Supplementary figures

for "Predicting Age by Mining Electronic Medical Records with Deep Learning Characterizes Differences between Chronological and Physiological Age" by Wang et al.

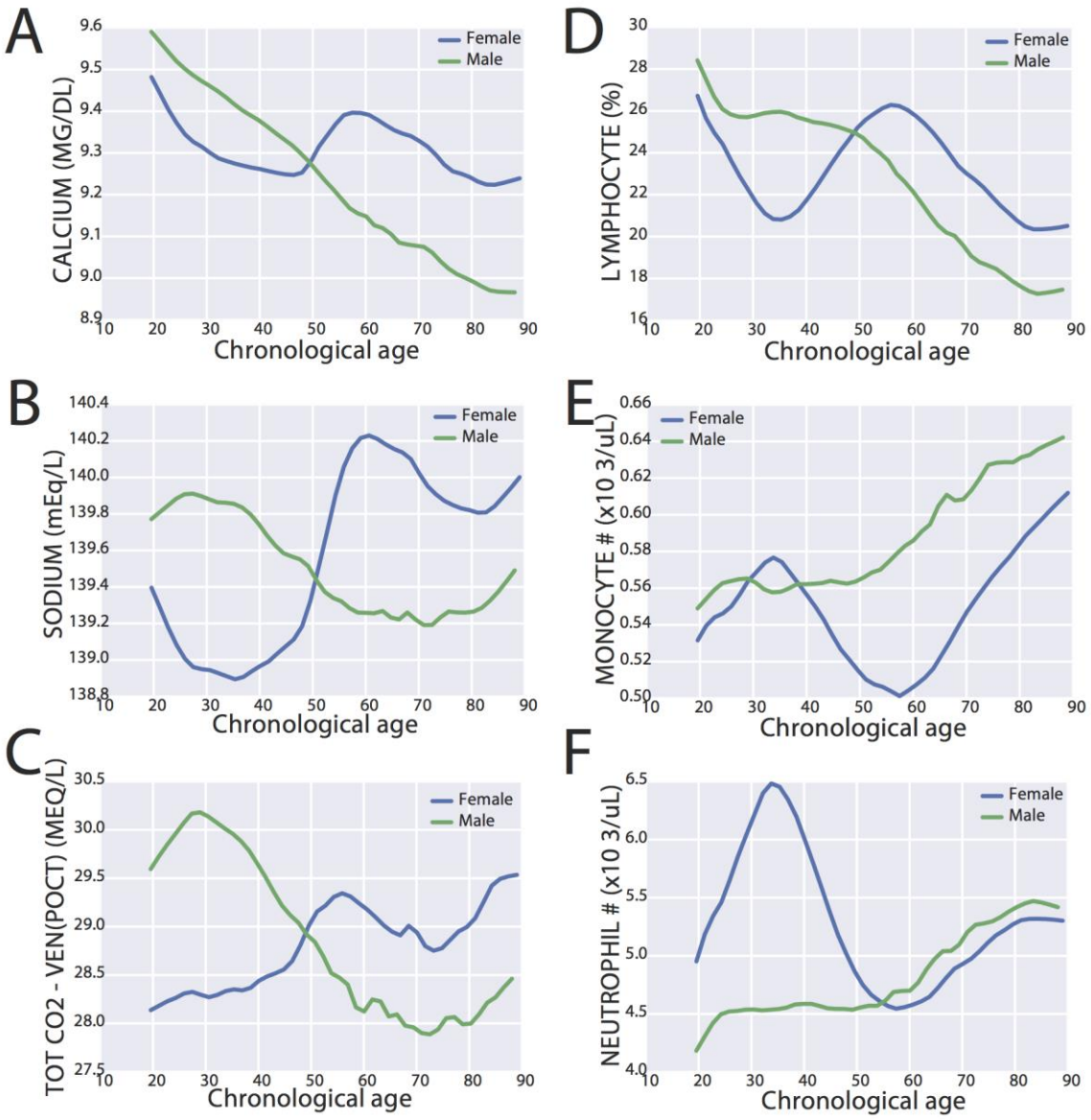

Figure S1. Examples of physiological measurements that have different trends across age between genders. The smoothened trends of representative lab tests across chronological age from female (blue) and male (green) patients.

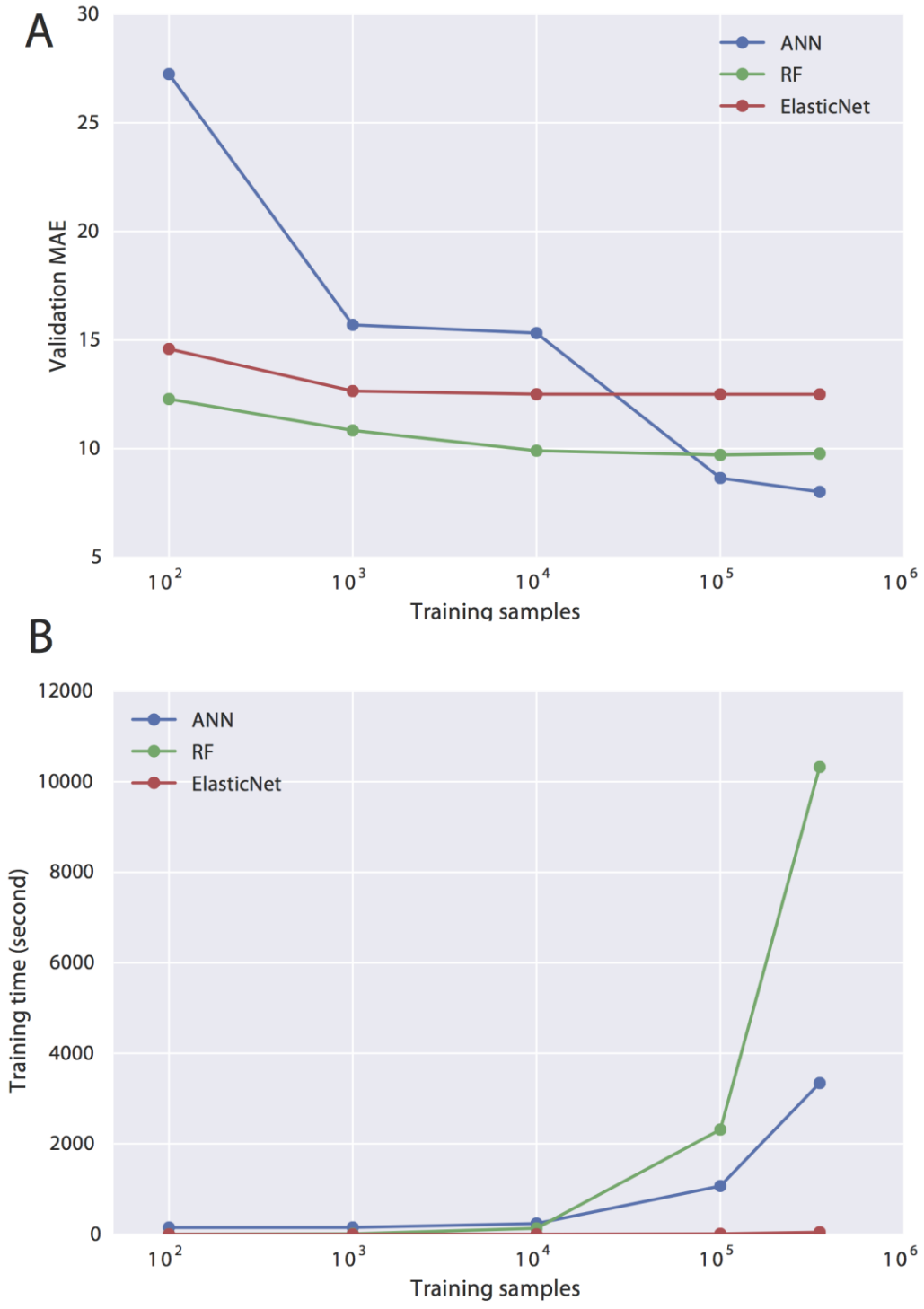

Figure S2. Performance comparison of different machine learning regressors. Plots show the number of training samples used versus (A) the mean absolute error (MAE) from the hold-out validation set, and (B) the training time for different machine learning models: Artificial Neural Network (ANN), Random Forest (RF) and Elastic Net.

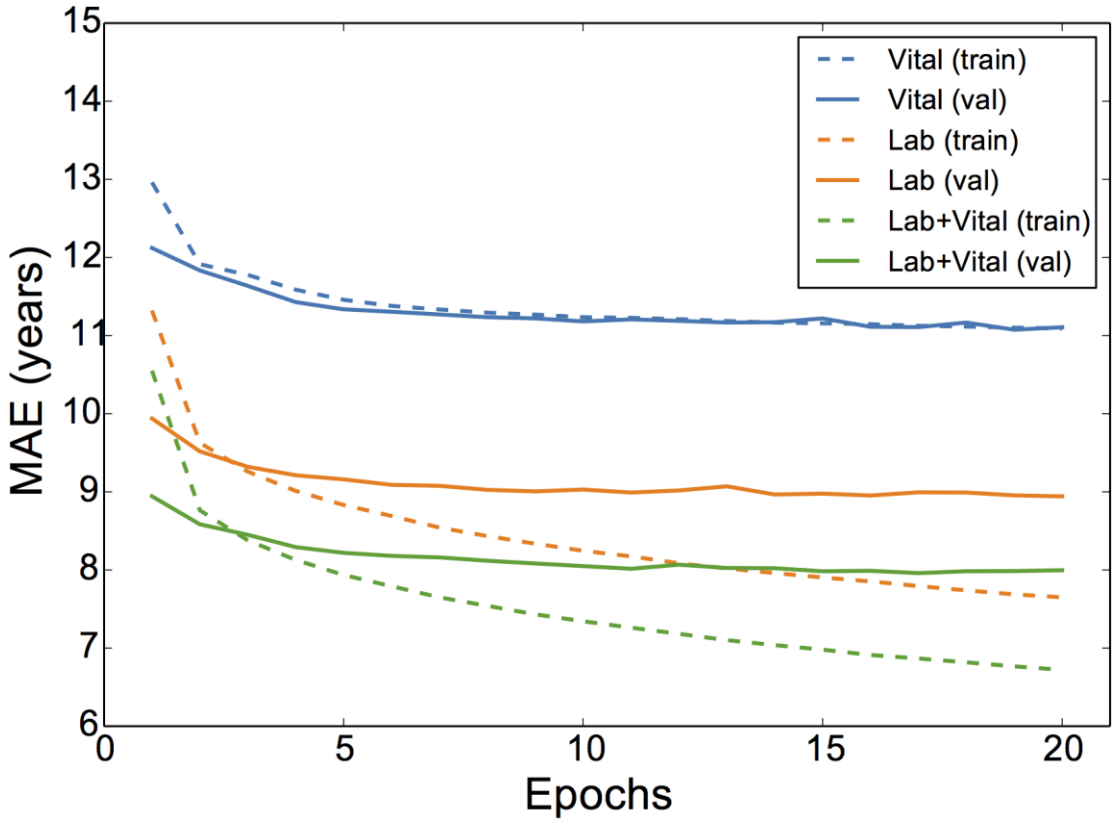

Figure S3. Predictive performances of regressors using different sets of EMR features over training epochs. EMR features, including vital signs alone, lab tests alone, and the combination of vital signs and lab tests used for training the regressor are plotted in blue, orange and green, respectively. Mean absolute errors (MAE) over the epochs are plotted in dashed and solid lines for training sets and validation sets, respectively.

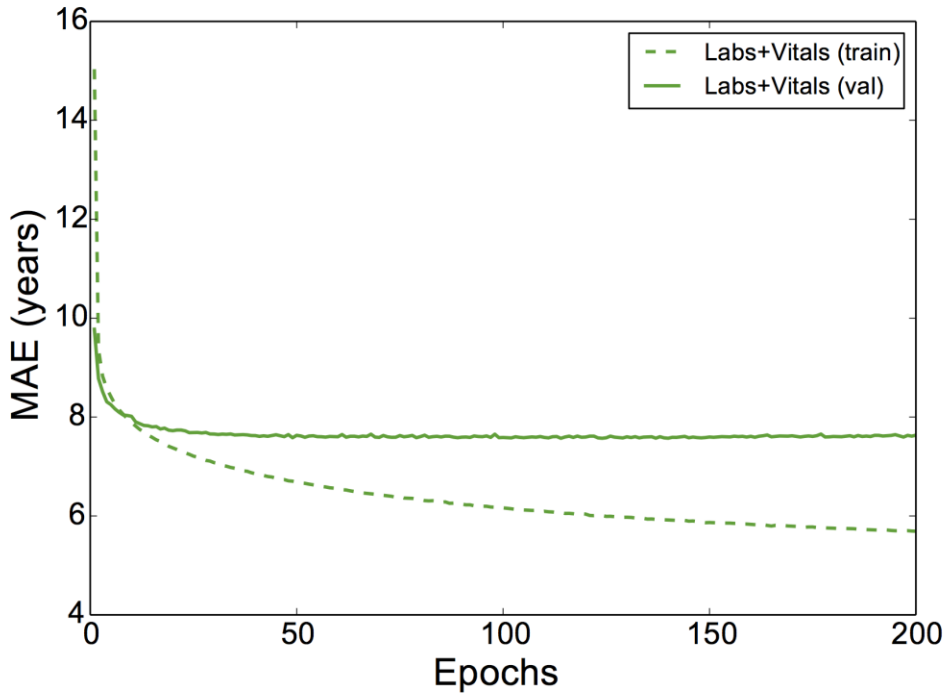

Figure S4. Predictive performances of the fine-tuned ANN regressors using the combination of lab tests and vital signs over training epochs. Mean absolute error (MAE) over epochs are plotted in dashed and solid lines for training sets and validation sets, respectively.

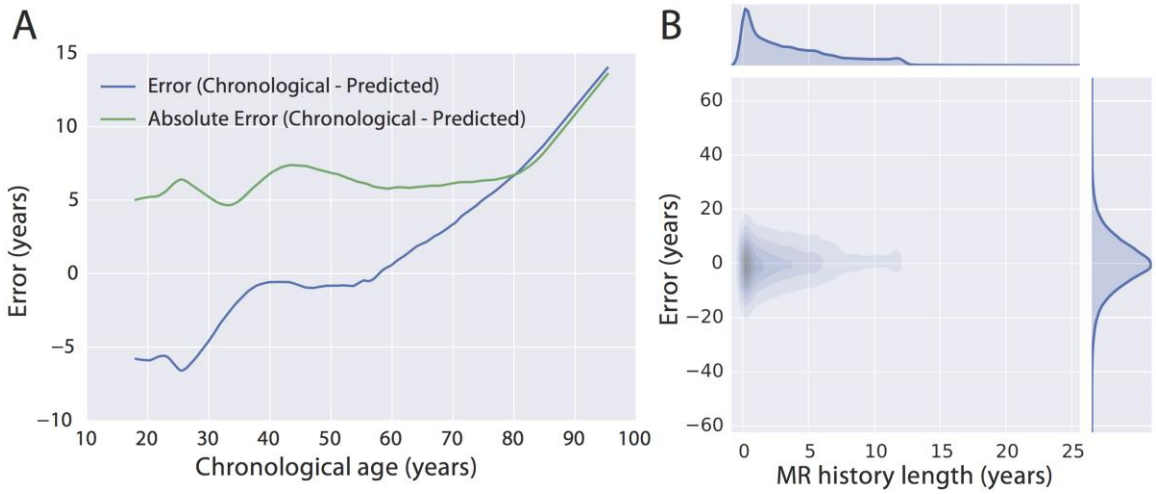

Figure S5. Prediction errors across chronological age for the entire cohort. The locally averaged error and absolute error of predicted age is plotted in blue and green lines, respectively. The errors were averaged using the locally weighted scatterplot smoothing (LOWESS) algorithm.

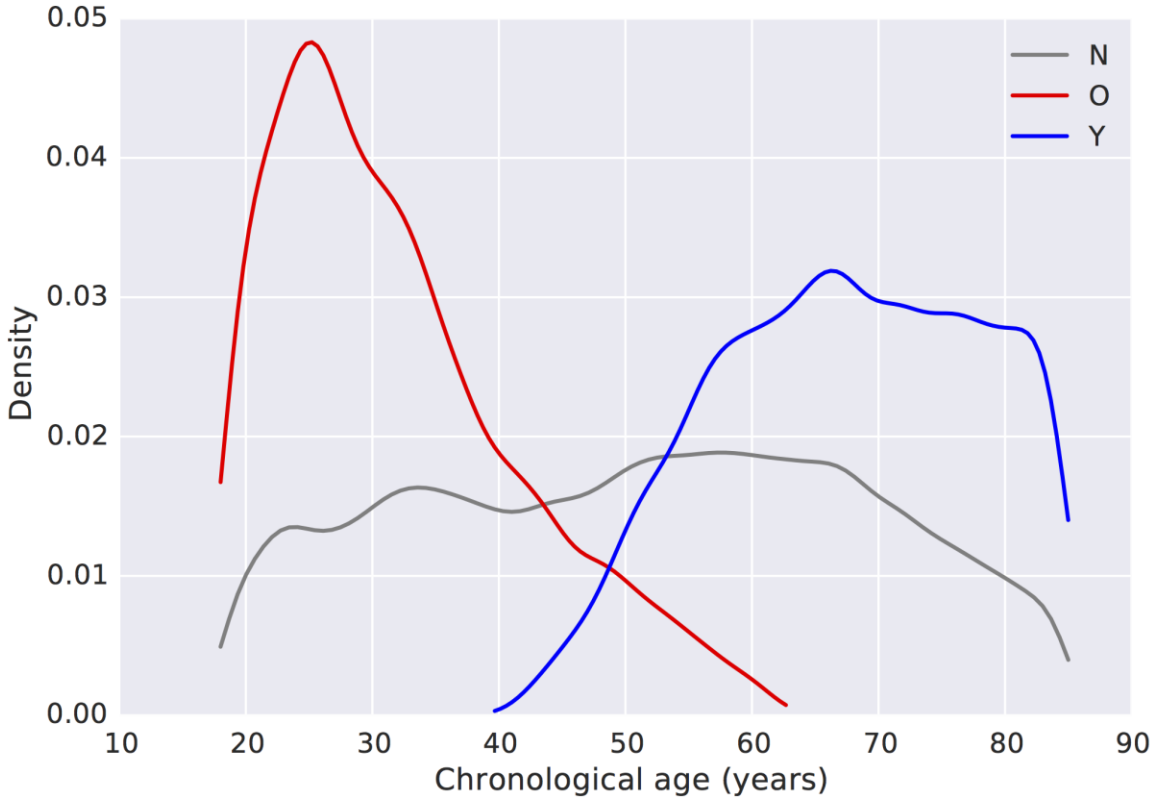

Figure S6. Distributions of chronological age for the three cohorts. The distributions of chronological ages for patients in the three cohorts: Young, Normal and Old are plotted in blue, grey and red, respectively.

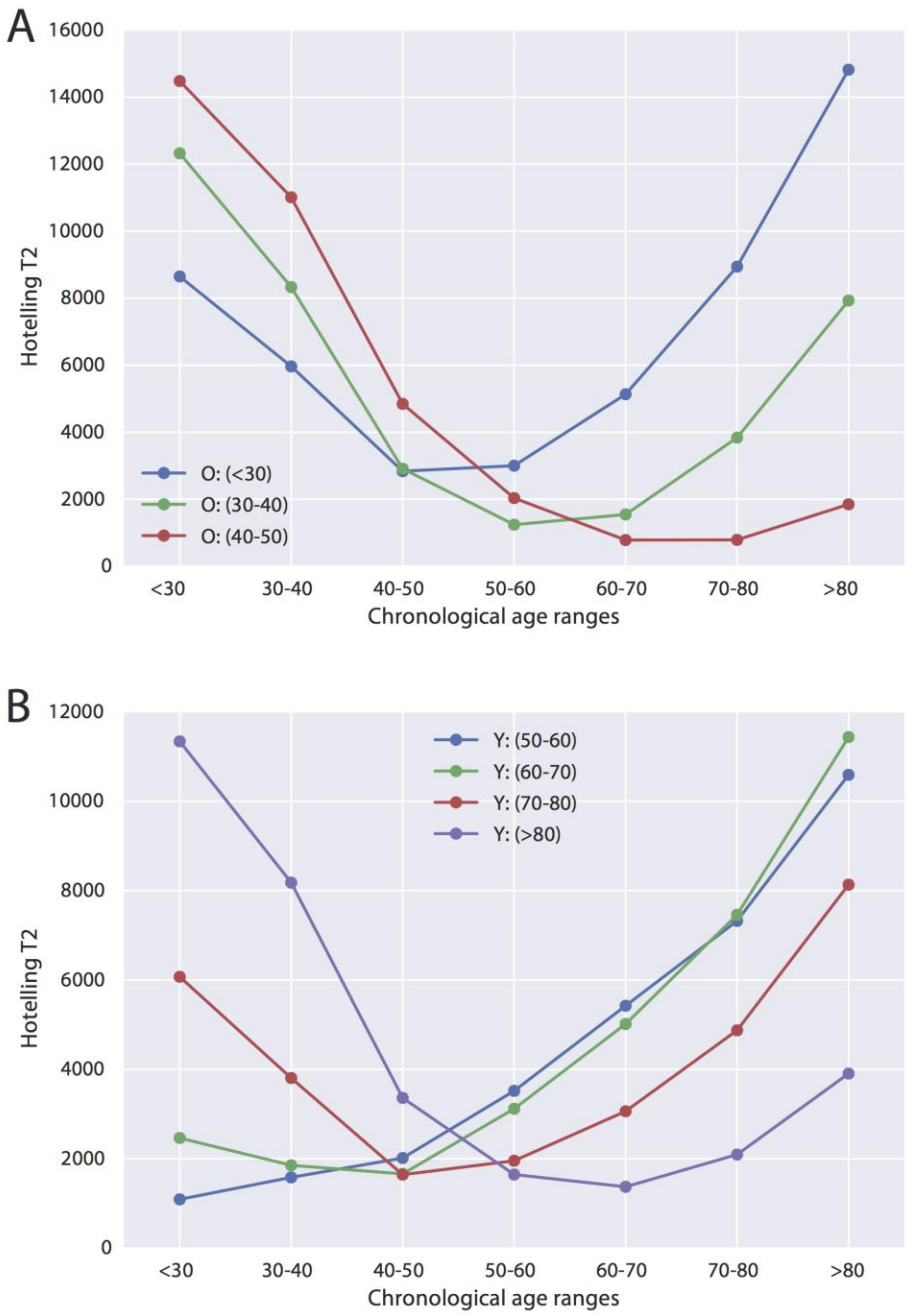

Figure S7. Hotelling's T-square statistics from multivariate hypothesis tests comparing physiological status of patients in different cohorts. (A) The Hotelling's T-square values from comparing the physiological status represented by the first 100 principal components (PCs) of all the vital signs and lab tests, and for patients in group O versus patients in group N across different chronological age ranges. (B) The Hotelling's T-square values from comparing the physiological status represented by the first 100 PCs of all the vital signs and lab tests, and for patients in group Y versus patients in group N across different chronological age ranges.

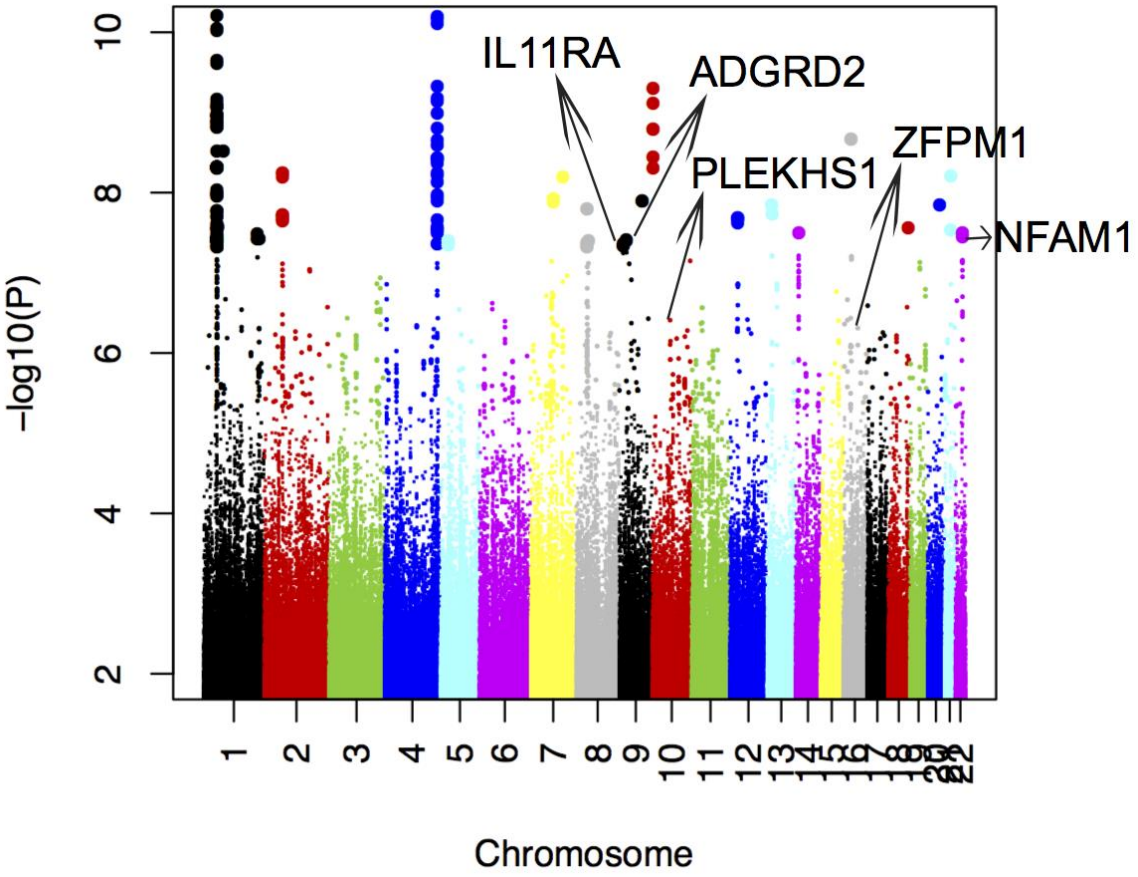

Figure S8. Manhattan plot of GWAS analysis comparing group O versus group N. Genes harboring significant SNPs causing coding missense mutations are highlighted.

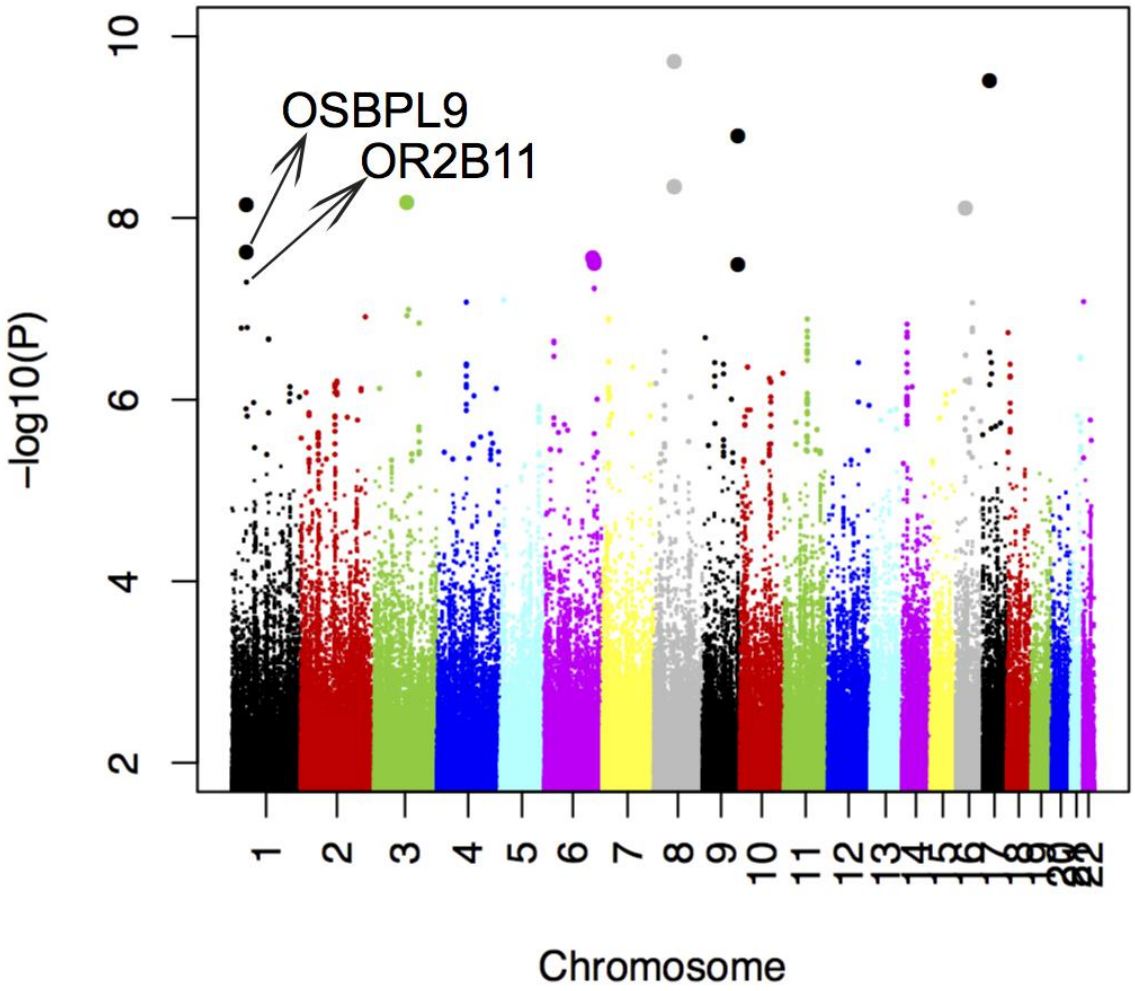

Figure S9. Manhattan plot of GWAS analysis comparing group Y versus group N cohorts. Genes harboring significant SNPs causing coding missense mutations are highlighted.

Legends for Supplementary Tables

Table S1 Comparisons of vital signs across the three cohorts. U denotes the statistic from Man-Whiteney U-test. Positive U values indicate the median of the group is larger than the median of the Normal group and negative U values indicate otherwise.

Table S2. Lab tests that are significantly different in the Old group and Young group.

Table S3. Diagnoses that are significantly enriched in the Old group and Young group.

Table S4. Medicines that are significantly different in the Old group and Young group.

Table S5. Enriched biological terms for GWAS significant genes from the Old versus Normal comparison.

Table S6. Enriched biological terms for GWAS significant genes from the Young versus Normal comparison.

Table S7. Enriched terms for significant up-regulated genes in the Young adipose tissue.

Table S8. Enriched terms for significant down-regulated genes in the Young adipose tissue.

Table S9. Top down-regulated genes in the Young adipose tissue.

Table S10. Enriched terms for significant down-regulated genes in the Old adipose tissue.