Population Genetic Inference for a Whole Genome Korean Sample

Daniel A. Vasco[1], Zhan Ye[2], Deukhwan Lee[3], Steven J. Schrodi[1,*], Simon Lin[2]

[1]Center for Human Genetics and [2]Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, WI, USA

[3]Department of Animal Life and Environment Sciences, Hankyong National University Seokjeong-Dong 67, Ansung City, Kyeonggi-Do, 456-749 Republic of Korea

*Correspondence: Dr. Steve Schrodi, 1000 N Oak Ave, Marshfield, WI 54449, USA, E-mail: SCHRODI.STEVEN@mcrf.mfldclin.edu

Keywords
Whole-genome sequences, population genetics, bioinformatics, parameter inference, historical population expansion

Although other East Asian populations have been interrogated using high throughput sequence data, little is known about the genome-wide patterns of variation in populations residing on the Korean peninsula. Using the 38 whole genome sequences available from the Korean Personal Genome Project (KPGP), we investigated phylogenetic relationships between the samples and several standard population genetic parameters and tests. In particular we infer that a rapid demographic expansion in effective population size occurred sometime in the recent past in the Korean population.

Estimation of ancestral and current effective population size using whole genomic data is a challenging problem using next generation sequence data. Here we show how genomic data sets can be used to estimate basic population parameters such as the effective population size and population growth rate. This can be challenging for computationally intensive methods based upon the full-likelihood. Instead of using a full likelihood we use an analogous function in which the data a replaced with a vector of summary statistics. This method has been implemented in the coalescence parameter inference package EVE (Vasco, 2008). We present preliminary results here which demonstrate that this method works well on the whole genome sample obtained for the KPGP and compares favorably in accuracy with other recent methods developed for analyzing NGS samples such as jPopGen Suite (Liu, 2012).

Overall molecular relationships mirrored important sample information for the KPGP. The phylogenetic trees for each chromosome demonstrated three outliers that are known to have European ancestry and recapitulated the monozygotic and dizygotic twin relationships in the sample set (see Figure 1, which shows a UPGMA tree for Chromosome 1 of the full sample).

In order to perform the population genetic analysis, SNPs were extracted from the variant calling format files for each of the samples offered by the CAMDA 2012 website. The ANNOVAR[1] is applied to generate the annotated file to obtain the final SNPs for each

sample. BEAGLE[2] was used to phase the genotypes extracted for each of the samples chromosome by chromosome. Figure 2 shows the distributions of the number of genotypes are shared among all the samples by chromosome. Only shared genotypes among all the samples are considered since our plan is to use the phased haplotypes estimated from all the samples to infer the population genetics properties of the KPGP samples. We did notice that there was one European and two admixture individuals within the KPGP sample, so that parameter estimation was performed excluding these individuals. All results shown here were obtained only from one strand of a phased pair of chromosomes.

Parameter estimation of important population genetic factors was performed using the phased KPGP sequence data. The critical parameter $\theta = 4N_e\mu$; where N_e is the effective population size and μ is the per site per generation scaled mutation rate, was estimated using several three constant population size estimators Watterson's, Tajima's and Fu's Blue (Wakeley, 2009). EVE fits the model of an exponentially population to whole genome sequence data (Vasco, 2008). This can be expressed as $N(t)=N_0e^{gt}$ where g is the growth rate and N_0 is the initial effective population size (i.e. the effective size at the time the sample was taken). In using the coalescent approach one looks backwards in time at N(-t) and from this vantage it appears that the population exponentially declines in size as the population experiences coalescence towards its most recent common ancestor. Therefore all EVE estimates of theta in per site substitutions are scaled in these units and allow taking into account demographic factors affecting the sequence data such as growth rate and change in N over time as individuals in the population evolved backwards in time to a most recent common ancestor. When solving for mutation rate in a rapidly expanding population one must explicitly take this process into account when interpreting the estimated parameters. For example, mutation rate is equal to theta divided by 4N(t) where N(t) must be estimated from data. Since N(t) may itself be quite large, theta itself becomes a function of t and expands as well. From this point of view the per site estimate of theta must keep expanding in order to scale properly with a fixed and small mutation rate. This inflation property for the per site EVE estimate of theta will actually be seen in the estimates and is an expected property of population undergoing a rapid expansion in its effective size. This likely to be a fundamental computational problem in interpreting population genetics estimates obtained from the 1000 genomes project data since it is thought that in recent human history a rapid population expansion in effective size occurred when humans first migrated out of Africa (Vasco, 2008; Wakeley, 2009).

The results from the parameter estimates may be summarized as follows:

1. There exists a direct relationship between average pairwise distance computed using a coalescent tree (computed here using EVE) and average pairwise distance computed using pairwise alignments graphed on the x-axis (obtained using jPopGen Suite). This is shown in Figure 3. Thus the genealogy of the Korean sample contains the complete information from the pairwise alignments and therefore maybe used to infer parameters using coalescent genealogies reconstructed from the aligned sequence data.

2. Figure 4 shows of per site theta (x-axis) and growth rate estimates (y-axis and scaled in 2N generations). There are three clusters. The first cluster to the far left are theta estimates obtained using Tajima's estimator using pairwise alignments from jPopGen Suite. The middle cluster represents theta estimates obtained using Watterson's estimator but from coalescent tree reconstructions from EVE. The theta estimates to the far right were obtained from EVE and represent a fit of an exponentially growing model to the whole genomes. It is clear that there exists substantial signal demonstrating a rapid change in effective population size in the Korean population. This is observed from the large growth rate estimates (scaled in 2N generations).
3. Figure 5 shows the relationship the EVE estimate of population growth rate and Tajima's D statistic, one of the most commonly used tests of selective neutrality. A negative Tajima's D statistic is a standard indicator of a rapid population expansion therefore it is reasonable that there should exist a strong correlation between estimated growth rate and Tajima's D as shown in the graph. We also found that compared to other populations studied the Korean samples exhibited an increased/decreased skewing of the site frequency spectrum towards rare variants (results not shown).

Future computational work involves incorporating sequence error models and analysis of bias. This will lead to a new class of fast and reliable bias-corrected coalescent estimators for efficient computational population genetic analysis of next generation sequence data. In future we will incorporate estimation of recombination, migration and selection using error models sequence data. Finally we are working on the incorporation of CNV and indels into coalescent estimators and using these to develop statistical tests of neutrality and parameter inference.

References

1. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res, 2010 sep;39(16):e164
2. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007 Nov;81(5):1084-97.
3. Wakeley, J. 2009. Coalescent Theory. Roberts and Co. Greenwood Village, Co.
4. Vasco, D.A. 2008. A fast and reliable computational method for estimating population genetic parameters. Genetics 179:951-963.
5. Liu, X., Y-X Fu, T.J. Maxwell and E. Boerwinkle. 2010. Estimating population genetic parameters and comparing model goodness-of-fit using DNA sequences with error. Genome Research.
6. Kang, C-J. and P. Marjoram. 2011. Inference of population mutation rate and detection of segregating sites from next-generation sequence data. Genetics
7. Liu, Xiaoming. 2012. jPopGen Suite: population genetic analysis of DNA

polymorphism from nucleotide sequences with errors. Methods in Ecology and Evolution
8. Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437-460.
9. Watterson, G.A. 1975. Number of segregating sites in genetic models without recombination. Theoretical Population Biology 7:256-276.
10. Hudson, R.R. 1990. Gene genealogies and the coalescent process. Oxford Surveys in Evolutionary Biology.

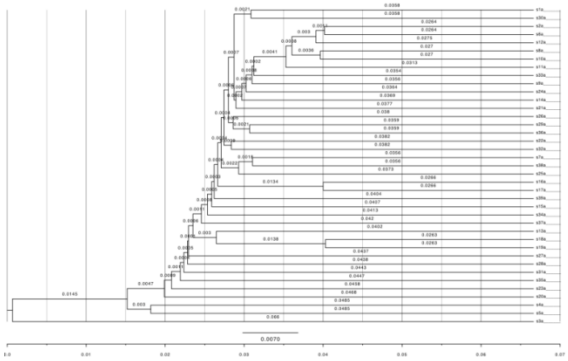

Figure 1. All phylogenetic trees for each chromosome (not shown here), such as this UPGMA tree which is shown for Chromosome 1, demonstrate that three outliers exist for the KPGP sample. One these is known to have European ancestry (labeled s3a here). Two others exhibit substantial admixture (labeled s4a and s5a here). The phylogenetics trees also demonstrate and recapitulate the monozygotic and dizygotic twin relationships in the sample set (not shown here except in Chromosome 1). The branch lengths are scaled in substitutions per site.

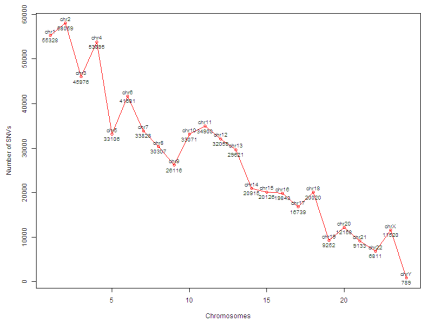

Figure 2. Distribution of the number of SNV shared among all the samples by chromosome.

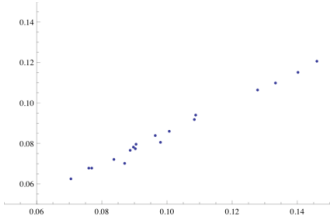

Figure 3. Graph showing the relationship between average pairwise distance computed using a coalescent tree reconstruction (such as shown in Figure 1), graphed on the y-axis obtained using EVE, and average pairwise distance computed using pairwise alignments graphed on the x-axis obtained using jPopGen Suite. This figure shows that the genealogy of the Korean sample contains the complete information from the pairwise alignments and therefore maybe used to infer parameters using coalescent genealogies reconstructed from the aligned sequence data.

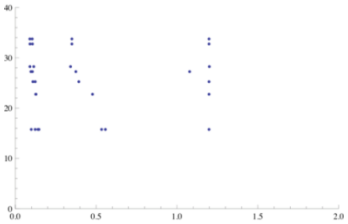

Figure 4. Graph of per site theta (x-axis) and growth rate estimates (y-axis and scaled in 2N generations). There are three clusters. The first cluster to the far left are theta estimates obtained using Tajima's and Fu's Blue estimator using jPopGen Suite. The middle cluster represents theta estimates obtained using Watterson's estimator but from coalescent tree reconstructions from EVE. The theta estimates to the far right were obtained from EVE and represent a fit of an exponentially growing model to the whole genomes. All EVE estimates of theta are scaled in $2N_0$ generations and obtained by fitting an exponential growth for change in effective size. While EVE estimates of theta in per site substitutions in the graph appear quite large these are scaled in these units representing rates of change in effective size and the least squares computational method automatically adjusts for growth rate and change in N over time as signal in the data since theta changes along side of N during the computation (see Vasco 2008 for more detail the mechanics of the estimation algorithm). Therefore when solving for mutation rate from these per site esimates for one must explicitly take this into account from the estimated parameters in terms of N(t) not a constant valued N. See text for a more detailed explanation.

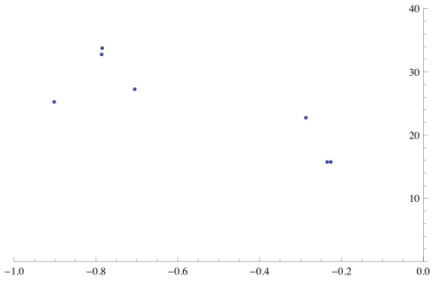

Figure 5. This graph shows the relationship the EVE estimate of population growth rate (y-axis) and Tajima's D statistic (x-axis).

Rare Haplotypes in the Korean Population

Sepp Hochreiter, Günter Klambauer, Gundula Povysil, Djork-Arné Clevert

Institute of Bioinformatics, Johannes Kepler University Linz, Austria

Knowledge of the haplotype structure of the human genome would improve genotype calls, increase the power of association studies, and shed light on the evolutionary history of humans. Common haplotypes are found by regions of linkage disequilibrium (LD) in genotype data. The advent of new sequencing technologies also facilitates the identification of rare haplotypes. However, LD-related methods fail to extract rare haplotypes because of the high variance of LD measures. Rare haplotypes can be inferred by a region of identity by descent (IBD) in two individuals. However, IBD detection methods require sufficiently long IBD regions to avoid high false positive rates and are computationally expensive as they consider all pairs of individuals. We propose identifying rare haplotypes by HapFABIA which uses biclustering to combine LD information across individuals and IBD information along the chromosome. HapFABIA significantly outperformed IBD methods at detecting rare haplotypes on simulated genotype data with implanted rare haplotypes.

To identify rare haplotypes in the Korean population, we applied HapFABIA to data from the Korean Personal Genome Project (KPGP) supplied via Critical Assessment of Massive Data Analysis (CAMDA). Genotyping data from the KPGP was combined with those from the 1000-Genomes-Project leading to 1,131 individuals and 3.1 million single nucleotide variants (SNVs) on chromosome 1 – we only analyzed chromosome 1 to comply with the Ft. Lauderdale agreement for the use of unpublished data for method development. For biclustering such large data sets, we developed a sparse matrix algebra for the FABIA biclustering algorithm.

HapFABIA identified 113,963 different rare haplotypes marked by tagSNVs that have a minor allele frequency of 5% or less. The rare haplotypes comprise 680,904 SNVs; that is 36.1% of the rare variants and 21.5% of all variants. The vast majority of 107,473 haplotypes is found in Africans, while only 9,554 and 6,933 are found in Europeans and Asians, respectively.

HapFABIA revealed a large number of genotyping errors in the KPGP data (e.g. Figure 3). The KPGP data comprises two twin pairs and a large Korean family that contains a Caucasian female from US. In particular, genotyping errors are found as SNV disagreements at twin haplotypes (e.g. Figure 5) and by haplotypes that were observed exclusively in KPGP samples including the Caucasian female (e.g. Figure 4). We corrected for these genotyping errors by removing haplotypes that are observed in just one population and removing all relations between individuals according to the pedigree information.

We characterized haplotypes by matching with archaic genomes. Haplotypes that match the Denisova or the Neandertal genome are significantly more often observed in Asians and Europeans. Interestingly, haplotypes matching the Denisova or the Neandertal genome are also found, in some cases exclusively, in Africans. Our findings indicate that the majority of rare haplotypes from chromosome 1 are ancient and are from times before humans migrated out of Africa.

The enrichment of Neandertal haplotypes in Koreans (odds ratio 10.6 of Fisher's exact test) is not as high as for Han Chinese from Beijing, Han Chinese from South, and Japanese (odds ratios 23.9, 19.1, 22.7 of Fisher's exact test) – see also Figure 7. In contrast to these results, the enrichment of Denisova haplotypes in Koreans (odds ratio 36.7 of Fisher's exact test) is is higher than for Han Chinese from Beijing, Han Chinese from South, and Japanese (odds ratios 7.6, 6.9, 7.0 of Fisher's exact test) – see also Figure 6 and examples in Figure 1 and Figure 2.

Data Analysis steps:

1. Combine the vcf genotyping data from KPGP with those from the 1000-Genomes-Project (3.1 million SNVs on chromosome 1 of 1,134 individuals – vcftools, samtools)
2. Remove common and private SNVs
3. Transform the genotyping data into the sparse matrix format of HapFABIA
4. Apply HapFABIA to extract haplotypes
5. Base calling of Denisova and Neandertal genome at the SNV positions of KPGP and 1000-Genomes-Project
6. Analyze and annotate the haplotypes

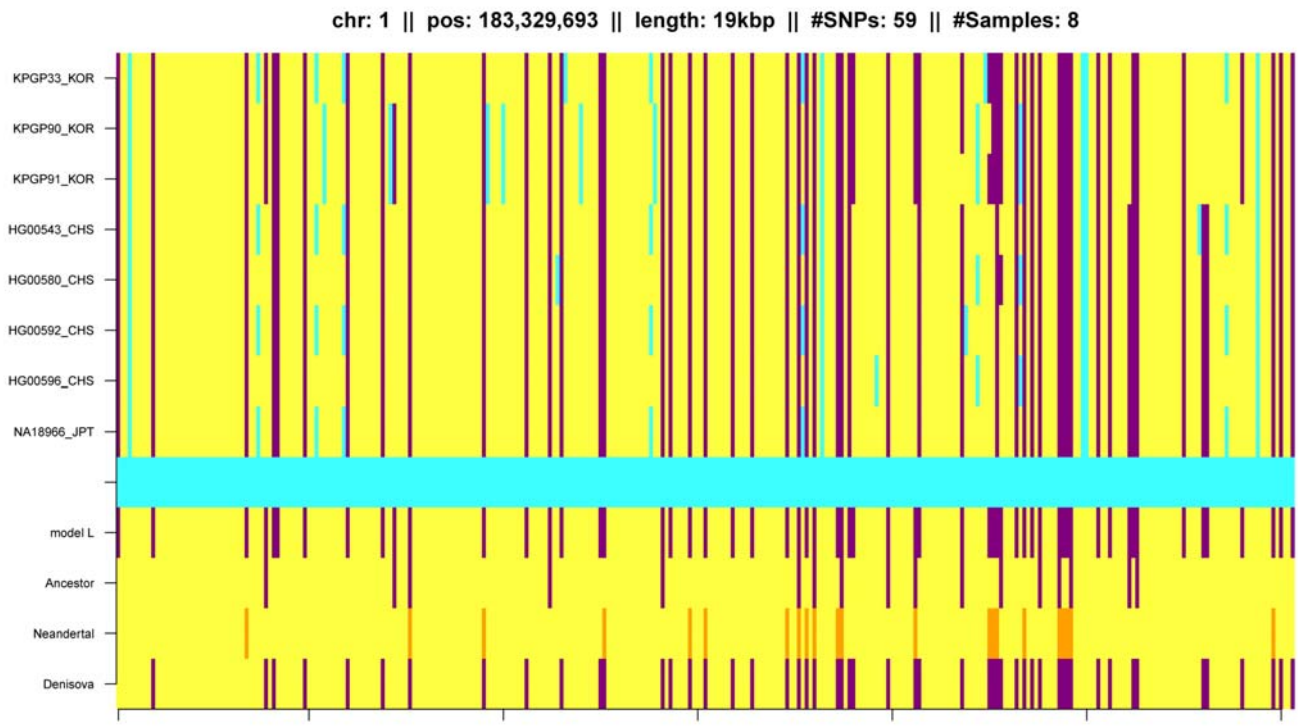

Figure 1: Example of a haplotype matching the Denisova genome found exclusively in Asians including Koreans. The y-axis gives all chromosomes that have the haplotype and the x-axis consecutive SNVs/Indels/SVs. Major alleles are shown in yellow, minor alleles of tagSNVs in violet, and minor alleles of other SNVs in cyan. The row labeled "model L" indicates tagSNVs identified by HapFABIA in violet. The rows "Ancestor", "Neandertal", and "Denisova" show bases of the respective genomes in violet if they match the minor allele of the tagSNVs (in yellow otherwise). Missing Neandertal tagSNV bases are shown in orange.

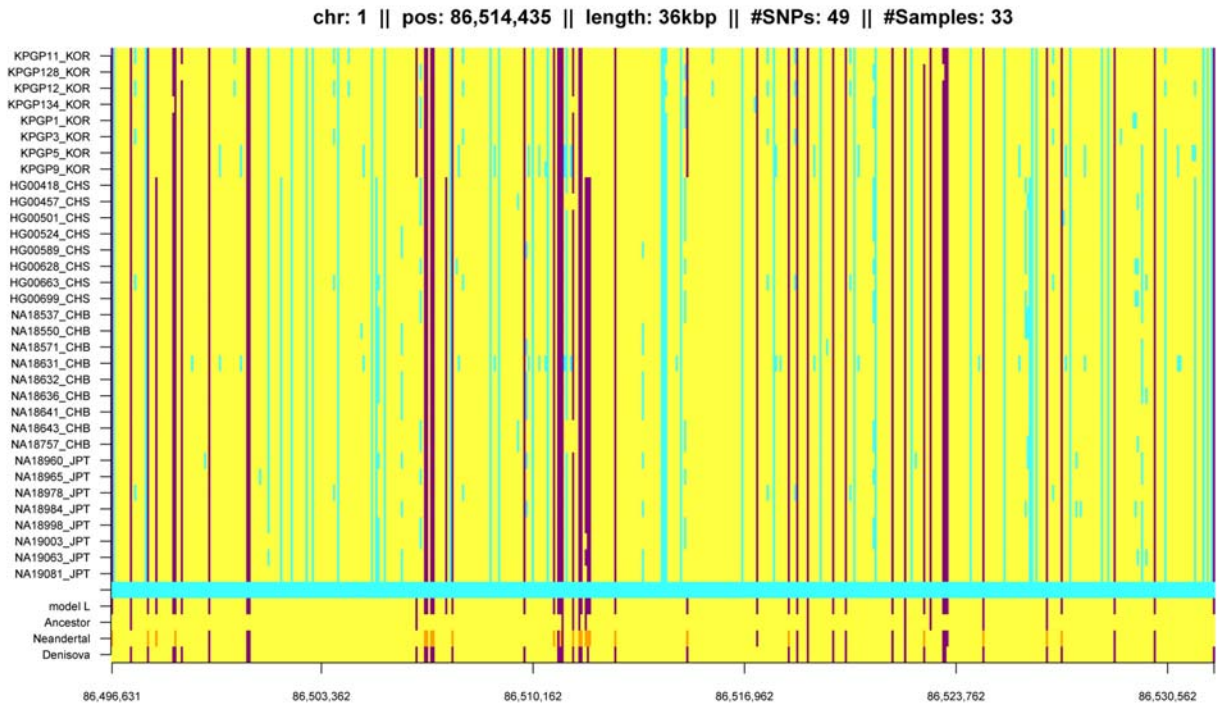

Figure 2: Another example of a haplotype matching the Denisova genome including Koreans.

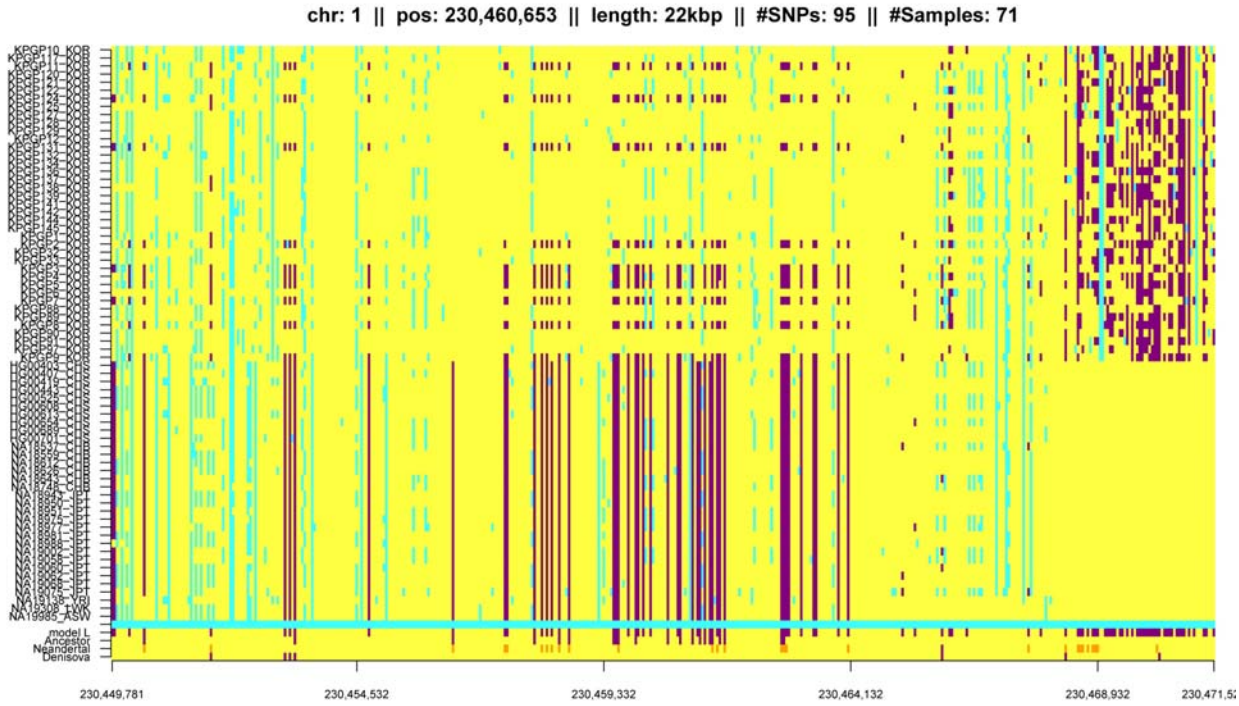

Figure 3: Example of a haplotype that contains genotyping errors (right border).

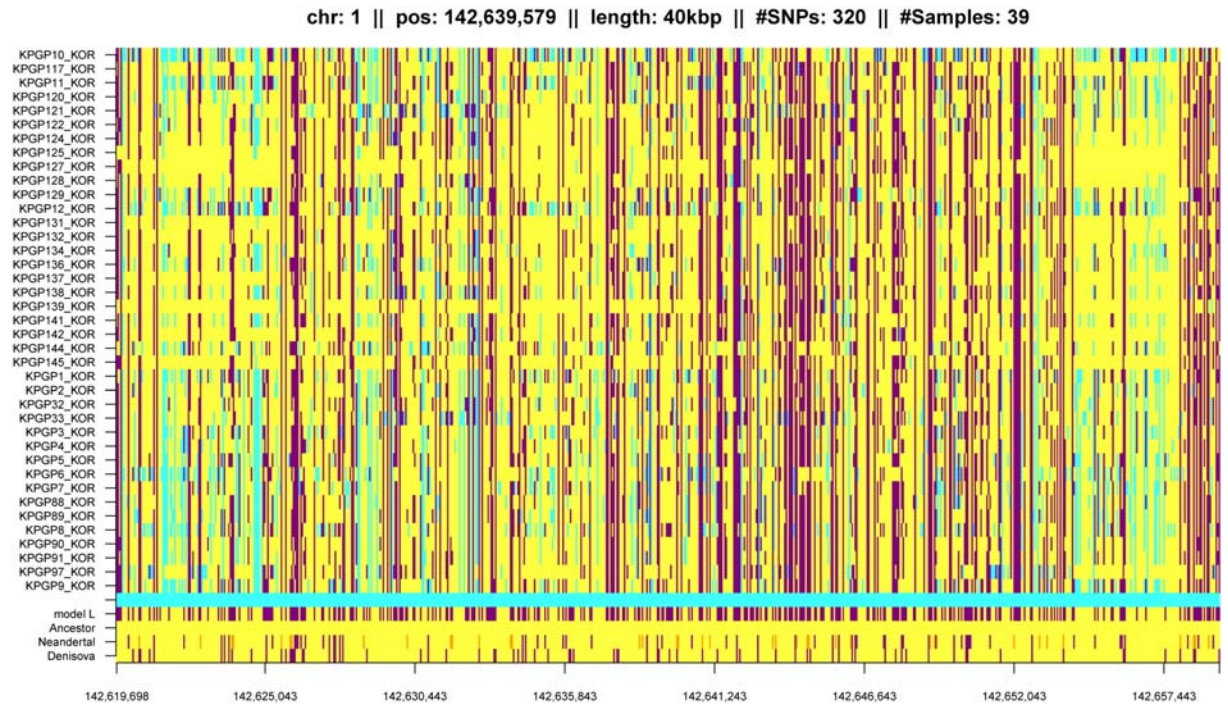

Figure 4: Example of a haplotype representing genotyping errors. The haplotype is exclusively observed in KPGP samples, however KPGP10 is an US American Caucasian female sequenced by KPGP. Many tagSNVs are inconsistent across samples.

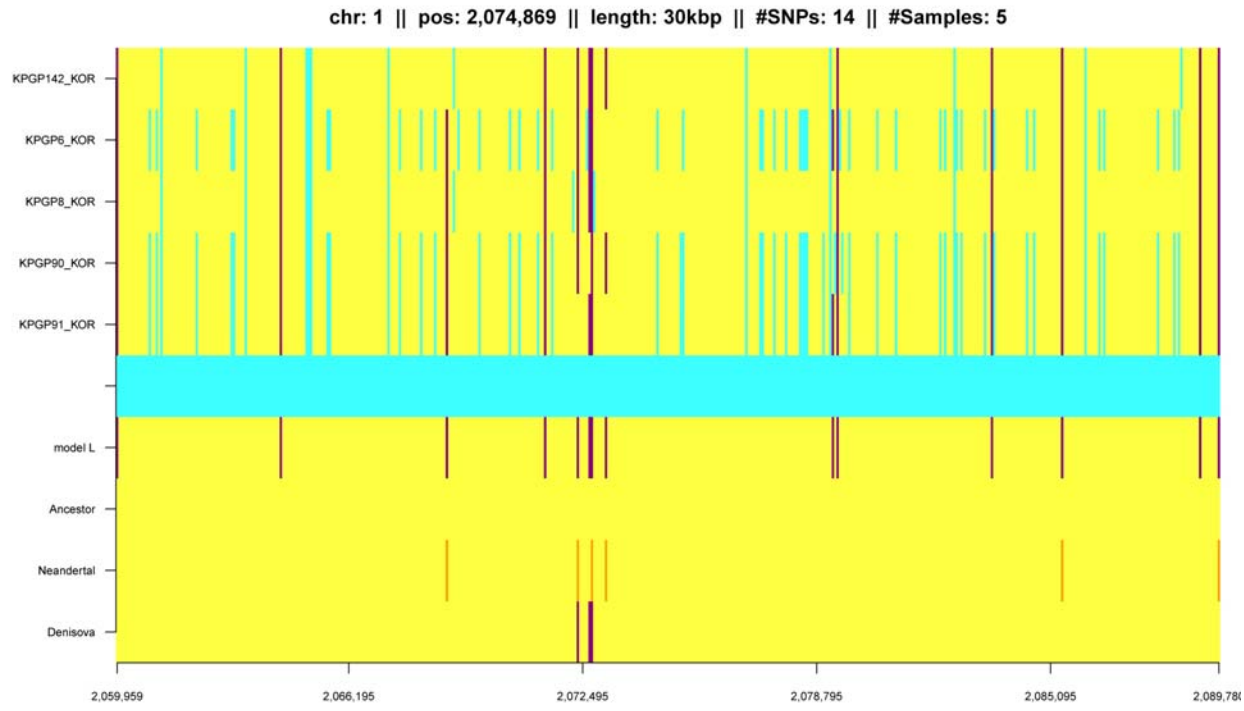

Figure 5: Example of a haplotype possessed by the twins KPGP90 and KPGP91. Differences in the twins are presumably genotyping errors.

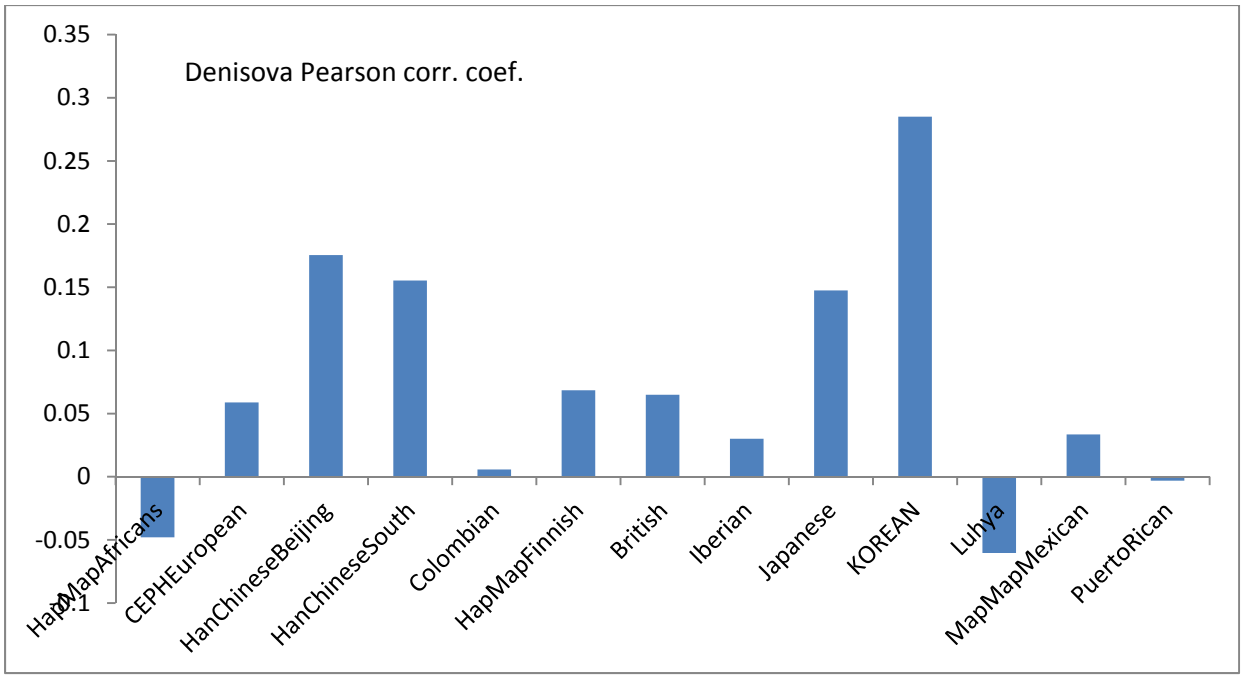

Figure 6: Persons correlation coefficient between Denisovian SNVs and subpopulations across all haplotypes of chromosome 1.

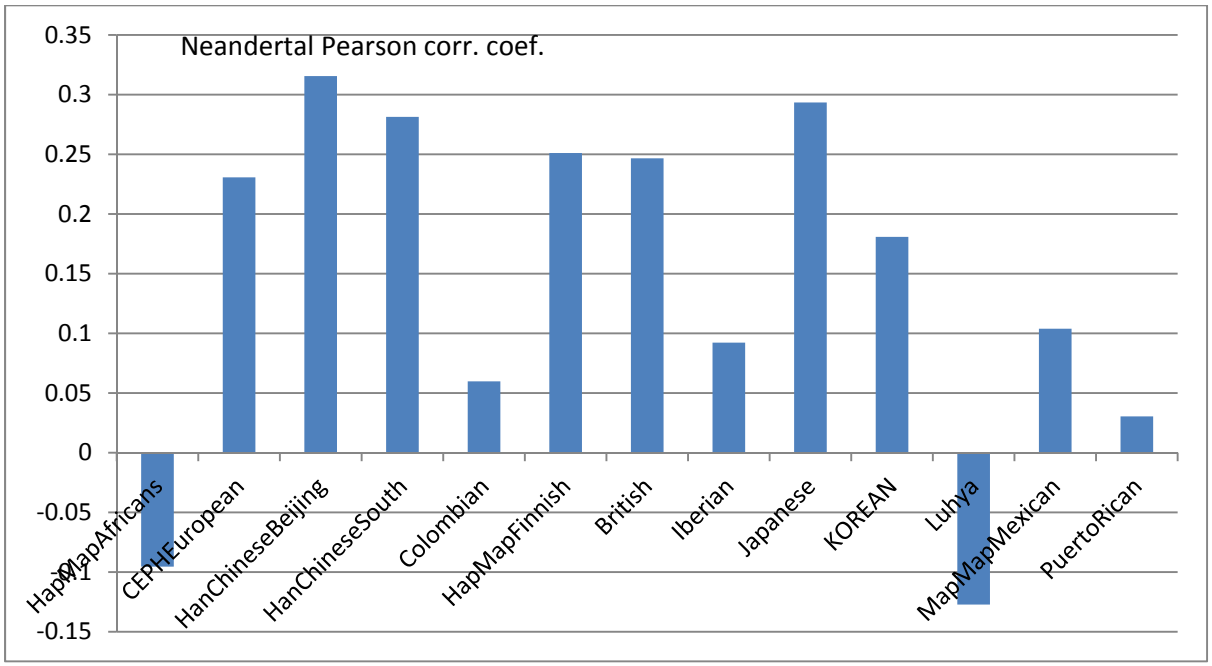

Figure 7: Persons correlation coefficient between Neandertal SNVs and subpopulations across all haplotypes of chromosome 1.

Characterization and Analysis of Korean Genomes

Zhan Ye[1], Daniel A. Vasco[2], Steven J. Schrodi[2], Simon Lin[1,*]

[1]Biomedical Informatics Research Center and [2]Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, WI, USA

*Correspondence: Dr. Simon Lin, 1000 N Oak Ave, Marshfield, WI 54449, USA, E-mail: Lin.Simon@mcrf.mfldclin.edu

Keywords
Whole-genome sequences, bioinformatics, polymorphisms, genetic variants

Large-scale, next generation DNA sequencing has increasingly become commonplace in numerous scientific activities, providing vital data fueling investigations into phylogenetic relationships, evolutionary models, agricultural traits, disease risk, and response to pharmaceuticals. While whole genome sequencing offers unprecedented opportunities, substantial challenges remain in the bioinformatics, statistical analysis and interpretation of results. In this study we analyzed the 39 human genomes from the Korean Personal Genome Project (KPGP) – all sequenced on the Illumina HiSeq 2000 Platform with 30-40x coverage – and were able to characterize genome-wide genetic variation patterns in the samples including the spatial distribution and inter-individual variation of (i) single nucleotide variants, (ii) indels, and (iii) frameshift mutations.
Our analyses are based on the variant calling format (VCF) files from the CAMDA 2012 website. For each of the sample, we have separate VCF files for single nucleotide variations (SNVs) as well as the indels (INDELs). All the VCFs are processed using by ANNOVAR[1] to get the final data sets for performing our downstream analyses.

Not unexpectedly, we found selected mapping errors occurring at highly repetitive regions. For example, all female subjects in the study exhibited sequence reads that mapped to the Y chromosome. Upon further analysis it was shown that these reads did not map uniquely in the genome. We suggest that data cleaning using raw fastq sequence files are very important for the follow-up analysis.

Interestingly, a positive correlation was observed for the rate of genetic variants in each of the different variant classes. In Fig. 1, we showed, for the INDELs counts from the NGS data, it follows a pattern where the average of the counts of INDELs from sample 1 to 20 are in general higher then sample 21 to 39 from chromosome 1 to chromosome 22, for the chromosome X and chromosome Y (17 female samples and 22 male samples), the random pattern are presented. This variation might be due to the sequencing for these 39 samples are not done at the same time or even not the same places.

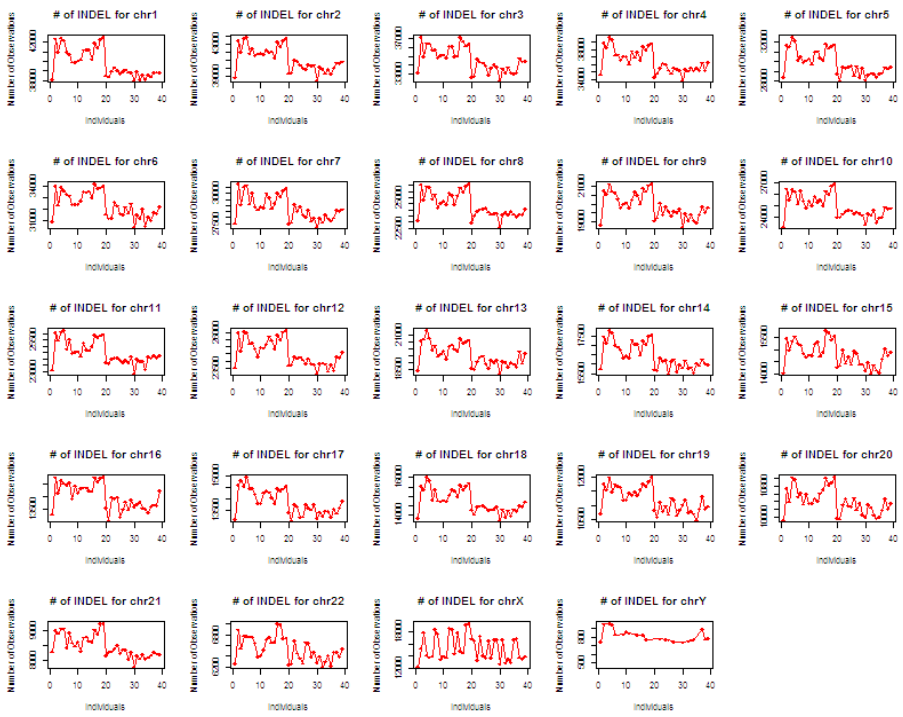

Fig. 1

In Fig. 2, we showed the graphs of variant frequency spectrum (the distribution for the number of variants having a given frequency in our sample) in each chromosome using the SNVs called. We calculated the allele frequencies for all the shared SNPs from all samples and reported then we reported the counts of SNPs which the MAF (minor allele frequency) is from 0 to 0.5. (0 out of 39*2 up to 39 out of 39*2).

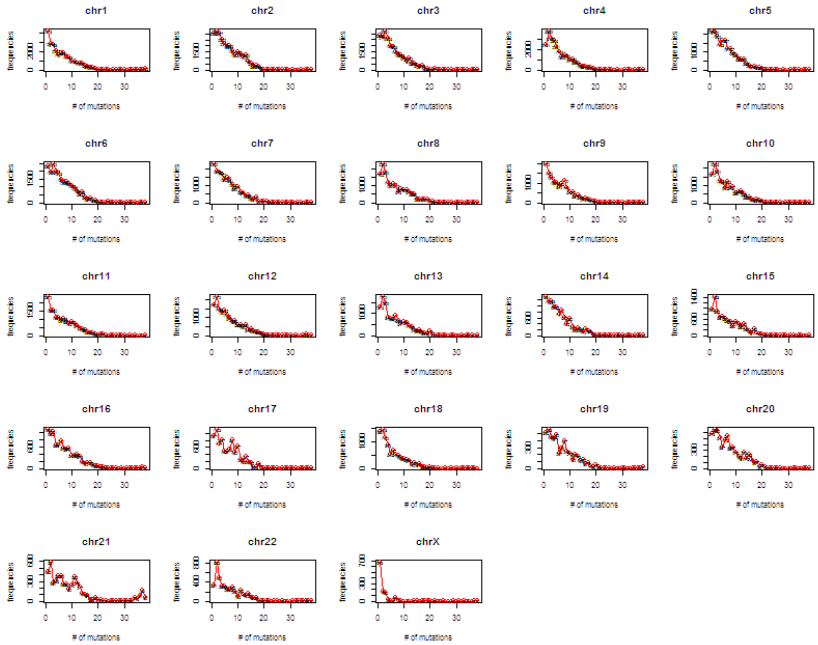

Fig. 2

In Fig. 3, SNVs called from each of the sample are summarized as the average number of the SNPs called for all samples and also the range of numbers of SNPs for each chromosome of all samples. The green lines show the average number of markers for each chromosome, where the blue and red lines represent the minimum and maximum numbers of markers among all the samples, respectively.

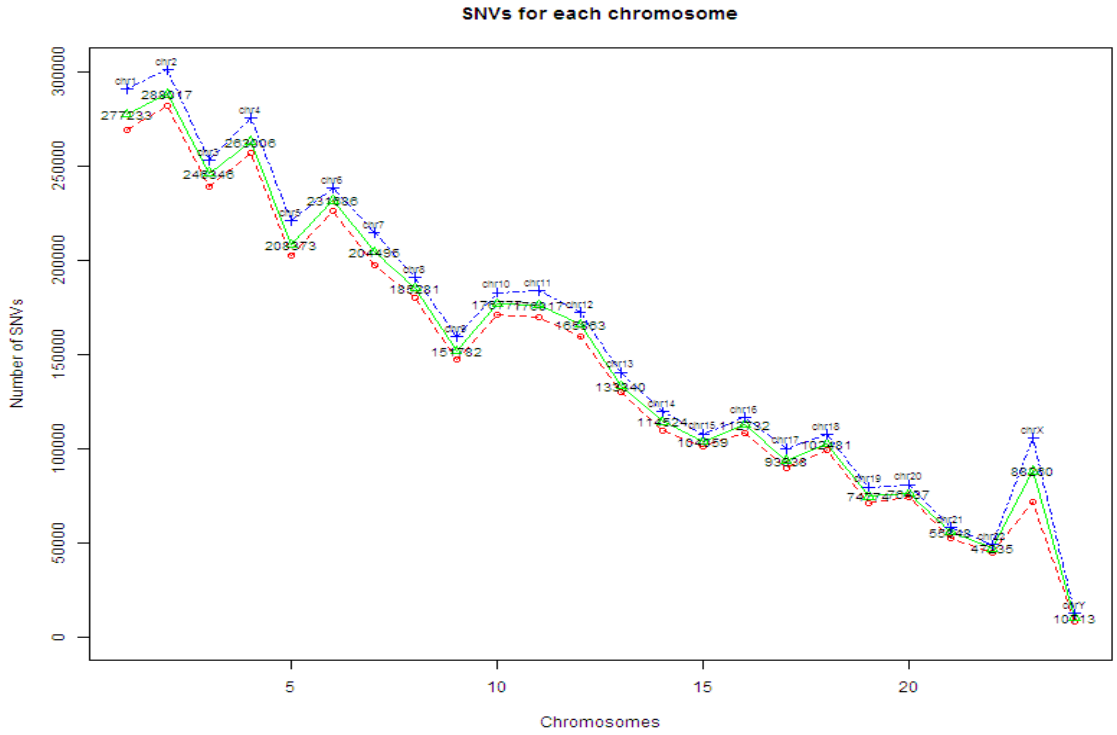

Fig. 3

In Fig. 4 and 5, we showed the graphs based on the counts of frameshift, insertions and deletions shared at least in 25% of the samples (10 samples). The red color is all the unrelated samples, the two blue samples marked as triangle are two monozygosity twins, the two green samples marked as diamond are dizygosity twins, the two light blue samples marked as cross are admixture of European with Korean samples and one black sample marked as circle is from European inheritance. All the frameshift insertions and deletions are summarized in gene levels. Further study using the quality scores as well as the read depth of the sequences would be conducted to confirm the individual insertions as well as deletions are valid. Genes found to have a high frequency of frameshift mutations in the KPGP sample set and their functional implications needs to be further discussed.

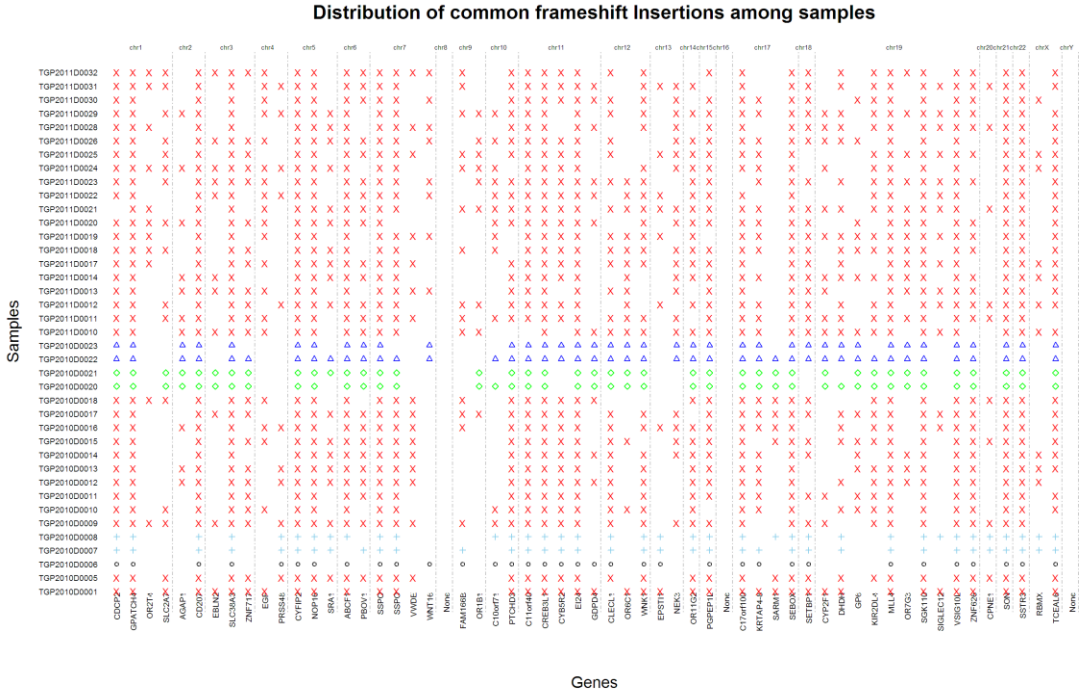
Distribution of common frameshift Insertions among samples
Samples
Genes

Fig. 4

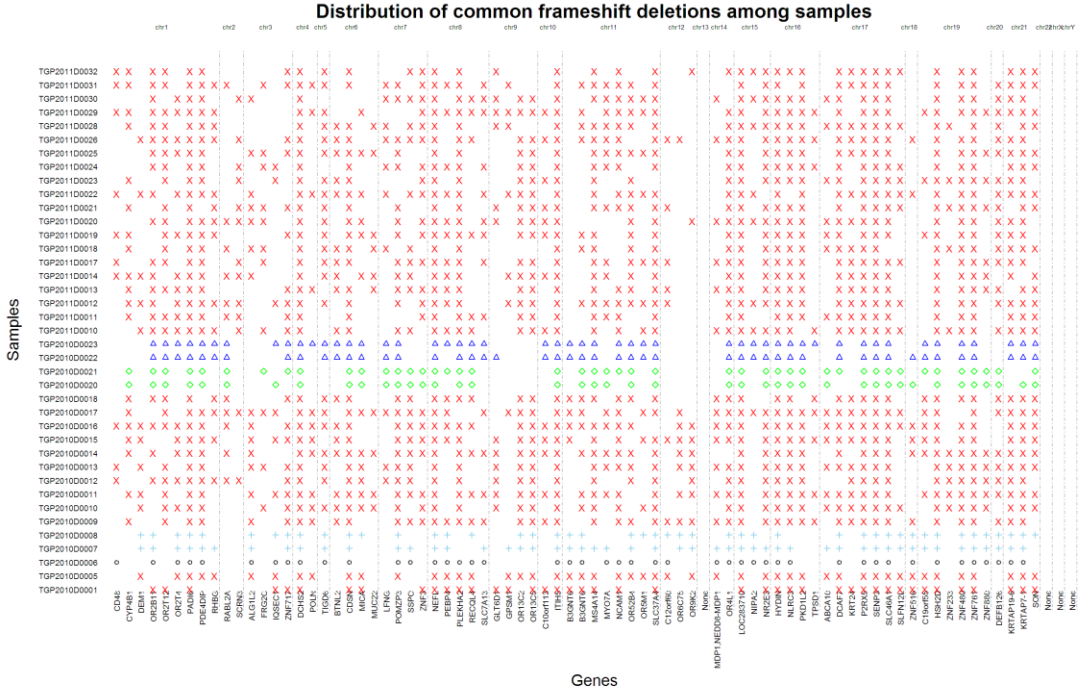
Distribution of common frameshift deletions among samples
Samples
Genes

Fig. 5

References

1. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res, 2010 sep;39(16):e164

CAGI: The Critical Assessment of Genome Interpretation, a community experiment to evaluate phenotype prediction

Susanna Repo[1,†], John Moult[2], Steven E. Brenner[1], CAGI Participants

[1] Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720. srepo@compbio.berkeley.edu, brenner@compbio.berkeley.edu
[2] IBBR, University of Maryland, Rockville, MD 20850. jmoult@umd.edu
[†]Currently at: EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK

The Critical Assessment of Genome Interpretation (CAGI, \'kā-jē\) is a community experiment to objectively assess computational methods for predicting the phenotypic impacts of genomic variation. In this assessment, participants are provided genetic variants and make predictions of resulting phenotype. These predictions are evaluated against experimental characterizations by independent assessors. The CAGI experiment culminates with a community workshop and publications to disseminate results, assess our collective ability to make accurate and meaningful phenotypic predictions, and better understand progress in the field. A long-term goal for CAGI is to improve the accuracy of phenotype and disease predictions in clinical settings.

The CAGI 2011 experiment consisted of 11 diverse challenges exploring the phenotypic consequences of genomic variation. In two challenges, CAGI predictors applied the state-of-the-art methods to identify the effects of variants in a metabolic enzyme and oncogenes. This revealed the relative strengths of each prediction approach and the necessity of customizing such methods to the individual genes in question; these challenges also offered insight into the appropriate use of such methods in basic and clinical research. CAGI also explored genome-scale data, showing unexpected successes in predicting Crohn's disease from exomes, as well as disappointing failures in using genome and transcriptome data to distinguish discordant monozygotic twins with asthma. Complementary approaches from two groups showed promising results in predicting distinct response of breast cancer cell lines to a panel of drugs. Predictors also made measurable progress in predicting a diversity of phenotypes present in the Personal Genome Project participants, as compared to the CAGI predictions from 2010.

CAGI is planned again for 2012 and we welcome participation from the community. Current information will be available at the CAGI website at http://genomeinterpretation.org.

Clinical Genomics: Genetic Prediction of Pharmacological Response to Lercanidipine and Risk of Hypertension

Zhan Ye[1], Daniel A. Vasco[2], Steven J. Schrodi[2], Simon Lin[1,*]

[1]Biomedical Informatics Research Center and [2]Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, WI, USA

*Correspondence: Dr. Simon Lin, 1000 N Oak Ave, Marshfield, WI 54449, USA, E-mail: Lin.Simon@mcrf.mfldclin.edu

Keywords
Whole-genome sequences, genetic prediction, hypertension, genetic variants

Genome-wide association studies have produced numerous marked association results, providing clues as to underlying gene networks responsible for common diseases. One of the more straightforward uses of these association data – particularly the small subset that enjoy moderate to large effect sizes – is in medical prognosis, diagnosis, and elucidation of pharmacogenetic effects. The use of sequence variants in clinical applications is not without substantial challenges, not the least of which include (i) accurate, error-free sequence data is needed for medical use, (ii) variant calling algorithms must produce reliable, replicable results, (iii) data storage and retrieval processes must be reliable and seamless for this massive data, (iv) previous studies must have produced clinically meaningful and actionable results in populations relevant to those individuals being tested, and (v) analysis techniques must be both statistically and genetically appropriate.

The Korean Personal Genome Project (KPGP) data available for the CAMDA 2012 competition consisted of 39 human genomes all sequenced on the Illumina HiSeq 2000 Platform with an average of 30-40x coverage. One individual in the KPGP was identified as being female and to have been prescribed the hypertensive pharmaceutical lercanidipine, presumably for the treatment of essential hypertension. In this study we sought to use the whole genome sequence data from this individual to address two clinical questions: First, given that a *CYP3A5*-linked polymorphism, rs776746, has been previously identified as modifying the pharmacologic response to lercanidipine, what clinically-relevant lercanidipine information can be derived from the sequence data at *CYP3A5* for this individual of interest? And second, how informative are previously identified GWAS polymorphisms for the calculation of a posterior probability of essential hypertension for this individual?

To answer the lercanidipine pharmacogenetics question, sequence information for individual TGP2010D0010 were obtained and all resulting *CYP3A5* variants
were analyzed for possible functional impact in regulation, protein function, and/or evolutionary conservation. Variants from single nucleoid variants (SNVs) as well as indels (INDELs) are generated using the variant calling format files from each of the KPGP samples obtained from CAMDA 2012 website. ANNOVAR[1] is used to generate the final annotated variants data both SNVs as well as INDELs for further study of

pharmacogenetics. Given the gene *CYP3A5*, Table 1 and 2 summarized the genetic variants detected from our NGS data.

To calculate a probability of essential hypertension in individual TGP2010D0010 given genetic factors, we identified a collection of SNPs that have been found to be significant in previous, highly-powered GWAS studies of essential hypertension and replicated in at least one additional, independent study. A handful of SNPs, obtained from 13 large-scale hypertension studies, were used as features in a predictive model for hypertension. We assumed conditional independence between SNPs and employed a Naïve Bayes modeling approach to obtain a posterior probability of essential hypertension (formula below). *M* is the number of SNPs selected from the previous studies.

$$P(HT|G_1,\ldots,G_M) = \frac{P(HT)\prod_{i=1}^{M} P(G_i|HT)}{P(HT)\prod_{i=1}^{M} P(G_i|HT) + [1-P(HT)]\prod_{i=1}^{M} P(G_i|\neg HT)}$$

The same approach could be repeated to generate posterior probability of essential hypertension for other samples in the KPGP. Our research is ongoing to find all the possible variants related to hypertension and evaluate in our model. The similar model would be easily adapted to other clinical conditions if the information for each of the KPGP samples is released.

References

1. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res, 2010 sep;39(16):e164

Sample ID	Type	Gene	RS	Chr	Ref	Obs	Zygosity	Qual	ReadDepth
TGP2010D0010	intronic	CYP3A5	rs186483446	chr7	T	C	het	12.3	32
TGP2010D0010	intronic	CYP3A5	rs2040992	chr7	G	A	hom	214	31
TGP2010D0010	UTR3	CYP3A5		chr7	T	G	het	24	36

Table 1: SNV detected with NGS data for individual TGP2010D0010

IND	Type	Gene	RS	Chr	Ref	Obs	Zygote	Qual	ReadDepth
TGP2010D0010	intronic	CYP3A5		chr7	-	AAAT	hom	176	18
TGP2010D0010	intronic	CYP3A5		chr7	GT	-	hom	192	30

Table 2: Indels detected with NGS data for individual TGP2010D0010

A heuristic framework for variant calling and de novo mutation detection in trios

Yongzhuang Liu[1,2] and Min He[1*]

1. Center for Human Genome Variation, Duke University School of Medicine, Durham, North Carolina 27708, USA.

2. School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China.

Correspondence author:
Min He, Ph.D.
Center for Human Genome Variation
Duke University School of Medicine
Durham, North Carolina
Email: min.he@duke.edu
Phone: (919) 668-1428

Recent advances in next-generation sequencing technologies make it possible to search for rare and functional variants for complex diseases systematically. Rare causal variants can be enriched in families with parents and an offspring (trio). Trio studies facilitate detection of de novo mutation and newly homozygous mutation events and provide an attractive resource in next generation sequencing studies. We developed a heuristic framework for variant calling and de novo mutation detection in trio samples. Our framework is able to accurately identify variant sites and assign individual genotypes for SNVs and INDELs, and can handle de novo mutation and newly homozygous mutation events. It increases the sensitivity and specificity of variant calling and de novo mutation detection. Unlike most current approaches, our framework reads data from three samples (parents and an offspring) simultaneously; a heuristic and statistical algorithm detects sequence variants and excludes false positive de novo mutations by testing the read counts supporting the reference allele vs. the read counts supporting a variant allele between the offspring and one of the parents respectively using a Fisher's exact test. If the larger p-value of the two tests is greater than 0.05, the candidate of de novo mutation could be false position and will be excluded from the candidate list of de novo mutations. Compared with the standard approach of ignoring relatedness, our methods identify and accurately genotype more variants, and have high specificity for detecting de novo mutations. The family-aware calling framework dramatically reduces Mendelian inconsistencies and is beneficial for trio analysis. We applied these methods to the analysis the two trios from the 1000 Genomes Project. Our results demonstrated the robust performance of the developed program for de novo mutation and newly homozygous mutation detection and shed new light on the landscape of detecting de novo mutations in trios.

The impact of collapsing data on microarray analysis and DILI prediction

Jean-François Pessiot* [1], Pui Shan Wong[1], Toru Maruyama[2], Ryoko Morioka[1], Sachiyo Aburatani[1], Michihiro Tanaka[3], and Wataru Fujibuchi†[3,1,2]

[1]Computational Biology Research Center, Advanced Industrial Science and Technology, Tokyo 135-0064, JAPAN
[2]Faculty of Science and Engineering, Waseda University, Tokyo 169-8555, JAPAN
[3]Center of iPS Cell Research and Application (CiRA), Kyoto University, Kyoto 606-8507, JAPAN

1 Introduction

In the field of toxicology, animal studies and *in vitro* experiments are frequently used as surrogates for human studies even though they have shown poor agreement so far. Besides, it is still unclear how the results obtained from one animal species, such as rats, can help important biomedical research areas for humans, such as the prediction of drug-induced liver injury (DILI). This work is an attempt to address both issues, using the toxicogenomics data provided by CAMDA.

First, we analyzed to what extent animal studies can be replaced by in vitro assays. We compared lists of differentially expressed probesets between rat *in vivo* and rat *in vitro* data, and found poor agreement between the two. This confirmed previous studies suggesting that probeset-level analysis has major limitations, and motivated us to consider higher levels of data abstractions. Thus, we present a data collapsing approach which improves the agreement between *in vivo* and *in vitro* data. We collapsed **probesets** and evaluated the *in vivo-in vitro* agreement using Gene Set Enrichment Analysis (GSEA). We also collapsed **time points** and evaluated the *in vivo-in vitro* agreement using the binary classification framework.

*jfk.pessiot@aist.go.jp
†w.fujibuchi@cira.kyoto-u.ac.jp

Second, we addressed the problem of predicting DILI using available microarray data. Intuitively, we would expect that unprocessed *in vitro* data is too noisy for DILI prediction and would need to be collapsed in order to achieve a better signal-to-noise ratio. In contrast, we would also expect unprocessed *in vivo* data to contain information that is important for predicting DILI. This information could be lost during data collapsing, resulting in a lower prediction performance. Our prediction results tend to confirm these assumptions, and suggest to use unprocessed *in vivo* data simultaneously with collapsed *in vitro* data to improve the prediction performance of DILI.

2 Comparison between *in vivo* and *in vitro* data

2.1 Differential expression results

We studied the similarity between *in vivo* and *in vitro* data at the probeset-level (using differential expression analysis) and at the probeset group level (using GSEA).

Probeset-level differential expression We applied empirical Bayes statistics [1] to test for agreement between *in vivo* and *in vitro* probesets at corresponding time points. The p-values obtained from statistical testing of *in vitro* probesets were used as scores to predict differentially expressed probesets *in vivo*. The classification performance was measured by the AUC score, averaged over all drugs and all doses. Ideally, a perfect agreement between *in vivo* and *in vitro* data would achieve AUC$= 1$. In our experiments, we obtained AUC$= 0.56$ at $t = 2$ hours, AUC$= 0.56$ at $t = 8$ hours and AUC$= 0.60$ at $t = 24$ hours. Overall, the agreement between *in vivo* and *in vitro* is poor, at the probeset-level, with respect to differential expression analysis.

Gene Set Enrichment Analysis GSEA [2] was used to compare expression patterns for *in vitro* and *in vivo* data. Each full list of *in vitro* probesets was used for enrichment testing against the top 1% absolute fold change (i.e. $\log_2$ ratio) list of *in vivo* probesets and vice versa. This was done for two pairs of *in vivo-in vitro* time points: (2 hours, 3 hours) and (24 hours, 24 hours). In short, four analyses were performed per drug. Fold change was calculated between control and low dosages versus medium and high dosages due to the low number of replicates. In this analysis, significant enrichment (p-value < 0.05) means that the high fold change probesets are expressed similarly, as a group, between *in vivo* and *in vitro* data.

Namely, about 55% of the 131 drugs had full agreement between *in vivo* and *in vitro* data, 34% of drugs had 75% agreement, 10% of drugs had 50% agreement and 2% of drugs had 25% agreement. To check that the lists of probesets were not simply correlated, Spearman's rank correlation was calculated between each pair of *in vivo* and *in vitro* lists. None of the pairs of lists had a higher correlation of 0.25 showing that there is little correlation between *in vivo* and *in vitro* data simply at the probeset-level.

[1] http://bioinf.wehi.edu.au/limma/

[2] http://www.broadinstitute.org/gsea/

GSEA using Gene Ontology on liver functions GSEA was used to test probesets associated with liver functions identified with gene ontology for enrichment between *in vivo* and *in vitro* data. AmiGO [3] was used to select 32 gene ontologies which were then ranked by the total number of child ontological nodes. The *in vivo* and *in vitro* data were then tested for enrichment against the probesets with the selected gene ontologies. About 50% of drugs had full agreement between enrichment analyses, 42% of drugs had half agreement and 8% of drugs had no agreement between the analyses. Thus, GSEA produces the same enrichment results between *in vivo* and *in vitro* about half the time and similar results most of the time with respect to liver functional enrichment analysis.

2.2 Collapsing probesets and time points

We now show that it is possible to improve the agreement between *in vivo* and *in vitro* data by appropriately collapsing probesets and time points. From the raw CEL files, we extracted the MAS5 probeset-level values using LIMMA. Then, we averaged those values over biological replicates. We computed the fold changes for each condition (drug, dose and time point), i.e., the $\log_2$ ratios between the sample values and the corresponding control values.

In Figure 1, we plotted the *in vitro* fold change (averaged over all drugs and all doses) as a function of the *in vivo* fold change, at $t = 24$ hours. This plot shows that while there is no obvious sign correlation between *in vivo* and *in vitro* data, there is a correlation between their absolute values. In other words, even if a gene has a highly positive fold change *in vivo*, we cannot always expect a highly positive fold change *in vitro*. However, a gene with a high *in vivo* absolute fold change tends to have a high *in vitro* absolute fold change as well.

In order to improve the signal-to-noise ratio of the data, we also considered a data collapsing strategy. We collapsed probesets into genes, by computing the average intensity of the probesets in each gene. We also collapsed each time series by computing their average absolute fold change. To evaluate our data collapsing strategy, we considered a binary classification problem where the top 1% genes with the highest *in vivo* average absolute fold change were defined as true positives, and the remaining genes were defined as the true negatives. The corresponding average absolute *in vitro* fold changes were used as prediction scores. The classification performance, which reflects the agreement between *in vivo* and *in vitro* data, was measured by the AUC. Our experiments showed that our data collapsing strategy achieved the highest average AUC score among all other tested pre-processings: **AUC=0.85 $\pm$ 0.04**.

Correlation matrix analysis Using the data collapsing approach defined previously, we evaluated the similarity between *in vivo* and *in vitro* data with respect to each gene. For each gene, we defined two correlation matrices characterizing the gene's responses to drugs *in vivo* and *in vitro*. If the Frobenius norm of the difference between these two correlation

[3] http://amigo.geneontology.org

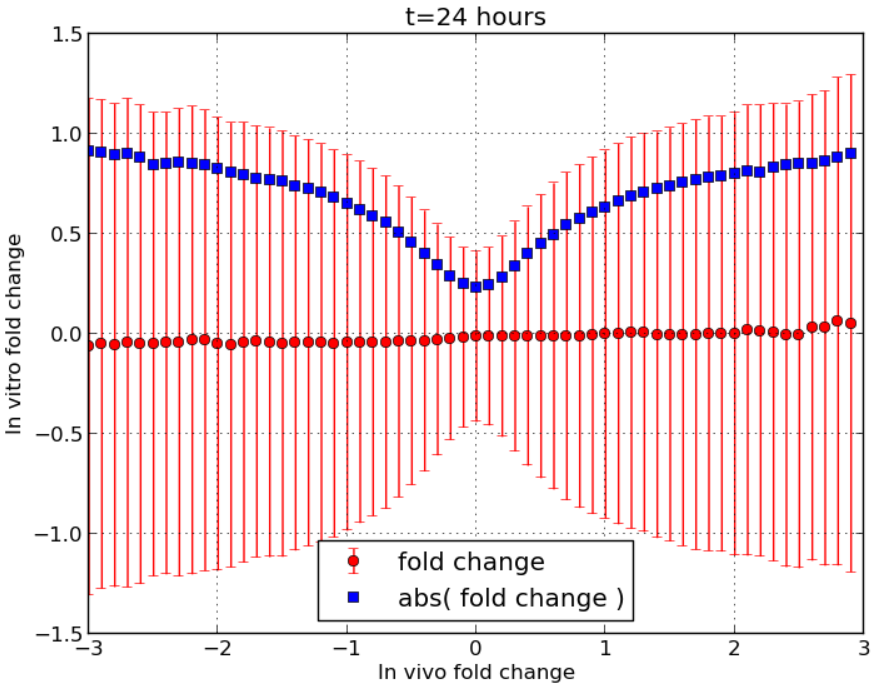

Figure 1: *In vivo* fold changes versus *in vitro* fold changes

matrices is small, then the corresponding gene behaves similarly *in vivo* and *in vitro*. When running downstream analysis of *in vitro* data, this approach can be used to filter out inconsistent genes, and keep the genes that show high correlation between *in vivo* and *in vitro* data. Table 1 shows the top 10 genes which show the most similar behaviours between *in vivo* and *in vitro* conditions.

3 Predicting drug-induced liver injury in humans

We considered the DILI prediction problem as a binary classification of "Most DILI" against "Less DILI or no DILI". For each available data source, we considered all DILI-

Gene Symbol	*In vivo-in vitro* dist.	Gene Symbol	*In vivo-in vitro* dist.
Dazap2	0.4884	Actr2	0.5808
Arf1	0.5446	Gdi2	0.5854
Aamp	0.5573	Cmpk1	0.6030
Ube2l3	0.5608	Morf4l1	0.6121
Cdc42	0.5629	Arpc2	0.6284

Table 1: Genes which behave similarly *in vivo* and *in vitro*, and their corresponding distance measures

Collapse probesets	Collapse time points	Absolute values	Human *in vitro*	Rat *in vitro*	Rat *in vivo* repeated dose	Rat *in vivo* single dose
False	False	False	0.52 ± 0.17	0.52 ± 0.14	$\mathbf{0.66 \pm 0.14}$	0.61 ± 0.12
False	False	True	$\mathbf{0.59 \pm 0.08}$	0.58 ± 0.08	0.61 ± 0.17	$\mathbf{0.67 \pm 0.15}$
False	True	True	$\mathbf{0.58 \pm 0.12}$	0.55 ± 0.20	0.52 ± 0.19	0.55 ± 0.10
True	False	False	0.50 ± 0.21	0.47 ± 0.18	0.64 ± 0.13	0.56 ± 0.20
True	False	True	$\mathbf{0.56 \pm 0.07}$	0.50 ± 0.16	0.55 ± 0.18	0.62 ± 0.18
True	True	True	$\mathbf{0.59 \pm 0.10}$	0.49 ± 0.12	0.59 ± 0.15	0.63 ± 0.17

Table 2: Average AUC scores for the DILI prediction problem

annotated drugs and doses with no missing data. The resulting human *in vitro*, rat *in vitro*, rat *in vivo* repeated dose, and rat *in vivo* single dose data contained 223, 303, 303, and 301 samples, respectively. Each sample corresponds to a (drug, dose) pair. The probeset space contained 54675 probesets for humans and 31099 probesets for rats. The gene space contained 20026 genes for humans and 13878 genes for rats. We used the linear SVM classifier and RBF kernel SVM classifier, and evaluated their classification performance using a ten-fold cross validation. Table 2 shows the AUC scores of the linear SVM.

Overall, AUC scores tend to be low, which shows that predicting DILI using expression data is a difficult problem. We notice that the rat *in vivo* repeated dose and rat *in vivo* single dose data reach high AUC scores when no data collapsing is applied ($0.61 \leq \text{AUC} \leq 0.67$). However, collapsing either the probesets or the time points tend to decrease the AUC scores. This suggests that *in vivo* data might contain important information related to DILI prediction that is partially lost during data collapsing.

In contrast, the human *in vitro* data achieves its lowest AUC score when no pre-processing is applied (AUC=0.52). Collapsing either the probesets or the time points tends to increase the AUC scores, although not as high as with the rat *in vivo* data. This suggests that even though the goal is to predict human DILI, using *in vivo* data from rats is more informative than using *in vitro* data from humans. In contrast, the rat *in vitro* data achieves the lowest AUC scores. This is not surprising as it combines the two limitations of the three other data sources: it is not human, and it is not *in vivo*.

In summary, we first showed that our data collapsing strategy (using gene-level representation, absolute fold changes, and average values over time points) achieved the highest average AUC score (0.85) between *in vitro* and *in vivo* among all other tested preprocessings. Second, we showed that for the DILI prediction problem, using *in vivo* data from rats is more informative than using *in vitro* data from humans. There could be further improvements in prediction performance by combining unprocessed rat *in vivo* data with processed (collapsed) human *in vitro* data.

Exploiting the Japanese Toxicogenomics Project for Predictive Modelling of Drug Toxicity

Djork-Arné Clevert[1†], Martin Heusel[1†], Andreas Mitterecker[1], Willem Talloen[2], Hinrich Göhlmann[2], Jörg Wegner[2], Andreas Mayr[1], Günter Klambauer[1], and Sepp Hochreiter[*1]

[1] Institute of Bioinformatics, Johannes Kepler University Linz, Linz, Austria; [2] Functional Genomics, Johnson & Johnson Pharmaceutical R&D, A Division of Janssen Pharmaceutica, Beerse, Belgium;
† Both authors contributed equally to this work.

Email: Djork-Arné Clevert - okko@clevert.de; Martin Heusel - mhe@gmail.com; Sepp Hochreiter[*]- hochreit@bioinf.jku.at;

[*]Corresponding author

Abstract

Motivation

In the last decade, surprisingly few drugs reached the market. Many promising drug candidates (approx. 80%) failed during or after Phase I, inter alia, due to issues with undetected toxicity [1]. The problem of undetected toxicity becomes even more apparent in the context of drug-induced illness which causes approximately 100,000 deaths per year solely in the USA [2]. Toxicogenomics tries to avoid such problems by prioritizing less toxic drugs over more toxic ones in early drug discovery. To this end, toxicogenomics employs high throughput molecular profiling technologies and predicts the toxicity of drug candidates. For this prediction, large-scale -omics studies of drug treated cell-lines and/or pharmacology model organisms are necessary. However, data exploitation of such large-scale studies requires a highly optimized analysis pipeline, that provides methods for correction of batch effects, noise reduction, dimensionality reduction, normalization, summarization, filtering and prediction.

In this work, we present a novel pipeline for the analysis of large-scale data sets in particular for transcriptomics data. Our pipeline was tested on the Japanese Toxicogenomics Project (TGP) [3], where we evaluated to what degree in vitro bioassays can be used to predict in vivo responses.

The evaluation tasks were to predict drug induced liver injury (DILI) concern [4] and the most prevalent *in vivo* pathological findings from the summarized *in vitro* gene expression values.

Methods and Material

The Japanese Toxicogenomics Project (TGP) is one of the most comprehensive efforts in toxicogenomics, including, among others, gene expression data, toxicological information and pathological data of 131 compounds *in vitro* and *in vivo* screened for toxicity in rat. In the course of the *in vitro* gene expression study, collagen cultured primary hepatocytes, isolated from Sprague-Dawley

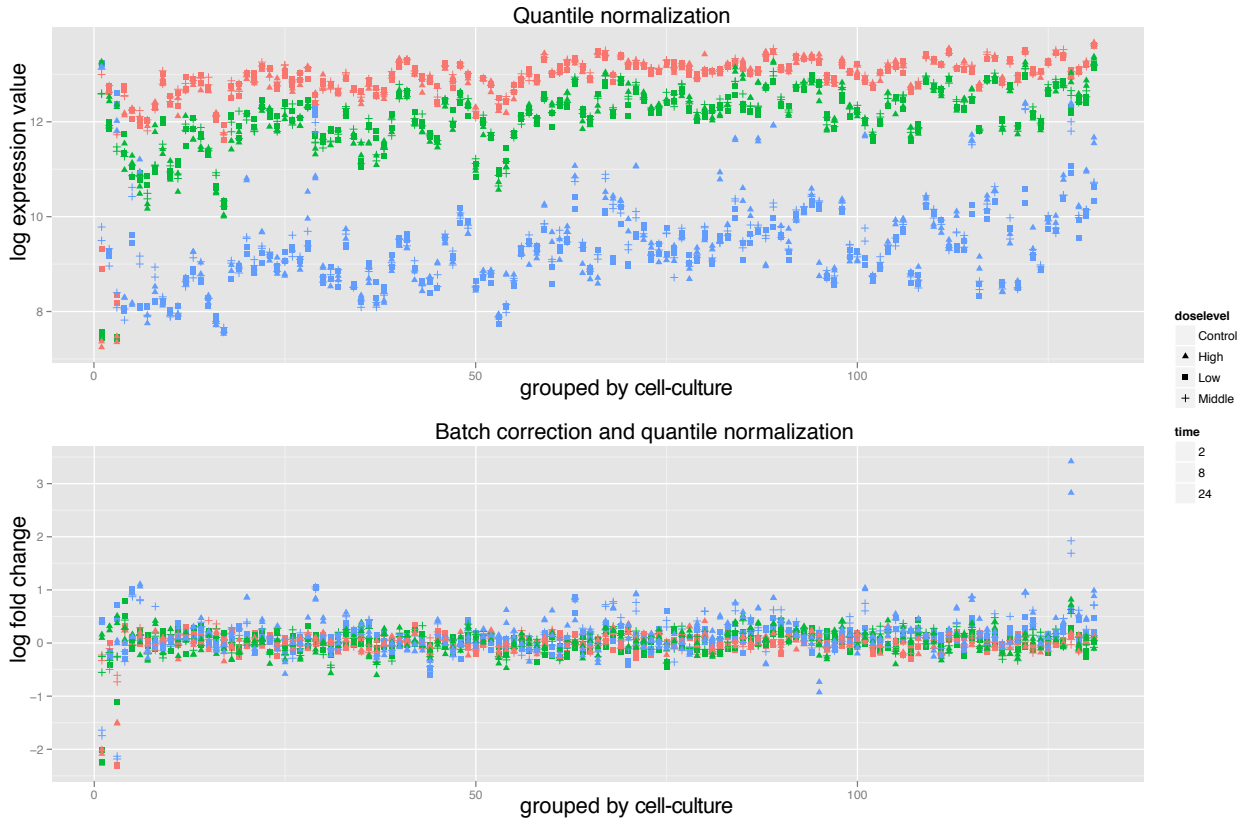

Figure 1: **(Upper panel)** The y-axis shows the log expression values of the fatty acid-binding protein 1 (Fabp1) estimated by FARMS after quantile normalization, while the grouped compounds are shown on the x-axis. The time points are encoded by orange, green and blue for 2h, 8h and 24h, respectively. The plot shows strong cell-culture effects, within the three time points and compounds, which could not be removed by the quantile normalization. **(Lower panel)** Same as upper panel but batch corrected before quantile normalization. The correction with the matched control within cell-culture clearly reduces the cell-culture effects, while compound induced expression changes are preserved.

rats, were treated with three compound concentrations (low, middle and high). To survey transcriptomic changes in response to the compound perturbation, mRNA was isolated at three time points (2h, 8h and 24h) and assayed in two replicates with Affymetrix RAE230_2.0 GeneChip microarrays.

The standard microarray preprocessing procedure consists of normalization, summarization and filtering. However, the standard preprocessing pipeline can not be applied to this data set, as the initial quality control of the microarray data revealed severe batch effects between the cell-cultures (see Figure 1). Therefore, we developed a three step normalization procedure which takes cell-culture batch effects into account. First, the probe-level data of the microarrays were baseline normalized to the same median. Secondly, a cell-culture batch correction was made by calculating probe intensity ratios using the corresponding control measurement for the cell-culture (only vehicle without compound) as reference. Finally, the probe intensity ratios were quantile-normalized [5] across all batches. For the next preprocessing step, summarization, we defined probe sets corresponding to genes using alternative CDFs (Version 15.1.0, ENTREZG) from Brainarray [6] and applied FARMS [7] for summarizing the intensity ratios at probe set level to obtain expression values per gene. Specially for this purpose, a new FARMS software package has been developed, that allows summarization of huge microarray data sets like those of the TGP. For the last preprocessing step, gene filtering, we applied the FARMS based informative/non-informative (I/NI) call [8–10] and excluded all non-informative probe sets.

After this data preprocessing, we predicted drug induced liver injury (DILI) concern and *in vivo* pathological findings for hypertrophy, vacuolization and ground glass appearance from the summarized *in vitro* gene expression values. We combined replicates by concatenating their features. Removing replicates is essential because otherwise the classification task leads to a trivial almost perfect solution in the LOO-CV task, by predicting one replicate by the other. For classification we used the Potential Support Vector Machine (P-SVM) [11] because it is well-suited for data sets with many samples. The PSVM is optimized for many samples since it isbased on a quadratic optimization problem in the number of features instead of the number of samples as standard SVMs. Therefore FARMS' gene filtering and P-SVM prediction are an ideal combination to process massive data sets.

Results

Leave-one-out cross-validation (LOO-CV) of the classifiers (see Table 1) for DILI concern, hypertrophy, vacuolization and ground glass appearance showed a sensitivity of 0.78, 0.81, 0.77 and 0.80 for a specificity of 0.63, 0.65, 0.58 and 0.56, respectively. These results are very promising, as due to the severe cell-culture effects we did not expect to obtain such a high classification performance.

Table 1: Summary of the LOO-CV for DILI concern and various pathological findings (first column "predicted finding"). The second column "# features" gives the number of features used for classification. The third column shows the "error rate" over all LOO runs, while the fourth and fifth columns report the "sensitivity" and "specificity", respectively.

predicted finding	# features	error rate	sensitivity	specificity
DILI concern	14	0.26	0.78	0.63
hypertrophy	47	0.32	0.81	0.65
vacuolization	39	0.38	0.77	0.58
ground glass appearance	56	0.42	0.80	0.56

References

1. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL: **How to improve R&D productivity: the pharmaceutical industry's grand challenge**. *Nature Rev. Drug. Discov.* 2010, **9**(3):203–214.
2. Lazarou J, Pomeranz BH, Corey PN: **Incidence of adverse drug reactions in hospitalized patients a meta-analysis of prospective studies**. *JAMA* 1998, **279**(15):1200–1205.
3. Uehara T, Ono A, Maruyama T, Kato I, Yamada H, Ohno Y, Urushidani T: **The Japanese toxicogenomics project: application of toxicogenomics**. *Mol. Nutr. Food. Res.* 2010, **54**(2):218–227.
4. Chen M, Vijay V, Shi Q, Liu Z, Fang H, Tong W: **FDA-approved drug labeling for the study of drug-induced liver injury.** *Drug Discov Today* 2011, **16**(15-16):697–703.
5. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185–193.
6. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, et al.: **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data**. *Nucleic Acids Res.* 2005, **33**(20):e175.
7. Hochreiter S, Clevert DA, Obermayer K: **A new summarization method for Affymetrix probe level data**. *Bioinformatics* 2006, **22**(8):943–949.
8. Talloen W, Clevert DA, Hochreiter S, Amaratunga D, Bijnens L, Kass S, Göhlmann HWH: **I/NI-calls for the exclusion of non-informative genes: a highly effective feature filtering tool for microarray data**. *Bioinformatics* 2007, **23**(21):2897–2902.
9. Talloen W, Hochreiter S, Bijnens L, Kasim A, Shkedy Z, Amaratunga D, Göhlmann HWH: **Filtering data from high-throughput experiments based on measurement reliability**. *Proc. Natl. Acad. Sci. USA* 2010, **107**(46):173–174.
10. Kasim A, Lin D, Sanden SV, Clevert DA, Bijnens L, Göhlmann H, Amaratunga D, Hochreiter S, Shkedy Z, Talloen W: **Informative or noninformative calls for gene expression: a latent variable approach**. *Stat. Appl. Genet. Molec. Biol.* 2010, **9**:1–29.
11. Hochreiter S, Obermayer K: **Support Vector Machines for Dyadic Data**. *Neural Comput* 2006, **18**(6):1472–1510.

Using Linear Models to bootstrap Non-Linear models for prediction of Drug Induced Liver Injury

Mike Bowles, Ron Shigeta, David Dehghan

Adv Machine Learning Seminar Hacker Dojo 140A South Whisman Road Mountain View, CA 94041(650) 898-7925

Prediction of the Drug Induced Liver Injury (DILI) is problematic. Given the number of inferences between rat liver expression data and the regulatory process which parses the DILI scores into grades, relying upon multiple indirect sources of reports on the drug is non linear. Figures 1. and 2. indicate the performance advantage of non-linear predictive models over linear ones. These models were trained on the CAMDA rat liver data set as follows.

We condensed the three DILI designations into a unified integer grade scale consistent with the FDA definitions [1] and trained supervised linear predictors [2]. We first trained a Lasso linear regression model to predict single integer DILI grade (SIDG) using Rat liver expression sets taken at the last (29 day) time point taken at high and zero dosage. This resulted in minimum, cross validated, root mean square prediction error of 1.91 over the range of DILI unified grades from 1 to 9 with a correlation of 0.74. Compared to actual DILI grades (Figure 1.) the predictions are compressed in scale and don't increase monotonically. A non-linear model might overcome these deficiencies. The sheer number of gene expression measurements in a micro array makes it difficult to employ the most powerful non-linear machine learning techniques (e.g. basis expansion or stochastic gradient boosted trees). The sparsity of the Lasso model suggests a way forward.

Lasso models give zero weight to most of the array intensities. The model shown in Figure 1. incorporates 1461 expression microarray probe sets. This is a small enough number of attributes to fit a boosted tree model. Figure 2. shows the performance of a cross validated gradient boosted tree model fit to the attributes isolated by lasso regression. These results are much improved, with the gradient boosted model predicted vs. actual DILI ratings was 0.90, with a root mean squared error of 1.81.

While the derived Gradient Boosted model shows drug specific patterning where predictions for drug data not used in testing do not predict as well, the priming of non linear models from linear models of expression data is significant for two reasons. Although there are many cases in bioinformatics where nonlinear models are needed to train adequate models, but because nonlinear models are searching through a space that is geometrically or exponentially larger than the number of linear degrees of freedom. In the case of the RAE230_2 arrays, stochastic gradient boosting would be searching through approximately a billion degrees of freedom. Limiting the number of probe sets for the gradient boost search makes a model trainable quickly enough to do iterative optimizations to improve the model. Since stochastic gradient boosting is generalizable to map-reduce and other big data treatments [5], we

see a promising path to broaden the opportunities to make more flexible models for expression data in the near future.

With approximately1500 probe sets incorporated to the model we have built a topic modeling system based on the Latent Dirichlet Allocation [8] method to convey a sense of the biological functions related to the distinguishing probe sets in our model. [Figure 3.]

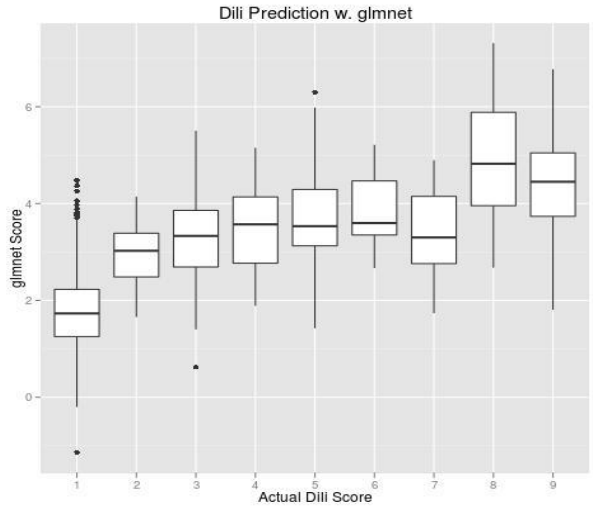

Figure 1: Lasso linear regression model predictions plotted against training unified DILI scores. Overall RMS error is 1.91.

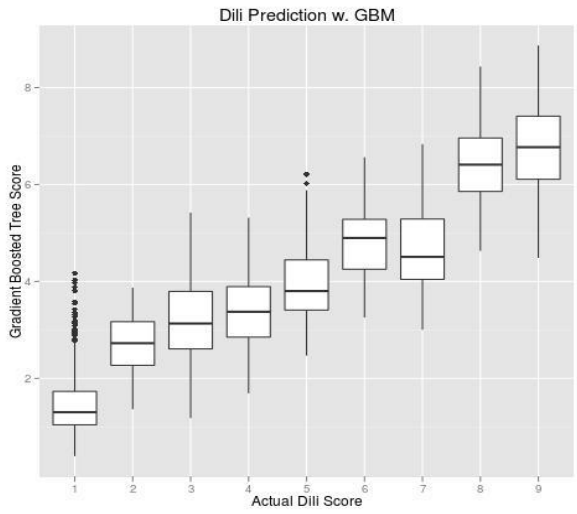

Figure 2: Gradient Boosted Tree Score predictions plotted against actual Unified DILI score shows a strong median correlation. Overall RMS error is 1.8.

```
 [1] "mutat syndrom caus famili disord diseas phenotyp clinic defect abnorm"
 [2] "dna methyl cell damag repair 2 beta status control cycl"
 [3] "protein interact complex membran bind subunit receptor domain local cytoplasm"
 [4] "gene express profil differenti tissu microarray develop analysi skin pattern"
 [5] "develop cell differenti regul transcript function embryon neural progenitor mous"
 [6] "cell express mrna marker differenti normal line tissu stain epitheli"
 [7] "cancer cell target line resist breast therapeut apoptosi treatment effect"
 [8] "patient 7 2 clinic 0 3 6 diabet 10 conclus"
 [9] "cell express transcript regul factor protein nuclear role phase gene"
[10] "associ genet variant gene studi individu risk cohort popul genome-wid"
[11] "gene identifi genom network data model predict set express approach"
[12] "activ kinas signal phosphoryl pathway growth inhibit regul factor receptor"
[13] "signal cell express prolifer mice immun activ human respons induc"
[14] "promot cell transcript express repress induc regul rna direct silenc"
[15] "activ enzym human format inhibit vitro substrat vivo recombin inhibitori"
```

Figure 3: Example topic models based from pubMed abstracts on 696 genes from the cross validated Gradient Boosted Tree Model.

To be submitted to CAMDA 2012 conference (http://www.camda.info/)

Pharmacogenomics in the Pocket of Every Patient? - A Prototype Card with Quick Response (QR) Code

Matthias Samwald[1], Zhan Ye[2], Daniel A Vasco[2], Steven Schrodi[2], Murray Brilliant[2], and Simon Lin[2]

[1]Medical University of Vienna, Austria and [2]Marshfiled Clinic Research Foundation, USA

The Institute of Medicine of the United States has recently envisioned an accelerated adoption of pharmacogenomics in clinical practice in the next five years. A major gap to be resolved is how to exchange the pharmacogenomics genotype information given the current disparate practice of paper-based and electronic-based medical records. In addition, how can a patient safekeep the genotype information?

We utilized the CAMDA 2012 Pharmacogenomics Data Set to demonstrate a conceptual design of what the future might be. The pharmcogenomics of KPGP001 patient is encoded into the “Medicine Safety Code” and printed on the back of a health insurance card.

Figure 1. An example health insure Card in the United States (left) and a conceptual design of the future of a health insurance card with the "Medicine Safety Code" (right).

The Medicine Safety Code is an international consortium promoting a simple, yet powerful format combined with software tools for making medical practice safer and more personalized (http://safe-code.org). After whole-genome sequencing, actionable pharmacogenomic traits of an individual patient can be encoded as a Medicine Safety Code, which can be represented as a 2D barcode (that can be printed and read by

common smartphones and other devices) and as a URL (that can be used by computer systems). This simplicity and flexibility makes the Medicine Safety Code universally available and applicable.

The pharmacogenomic traits can be used to improve the selection and dosing of many common pharmaceuticals, thereby lowering the rate of adverse drug events and healthcare costs -- while improving the efficiency of treatment and the quality of life of patients. Moreover, the proposed framework can be extended to include adverse drug reactions of the patient.

In terms of privacy and data security, the Medicine Safety Code offers the same level of security of paper-based medical information. As such, it fits nicely into the current legal framework and social acceptance.