on known pharmacokinetic, pharmacodynamics and/or immuno-pharmacogenomics candidate genes. However, polymorphisms in other genes or rare gene variants, either alone or in combination, may, also modulate drug response but have yet to be thoroughly investigated. With the advent of high throughput genomic tools, it is now possible to explore the association of other genes with drug response using GWAS (Genome Wide Association Study), exome sequencing or even WGS (Whole Genome Sequencing). Current genomic approaches which mainly focused on tag-SNPs in GWAS only represent a very small proportion of all potentially functional SNPs (pSNPs) in the human genome with likelihood to be causative. While WGS would be the most ideal approach to examine all SNPs including pSNPs, exome sequencing is a cost-effective way to facilitate the interrogation of all pSNPs in the most informative functional region of the genome, namely, the coding region.

Thus far, traditional statistical methods have been the primary tool to associate genetic and other features with drug response and/or disease susceptibility. However, these statistical approaches have some limitations. This includes the requirement for large sample sizes, which can potentially be mitigated by machine learning (ML) which are suited for high dimensionality complex problems, including the consideration for non-linear interactions between features. Nonetheless, a major limitation of biomedical datasets, even for ML, is that their high dimensionality is often coupled with limited labelled sample size, which pose a challenge for learning models to predict individualized response to drugs. ML-based dimensionality reduction and feature selection strategies can help reduce the feature space and select the most informative features that can accurately predict an outcome. Nonetheless, to achieve acceptable accuracy in pharmacogenomics, careful data pre-processing and feature handcrafting with strong domain knowledge is necessary.

Here, we introduce a novel biologically meaningful, feature pre-processing/engineering strategy focused on haplotypes of SNPs in the coding regions of genes with the potential to be functional (pfcHap). By integrating
pfCHap together with ML feature elimination and selection, the strategy identifies a signature of potentially causative genetic and non-genetic factors that can robustly predict response to the methotrexate (MTX) drug in Rheumatoid Arthritis (RA) patients across diverse ML models.

RA was proposed to be particularly appropriate for personalized therapy because of the costly therapy due to prolonged disease duration, low response to the conventional therapy, trial-and-error nature of therapy prescription, and the risk of serious drug-induced side effects. The disability brought on by comorbidities such as coronary artery disease and hyperlipidaemia in combination with poorly controlled RA adds to the healthcare costs and further strains the health care system.

RA is a chronic inflammatory disease involving primarily the joints with a prevalence of an average of ~5 per 1000 people that varies across different populations. It imposes huge socioeconomic burden on both the patient and society as it commonly affects middle-aged adults at their economic peak. Inadequate treatment of RA leads to irreversible joint damage resulting in potential disabilities that affects the patient’s quality of life and work productivity, and even premature death. Hence, timely and appropriate control of this condition is critical to minimize the morbidity and mortality. Amongst the disease-modifying antirheumatic drugs (DMARDs) in RA, MTX is the anchor agent and the recommended first-line choice for the majority of RA cases. Approximately 25 to 40 percent of patients improve with MTX monotherapy, which is further increased to 50 percent for patients receiving combination therapy with glucocorticoids. Patients with inadequate response to MTX monotherapy are offered alternative biologic and targeted synthetic DMARDs. The current state-of-the-art management of RA is still primarily based on trial-and-error with recommendations from the European League Against Rheumatism (EULAR) being to assess the effectiveness of MTX therapy between three to six months of administering the drug and re-evaluating the treatment approach of poor responders thereafter. This suggests that rheumatologists only know the effectiveness of MTX after the patient is already on the drug for 3-6 months. Earlier identification of poor responders of MTX, preferably even before drug administration, will enable prompt initiation of alternative treatment which could mitigate disease progression.

To date, the study of MTX response in RA have been focused on the genetic variability in specific genes, often involving those in MTX-related pathways or RA susceptibility or Genome wide association study (GWAS) interrogating individual SNPs. A recent review summarised 125 SNPs from 34 genes involved with MTX metabolism, transport or RA progression/pathogenesis were previously evaluated for associations with MTX response. However, some of these studies have reported contradictory results, including conflicting reports of associations of polymorphism rs1045642 (3435C > T) in the ATP-binding cassette B1 (ABCB1) transporter gene, with MTX efficacy in two separate Japanese cohort studies.

While there is recent increasing interest to employ ML for electronic diagnosis, prediction of disease progression and drug response of RA patients, these methods remain at its infancy, with few studies exploring predictive models to evaluate MTX drug response. They mainly focus on specific subsets of SNPs with non-genetic factors or other molecular signatures (e.g. transcription/epigenetic-based signatures). Most of the methotrexate predictive models employed simple machine learning models such as logistic regression and mainly focused on electronic medical records, or juvenile RA. Thus far, more complex ML models with careful domain-based, biologically meaningful feature handcrafting have yet to be applied comprehensively to improve predictive performance of response to RA drugs.

Previous haplotype-based studies for other diseases mainly examined haplotype of SNPs within specific window sizes or employed to account for familial correlation for association between rare haplotypes and complex disease. Here, we report a novel, cost-effective approach through exome sequencing to interrogate haplotypes of SNPs in the coding regions of genes (pfCHap) of the entire human genome, since coding regions, which are translated into proteins represent one of the most functional regions of genes. As complex phenotypes are likely due to the interaction of multiple SNPs in a functional unit (e.g. coding region) rather than single SNP acting in isolation, and SNPs altering amino acids in the same protein may interact with each other to alter the folding or function of protein (e.g. binding to substrates, etc), this approach of focusing on pfCHap has the potential for identifying signatures of pfCHap that not only account for SNP-SNP interactions but also a higher likelihood of being the causative combination of variants.

**Methods**

Fig. 1 illustrates our strategy to identify biologically meaningful features with good prediction performance for MTX response in RA patients.

**Study cohort**

This study examined 349 subjects of Chinese ethnicity receiving MTX treatment for RA. Patients were at least 18 years old and satisfy the 1987 American College of Rheumatology revised criteria or the 2010 American College of Rheumatology/European League Against Rheumatism criteria for RA. This study was endorsed.
by the National Healthcare Group Domain Specific Review Board (DSRB 2015/00582). All protocols were carried out according to the Declaration of Helsinki and informed consent was collected from all patients. All patients received at least 3 months of MTX treatment at 15 mg per week and >90% of the patients completed 2 years of treatment. MTX drug response is defined as remission within/at two years post-treatment, which is determined by evaluation of Disease Activity Score in 28 joints (DAS28). DAS28 is a composite score representative of RA activity that includes the number of tender joints and swollen joints, erythrocyte sedimentation rate, and a global assessment of health. Responders to MTX were classified as patients with DAS28 <2.6 following MTX treatment. Various non-genetic features (including demographic/clinical characteristics and medication status) of patients were also recorded and summarised in Table S1.

**Figure 1.** Pipeline employed to identify the predictors of MTX response.

349 patient samples were first divided into training (n=279, 70%) and test (n=70, 30%) sets using a stratified split, such that datasets consist of the proportion of responders and non-responders that is representative of the original dataset. The training set was then further split into eight subsets consisting of different sample size ranging from 30% to 100% of the samples in the training set. Within each subset, the important features (coding haplotypes or integration of coding haplotypes with non-genetic features) were selected using recursive feature elimination with cross-validation (RFECV), applied with Random Forest Classifier as the estimator of choice. The important features that were commonly identified in all eight subsets were then shortlisted and identified as the set of important features that are predictive of MTX response. The predictive performance of these features was assessed in six different machine learning models, using cross-validation within the training set and the unseen test dataset.

**Exome sequencing, sequence alignment, and single variant analysis**

Enrichment of the exome region of genomic DNA was performed with the Nimblegen SeqCap EZ kit (Roche) and with Agilent SureSelect Human All Exon kit (Agilent Technologies, CA). Products were then purified using AMPure XP system (Beckman Coulter, Beverly, USA) and quantified using the Agilent high sensitivity DNA assay on the Agilent Bioanalyzer 2100 system. The exome sequencing was performed by commercial providers using the IlluminaHiSeq2000 100PE platform.

Using the BWA-MEM algorithm, the sequenced data was aligned to the hs37d5 human reference genome. Following, PicARD was used to remove duplicated reads. Each sample was processed separately using the base recalibration and haplotypcaller modules of GATK v3. Thereafter, using genotypeGVCF,
variants were called on all the samples together. As
per GATK best practice for hard filtering, SNP filtering
was performed with "QD < 2.0 || FS > 60.0 || MQ
< 40.0 || MQRankSum < -12.5 || ReadPosRankSum <
-8.0" as the criteria and indel using "QD < 2.0 || FS >
200.0 || ReadPosRankSum < -20.0".

Training and test data
The total dataset from 349 subjects were split into a
training (80%, N = 279) and an unseen test set (20%, N = 70) in a stratified manner to maintain the ratio
between MTX responders and non-responders. Splitting
of dataset was performed within the Python (3.8.3) envi-
ronment using the Scikit-learn module. To avoid data
leakage, subsequent processing steps were performed
on the training and test datasets separately. The training
dataset (80%) is further processed into 8 subsets of vari-
able sample sizes through stratified random sampling
with replacement.

Genetic/Biological-based feature engineering: Data
processing and defining haplotypes
Biallelic variants with >10% genotype missingness or
deviate from Hardy-Weinberg equilibrium (p-value >
0.001) were removed before downstream analyses.
Thereafter, using the UCSC table browser as refer-
ence genome. Genotype phasing was then per-
duced using ANNOVAR based on the hg19
reference genome. Haplotypes carrying all reference SNPs were
constructed using the BEAGLE 5.1 software together with
HapMap Phase II recombination maps for each respec-
tive chromosome. All variants within coding regions were
selected using BEDTools Suite. These variants were annotated using ANNOVAR based on the hg19 reference genome. Genotype phasing was then performed using the BEAGLE 5.1 software together with HapMap Phase II recombination maps for each respective chromosome. All variants within coding regions of a gene were used for haplotype construction. Only haplotypes with a minor haplotype frequency > 0.01 were further analysed. Haplotypes carrying all reference SNPs were removed. The final total of ~39,000 haplotypes were combined with 30 non-genetic features for downstream analyses.

Selecting features/predictors that are predictive for
MTX responses in RA patients
Within each training subset, features with near zero var-
iance (i.e., features which are almost constant across all
samples) or display > 95% correlation with other fea-
tures were excluded. To identify a signature of robust
features that are important for prediction, we utilised
recursive feature elimination with cross-validation
(RFECV), as implemented in the Scikit-learn module, incorporating a Random Forest classifier as the estima-
tor. The process was performed using a 5-fold cross-vali-
dation until an optimal number of features was
selected. To obtain features with high stability of impor-
tance, features that are common across all 8 training
subsets were selected for further evaluation of their pre-
dictive performance.

Evaluating predictive performances of selected
features
Selected features were evaluated across six ML diverse
algorithms: Neural Networks, Support Vector Machines, two regression-based algorithms (Logistic
regression and Elastic nets), two tree-based algorithms (Random Forests and Boosted Trees) for a broad
representation, as there is currently no consensus as to which ML algorithm is the most appropriate for genomics.
Predictions were performed using the Python Scikit-
learn module with default parameters. Performance
evaluation of each classifier was conducted using a 5-
fold cross-validation of the training set (n=279), and the
area under the curve (AUC) of a receiver operating char-
acteristic (ROC) curve was generated. Since a smaller
number of features is generally preferable for facilitating
clinical implementation, the minimum number of features that achieve good predictive performance in the
training set was selected as the best predictors. These
predictors were then tested in the unseen test set using
the six ML models, and the AUC of a ROC curve was generated.

Analysis for potential functions of SNPs within
selected haplotypes
To explore for possible functional mechanisms underlying
the differences in MTX responses associated with
selected haplotypes, SNPs within selected haplotypes
were evaluated for their potential or predicted functional-
ity, through interrogation of an updated potentially
functional SNP (pfSNP) resource developed by our lab-
atory which has included data from more recently
published prediction databases including expression-
associated SNPs or (expression quantitative trait loci
eQTLs). The functionality of SNPs was assessed
for their potential to alter important functional regions
(e.g., transcription factor binding sites, miRNA binding
sites, exonic splice enhancer/silencer (ESE/ESS), etc) or
induce nonsense-mediated decay (NMD). Additionally,
non-synonymous SNPs were also evaluated for their
potential deleterious effects while synonymous SNPs
were identified for possible codon usage bias.

To further explore the functions of these haplotypes,
pathway enrichment analyses were performed on the
genes corresponding to the most predictive haplotypes
using ConsensusPathDB resource.

Ethics
This study was endorsed by the National Healthcare
Group Domain Specific Review Board (DSRB 2015/
00582). All protocols were carried out according to the
Declaration of Helsinki and informed consent was collected from all patients.

Role of funders
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Any opinions, findings, or recommendations expressed in this material are those of the authors and do not reflect the views of the funders.

Results
Combination of biological and ML approaches reduces complexity of dataset and identifies a small set of features predictive for MTX responses
A total of 114,000 SNPs were identified from exome sequencing of blood DNA. With such high-dimensional data where the potential predictors far outnumber the number of samples, the identification of features with good predictive performance in unseen datasets becomes very challenging. Here, two consecutive steps were employed to mitigate this ‘curse of dimensionality’. The first step was a biologically meaningful, feature handcrafting, serving as a genetic approach of dimensionality reduction. This involved the derivation of haplotypes from SNPs in the coding regions of genes (pfcSNPs) as these are functional regions and are more likely to be causative, potentially altering protein structure and/or function. A total of 52,331 pfcHaps were derived from ~13,000 genes, which represented a ~54.1% reduction in the number of potential predictors. The exclusion of pfcHaps that comprises only reference alleles further reduced the number to 39,160 leading to a ~65.6% reduction in potential predictors. The second step utilised ML recursive feature elimination (RFECV) with Random Forest classifiers. Repeated eight times, RFECV on each training subset reduced the set of 39,160 pfcHaps, and identified between 411 to 2602 important pfcHaps, of which 120 were commonly identified in all eight subsets and used for training with cross-validation of the six ML models. These 120 pfcSHaps displayed reasonable cross-validation AUCs of between 0.794 and 0.900 (Fig. 2).

Integrating coding haplotypes with non-genetic factors identifies a smaller set of features with similar predictive capacity for MTX response
To determine whether the integration of pfcHaps with non-genetic features from medical records can improve classifier accuracy and/or reduce the number of features required for accurate prediction, we combined the 39,160 pfcHaps with 30 non-genetic factors (Table S1) and re-performed the same RFECV with Random Forest classifier in the eight training subsets. Between 363 to 3612 important features were identified in the eight training subsets (Fig. 3), of which 100 (95 pfcHaps and 5 non-genetic features) were common across all eight training subsets. The 5 non-genetic features were platelet count, haemoglobin levels, duration of morning stiffness, age, and presence of anti-cyclic citrullinated peptides antibody (anti-CCP). These 100 features were then trained on the six ML models, and exhibited improved predictive performance as compared to using haplotypes alone, with cross-validation AUCs between 0.822 and 0.906, sensitivity between 0.744 and 0.817 and specificity between 0.766 and 0.866 (Fig. 4). In addition to the improvements in predictive performances, there was a 16.7% reduction in the number of features, from 120 to 100, compared to the analyses focused on pfcHaps only. As such, these 100 features were noted to be the best predictors and were tested in an unseen test set. Significantly, the robustness of these 100 features to predict MTX in the unseen dataset was evident from the good predictive performance with AUCs between 0.775 and 0.828, sensitivity between 0.656 and 0.813 and specificity between 0.684 and 0.868 (Fig. 5) across all six ML models.

Majority of SNPs within selected haplotypes are associated with changes in gene expression
Of the 100 features determined to be the best predictors, 95 were pfcHaps derived from 142 unique SNPs. 93.0% [132] of these SNPs are eQTL SNPs[54,55] (Table 1, black box). Approximately 40.8% (58) are non-synonymous while 59.2%(84) are synonymous SNPs. Majority(45) of the non-synonymous SNPs are predicted to be benign, while 5 and 8 SNPs are predicted to be possibly damaging and deleterious, respectively. 18.3% (26) of the 142 SNPs are also predicted to potentially alter transcription factor binding sites while 2.1%(3) can potentially affect miRNA binding sites and 52.1%(74) can potentially modify ESE/ESS sites. The SNPs and their potential functions are summarised in Table 1.

Genes of the 95 predictive haplotypes are involved in a variety of pathways
To gain mechanistic insights into these 95 predictive pfcHaps, a pathway enrichment analysis was performed. Genes of these pfcHaps are enriched in diverse pathways, including Rho activation of PAKs, ROCKs, and CITs, complement and coagulation cascade, beta-2 cell surface interactions, ion channel transport, transcriptional activation of mitochondrial biogenesis and rheumatoid arthritis (Figure S1).

Discussion
The current study aims to identify biologically meaningful features that are predictive for MTX response in RA patients. Functional coding haplotypes (pfcHaps)
Figure 2. Predictive performance of 120 haplotypes (from Haplotype-only analysis) in the training set using 5-fold cross-validation. ROC curves of 120 haplotypes using (a) Random Forest, (b) Logistic Regression, (c) Support Vector Machine, (d) Boosted Trees, (e) Elastic Net, and (f) Neural Network.
present as biologically meaningful predictors and provide us with an opportunity to better understand the combined role that multiple SNPs within a single functional unit have in influencing the function of genes to explain the variation in phenotypes between individuals. Interrogation at a haplotype level has the potential to alleviate problems commonly faced in current genome wide association studies (GWAS) involving large number of SNPs which probe tag-SNPs that are in linkage disequilibrium with the causal SNP but are not necessarily the causal SNPs. More importantly, by focusing on haplotypes we have greatly reduced the complexity of the study and partially mitigated the curse of dimensionality. As such, this approach of feature handcrafting can be viewed as a biologically meaningful, genetic dimensionality reduction strategy. We supplemented the genetic dimensionality reduction strategy with ML feature elimination/selection using RFECV with Random Forest classifiers. Random Forest classifiers were chosen for our study due to its robustness to over-fitting, its computational efficiency, and the provision of feature importance scores. When coupled with RFECV, Random Forest is able to identify a small subset of features that produces the highest accuracy in the specific classifier model. To ensure that the selected features are stable and robust to differences in sample size, RFECV feature selection was performed on eight random subsets of training data with variable sample sizes. Only features that are common across all eight random, variable sized training subsets were selected for further evaluation of their predictive performance.

Overall, our proposed strategy allowed us to overcome issues of redundancy and irrelevancy of information that are commonly faced when handling high dimensional data which would have led to reduced efficiency and accuracy of ML models trained.

Our strategy and the incorporation of non-genetic factors identified 95 coding haplotypes and 5 non-genetic factors for training. Notably, the identified features displayed good predictive performance classifying the MTX response of 70 patients whose data had been placed aside as the unseen test dataset. The overall performance (AUC, sensitivity, specificity) of the unseen test set suggested that the performance obtained during
Figure 4. Predictive performance of 95 haplotypes and 5 non-genetic factors in the training set using 5-fold cross-validation.

ROC curves of 95 haplotypes and 5 non-genetic factors using (a) Random Forest, (b) Logistic Regression, (c) Support Vector Machine, (d) Boosted Trees, (e) Elastic Net, and (f) Neural Network.
Figure 5. Predictive performance of 95 haplotypes and 5 non-genetic factors in the unseen test set.

ROC curves of 95 haplotypes and 5 non-genetic factors using (a) Random Forest, (b) Logistic Regression, (c) Support Vector Machine, (d) Boosted Trees, (e) Elastic Net, and (f) Neural Network.
| Gene | Haplotype | mChromosome | Start | End | Genomic Interval | Strand | TF | TF binding | mChIP Binding | mRNA Binding | AA Change |
|------|-----------|-------------|-------|-----|-----------------|-------|----|------------|---------------|--------------|-----------|
| SOX8  | SOX8_1.0  |            |       |     |                 |       |    |            |               |              |           |
| ZFRA1 | TFRA1_0.0 |            |       |     |                 |       |    |            |               |              |           |
| NKX2.5 | NKX2.5_0.0 |            |       |     |                 |       |    |            |               |              |           |
| HNF4A | HNF4A_0.0  |            |       |     |                 |       |    |            |               |              |           |
| HNF4B | HNF4B_1.0  |            |       |     |                 |       |    |            |               |              |           |
| LMX1  | LMX1_1.0   |            |       |     |                 |       |    |            |               |              |           |
| LMX2  | LMX2_1.0   |            |       |     |                 |       |    |            |               |              |           |
| SRY   | SRY_0.0    |            |       |     |                 |       |    |            |               |              |           |
| UBCD1 | UBCD1_1.0  |            |       |     |                 |       |    |            |               |              |           |
| ARAP1 | ARAP1_1.0  |            |       |     |                 |       |    |            |               |              |           |
| IL27  | IL27_0.0   |            |       |     |                 |       |    |            |               |              |           |
| IL6R  | IL6R_1.0   |            |       |     |                 |       |    |            |               |              |           |
| YWHAU | YWHAU_1.0  |            |       |     |                 |       |    |            |               |              |           |
| YWHAE | YWHAE_2.0  |            |       |     |                 |       |    |            |               |              |           |
| YWHAZ | YWHAZ_3.0  |            |       |     |                 |       |    |            |               |              |           |

Table 1 (Continued)
| Article ID | Description | Expression | Fold Change | Significance | Notes |
|-----------|-------------|------------|-------------|--------------|-------|
| MAXIP4 | MAXIP4_1.0 | 0.0073 | Y | 6 | p53, GATA-1, GATA-2, GATA-3 |
| MAXIP4 | MAXIP4_2.0 | 0.0068 | Y | 2 | |
| MAXIP4 | MAXIP4_3.0 | 0.0062 | Y | 1 | |
| MAXIP4 | MAXIP4_4.0 | 0.0053 | Y | 2 | |
| MAXIP4 | MAXIP4_5.0 | 0.0053 | Y | 3 | |
| MAXIP4 | MAXIP4_6.0 | 0.0053 | Y | 2 | |
| MAXIP4 | MAXIP4_7.0 | 0.0053 | Y | 1 | |
| MAXIP4 | MAXIP4_8.0 | 0.0053 | Y | 1 | |
| MAXIP4 | MAXIP4_9.0 | 0.0053 | Y | 2 | |
| MAXIP4 | MAXIP4_10.0 | 0.0053 | Y | 3 | |
classifier training was not simply a case of overfitting and must have had some genuine ability to distinguish MTX response. Furthermore, the observed robustness of our features across diverse ML models indicated a non-dependency of ML strategy for effective classification prediction.

The 5 non-genetic predictive features were previously reported to be linked to RA/MTX. Platelet count, haemoglobin value, duration of morning stiffness and presence of anti-CCP are well-known markers for both the diagnosis of RA and the evaluation of disease activity.\(^6^2\)\(^{-}\)\(^{-}\)\(^6^5\) Conversely, a patient’s age is also generally an underlying factors influencing their response to drugs due to changes in pharmacokinetics and pharmacodynamics with age.\(^6^6\)

We further investigated the significance of the 142 non-reference SNPs within these 95 predictive coding haplotypes (Table 1). Curiously, although coding regions of the genes, which should affect structure and function, were interrogated, majority (93.0%) of these SNPs were previously predicted as eQTLs (i.e., associated with changes in gene expression) suggesting that perhaps even polymorphisms within coding regions may influence gene expression. Of note, among the non-

| Non-Genetic Features | Value |
|----------------------|-------|
| Platelet Count (X)   | 0.1255 |
| Haemoglobin Value (g/dL) | 0.0784 |
| Duration of Morning Stiffness (hrs) | 0.0564 |
| Age (Years)          | 0.0298 |
| Anti-citrullinated Peptides (anti-CCP) | 0.0786 |

**Table 1:** Potential function of 142 SNPs in 95 coding haplotypes which are identified as potential predictors of methotrexate response.

---

\(^{62}\)\(^{-}\)\(^{-}\)\(^6^5\) Articles
synonymous SNPs, most are predicted to be benign with only a few predicted to be possibly damaging or have deleterious effect. Several SNPs were also predicted to alter consensus binding sites including transcription factor/miRNA binding sites as well as ESE/ESS. Taken together, these SNPs in the predictive pfcHap may alter the expression, structure and/or activity/function of the gene/gene product, either individually or through interaction with other SNPs within the same coding haplotype. Overall, the enrichment of potentially functional SNPs within the predictive pfcHaps highlights the ability of our approach to select haplotype of SNPs that are likely to be functional. Identifying predictive pfcHap as haplotype of SNPs in functional coding region offers us an opportunity to examine how multiple SNPs altering amino acids in the same protein may alter the structure or function (e.g. binding to substrates) of the protein differently from isolated SNPs. An example is the predictive pfcHap STING1_1.0 (HAQ) (Table 1) which comprises 3 non-synonymous SNPs (rs7380824 (R71H), rs78233829 (G230A), rs11554776 (R293Q)) within the STING1 (Stimulator of Interferon Gene 1) gene that was involved in both MTX and RA (Figure S2). Specifically, STING1 is implicated in the cGAS-STING pathway that contributes to inflammatory response in RA, and MTX has been reported to inhibit the activation of STING and downstream effects. Individuals harbouring the HAQ haplotype of the predictive pfcHap STING1_1.0 were previously reported to be more susceptible to viral infection and less responsive to DNA vaccines. An incomplete STING1 crystal structure (4KSY) encompassing the region of only two of the 3 polymorphisms, namely G230A (rs78233829) and R293Q (rs11554776) is available on RCSB PDB71 (www.rcsb.org) and the PyMol72 software (www.pymol.org/pymol) was employed to predict the potential changes in STING1 protein structure associated with these 2 aa changes (G230A and R293Q), either alone or together. As shown in Figure S3 and Video S1 the G230A polymorphism not only potentially altered the beta strand at the site of the polymorphism (Region i) but it was also predicted to alter the alpha helix structure at a distant site (~a2653; Region ii) that is closer to the R293Q (Region ii). As these 2 polymorphisms reside within the cyclic dinucleotide binding region (aa153-aa340) of the STING1 protein,73 it is thus likely that the AQ haplotype will modulate the binding to cyclic dinucleotide differently from either of the single polymorphisms (either G230A or R293Q). Apart from this, the positive charge contributed by R293 has additionally been proposed to be important for disulfide bond formation or related modifications in relation to the neighbouring conserved C292. Thus, the absence of a positive charge from Q293 may possibly hinder functions or interactions associated with C292 that are needed for gene function. Hence, identifying predictive pfcHaps rather than isolated SNPs provides opportunities to explore interactions between SNPs within the gene (pfcHaps) and whether these interactions may modulate protein folding/structure and/or function.

To our knowledge, none of the SNPs within the pcfHaps were previously reported to be associated with MTX response in RA patients. Hence, we probed for previously reported associations with the genes of these predictive pfcHaps. Using the Python PyMed library, we performed a batch query interrogating the PubMed database using names of genes of these predictive haplotypes together with key terms including “Rheumatoid Arthritis”, “Methotrexate”, and “MTX”. Our search identified several publications where these genes were mentioned together with the key terms (Figure S2). Ten genes have been reported to be associated with either MTX or RA with 4 of them (CR1, ICAM1, MMP3, MTR) being mentioned in numerous publications (28 – 63 publications) (Figure S2). Although majority (84) of the genes of the predictive pfcHaps had not been previously reported to be associated with RA and/or MTX, our pathway enrichment analyses (Figure S1) highlighted the relevance of the predicted pfcHaps to those pathways reported to be associated with MTX and/or RA (summarized in Table S2).

Thus far, studies that examined MTX response in RA patients mainly focussed on statistical association of non-genetic factors and/or genetic variants employing either the very popular genome-wide association studies (GWAS) or gene specific association analyses interrogating specific genes in the MTX/RA pathways with MTX response. This popular classical statistical approach employs Raw P-Value Thresholding (RPVT) where a P-value is assigned to each SNP; and the inferred confidence of a variant in accounting for the phenotype in the dataset is assessed by its statistical significance via comparison to a predefined threshold that considers a balance between Type I and Type 2 errors. Furthermore, as statistical approach evaluates individual SNP independently and in parallel, it does not consider potential higher order interactions amongst SNPs and is less able to identify variants with small effects because of statistical power constraints due to excessive multiple testing. Additionally, as classical statistical approach was originally designed for datasets with limited dependent and independent variables, statistical inferences are less precise with large number of variables as observed in GWAS studies since the possible associations among the many variables also increase drastically leading to more complex relationships.
On the other hand, the Machine Learning (ML) approach employed in this study is particularly suited for dealing with rich, unwieldy, ‘wide’ data where the independent variables (e.g. SNPs) exceeds the number of samples, since it ‘makes minimal assumptions about the data-generating systems’ and is effective even with data from less well-controlled experimental design or with ‘complicated nonlinear interactions’. Being a ‘statistic-free’ approach, type 1 error no longer poses an issue as there is no necessity to determine the population distribution or evaluate the P-value / confidence intervals or test the null hypothesis. As such, ML approach was reported to be more robust in identifying SNPs with small effects or SNP sets with more complex epistasis. While statistical approach concentrates on inference of relationship, ML focuses on making generalizable predictions using general algorithms to identify patterns in complex data based on its empirical capabilities. It does so by training different models to achieve the best predictive performance on an unseen test set which is required to minimize overfitting to the training dataset.

Nonetheless, ML approaches suffer from 2 major limitations. Similar to the statistical approach, the first limitation of the ML approach is the ‘curse of dimensionality’ where there are too many features in the dataset complicating the training of ML models. When models are trained with too many variables, they may not capture all possible combinations leading to high-variance and overfitting of the training model, resulting in the model being unable to predict accurately when less frequently occurring combinations in the test set are fed into the model. Another aspect of the ‘curse of dimensionality’, which greatly affects clustering or nearest neighbours-based ML methods, is ‘distance concentration’ with convergence of all pairwise distances between different samples/points as the number of variables increases, leading to difficulty in clustering high dimension data. Several machine learning tools of feature selection / reduction / extraction have been developed to mitigate this ‘curse of dimensionality’. In this study, we included a genetic approach to dimensionality reduction by focusing on haplotypes of coding SNPs, which greatly reduced the complexity from 114,000 SNPs to 52,331 pfCHaps, partially mitigating the curse of dimensionality, before ML feature selection approach was employed to further reduce the dimensionality and identify features for model training. Furthermore, focusing on haplotypes of coding SNPs (pfCHaps) partially addresses the clustering issue of ML by clustering SNPs in the coding regions of genes in a biologically meaningful way. The second limitation of the ML approach is its low interpretability where the underlying mechanism is less clear, since ML sacrifice interpretability for predictive power, leading to less acceptance of these approaches by many biomedical scientists. In this study, by focusing on pfCHaps, the interpretability issue is also partially addressed, since the coding region of genes represents one of the most functional regions of genes that is translated into protein and are thus most likely to be biologically relevant. In fact, majority of the SNPs within the predictive pfCHaps were found to be potentially functional (e.g., altering important functional sites including transcription factor binding, splicing, microRNA, etc and/or associated with changes in gene expression and/or predicted to have deleterious effect on the protein) (Table 1). The interpretability and biological significance of the features selected through our algorithm is further highlighted by previous reports of the association of some of the genes of the predictive pfCHaps with either MTX and/or RA (Figure S2) as well as the inferred enriched pathways of some other predicted pf genes being associated with MTX/RA (Figure S1, Table S2).

In summary, our approach not only uncovered known non-genetic factors that were previously associated with MTX/RA, but also novel genetic features, namely haplotype of coding SNPs (pfCHap) that can predict MTX response in RA patients. Some of these predictive pfCHap resides in genes that were previously reported to be associated with MTX/RA or in pathways associated with MTX/RA, highlighting novel connections that have yet to be investigated, providing us with an opportunity to gain new mechanistic insights into the contributing factors that may account for differences in MTX response in RA patients. Our findings thus serve to complement currently reported non-genetic and genetic features that are associated with MTX response by providing a totally different set of effective genetic features, pfCHap, that can predict MTX response. It would be worthwhile to evaluate if these MTX predictive features that we identified will also be able to predict response to other small molecule anti-arthritic drugs and/or biological agents used for the treatment of RA.

Further optimization and sensitivity analyses of the current model can be performed to fine tune the model as well as assess alternative models in the unseen test dataset. Another worthwhile future direction is to focus on examining whole genome sequencing to identify more predictive haplotypes in all functional regions, including the promoter, intron, 3’ and 5’ UTRs, non-coding regions, etc, to further improve the predictive performance in unseen data, not only in other patients with similar profiles but also in patients of different profiles (e.g., different ethnicity, etc) so that the predictive features identified can be generalizable. The effectiveness of other ML feature engineering algorithms that have been successfully applied in biomedical studies could also be explored.

While further validation and refinement of these predictors will be necessary, we can remain hopeful that these predictors have the potential to be used in clinical practice to facilitate decision-making for the treatment
of RA patients. This approach may also be applicable for
the identification of predictive features associated with
response to other drugs and even disease susceptibility
or traits.

Declaration of Competing Interest
CGL, KPL, CCK, SSC, AJWL, and LJL declare that they
have a pending Coversheet IP application.

Contributors
CGL and KPL conceived the study, directed the
research, and edited the manuscript. AJWL, LJL, and
BNSO designed and performed the bioinformatics anal-
ysis. AJWL drafted the manuscript. ETK, JWLT,
TRASG, KPL provided the clinical samples and clinical
insights. CCK analysed the sequencing data and pro-
duced scientific insights. SSC and LTK provided sci-
tific insights and edited the manuscript. All authors
read and approved the final manuscript.

Acknowledgments
This work was supported by grants from the Singapore
Ministry of Health’s National Medical Research Council
(NMRC) [NMRC/CBRG/0095/2015 (to NCC),
CG12Aug17 (to TTSH), CGAug16Mo12 (to TTSH) and
NMRC/C117/2013 (to TTSH)]; National Cancer Cen-
ter Research Fund and block funding Duke-NUS Medi-
cal School to A/P Caroline G.L. LEE.; Singapore
Ministry of Education Academic Research Fund Tier 2 grant
MOE2019-T2-1-138 to LTK. The funders had no role in
study design, data collection and analysis, decision to
publish, or preparation of the manuscript.

Supplementary materials
Supplementary material associated with this article can
be found in the online version at doi:10.1016/j.
ebiom.2021.103800.

Data Sharing Statement
The data that support the findings of this study are avail-
able from the corresponding author, but restrictions
apply to the availability of these data, which were used
under the license for the current study, and so are not
publicly available. Data are however available from the
authors upon reasonable request and with permission
of the corresponding author.

References
1. Relling MV, Klein TE, CPIC. Clinical pharmacogenetics implemen-
tation consortium of the pharmacoepidemiology network research.
Clin Pharmacol Ther 2011;89:654–7. https://doi.org/10.1038/cpt.2010.276.
2. Relling MV, Klein TE, Gammal RS, Whirl-Carrillo M, Hoffman JM,
Caudle KE. The clinical pharmacogenetics implementation consorti-
um: 10 years later. Clin Pharmacol Ther 2020;107:171–5. https://doi.
3. Roden DM, McLeod HL, Relling MV, Williams MS, Mensah GA,
Peterson JF, et al. Pharmacogenomics HHS public access. Lancet
2019;394:321–3. https://doi.org/10.1016/S0140-6736(19)31276-0.
4. Poldrack RA, Huckins G, Varoquaux G. Establishment of best practi-
ces for evidence for prediction: a review. JAMA Psychiatry
2020;77:354–40. https://doi.org/10.1001/jamapsychiatry.2019.3671.
5. Varga TV, Niss K, Estampador AC, Collin CB, Moseley PL. Associa-
tion is not prediction: A landscape of confused reporting in diabetes
– A systematic review. Diabetes Res Clin Pract 2020;190:103497.
6. Goh WW, Bitt, Wong L. Dealing with confounders in omics analy-
sis. Trends Biotechnol 2018;36:488–98. https://doi.org/10.1016/j.
tibtech.2018.01.013.
7. Chattopadhyay A, Lu T-P. Gene-gene interaction: the curse of
dimensionality. Ann Transl Med 2019;7:813. https://doi.org/
10.21037/atm.2019.12.87.
8. Botta V, Louppe G, Geurts P, Wehenkel L. Exploiting SNP correla-
tions within random forest for genome-wide association studies.
PLoS One 2014;9:e93379. https://doi.org/10.1371/JOURNAL.
PONE.0093379.
9. Ching T, Himmelestein DS, Beaulieu-Jones BK, Kalinin AA, Do BT,
Way GP, et al. Opportunities and obstacles for deep learning in biol-
ogy and medicine. J R Soc Interface 2018;15. https://doi.org/
10.1098/rsif.2017.0187.
10. Kardal MA, Bay-Jensen AC, Henriksen K, Christiansen C, Gennet
HK, Chamberlain C, et al. Rheumatoid arthritis: A case for personal-
ized health care? Arthritis Care Res 2014;66:1279–80. https://doi.
.org/10.1002acr.22289.
11. Dougados M, Soubrier M, Antunez A, Balint P, Balas A, Buch MH,
et al. Prevalence of comorbidities in rheumatoid arthritis and evalua-
tion of their monitoring: results of an international, cross-sectional
study (COMORA). Ann Rheum Dis 2014;73:62–8. https://doi.org/
10.1136/annrheumdis-2013-204243.
12. Aletaha D, Smolen JS. Diagnosis and management of rheumatoid
arthritis: a review. JAMA - J Am Med Assoc 2018;320:1569–72.
https://doi.org/10.1001/jama.2018.11910.
13. Silman AJ, Pearson JE. Epidemiology and genetics of rheumatoid
arthritis. Arthritis Res 2002;4:5265–72. https://doi.org/10.1186/
art78.
14. Nikiforou E, Guh D, Bansback N, Zhang W, Dixey J, Williams P,
et al. Work disability rates in RA. Results from an inception cohort
with 24 years follow-up. Rheumatology 2012;51:385–92. https://doi.
.org/10.1093/rym/kr2401.
Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, et al. High-affinity ligand for STING. AMP containing mixed Phosphodiester linkages is an endogenous structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res* 2021;49:e437–51. https://doi.org/10.1093/nar/gkaa1058.

Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTeX) project. *Nat Genet* 2015;47:580–5. https://doi.org/10.1038/ng.3245.

Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, Herrig R. ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res* 2013;41:D712–7. https://doi.org/10.1093/nar/gks1126.

Vay S, Patel N, Turcott M, Bosse Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nat Rev Genet* 2012;13:457–62. https://doi.org/10.1038/nrg3240.

Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, Herrig R. ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res* 2013;41:D712–7. https://doi.org/10.1093/nar/gks1126.

Vay S, Patel N, Turcott M, Bosse Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nat Rev Genet* 2012;13:457–62. https://doi.org/10.1038/nrg3240.

Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, Herrig R. ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res* 2013;41:D712–7. https://doi.org/10.1093/nar/gks1126.

Vay S, Patel N, Turcott M, Bosse Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nat Rev Genet* 2012;13:457–62. https://doi.org/10.1038/nrg3240.

Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, Herrig R. ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res* 2013;41:D712–7. https://doi.org/10.1093/nar/gks1126.

Vay S, Patel N, Turcott M, Bosse Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nat Rev Genet* 2012;13:457–62. https://doi.org/10.1038/nrg3240.

Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, Herrig R. ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res* 2013;41:D712–7. https://doi.org/10.1093/nar/gks1126.

Vay S, Patel N, Turcott M, Bosse Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nat Rev Genet* 2012;13:457–62. https://doi.org/10.1038/nrg3240.

Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, Herrig R. ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res* 2013;41:D712–7. https://doi.org/10.1093/nar/gks1126.