Bioinformatics

doi.10.1093/bioinformatics/xxxxxx

Advance Access Publication Date: Day Month Year

Manuscript Category

Supplementary Information

Gene Prioritization Using Bayesian Matrix Factorization with Genomic and Phenotypic Side Information

Pooya Zakeri [1,*], Jaak Simm [1],, Adam Arany [1], Sarah Elshal [1], and Yves Moreau [1,*]

[1] STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics Department, Department of Electrical Engineering, KU Leuven, Leuven, 3001, Belgium.

*To whom correspondence should be addressed.

Abstract

Section 1: Supplementary notes on Probabilistic Matrix Factorization, our proposed Gibbs sampler, and our assessment strategy.

Section 2: A detailed discussion of our results and supplementary data.

Contact: pooya.zakeri@esat.kuleuven.beyves.moreau@esat.kuleuven.be

Supplementary information: Supplementary data are available at online.

1 Supplementary notes on Approach

1.1 Probabilistic Matrix Factorization

As discussed earlier, the main notion behind the probabilistic matrix factorization (PMF) (Mnih and Salakhutdinov, 2007) is to find a factorization that minimizes the mean square error (MSE) on the observed data, and maintain good performance on those observed data considered for test set, with the

assumption of Gaussian noise in the data. For example, to factorize the OMIM matrix (Amberger *et al*., 2011) using PMF, it represents each row (gene) and each column (phenotype) by a latent vector of size D, and then find the

$$\min_{G,P} \sum_{(i,j)\in I_{\mathrm{OMIM}}} \mathrm{OMIM}_{i,j} - G_i^T \times P_j + \lambda_G \parallel G \parallel_F^2 + \lambda_P \parallel P \parallel_F^2 \tag{S1}$$

where $\mathrm{OMIM}_{i,j} \in$ OMIM matrix is the observed data, λ_G and λ_P are regularization parameters and greater than zero, and $\| \cdot \|_F$ denotes the Forbenius norm, which is defined as:

$$\parallel R \parallel_F = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{M} \mid r_{i,j} \mid^2} = \sqrt{trace(RR^T)} \tag{S2}$$

PMF, Indeed, use a linear model with Gaussian observation noise. Accordingly, last two terms in the PMF optimization problem (1) are derived from the Gaussian prior with zero mean on both latent variables, G_i and P_j , and a Gaussian noise model on $\mathrm{OMIM}_{i,j}$.

1.2 A Detailed on our proposed model

As we discussed in the manuscript, Bayesian PMF (BPMF) (Salakhutdinov and Mnih, 2008), proposes a fully Bayesian treatment of the PMF approach by introducing common multivariate Gaussian priors for latent variables. In order to complete the OMIM matrix using BPMF, similarly to PMF, BPMF also uses a linear model with Gaussian observation noise. As we shown, we propose an extended version of BPMF with the ability to work with multiple auxiliary Genomic and Phenotypic information. To incorporate the geneâĽ™s features $x_i \in \mathbb{R}^{F_{\mathrm{gene}}}$ (the phenotypeâĽ™s features $z_i \in \mathbb{R}^{F_{\mathrm{phen}}}$), we integrate a term $\beta_{\mathrm{gene}}^T x_i$ ($\beta_{\mathrm{phen}}^T z_i$) into the Gaussian mean μ_G (μ_P). Then, it was shown that the equations (3) and (4), in the original manuscript are expressed as

$$P(G, \mid x_i, \mu_G, \Lambda_G) = \mathcal{N}(G_i \mid \mu_G + \beta_{\mathrm{gene}}^T x_i, \Lambda_G^{-1}) \tag{S3}$$

$$P(P, \mid z_j, \mu_P, \Lambda_P) = \mathcal{N}(P_j \mid \mu_P + \beta_{\mathrm{phen}}^T z_j, \Lambda_P^{-1}) \tag{S4}$$

Where $\beta_{\mathrm{gene}} \in \mathbb{R}^{F_{\mathrm{gene}} \times D}$ and $\beta_{\mathrm{phen}} \in \mathbb{R}^{F_{\mathrm{phen}} \times D}$ have been introduced as the link matrices for the gene's (phenotype's) features. F_{gene} (F_{phen}) is the dimension of the gene's (phenotype's) features. Theses equations above (S3 and S4) offers the linear model for latent vectors. Technically, our proposed model learns the link matrices β_{genes} and β_{phen} to predict latent variables G_i and P_j from x_i and z_j, respectively.

To have a fully Bayesian treatment for β_{gene} (β_{phen}), we consider a zero mean multivariate normal as its prior. However, our proposed prior on β_{gene} (β_{phen}) scales with the precision of latent variables in order to incorporate the auxiliary information to the factorization process.

$$\begin{aligned} P(\beta_{\mathrm{gene}}, \mid \Lambda_{gene}, \lambda_{\beta_{\mathrm{gene}}}) &= \mathcal{N}(vec(\beta_{\mathrm{gene}}) \mid 0, \Lambda_G^{-1} \otimes (\lambda_{\beta_{\mathrm{gene}}} I)^{-1}) \\ &= \lambda_{\beta_{\mathrm{gene}}}^{F_{\mathrm{gene}} D/2} \mid \Lambda_G \mid^{\frac{D}{2}} exp(\frac{1}{2}\lambda_{\beta_{\mathrm{gene}}} tr(\beta_{\mathrm{gene}} \Lambda_G^{-1} \beta_{\mathrm{gene}}^T) \end{aligned} \tag{S5}$$

where $\otimes$ denotes the Kronecker product, and $vec(\beta_{\mathrm{gene}})$ denotes the vectorization of β_{gene}. And $\lambda_{\beta_{\mathrm{gene}}} \geq 0$ is the diagonal element of the precision matrix, and Λ_G is the precision matrix of the latent variable for genes. This prior is natural as the scale of G is not predetermined. Accordingly, we place

a gamma distribution hyperprior [1] on $\lambda_{\beta_{\text{gene}}}$, owning the fact that the choice of $\lambda_{\beta_{\text{gene}}}$ is problem dependent, as discussed in similar context(Porteous *et al*., 2010).

$$P(\lambda_{\beta_{\text{gene}}}, \mid \mu, \nu) = Gamma(\lambda_{\beta_{\text{gene}}} \mid \mu, \nu) \propto \lambda_{\beta_{\text{gene}}}^{\nu/2-1} exp(-\frac{\nu}{2\mu}\lambda_{\beta_{\text{gene}}}) \tag{S6}$$

where μ and ν are fixed hyperparameters, which are both set to 1 in the experiments. The same full Bayesian treatment is developed for β_{phen}.

1.3 A Detailed on our proposed Gibbs sampler

In this section we present the conditional distributions of the Gibbs sampler for all variables except for β_{gene} and β_{phen}. The outline of Gibbs sampling for our proposed model is as follows: Based on (4) and (5), the conditional probability for G_i and P_j are

$$\begin{aligned} P(G_i \mid \text{OMIM}, P, \Theta_G, x, \beta, \lambda, \alpha, \Lambda_G) &= \mathcal{N}(G_i \mid \mu^*_{G_i}, [\Lambda^*_{G_i}]^{-1}) \\ &\propto \prod_{j \in I_{\text{OMIM}_{(i,)}}} \mathcal{N}(\text{OMIM}_{(i,j)} \mid G_i^T P_j, \alpha^{-1}) \\ &\times \mathcal{N}(G_i \mid \mu_G + \beta_{\text{gene}}^T x_i, \Lambda_G^{-1})) \end{aligned} \tag{S7}$$

where

$$\Lambda^*_{G_i} = \Lambda_G + \alpha \textstyle\sum_{(i,j) \in I_{\text{OMIM}_{(i,)}}} P_j P_j^T \tag{S8}$$

$$\mu^*_{G_i} = [\Lambda^*_{G_i}]^{-1} + (\Lambda_G(\mu_G + \beta_{\text{gene}}^T x_i) + \alpha \textstyle\sum_{(i,j) \in I_{\text{OMIM}_{(i,)}}} P_j \text{OMIM}_{(i,j)}) \tag{S9}$$

From the above equations, we see that if gene i does not have features then $\beta_{\text{gene}}^T x_i$ is 0. We also place the Normal -Wishart hyperprior for μ_G and λ_G, same as being used in BPMF.

$$P(\mu_G, \Lambda_G \mid \Theta_0) = \mathcal{N}(\mu_G \mid \mu_0^*, (\beta_0^* \Lambda_G)^{-1}) \mathcal{NW}(\Lambda_G^{-1} \mid W_0, \mu_0) \tag{S10}$$

Where $\mathcal{NW}$ denote the normal Wishart distributions. In this study, we set $\mu_0 = 0$, $\beta_0 = 2$ and $\nu_0 = D$ and ,W_0 to the identity matrix for both gene and phenotype. Combining the hyperprior in S8 with S1 we get the following conditional probability:

$$P(\mu_G, \Lambda_G \mid G, x, \beta_{\text{gene}}, \Theta_0) = \mathcal{N}(\mu_G \mid \mu_0^*, (\beta_0^* \Lambda_G)^{-1}) \mathcal{NW}(\Lambda_G^{-1} \mid W_0^*, \mu_0^*) \tag{S11}$$

where

$$\mu_0^* = \frac{\beta_0 \mu_0 + N_G \bar{G}}{\beta_0 + N_G} \qquad \beta_0^* = \beta_0 + N_G \qquad \nu_0^* = \nu_0 + N_G \tag{S12}$$

$$[W_0^*]^{-1} = W_0^{-1} + N_G \bar{S} + \frac{\beta_0 N_G}{\beta_0 + N_G}(\mu_0 + \bar{G})(\mu_0 + \bar{G})^T \tag{S13}$$

[1] A hyperprior is a prior distribution on a hyperparameter, that is, on a parameter of a prior distribution.

$$\bar{U} = \frac{1}{N_G}\sum_{i=1}^{N_G}(G_i - \beta_{\text{gene}}^T x_i) \qquad \bar{S} = \frac{1}{N_G}\sum_{i=1}^{N_G}(G_i - \beta_{\text{gene}}^T x_i)(G_i - \beta_{\text{gene}}^T x_i)^T \tag{S14}$$

While in BPMF the Gaussian priors model the latent variables G_i , in our proposed work the Gaussian priors model the residual $G_i - \beta_{\text{gene}}^T x_i$ instead; this being the key difference in comparison with BPMF. The conditional probability for precision parameter if the weight vector, $\lambda(\beta_{\text{gene}})$, can be obtained from (14) and (15)

$$P(\lambda_{\beta_{\text{gene}}} \mid \beta_{\text{gene}}, \Lambda_G, \mu, \nu) = gamma(\lambda_{\beta_{\text{gene}}} \mid \bar{\mu}, \bar{\nu}) \tag{S15}$$

where

$$\bar{\nu} = F_{\text{gene}} D + \nu \qquad \bar{\mu} = \frac{(F_{\text{gene}} D + \nu)\mu}{\nu + \mu tr(\beta_{\text{gene}} \Lambda_G^{-1} \beta_{\text{gene}}^T)} \tag{S16}$$

Note that the conditional probability for $\lambda_{\beta_{\text{phen}}}$ has exactly the same form.

1.4 A Detailed on our Noise Injection Sampler

Based on equations S12 and S14, the conditional probability for β_{gene} can be reformulated as

$$\begin{aligned} &P(\beta_{\text{gene}} \mid \mu_{gene}, \Lambda_G, G, x, \lambda_{\beta_{\text{gene}}}) \\ &\propto exp(-\frac{1}{2}\sum_{i=1}^{N_G}(G_i - \mu_G - \beta_{\text{gene}}^T x_i)^T \Lambda_G (G_i - \mu_G - \beta_{\text{gene}}^T x_i)) \\ &- \frac{1}{2}\lambda_{\beta_{\text{gene}}} tr(\beta_{\text{gene}} \Lambda_G^{-1} \beta_{\text{gene}}^T)) \end{aligned} \tag{S17}$$

Let X represents $X = [x_1, \cdots, x_N]$ and $U = [G_1 - \mu_G, \cdots, G_N - \mu_G]$. As both the likelihood and prior terms contain Λ_G, thus it can be factored out:

$$\begin{aligned} &P(\beta_{\text{gene}} \mid \mu_{gene}, \Lambda_G, G, x, \lambda_{\beta_{\text{gene}}}) \\ &\propto exp(-\frac{1}{2} tr[(G - X\beta_{\text{gene}})^T (G - X\beta_{\text{gene}}) + \lambda_{\beta_G}(\beta_{\text{gene}} \beta_{\text{gene}}^T)\Lambda_G]) \end{aligned} \tag{S18}$$

Then, the parameters of the Gaussian (mean and precision) can be reformulated as:

$$\begin{aligned} &P(\beta_{\text{gene}} \mid \mu_{gene}, \Lambda_G, G, x, \lambda_{\beta_{\text{gene}}}) \\ &\propto exp(-\frac{1}{2} vec(\beta_{\text{gene}} - \hat{\beta}_{gene})^T (\Lambda_G \otimes (X^T X + \lambda_{\beta_{\text{gene}}} I)) vec(\beta_{\text{gene}} - \hat{\beta}_{\text{gene}})) \end{aligned} \tag{S19}$$

where $\beta_{\text{gene}} = (X^T X + \lambda_{\beta_{\text{gene}}} I) X^T G$ is the mean and $\Lambda_G \otimes (X^T X + \lambda_{\beta_{\text{gene}}} I)$ is the precision of the posterior.

1.5 Supplementary details of our assessment strategy

Table S1 shows the details of our first assessment strategy. As explained in the paper among 2625 disease-gene annotations extracted from OMIM database (Amberger *et al*., 2011) , six random splits into 90% training (2363 annotations) and 10% (262 annotations) test data prepared. The generated test sets includes 131, 138, 148, 133, 150 and 139 diseases.

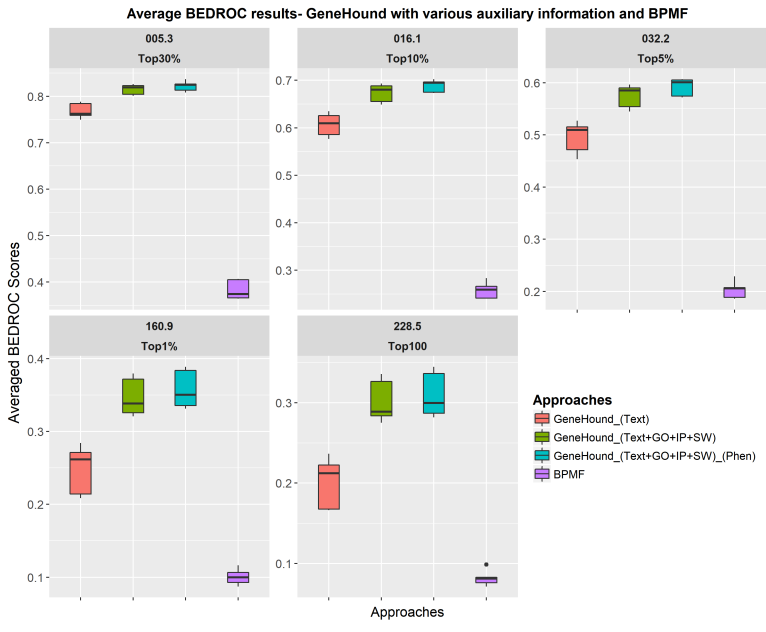

Fig. S1. Average BEDROC scores result: GeneHound with various side information vs BPMF. The performance of GeneHound with various side information, and BPMF are evaluated on OMIM1 benchmark. We use 40 latent dimensions for all proposed models. The label of each panel corresponds to the value of α used to evaluate the model. A greater α, emphasizes more on early recognition. BPMF does not use side information, and consequently it fails to provide an accurate matrix completion. OMIM1 benchmark was designed to investigate the advantage of incorporating side information into matrix factorization process. The number of training genes on this benchmark for some diseases are too low to have a fair and accurate comparison with Endeavour. Accordingly, the OMIM2 benchmark was designed to have a fair comparison between Endeavour and our proposed method.

2 A detailed discussion of results

Table S1 shows 65 diseases that we use to evaluate our Bayesian data fusion model for gene prioritization. For each disease, the total number of available annotations in OMIM for theses diseases used for both training and test are also reported in the Table S1.

Table S1: The 65 diseases that we investigated in this study. The total number of disease-linked genes (including the training and test) for each disease investigated in this work are listed. The OMIM diseases investigated in this study are grouped based on the 10th version of the International Statistical Classification of Diseases and Related Health Problems (ICD-10) (Icd, 2010).

Diseases Name	Number of genes	BEDROC scores $\alpha = 160.9$	ICD-10 chapter
AIDS	45	0.49	A
Malaria	16	0.28	A
Mycobacterium tuberculosis	11	0.31	A
Colorectal cancer	55	0.34	C
Gastric cancer	12	0.44	C
Leukemia	56	0.34	C
Lipoma	14	0.25	C
Lung cancer	15	0.32	C

	Pancreatic cancer	10	0.40	C
	Prostate cancer	15	0.19	C
	Renal cell carcinoma	15	0.26	C
	Diamond-Blackfan anemia	10	1.00	D
	Fanconi anemia	16	0.85	D
	Immunodeficiency	15	0.30	D
	Neutropenia	13	0.09	D
	Spherocytosis	16	0.31	D
	Thrombocytopenia	11	0.40	D
	Ceroid lipofuscinosis	10	0.74	E
	Combined oxidative phosphorylation deficiency	15	0.44	E
	Congenital disorder of glycosylation	15	0.76	E
	Glycogen storage disease	16	0.74	E
	Lipodystrophy	62	0.42	E
	Obesity	24	0.33	E
	Peroxisome biogenesis disorder	10	0.70	E
	Premature ovarian failure	10	0.27	E
	Alzheimer diseaser	14	0.37	F
	Mental retardation	82	0.59	F and G
	Parkinson disease	16	0.60	F and G
	Amyotrophic lateral sclerosis	41	0.58	G
	Centronuclear myopathy	30	0.46	G
	Charcot-Marie-Tooth disease	30	0.67	G
	Dystonia	13	0.24	G
	Epilepsy	29	0.55	G
	Epileptic encephalopathy	14	0.47	G
	Leigh syndrome	19	0.45	G
	Miller syndrome	14	0.31	G
	Neuronopathy	23	0.49	G
	Spastic paraplegia	11	0.57	G
	Spinal muscular atrophy	10	0.50	G
	Cataract	28	0.57	H and Q
	Retinitis pigmentosa	122	0.58	H
	Usher syndrome	11	0.86	H
	Deafness	74	0.79	H2
	Cardiac arrhythmia	12	0.58	I
	Coronary artery disease	11	0.27	I

Left ventricular noncompaction	46	0.55	I
Myocardial infarction	13	0.19	I
Asthma and nasal polyps	16	0.20	J
Ciliary dyskinesia	18	0.72	J
Rheumatoid arthritis	10	0.25	M
Systemic lupus erythematosus susceptibility	12	0.35	M
Spermatogenic failure	30	0.25	N
Bardet-Biedl syndrome	17	0.84	Q
CHARGE syndrome	14	0.02	Q
Ectodermal dysplasia	10	0.35	Q
Ehlers-Danlos syndrome	10	0.52	Q
Ichthyosis	14	0.31	Q
Joubert syndrome	27	0.76	Q
Meckel syndrome	16	0.66	Q
Metaphyseal anadysplasia	13	0.34	Q
Microcephaly	15	0.22	Q
Osteogenesis imperfecta	21	0.44	Q
Osteopetrosis	14	0.46	Q
Hypogonadotropic hypogonadism with or without anosmia	19	0.40	R
Mitochondrial complex deficiency	28	0.64	NA

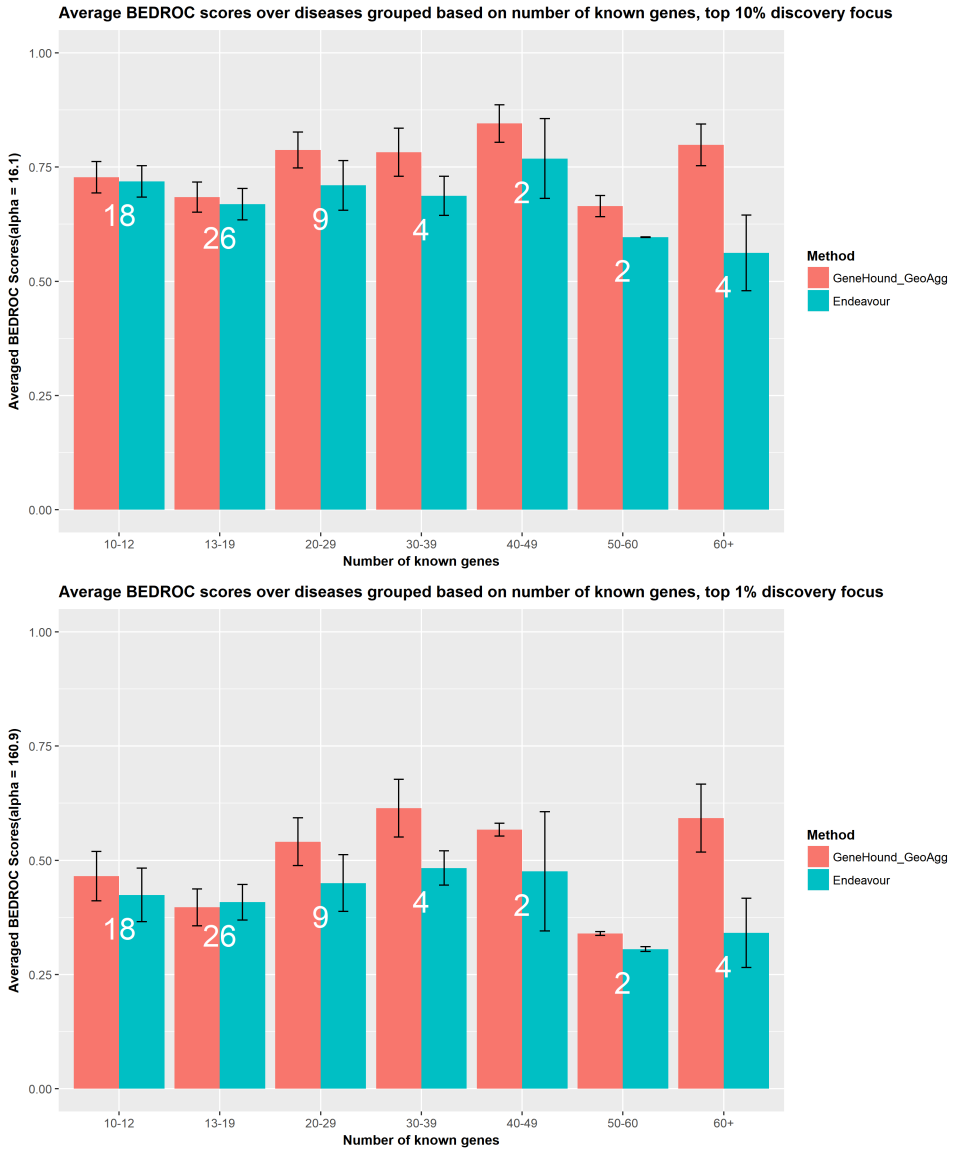

Fig. S2. The average BEDROC scores over diseases grouped based on the number of known genes for 2 different enrichment focuses. GeneHound and Endeavour are evaluated on OMIM1 benchmark. The label of each bar corresponds to the number of diseases in each group. A greater α, emphasizes more on early recognition.

Figure S2 shows that GeneHound improves the predictive performance at top 1% enrichment focus by 9.68% on diseases with 10, 11, and 12 known genes, 20.11% on diseases with 20 to 29 known genes, 27.13% on diseases with 30 to 39 known genes, 19.27% on diseases with 40 to 49 known genes, 11.12% on diseases with 50 to 59 known genes, and 73.74% on diseases with at least 60 training genes, whereas the largest performance drop was 2.73% on diseases with 13 to 19 known genes. These results indicate that our proposed gene prioritization method can enhance the predictive performance of gene prioritization task by employing the multi-task approach.

Author's contributions

PZ, JS, AA and YM conceived, designed and developed the study and models. PZ developed and performed the experiments. PZ and JS analyzed the results. PZ and SE developed data sources and benchmarks. PZ wrote the the manuscript and all authors edited and improved the manuscript. All authors read and approved the final manuscript.

Table S2. Early discovery improvements challenge: GeneHound vs Endeavour.It is observed that GeneHound improves the BEDROC score (at $\alpha = 228.5$ which corresponds to top 1% early discovery focus) obtained using Endeavour by more than 50% and 100% for twelve and three diseases, respectively. This compares to three and zero BEDROC score improvements of diseases in OMIM2 benchmark using Endeavour.

Methods	**Disease names**	
Early enrichment improvement	**more than** 50%	**more than** 100%
GeneHound improves Endeavour	Ceroid lipofuscinosis (E) Charcot-Marie-Tooth (H2) Deafness(H2) Osteogenesis imperfecta (Q) Epilepsy(G) Gastric cancer(C) Lipoma(C) Spinal muscular atrophy (G) Amyotrophic lateral sclerosis (G) Retinitis pigmentosa (H) Mental retardation(F) Coronary artery disease(I)	Osteogenesis imperfecta (Q) Lipoma(C) Spinal muscular atrophy (G) Mental retardation(F)
Endeavour improves GeneHound	Ectodermal dysplasia (Q) Microcephaly (Q) Spherocytosis (D)	none
Final Results	12-3	4-0

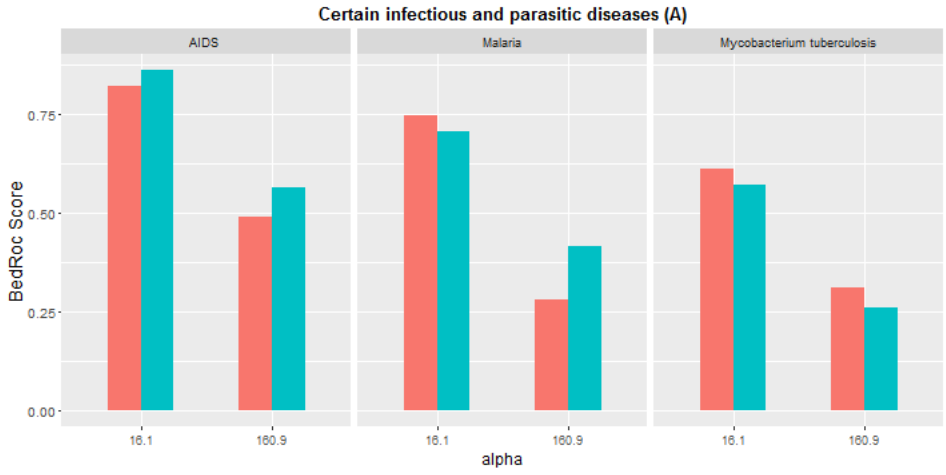

Fig. S3. BEDROC scores result for the Certain infectious and parasitic diseases(A). The left bar shows the result of the GeneHound, and the right bar indicates the result of Endeavour (Auhtor et al. , 2006; Tranchevent *et al*., 2016)

.

References

Mnih A, Salakhutdinov R. Probabilistic matrix factorization. In Advances in neural information processing systems (NIPS). 2007;1257-1264.

Amberger J, Bocchini C, Hamosh A. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM). Hum Mutat. 2011;32(5):564-567.

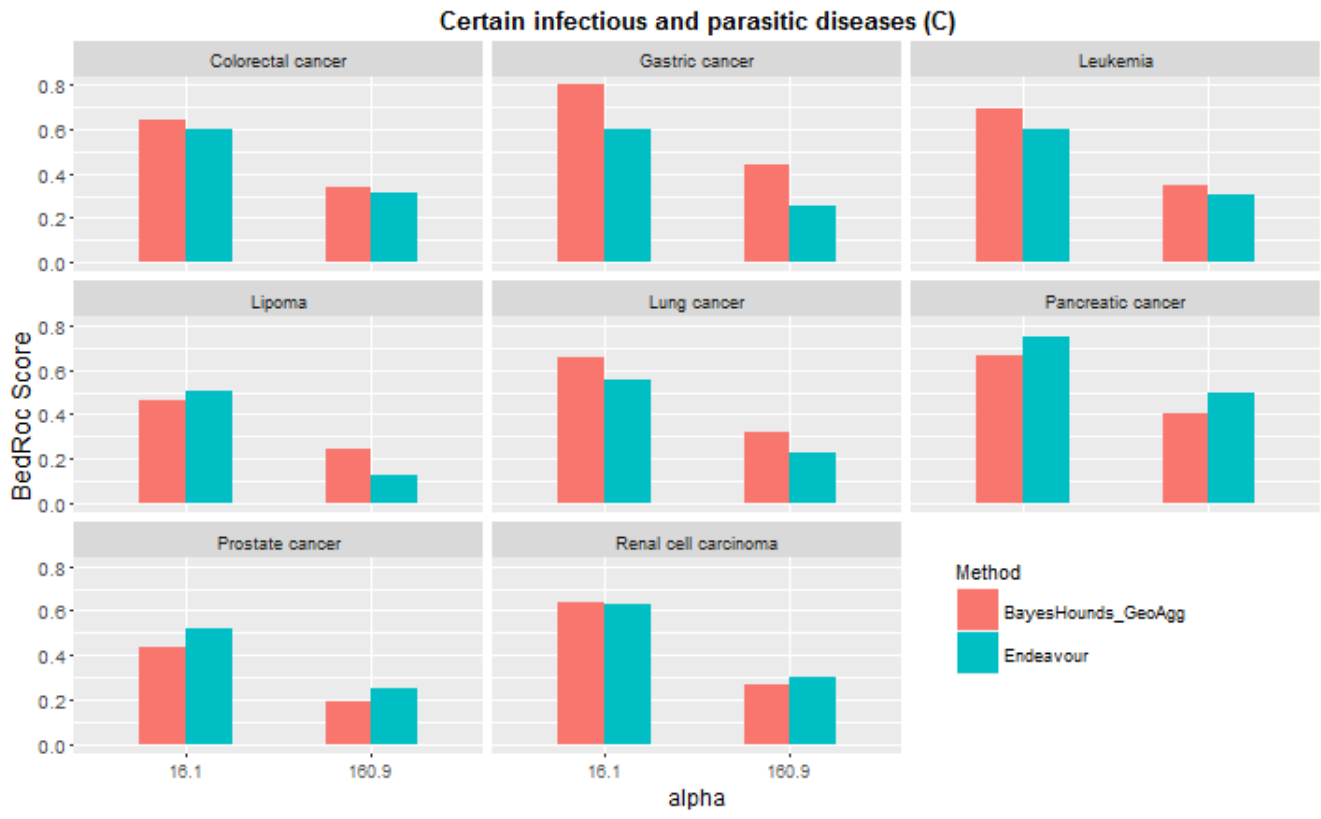

Fig. S4. BEDROC scores result for diseases classified as the Neoplasm (C) in ICD-10.

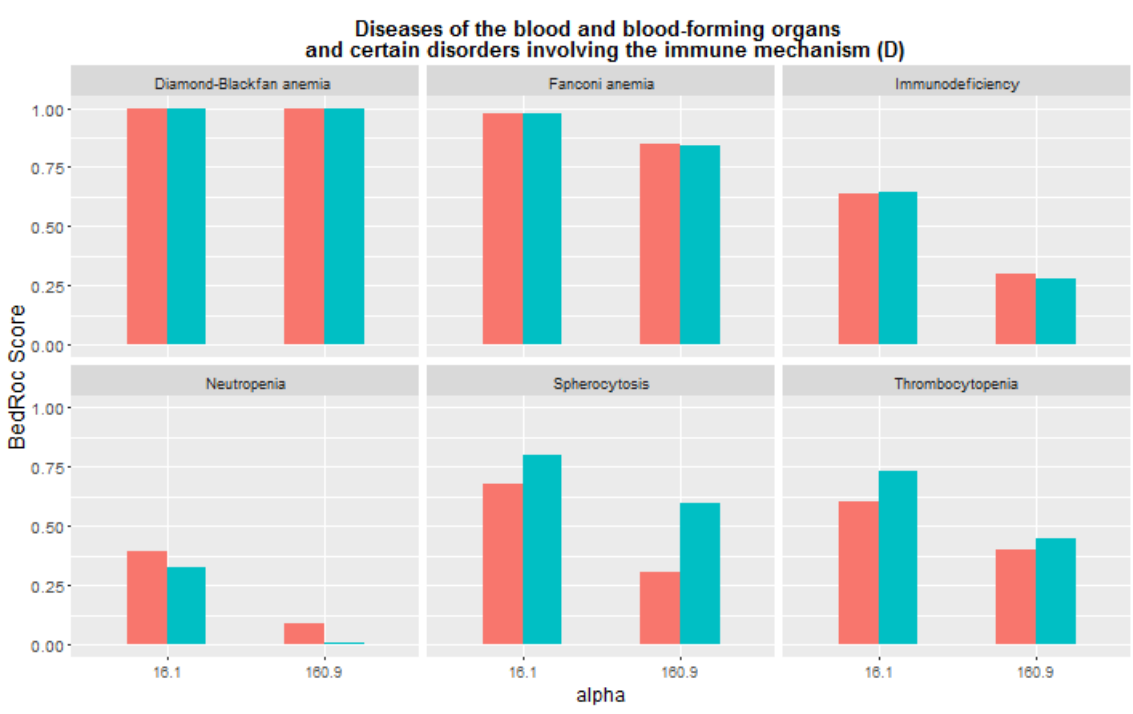

Fig. S5. BEDROC scores result for the disease of the blood and blood-forming organs and certain disorders involving the immune mechanism (D).

Aerts S, et al. Gene prioritization through genomic data fusion. Nat Biotech, 2006;24(5):537-544.

Porteous I, Asuncion AU, Welling M. Bayesian matrix factorization with side information and dirichlet process mixtures. in AAAI Conference on Artificial Intelligence, 2010.

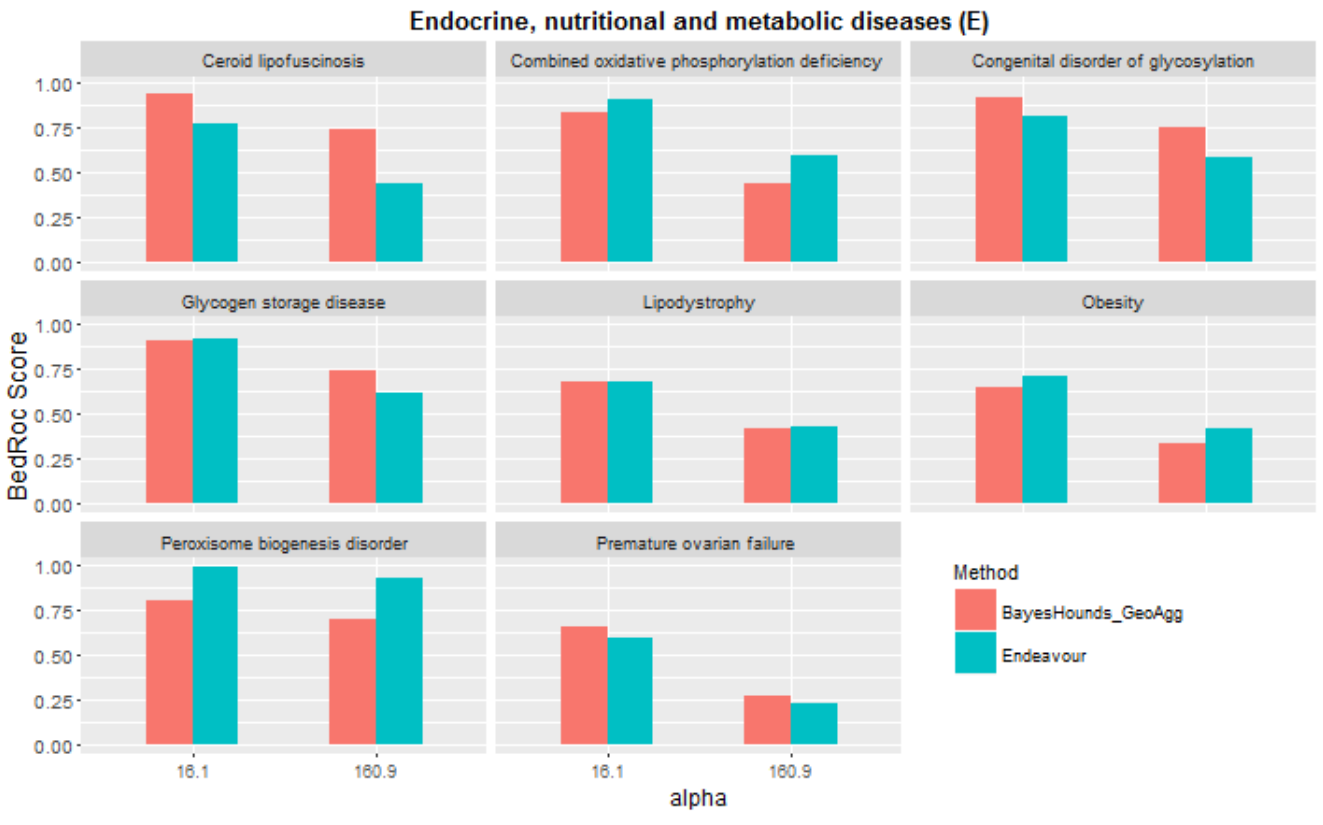

Fig. S6. BEDROC scores result for the endocrine, nutritional and metabolic diseases(E).

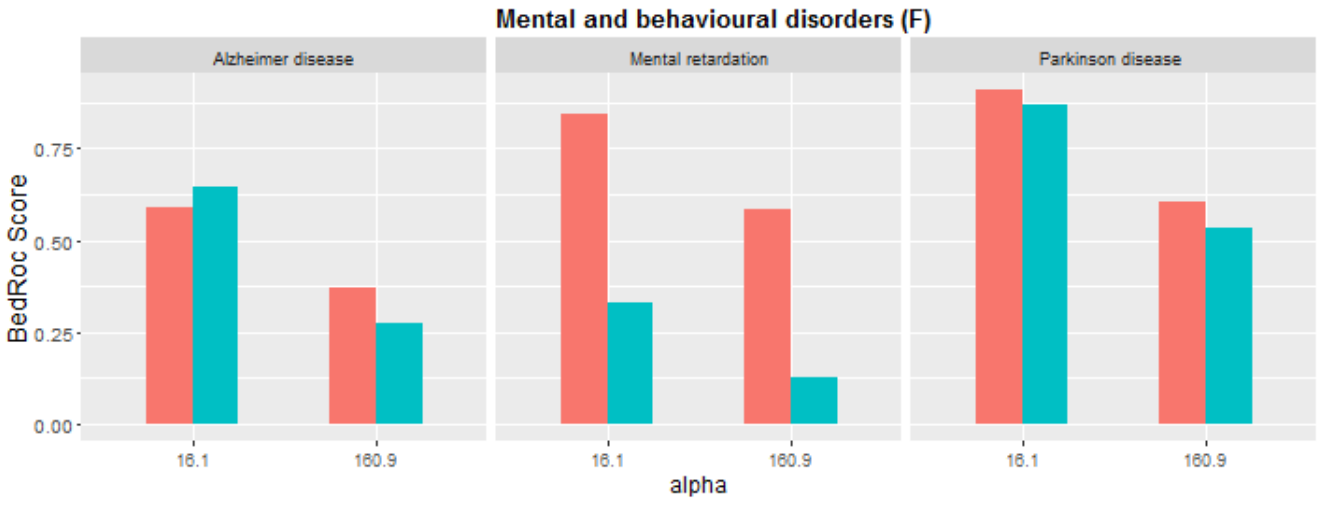

Fig. S7. BEDROC scores result for the disease of Mental and behavioural disorders (F).

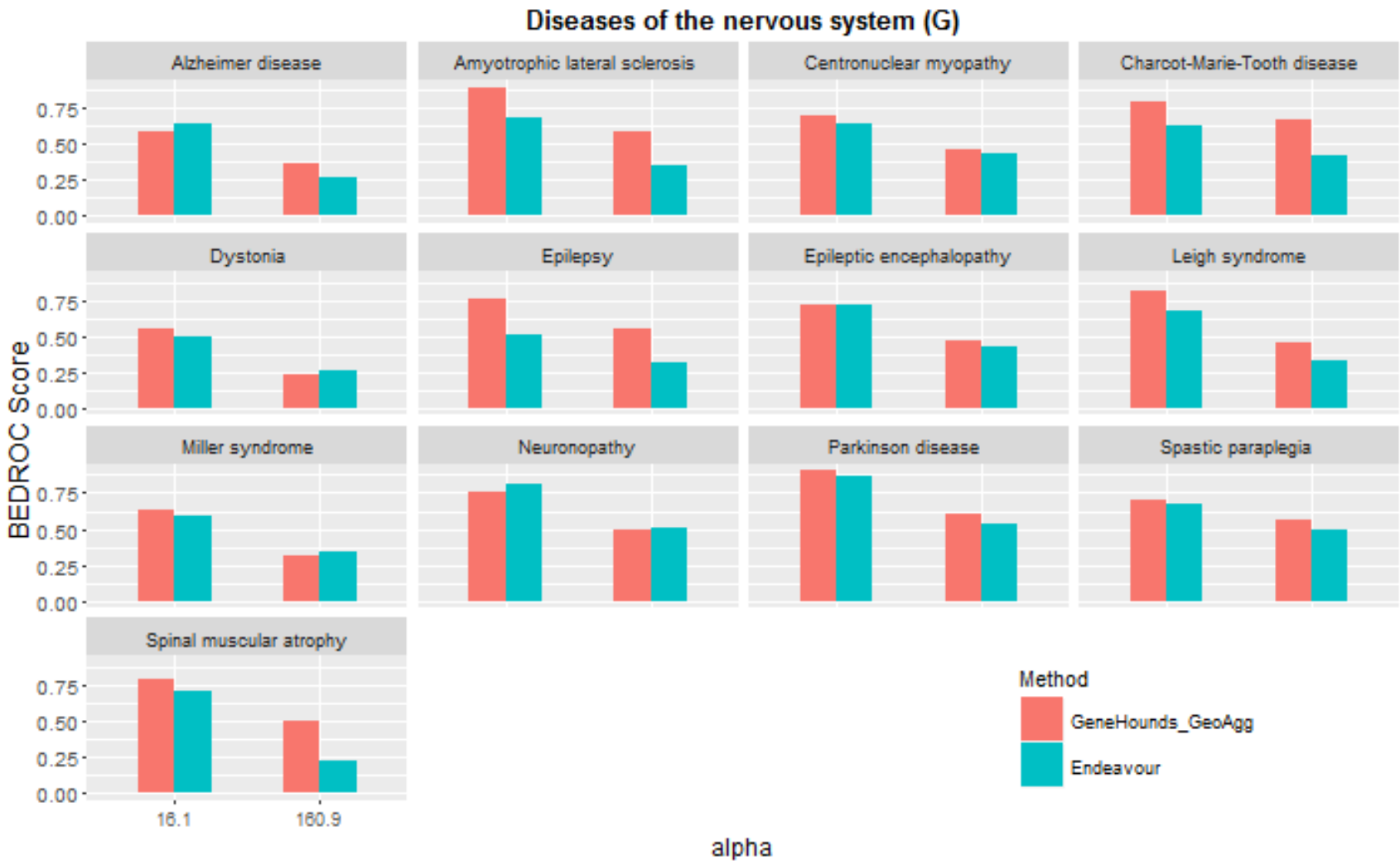

Fig. S8. BEDROC scores result for diseases of the nervous system (G).BEDROC score for each disease is calculated over all cross-validated genes. The α are set to 16.1 and 160.9 which correspond to 80% of the BEDROC being assigned to the top 10% and 1% prioritized genes. The left bar shows the result of $GeneHound\_GeoAgg$, and the right bar indicates the result of Endeavour, $GeneHound\_GeoAgg$ offers the best average BEDROC score of 0.73 and 0.48 overall diseases of the nervous system score at 10% and 1% early discovery focuses, respectively. This compares to 0.66 and 0.37 using Endeavour.

Salakhutdinov R, Mnih A. Bayesian probabilistic matrix factorization using markov chain monte carlo. in Proceedings of the 25th international conference on Machine learning (ACM). 2008;pp. 880-887.

Tranchevent LC, et al. Candidate gene prioritization with Endeavour. Nucleic Acids Res. 2016;44(W1):W117-W121.

International Classification of Diseases, Tenth Revision, Clinical Modification. Retrieved 2015, https://www.cdc.gov.

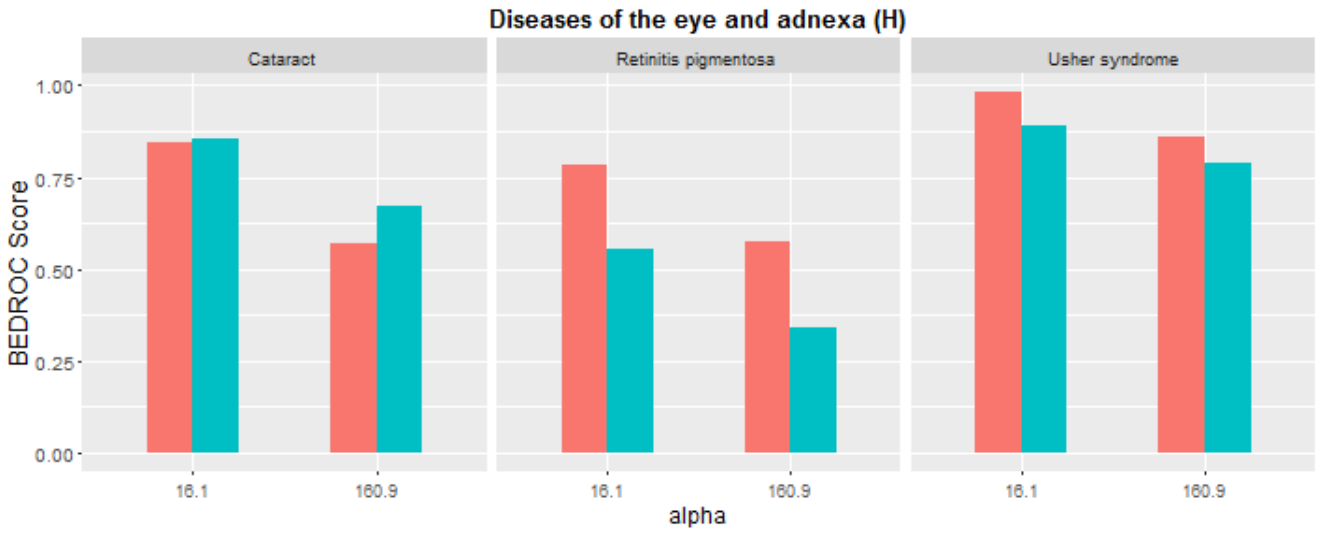

Fig. S9. BEDROC scores result for diseases of the eye and adnexa (H).BEDROC score for each disease is calculated over all cross-validated genes. $GeneHound\_GeoAgg$ offers the very promising average BEDROC scores of 0.86 and 0.66 overall diseases of the eye and adnexa score at 10% and 1% early discovery focuses, respectively. These are versus 0.66 and 0.37 using Endeavour.

enrichment focuses.

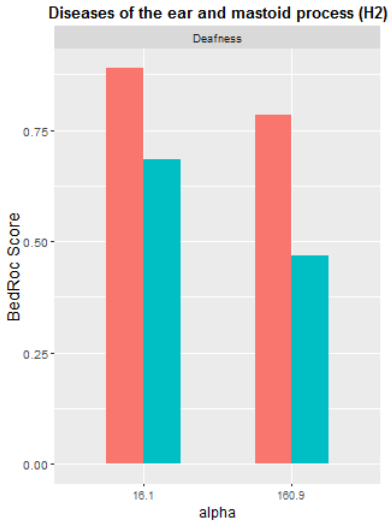

Fig. S10. BEDROC scores result for the disease of ear and mastoid process (H2).

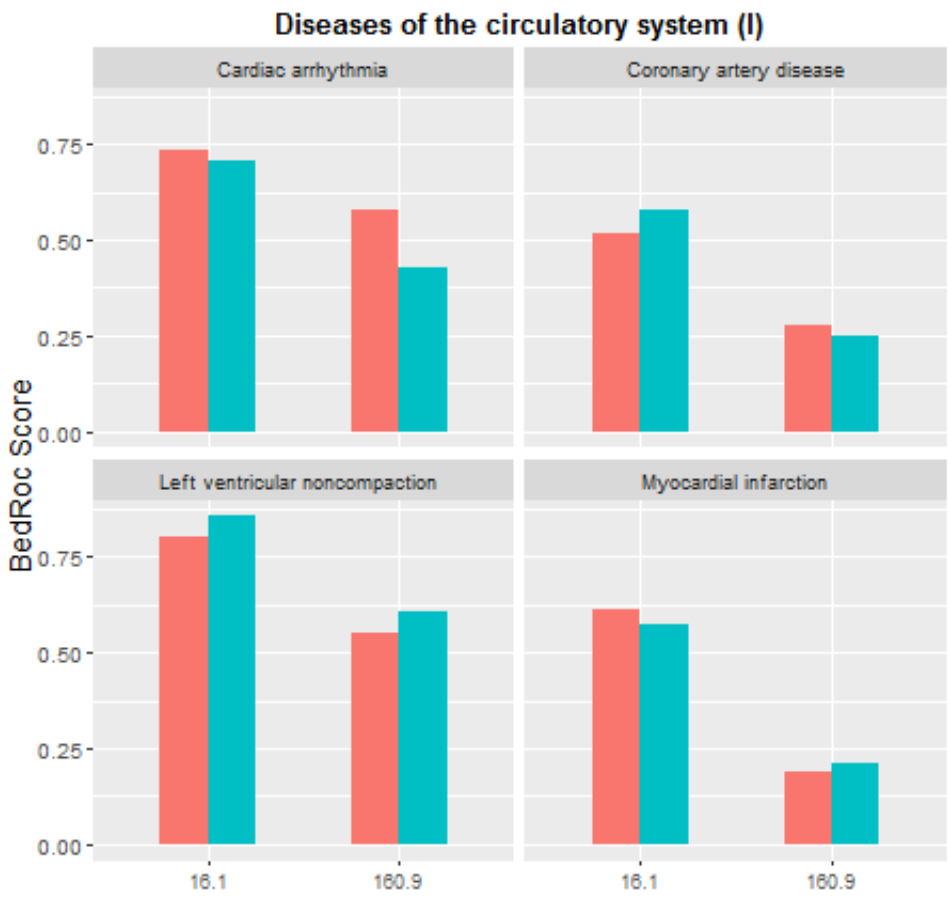

Fig. S11. BEDROC scores result for the disease of the circulatory system (I).

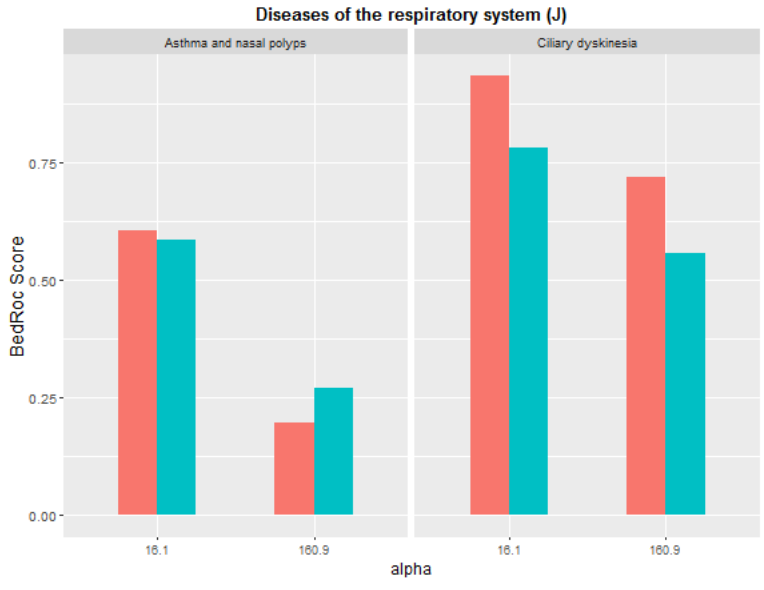

Fig. S12. BEDROC scores result for the disease of the respiratory system (J).

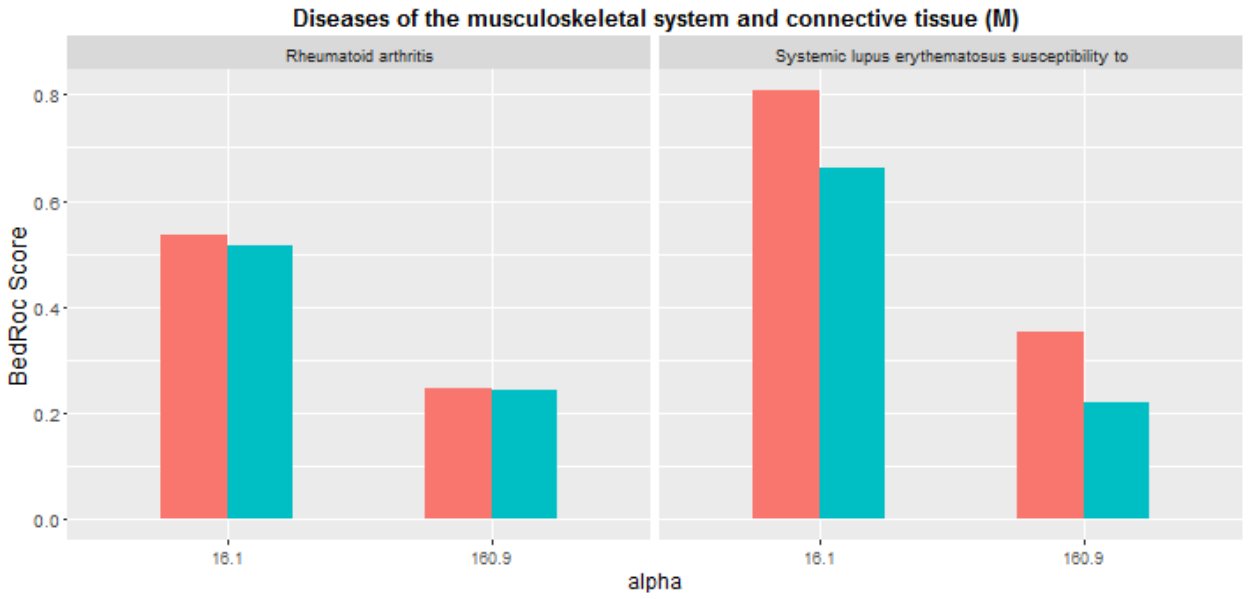

Fig. S13. BEDROC scores result for the disease of musculoskeletal connective tissue (M).

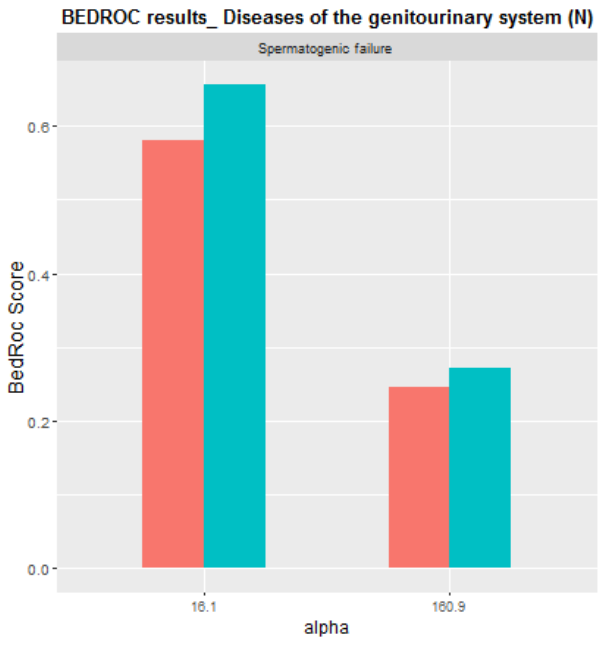

Fig. S14. BEDROC scores result for the disease of the genitourinary system (N).

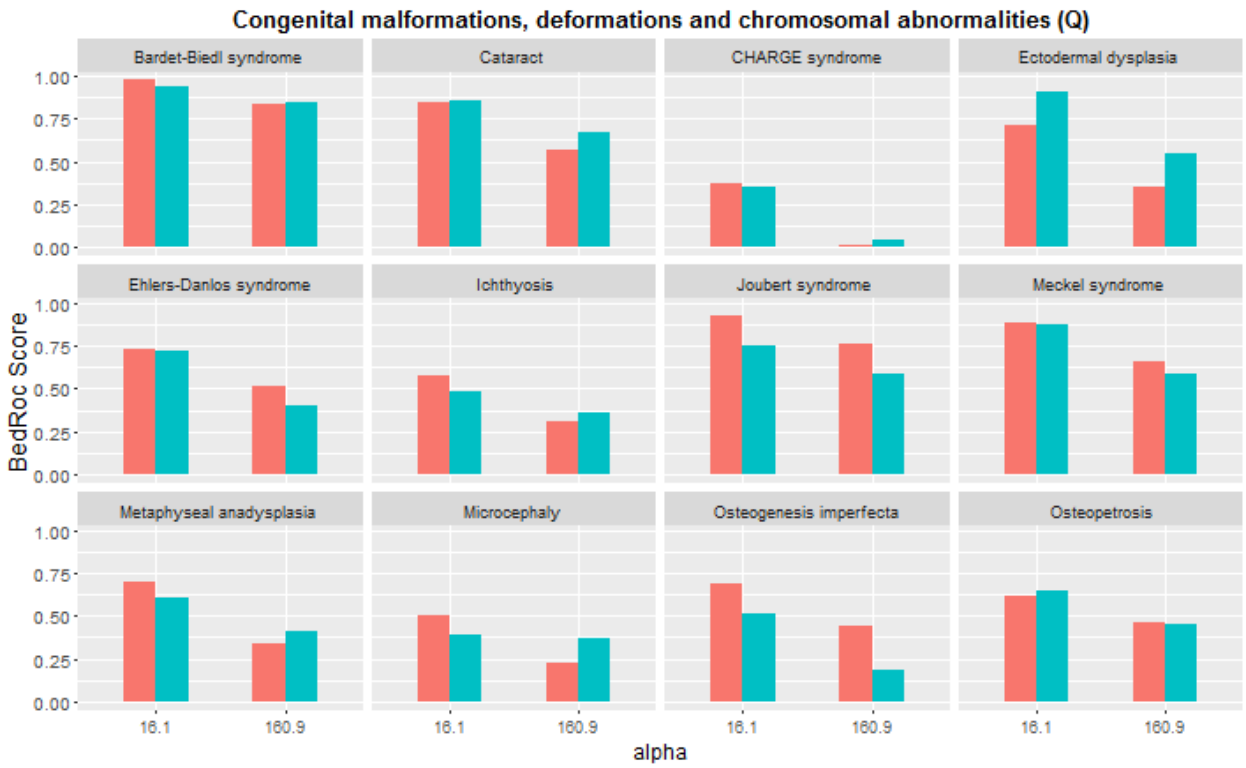

Fig. S15. BEDROC scores result for the congenital malformations, deformations and chromosomal abnormalities (Q).

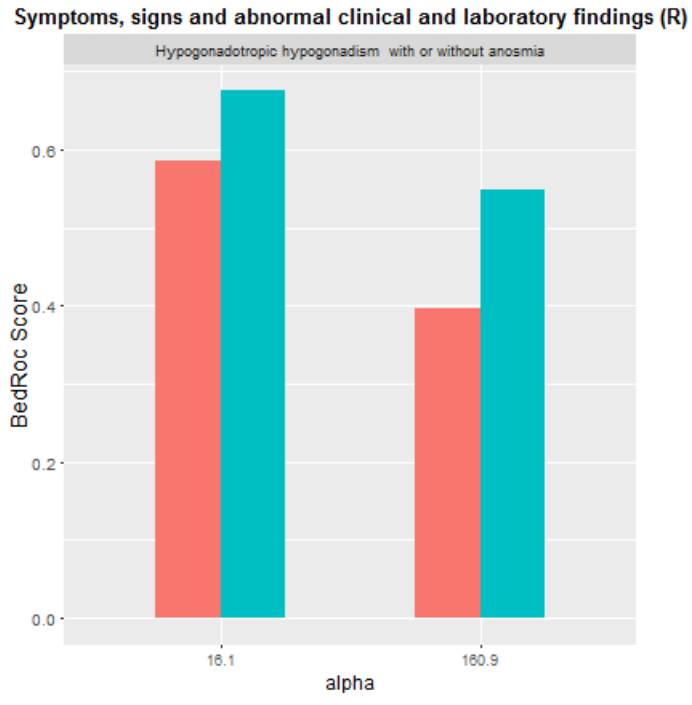

Fig. S16. BEDROC scores result for the disease of symptoms, signs and abnormal clinical and laboratory findingg (R).

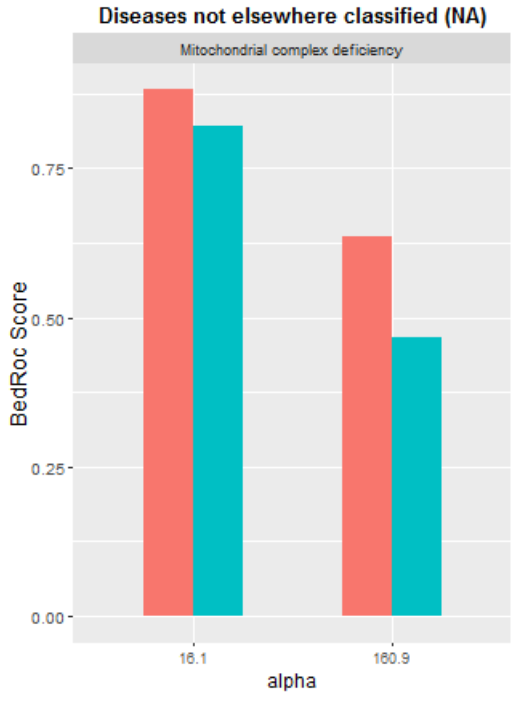

Fig. S17. BEDROC scores result for the Mitochondrial complex deficiency (NA) Mitochondrial complex deficiency is not classified in ICD-10.