Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

*IEEE/ACM Trans Comput Biol Bioinform*. Author manuscript; available in PMC 2017 August 18.

Published in final edited form as:

*IEEE/ACM Trans Comput Biol Bioinform*. 2015 ; 12(1): 30–39. doi:10.1109/TCBB.2014.2359446.

# Phenotype-Dependent Coexpression Gene Clusters: Application to Normal and Premature Ageing

**Kun Wang**,
Center for Bioinformatics and Computational Biology, 3111 Biomolecular Sciences Building, University of Maryland, College Park, MD 20742

**Avinash Das**,
Department of Computer Science, 3111 Biomolecular Sciences Building University of Maryland, College Park, MD 20742

**Zheng-Mei Xiong**,
Department of Cell Biology and Molecular Genetics, 2219 Bioscience Research Building, University of Maryland, College Park, MD 20742

**Kan Cao**, and
Department of Cell Biology and Molecular Genetics, 2114 Bioscience Research Building, University of Maryland, College Park, MD 20742

**Sridhar Hannenhalli**
Center for Bioinformatics and Computational Biology, 3104G Biomolecular Sciences Building, University of Maryland, College Park, MD 20742

## Abstract

Hutchinson Gilford progeria syndrome (HGPS) is a rare genetic disease with symptoms of aging at a very early age. Its molecular basis is not entirely clear, although profound gene expression changes have been reported, and there are some known and other presumed overlaps with normal aging process. Identification of genes with aging- or HGPS-associated expression changes is thus an important problem. However, standard regression approaches are currently unsuitable for this task due to limited sample sizes, thus motivating development of alternative approaches. Here, we report a novel iterative multiple regression approach that leverages co-expressed gene clusters to identify gene clusters whose expression co-varies with age and/or HGPS. We have applied our approach to novel RNA-seq profiles in fibroblast cell cultures at three different cellular ages, both from HGPS patients and normal samples. After establishing the robustness of our approach, we perform a comparative investigation of biological processes underlying normal aging and HGPS. Our results recapitulate previously known processes underlying aging as well as suggest numerous unique processes underlying aging and HGPS. The approach could also be useful in detecting phenotype-dependent co-expression gene clusters in other contexts with limited sample sizes.

### Index Terms

Algorithms; parameter learning; models; biology and genetics

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

## 1 Introduction

HGPS is a genetic disorder that is identified in approximately one out of four million live births. Although rare, this autosomal dominant disease has severe effects-children with HGPS undergo accelerated aging and face average life expectancy of just 13.4 years. Visible symptoms can include a pronounced forehead, short stature, receding mandible, conspicuous veins in the scalp, hair loss, a “pinched” nose, and extreme lipodystrophy. Internally, patients with HGPS undergo accelerated organ degeneration, and death typically results from coronary artery disease or stroke [1]. Approximately 90 percent of the HGPS cases are caused by a de novo mutation at position 1,824 of the lamin A gene LMNA (C1824T, G608G) [2]. Lamin A is a major nuclear structural component and has several other important functions, including its support and regulation of protein complexes that participate in nuclear positioning, DNA replication, gene expression, transcription, and repair [3]. The HGPS G608G mutation activates a cryptic splicing site in exon 11 and produces a lamin A deletion mutant named progerin [4].

A number of hypotheses have been developed to explain the mechanisms leading to the clinical manifestations of HGPS. Among them, the “gene expression” model, which proposes that progerin alters the nuclear structure and subsequently affects gene expression, has been supported by various lines of evidence [5]. A general loss of heterochromatin and dislocation of epigenetic marks have been observed in HGPS cells [6], [7], [8]. In addition, it has been shown that lamin A interacts with transcription regulatory proteins (e.g., retinoblastoma protein pRb), signaling molecules (e.g., protein kinase C), and chromatin proteins (e.g., histones and barrier-to-autointegration factor (BAF)), implicating its direct involvement in gene expression and signaling [9], [10], [11].

Accordingly, changing expression levels of various genes have been observed in HGPS cells. To date, four independent HGPS microarray studies have been published. Park et al. examined 384 known genes and reported four genes with more than twofold changes [12]. Ly et al. monitored the expression of approximately 6,000 genes and found 61 altered in HGPS [13]. Csoka et al. analyzed approximately 33,000 genes and found 361 genes that showed statistically significant change [14], and more recently, Marji et al. compared 4 HGPS fibroblast lines with four age-matched controls, and suggested that a lamin A-Rb signaling is a major defective signaling pathway in HGPS cells [15]. While these microarray studies are not in complete agreement with each other, transcription factors, extracellular proteins, and cell cycle regulators appear to be the largest affected functional category.

As the relationship between nuclear lamins and gene expression is continued to be explored, we are optimistic that the gene expression model may help to shed light on the causes of the premature aging phenotypes associated with HGPS. On the other hand, it is of great interests to determine how the gene expression pattern in this disease resembles and is distinct from the pattern observed in normal aging. A detailed comparative investigation of genome-wide gene expression patterns associated with HGPS and normal cellular aging has not yet been reported and may reveal common and distinctive biological pathways underlying these two conditions. Comparative exploration of gene expression changes in normal aging and HGPS has not been possible thus far due to unavailability of genome-wide expression profiles in

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

HGPS samples at different cellular ages. Thus, in this study, we have collected RNA samples from a HGPS primary fibroblast cell line and from a genetic background matched normal control at early, middle and late cellular passages, and conducted genome-wide RNA-seq.

Although, we have generated the first whole genome RNA-seq based transcriptomic profile in cell cultures at three different "ages" in both normal and HGPS samples, the number of samples (n = 6) is not sufficient to assess individual genes with respect to their co-variation with age or HGPS using standard regression approaches, such as those used for eQTL studies, with hundreds of samples [16]. At the same time genes are known to form co-expression clusters reflecting common or interdependent regulatory mechanisms, and the traditional gene-centric regression approach does not leverage this fact. To address the limitations in the sample size, we have developed an iterative procedure that leverages co-expressed gene clusters while iteratively refining the cluster based on a cluster-centric multivariate regression's goodness of fit criteria. We have performed a number of tests to show the robustness and efficacy of the approach.

We have applied our approach to the RNA-seq profiles in six samples—three healthy and three HGPS samples at three different ages, as approximated by the number of passages. We then comparatively investigated the various clusters whose expression significantly co-varied with age, or disease, or both. Our results recapitulated previous findings on biological processes involved in the aging process, and while revealing some parallels between the aging process and HGPS revealed several important differences. Overall, we found that the HGPS gene expression profiles to be substantially different from those for normal fibroblast passaged into cellular senescence. Gene clusters showing decreasing expression with aging showed a significant enrichment of genes are related to cell cycle regulation, as noted previously, despite major differences in samples, methods, and data analysis [13]. Also, consistent with previous reports, gene clusters with age-associated increase in expression we enriched for genes involved in programmed cell death regulation and ECM organization. Similarly, gene clusters with age-associated decrease in expression were related to cell cycle regulation. For instance, Forkgead box protein M1 (FOXM1), a key transcriptional regulator of a large group G2-M specific genes was down-regulated in older samples. This gene has been shown to regulate a large group of G2-M specific genes [19], including a key mitotic cyclin, cyclin B1. In summary, based on a novel, broadly applicable approach, our work establishes a first benchmark study directly comparing the transcriptomic changes underlying two related phenotypes—normal aging and premature aging.

## 2 Material and Method

### 2.1 Cell Culture, RNA Preparation, and RNA-Seq Experiment

Normal (HGADFN168, Father) and HGPS (HGADFN167, son) primary fibroblasts were obtained from the Progeria research foundation. All fibroblast cell lines were cultured in MEM (Life Sciences) supplemented with 15 percent FBS (Gemini Bio-Products) and 2 mM L-glutamine (Life Sciences) at $37°C$ with 5 percent $CO_2$. RNA samples were collected from these two cell lines at early (passage 11), middle (passage 16) and late stages (passage 20 for HGPS, and passage 23 for normal) during replicative senescence. Total RNA from different cell lines was extracted with Trizol (Life Sciences) and purified using the RNeasy Mini Kit

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(Qiagen) according to the manufacturer's instructions. The RNA yield was determined using the NanoDrop 2000 spectrophotometer. The RNA-seq sample preparation and sequencing were conducted according to the illumina Truseq RNA sample preparation V2 guide by the IBBR sequencing Core facility at the University of Maryland.

### 2.2 RNA-Seq Data Processing

We processed each of the six samples identically using the Cufflinks suite of tools following the recommended protocol [17] yielding RNA expression value (FPKM) for ~14,000 human genes in each of the six samples. In addition, to guide the iterative cluster refinement procedure (see below), we obtained the RNA-seq profiles for 15 independent tissue types from Gene Expression Omnibus (GEO) [18]. There were 9,453 genes in common between our samples and the GEO samples, which were ultimately used for all follow up analyses.

### 2.3 Joint Regression Clustering

**Workflow—**Fig. 1 shows the complete workflow of proposed joint regression clustering method. We start with initial clustering of genes (k-mean clustering) based on age-progeria data. We refine the clusters iteratively to minimize the average error of predicted (from cluster regression) gene expression till convergence. It is computationally expensive to calculate the average error for all possible refinements. Therefore, we approximated the average error by its maximum likelihood (ML) estimate. After every k rounds if actual average error increases, we randomly reverse some of the gene reassignments.

**Linear regression model—**Linear regression is widely used method to study the effect of covariates on expression variance between samples. The linear regression model for gene expression with aging and HGPS as covariates can be expressed as:

$$g_{ij}=\mu_j+\beta_j^1 a_i+\beta_j^2 d_i+\beta_j^3 a_i d_i. \quad (1)$$

Where, $g_{ij}$ is expression of $j$th gene in $i$th sample. $\mu_j$ is the basal expression of $j$th gene. $\beta_j$ is vector of the regression coefficients of $j$th gene for covariates age $a_i$ (1: young, 2: middle age, 3: old) and HGPS state $d_i$ (0: normal, 1: HGPS), and interaction term $a_i d_i$. A vector of coefficients must be estimated for each gene separately for model (1). This is clearly limiting as we have only six samples. Moreover, there are thousands of genes that will be differentially expressed in different sample. To learn regression coefficient separately for each gene is not an effective approach because small sample size will have low statistical power (see results) and many genes are expected to vary in a similar manner with respect to the covariates. Additionally, it will be hard to extract meaningful result and visualize the effect of covariates from separate regression coefficients for thousands of genes. We can cluster genes based on its expression variance w.r.t its covariates and estimate coefficients jointly for a cluster of co-varying genes.

**RegressionClust model—**To overcome above limitations and to leverage clusters of potentially co-varying genes we propose the following model, RegressionClust:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$g_{ijc}=\mu_{ic}+\beta_c^1 a_i+\beta_c^2 d_i+\beta_c^3 a_i d_i \quad g_{ij}-w_{jc}=g_{ijc}. \quad (2)$$

Where, $g_{ijc}$ is imputed expression of $j$th gene belonging to cth gene-cluster in ith sample. $\mu_{ic}$ is the basal expression of genes in $c$th gene-cluster in ith sample. $\beta_c$ is cluster specific vector of the regression coefficients of the covariates. $w_{jc}$ is distance of $j$th gene from its cluster center and is computed as the difference between the mean expression value of the gene and that of all genes in the cluster that the gene is assigned to. $w_{jc}$ is updated after each reassignment.

This is a dual optimization problem—fit a regression model for each cluster and refine clusters to maximize overall explained variance. The objective functions are:

1. Regression: find optimal regression coefficients such that gene expression variance within cluster explained by covariates is maximized,

   i.e. $argmin_{\beta_c} Q_c^2=\sum_i\sum_i (g_i(\beta_c)-g_{ijc})^2$.

   Where, $g_i(\beta_c)=\mu_{ic}+\beta_c^1 a_i+\beta_c^2 d_i+\beta_c^3 a_i d_i$

   This is equivalent to

   $$I=argmax_{\beta_c} R_c^2=1-\frac{\sum_j\sum_i (g_i(\beta_c)-g_{ijc})^2}{\sum_j\sum_i (\bar{g}-g_{ijc})^2} \quad (3)$$

   $\bar{g}$ is the mean gene expression value in cluster c.

2. Cluster refinement: find optimal set of clusters (or, clustering) such that each cluster is tight (maximize overall explained variance), i.e. minimize

   $$F^2=\sum_c F_c^2=\sum_c\sum_{j\in c}\sum_i |g_{ij}-w_{jc}-g_i(\beta_c)|^2. \quad (4)$$

**Inference—**Independent maximization of $R^2$ is equivalent to linear regression, while independent minimization of $F^2$ is clustering. We estimate the parameters of Regression-Clust model by iteratively optimizing the two objective functions. It is important to note that $w_{jc}$ should be independent of the expression variance due to the covariates, therefore we estimate them from gene expression of 15 independent normal expression samples collected from GEO database. As a side note, this iterative inference is similar to expectation maximization (EM) algorithm [22]. In particular, if instead of hard assignment of gene cluster, fuzzy assignment to cluster is used, it can be proved that it is equivalent to EM algorithm. However, we chose to use hard assignment because we found that fuzzy clustering increases computational cost without significant gain in the overall performance. Maximization of $R^2$ is explained next.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

**Initializing the cluster set—**We initialized the clusters using k-means clustering on the sample data, for different number of clusters (100, 200, 300, 400, 500, 600).

**Greedy maximal matching cluster refinement—**To greedily refine the clusters, a change in $F^2$ should be calculated for each possible reassignment (move of gene from each cluster to another). Each possible reassignment changes $\beta_c$ and $g_i(\beta_c)$. Running linear regression for each possible reassignment is clearly computationally limited. We therefore use maximum likelihood estimate of $g_i(\beta_c)$ to estimate change of $F^2$. The maximum likelihood estimate of $g_i(\beta_c)$ can be determined by differentiating equation (3) w.r.t to $g_i(\beta_c)$:

$$\begin{aligned} \frac{dI}{dg_i(\beta_c)} &= -\frac{2(\sum_j \sum_i (g_i(\beta_c) - g_{ijc}))}{\sum_j \sum_i (\overline{g} - g_{ijc})^2} = 0 \\ g_{iML}(\beta_c) &= \frac{1}{J_c} \sum_j gijc. \end{aligned} \quad (5)$$

Where $J_c$ is total number of genes in cluster c. Now, we can replace this in equation (4) to obtain its ML estimate:

$$\tilde{F}^2 = \sum_c \sum_{j \in c} \sum_i |g_{ij} - w_{jc} - g_{iML}(\beta_c)|^2 \quad (6)$$

$g_{iML}(\beta_c)$ as well as $w_{jc}$, can be locally updated efficiently for each possible single gene moves.

To allow at most (a single gene) change to a cluster in an iteration, we construct a graph with all clusters of current clustering as nodes and each single gene move as a directed edge between originating and destination cluster and change in $\tilde{F}^2$ due to the move as the edge weight. We then performed maximal matching on this graph to minimize $F^2$ and allowed single change to a cluster. We then proceed with single gene moves corresponding to maximally matched edge as our cluster refinement. The maximal matching also ensures that same cluster will not be source of one gene and destination for some other gene as shown in Fig. 2. Box 1 shows pseudo code for this greedy cluster refinement. c_vector, real_gene_exp, geo_gene_exp, num_c respectively are cluster membership vector, age-progeria data and GEO data. Although we use $\tilde{F}^2$ for cluster refinement, but after every k iterations if $F^2$ increases for the selected gene moves, we randomly reverse some of those moves. We chose k = 10 for all analysis.

**Box. 1**

**The greedy maximum matching clusters refinement algorithm**

**Greedy maximal matching cluster refinement algorithm:**

```
Greedy_Cluster_Refinement(c_vector,real_gene_exp,geo_gene_exp,num_c)
    G=Initialize_Graph(c_vector,real_gene_exp,geo_gene_exp,num_c);
    while |unmarked_E| > 0
```

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

**Greedy maximal matching cluster refinement algorithm:**

```
    max(w(e(i, j, gene_id) ∈ unmarked_E));
    marked_E ← e(i, j, gene_id);
    marked_V ← i, j;
  for each e(m, n, gene_id') ∈ unmarked_E
    if m or n ∈ marked_V || gene_id=gene_id'
      delete e(m,n,gene_id');
  for each e(m, n, gene_id)
    c_vector[gene_id] = n;
  return c_vector;
Initialize_Graph(c_vector, real_gen_exp, geo_gene_exp, num_c)
  Graph G = (V, E);
  for each cluster c(i)
    V ← n(i);
  for each gene gene_id
    for t in {1···num_c - 1}
      update w_jc;
      calculate obj_diff;
      if var_diff < 0
        E ← e(c_vector[gene_id], t, gene_id)
        w(e(c_vector[gene_id], t, gene_id) = obj_diff;
  return Graph G;
```

Note that $w_{jc}$ is re-calculated after every cluster update and multiple changes to a cluster can in fact result in overall increase in $F_c^2$. Our matching strategy involving at most one change to each cluster ensures overall reduction of square errors. In addition, by virtue to selecting a maximal matching we maximize the improvement in square errors. The steps of calculating $w_{jc}$, $F_c^2$ and maximal matching cluster refinement are repeated until convergence.

**Adjusted $R^2$—**The quality of regression fit is generally estimated using the $R^2$ statistic as defined above. However, to account for the varying number of clusters and the number of parameters, we instead use adjusted $R^2$ (Adj-$R^2$) for a cluster computed as:

$$R_{adj}^2 = 1 - \left( \frac{(1 - R_c^2)(n-1)}{n-k-1} \right).$$

Where n is the cluster size, and k is the number of coefficients in the multiple linear regressions. $R_c^2$ is defined in equation (3).

## 2.4 GO Analysis

We assessed enrichment of GO biological processes and KEGG pathways in co-expressed gene clusters whose expression co-varied with age and/or HGPS using R's GOstats package.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The significance was corrected for multiple testing using the Benjamini-Hochberg procedure. An FDR threshold of 0.05 was used.

# 3 Results

## 3.1 Method Performance and Efficacy

We first cluster the 9,453 genes into 200 clusters (see M&M) and applied the regression model within each cluster independently. Fig. 3 shows ("Initial FG" plot) the goodness of fit as represented by Adj-$R^2$. We then iteratively refine the clustering with the explicit goal to improve the cluster "tightness" which interestingly has an indirect effect on the Adj-$R^2$. As shown in "Final FG" plot in Fig. 3, the Adj-$R^2$ distribution shifts to higher values. As a control when we randomly permuted the initial expression data and repeat the entire procedure, the final refined clusters show a much inferior distribution as shown in "Final BG" plot in Fig. 3. This supports the efficacy of the refinement step and indicates a substantial pattern in the expression data. The refinement step took 143 iterations until convergence.

## 3.2 Convergence

The maximal matching clustering refinement (see M&M) monotonically decreases the value of $\tilde{F}^2$. Intuitively, the algorithm will converge because if $\tilde{F}^2$ is minimized the cluster will contain co-expressing genes which will have similar regression coefficient vectors. Fig. 4 shows changes in $\tilde{F}^2$ and concomitant changes in total squared residuals due to regression through successive iterations.

## 3.3 Method Robustness

Next we assessed the extent to which the quality of final clustering depends on the initial clustering step. To do so, starting with initial clustering, we perturb the clustering to various extent (defined by parameter $a$) and quantify the quality of final clustering. For instance, we randomly select a fraction of genes and randomly assign to existing clusters. We first noticed that as we increase $a$, it takes longer for the clustering to converge—roughly an eight-fold increase in real run time when using a random clustering compared with co-expression cluster as described in M&M. We compared the Adj-$R^2$ distributions for increasing values of $a$. As shown in Fig. 5, the overall Adj-$R^2$ distribution shows modest reduction in quality (compare plots for $a > 0$ with $a = 0$), which nevertheless is better than initial clustering (compare plots for $a > 0$ with initial clustering). A direct comparison of Adj-$R^2$ between each of the perturbed data and the unperturbed data using Wilcoxon test shows no significant difference, thus supporting robustness of the iterative procedure.

## 3.4 Effect of Cluster Size

We have arbitrarily assumed the number of clusters to be 200, corresponding to an average cluster size of 50 which seems reasonable. However, we tested the effect of number of clusters on the quality of clustering. Fig. 6 shows the distributions of Adj-$R^2$ for different cluster size. The overall quality seems to saturate at around 300 clusters. However, with increasing number of clusters and thus decreasing cluster sizes, the power to detect

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

functional enrichment is compromised. We have therefore performed the follow up functional enrichment analysis with 200 clusters.

### 3.5 Performance

To illustrate the advantage of the proposed RegressionClust model in small sample size, we compared its p-values of regression coefficients with those of single gene regression model. Figs. 7 and 8 show the p-value distributions of regression coefficients of both models for six samples of age-progeria data. We observed coefficients estimated from single gene model are not significant. On the other hand most of coefficients generated from the RegressionClust model are significant.

### 3.6 Identification of Gene Clusters Whose Expression Co-Vary with Age and/or HGPS

We applied our approach to our in house RNA-seq gene expression data for six fibroblast samples—three normal at different cellular ages and three from HGPS at different ages. Using 200 initial co-expression clusters, we iteratively refined the clusters based on tightness of the cluster criterion (see M&M) while estimating cluster specific regression coefficients $\beta_1$, $\beta_2$, and $\beta_3$ for each final cluster along with the p-value for the null hypothesis that the coefficient is zero. We corrected all p-values thus obtained using Benjamini-Hochberg procedure. Next we examined the normalized coefficient values (effect size). Also, we used normalized coefficients as follows: $\beta' = \beta \sqrt{\frac{\sum_j X^2}{\sum_j g_{jc}^2}}$, where X is the input vector of the covariates (age, HGPS status, and interaction) and $g_{jc}$ is gene expression vector for the genes in a specific cluster.

In clusters where only the age and interaction coefficients were significant, the age alone tended to have larger effect on gene expression relative to interaction (Fig. 9). The gene expressions increased with age while the interaction terms in general had negative effect on gene expressions (Fig. 9). Likewise in clusters where only progeria and interaction coefficients were significant, the interaction terms had negative effect compared to progeria (Fig. 10). In addition we also specifically examined the clusters where there is only one significant term, it seems that a greater proportion of gene clusters which are only affected by age are biased toward up-regulation, however the effect on the clusters which significantly affected only by progeria is unbiased (Fig. 11).

### 3.7 Functional Analysis of Specific Gene Clusters Whose Expression Co-Vary with Age and/or HGPS

Considering three coefficients $\beta_1$, $\beta_2$, and $\beta_3$, (normalized values) and their signs, there are eight possibilities for various combinations of these coefficients being significant (FDR ≤ 0.05). For instance, 1+2− represents the clusters for which both $\beta_1$ and $\beta_2$ were significant and $\beta_3$ was not. We further selected the significant coefficients whose absolute value was at least 0.5, to exclude extremely small effect sizes. Table 1 shows the number of clusters with different combinations of significant coefficients with relative large effect sizes.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

We performed functional enrichment analysis [20], [21] using GO Biological processes and KEGG pathways based on FDR threshold of 0.05. For ease of interpretation, we consider only the clusters in six categories: 1+: the expression increases with age (and no other effect), 1−: the expression decreases with age (and no other effect), 2+: the expression increases with HGPS (and no other effect), 2−: the expression decreases with HGPS (and no other effect), 1−2−: the expression decreases both with age and HGPS (no interaction), and finally 3+: the expression increases with age only in HGPS patients (there were no significant clusters in 3− category). To underscore the relative advantage of our specific approach that can potentially distinguish between mechanisms mediated by increase or decrease in gene expression, and also age-related and HGPS-related mechanisms, we directly compare the functions enriched in specific categories.

When we compared functions enriched in 1+ (43 terms enriched) and 1− (13 GO terms enriched) category, only one was common between the two—“growth”. Among 42 Terms unique to 1+ included “extracellular matrix organization” and several referred to “cell death” but others were admittedly harder to interpret. Interestingly, all 12 terms unique to 1− referred to “cell cycle”, “regulation of growth”, “cytoskeleton/spindle organization”, and “chromosome segregation”, which according to our analysis are suppressed with age.

The number of enriched terms in 2+ and 2− clusters were 157 and 215 respectively. To reduce this to a manageable list we only considered GO terms with at most 20 genes annotated, bringing the numbers down to 17 and 19 with 3 in common. The common referred to “Interferon response”. The 14 terms unique to 2+ referred to “cell aging”, “response to unfolded proteins”, “negative regulation of cell adhesion”, and “metabolism”. The 16 terms unique to 2− were clearly different and referred to “development”, “response to oxidative stress”, and “regulation of angiogenesis”. Notably only 2+ clusters were enriched for numerous KEGG pathways related to kinds of cancers.

Interesting we did not detect any gene cluster in “1+2+” category, but we did detect clusters in “1−2−” category and the main theme among the enriched terms in these clusters were “cell signaling” with additional KEGG pathways—“pancreatic cancer” and “gap junction” which again refers to cell signaling.

Finally, category 3+ which refers to clusters whose expression increases with age, particularly in HGPS population has several enriched terms in common with 1+ category, which include “signal transduction”, “wound healing” and several cancer related pathways. The 3+ category includes numerous enriched terms not detected for 1+ category. These include “ossification”, several infectious disease related pathways and “rheumatoid arthiritis”.

### 3.8 Functional Comparison of the Initial Clusters and Refined Clusters

To assess whether our joint regression clustering approach improves functional enrichment of the clusters, we compared the enrichment of functional GO terms in the initial and final clusters for a few selected genes previously known to be involved in aging or Progeria. For some genes, we found their initial cluster and final assigned cluster have different biological functions. As an illustrative example when we compared the clusters containing FOXM1

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

gene, a key transcriptional regulator of cell cycle progression, the final refines cluster containing FOXM1 was specifically enriched for Cell Cycle while the initial cluster containing FOXM1 was enriched in general terms such as transcription activity and DNA binding but not for Cell Cycle. However we acknowledge that it is difficult to quantify the functional enrichment difference between the initial clusters and the final reassigned clusters because the clusters compared may have different size that will influence the enrichment score.

## 4 Discussion

Our works make three main contributions which pertain to data, method, and application to make new biological discoveries. With regards to data we have generated the first RNA-seq data in a controlled fashion for aging HGPS primary cells and passage and genetic background matched normal control cells. Methodologically, here we have presented a regression based approach that leverages clusters of genes with covarying expression to robustly estimate regression coefficients representing dependence on age, HGPS and the interaction between the two. Our approach iteratively refines the clusters using a cluster “tightness” criterion which, as we show analytically, simultaneously improves the goodness of fit while increasing the computational efficiency substantially. The proposed method should be useful in several other contexts with limited number of samples. Finally, application of our method to the data recapitulates previous discovery of age-dependent gene expression changes as well as makes several important observations in a comparison between age and HGPS. In the following, we elaborate on these observations.

Previous microarray studies of HGPS and aged normal fibroblasts have revealed some insights into the gene expression changes during the normal and the premature aging. Ly et al. used fibroblast cells from young, middle and aged normal donors as well as from a HGPS patient, and identified 61 differentially expressed genes out of the 6,000 genes monitored, among which there are two major functional groups: (1) genes involved in cell cycle progression and (2) genes involved in maintenance and remodeling of the extracellular matrix (ECM) [13]. Interestingly, most of the cell cycle genes showed down-regulation in aged cells and HGPS cells, and the ECM genes are affected in both directions in aged and HGPS cells. Using genome-wide affymatrix microarrays, Csoka et al. defined 361 differentially expressed genes in HGPS fibroblast cells (out of 33,000 genes on the array), and found that the two most prominent categories encoded transcriptional factors and ECM proteins [14].

Because of our specific methodology we were able to identify gene clusters whose expressions co-vary exclusively with age, or disease, or in specific combinations of age and disease. We identified several predominant gene clusters, whose expressions were altered either under the disease condition HGPS and/or during the normal cellular senescence. Of a particular note, our analysis indicated that the HGPS gene expression profiles show important differences from the profiles of normal fibroblast passaged into cellular senescence. In the “1−” clusters, we found that the majority of genes are related to cell cycle regulation, which is in highly significant agreement with the results from Ly et al. despite major differences in samples, methods, and data analysis. For example, Forkhead box

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

protein M1 (FOXM1), a key transcriptional regulator of cell cycle progression, was found to be down regulated in both studies. This gene has been shown to regulate a large group of G2-M specific genes [19], including a key mitotic cyclin, cyclin B1, which was also identified by both our analysis and Ly et al. In addition, consistent with previous reports on ECM proteins, we found that the "1+" clusters are enriched with functional categories of programmed cell death regulation and ECM organization.

However, the responses in HGPS cells differ: In the "2+" clusters that positively associate with HGPS disease condition, the prominent categories are related to metabolic functions, implying an activation (or at least an attempted activation) of the biological processes involved on various cellular metabolic activities in HGPS cells. To date, the metabolisms in HGPS cells have not been systematically examined, nor any metabolite profilings in HGPS cells have been conducted. Our study provides the first genome-wide evidence of the affected metabolisms in HGPS cells, and points to a potentially important direction for future HGPS research. Interestingly in the "2−" clusters whose expression is reduced in HGPS, we found gene clusters including protein transcription and translation and protein biosynthesis, reflecting an overall slow down in cellular growth in the prematurely aged HGPS cells. Fig. 12 shows the cluster of terms significantly enriched among these gene clusters computed by NIH DAVID tool [20], [21].

Because lamin A/progerin resides in the inner nuclear rim, and plays a role in organizing chromatin, it is not surprising to identify wide spreading changes in gene expression in HGPS cells. The challenge is to determine the specificity of progerin-related changes and of the age-related changes, and illuminate their potential interplays. In an attempt, we examined the functional groups in the genes whose expression increases with age specifically in HGPS cells (the "3+" clusters). Interestingly, a prominent functional gene group is related to signal transduction, including transmembrane receptors (e.g. insulin-like family peptide receptor 1 and stannin), protein kinases (e.g. membrane associated guanylate kinase and protein kinase C), and transcriptional regulators (e.g. ADP-ribosylarginine hydrolase and microphthalmia-associated transcription factor). Additional studies, especially those conducted in cell types other than fibroblasts, are required before we can understand the contributions of progerin/lamin A and cellular aging to gene expression in complex organisms. The study reported here provides a first genome-wide, multi-stage RNA-seq experiment with a novel iterative multiple regression approach to examine this important mechanism.

## 5 Conclusion

We have performed RNA-seq profiling in fibroblast cell cultures at three different cellular ages, both from HGPS patients and matched normal samples. We then performed gene expression analysis by using Cufflinks suite. To address the issue of small sample size we developed a novel joint regression clustering approach that leverages co-expressed gene clusters to identify gene clusters whose expression changes significantly with age and/or disease state. Finally we performed functional analysis on resulting clusters. Based on our approach applied to novel RNA-seq data in HGPS and aging the results recapitulate the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

previously known processes underlying aging while at the same time suggests numerous unique processes underlying aging and HGPS.

## Acknowledgments

Kan Cao and Sridhar Hannenhalli conceived the project. Expression data was generated by Kan Cao and Zheng-Mei Xiong, Kun Wang and A.S. developed the method under the supervision of Sridhar Hannenhalli, Kun Wang, A.S., Kan Cao and Sridhar Hannenhalli wrote the manuscript. The authors would like to thank Justin Malin for his comments and helpful discussions. This work is funded by NIH R01GM100335 to Sridhar Hannenhalli and Ellison Medical Foundation New Scholar Award to Kan Cao.

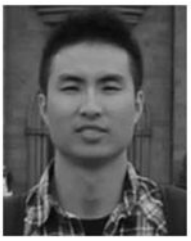
## References

1. Merideth MA, et al. Phenotype and course of Hutchinson-Gilford progeria syndrome. New Engl J Med. 2008; 358(6):592–604. [PubMed: 18256394]
2. Eriksson M, et al. Recurrent de novo point mutations in lamin A cause Hutchinson-Gilford progeria syndrome. Nature. 2003; 423(6937):293–298. [PubMed: 12714972]
3. Zhang H, Kieckhaefer JE, Cao K. Mouse models of laminopathies. Aging Cell. 2013; 12(1):2–10. [PubMed: 23095062]
4. Gordon LB, Cao K, Collins FS. Progeria: Translational insights from cell biology. J Cell Biol. 2012; 199(1):9–13. [PubMed: 23027899]
5. Mounkes L, et al. The laminopathies: Nuclear structure meets disease. Current Opinion Genetics Develop. 2003; 13(3):223–230.
6. Shumaker DK, et al. Mutant nuclear lamin A leads to progressive alterations of epigenetic control in premature aging. Proc Natl Acad Sci USA. 2006; 103(23):8703–8708. [PubMed: 16738054]
7. Goldman RD, et al. Accumulation of mutant lamin A causes progressive changes in nuclear architecture in Hutchinson-Gilford progeria syndrome. Proc Natl Acad Sci USA. 2004; 101(24): 8963–8968. [PubMed: 15184648]
8. McCord RP, et al. Correlated alterations in genome organization, histone methylation, and DNA-lamin A/C interactions in Hutchinson-Gilford progeria syndrome. Genome Res. 2013; 23(2):260–269. [PubMed: 23152449]
9. Somech R, et al. The nuclear-envelope protein and transcriptional repressor LAP2beta interacts with HDAC3 at the nuclear periphery, and induces histone H4 deacetylation. J Cell Sci. 2005; 118(17): 4017–4025. [PubMed: 16129885]
10. Gotzmann J, Foisner R. A-type lamin complexes and regenerative potential: A step towards understanding laminopathic diseases? Histochem Cell Biol. 2006; 125(1–2):33–41. [PubMed: 16142451]
11. Wilson KL, Foisner R. Lamin-binding proteins. Cold Spring Harb Perspect Biol. 2010; 2(4):a000554. [PubMed: 20452940]
12. Park WY, et al. Gene profile of replicative senescence is different from progeria or elderly donor. Biochem Biophys Res Commun. 2001; 282(4):934–939. [PubMed: 11352641]
13. Ly DH, et al. Mitotic misregulation and human aging. Science. 2000; 287(5462):2486–2492. [PubMed: 10741968]
14. Csoka AB, et al. Genome-scale expression profiling of Hutchinson-Gilford progeria syndrome reveals widespread transcriptional misregulation leading to mesodermal/mesenchymal defects and accelerated atherosclerosis. Aging Cell. 2004; 3(4):235–243. [PubMed: 15268757]
15. Marji J, et al. Defective lamin A-Rb signaling in Hutchinson-Gilford progeria syndrome and reversal by farnesyltransferase inhibition. PLoS One. 2010; 5(6):e11132. [PubMed: 20559568]
16. Stranger BE, et al. Population genomics of human gene expression. Nature Genetics. 2007; 39(10): 1217–1224. [PubMed: 17873874]
17. Trapnell C, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols. 2012; 7(3):562–578. [PubMed: 22383036]
18. Barrett T, et al. NCBI GEO: Archive for functional genomics data sets–10 years on. Nuclear Acids Res. 2011; 39:D1005–1010.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

19. Myatt SS, Lam EW. The emerging roles of forkhead box (Fox) proteins in cancer. Nature Rev Cancer. 2007; 7(11):847–859. [PubMed: 17943136]
20. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocol. 2009; 4(1):44–57.
21. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009; 37(1):1–13. [PubMed: 19033363]
22. Dellaert F. The expectation maximization algorithm. Georgia Inst Technol, Tech Rep GIT-GVU-02-20. 2002

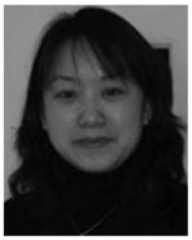
## Biographies



**Kun Wang** received the BSc degree in computer science from Shandong University in China in 2008 and the MSc degree in computer science from Arkansas State University in 2011. Currently, he is working towards the PhD degree in computational biology and bioinformatics program in the University of Maryland at College Park. His research interests include comparative genomics, gene regulation, splicing regulation, and normal aging.

**Avinash Das** received the B-Tech degree from the National Institute of Technology, Trichy, 2007 in electronics and communication. He received the MSc degree in computer science and statistics from Ecole Polytechnique Fdrale de Lausanne in 2010. He finished the MSc scholar thesis from the Genome Science Department, University of Washington, Seattle, in 2011. He is a doctoral student in the Computer Science Department at the University of Maryland College Park and National Human Genome Research Institute. His primary research interest include Bayesian modelling of developmental and disease association studies in human.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



**Zheng-Mei Xiong** received the BSc degree in clinical medicine from Guiyang Medical University, China, in 1994, and the PhD degree in pharmacology from Kansai Medical University, Japan, in 2009. She joined Dr. Kan Cao's lab as a postdoctoral fellow in 2010 and focused on study of molecular mechanisms underlying Hutchinson Gilford progeria syndrome (HGPS) by utilizing iPS cells as a system. She recently is interested in studying mitochondrial defects in HGPS and in developing novel medicines in HGPS therapeutics.

**Kan Cao** received the BSc degree in biology from the Nanjing University China in 1997, and the PhD degree in biology from the Johns Hopkins University in 2005. She did her postdoctoral fellowship in genomics at the US National Institutes of Health from 2005 to 2010. She is an assistant professor of cell biology and molecular genetics at the University of Maryland College Park. She studies molecular and cellular mechanisms underlying Hutchinson Gilford progeria syndrome, a rare premature aging disease, and the normal human aging process. She was named the New Scholar in Aging by the Ellison Medical Foundation in 2011, and received Board of Visitors junior faculty award from the University of Maryland in 2013.

**Sridhar Hannenhalli** received the PhD degree in computer science from The Pennsylvania State University in 1996, where he developed combinatorial algorithms for genome rearrangement problems. He is an associate professor in the Department of Cell Biology and Molecular Genetics at the University of Maryland and also in the Center for Bioinformatics and Computational Biology. He has previously held research positions at Glaxo Smithkline, Celera Genomics and as assistant and then associate professor at the University of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Pennsylvania. His current research focus is on comparative genomics, transcriptional regulation and evolution.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

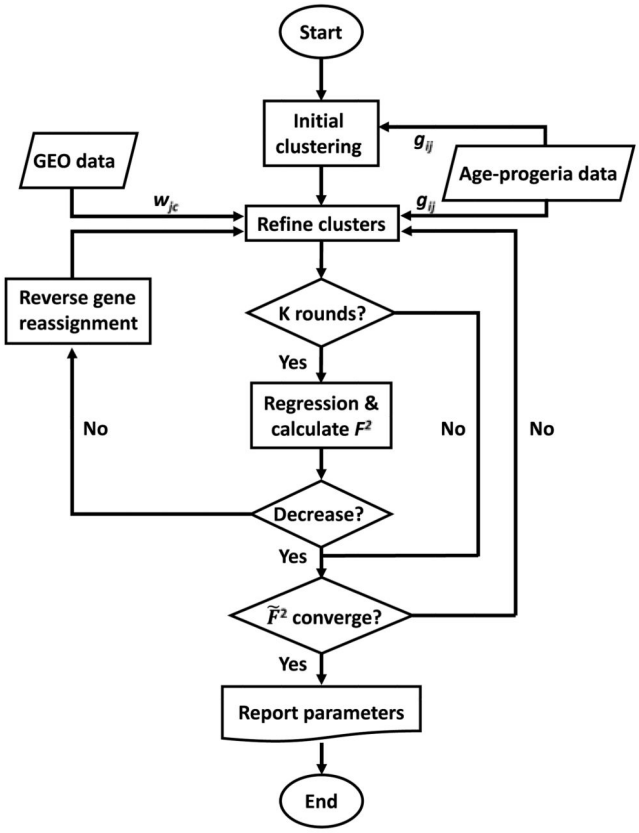



**Fig. 1.**
Method workflow. See text for details.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

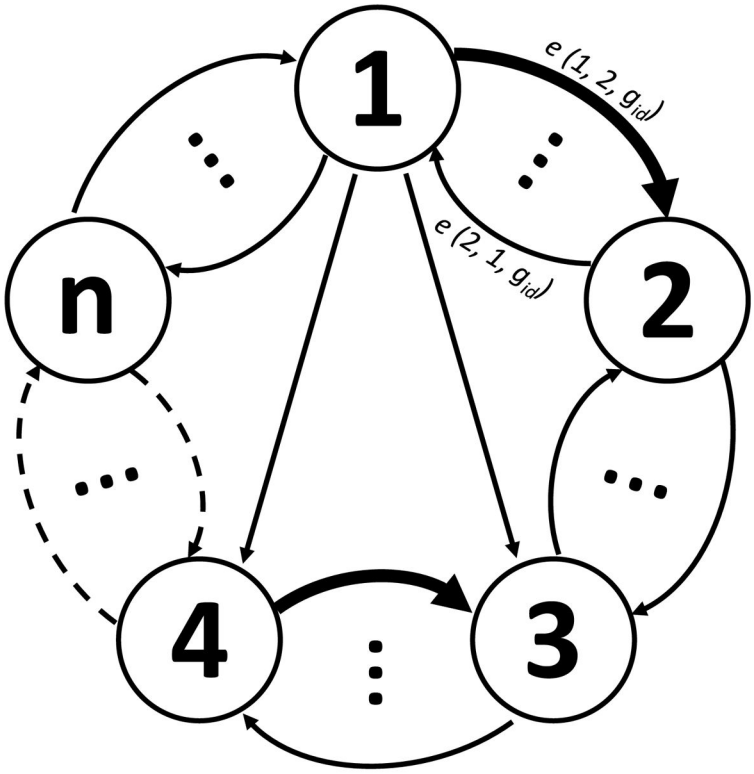



**Fig. 2.**

The Graph constructed for greedy maximum matching cluster refinement. There are n clusters, we construct a node for each cluster, and the edge is added if the possible reassignment will decrease the objective function score. The Edge e(1, 2, $g_{id}$) is the gene reassignment that moves gene $gene_{id}$ from cluster 1 to cluster 2. The bold edges are greedily selected edges, based on which we perform the final genes reassignment.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

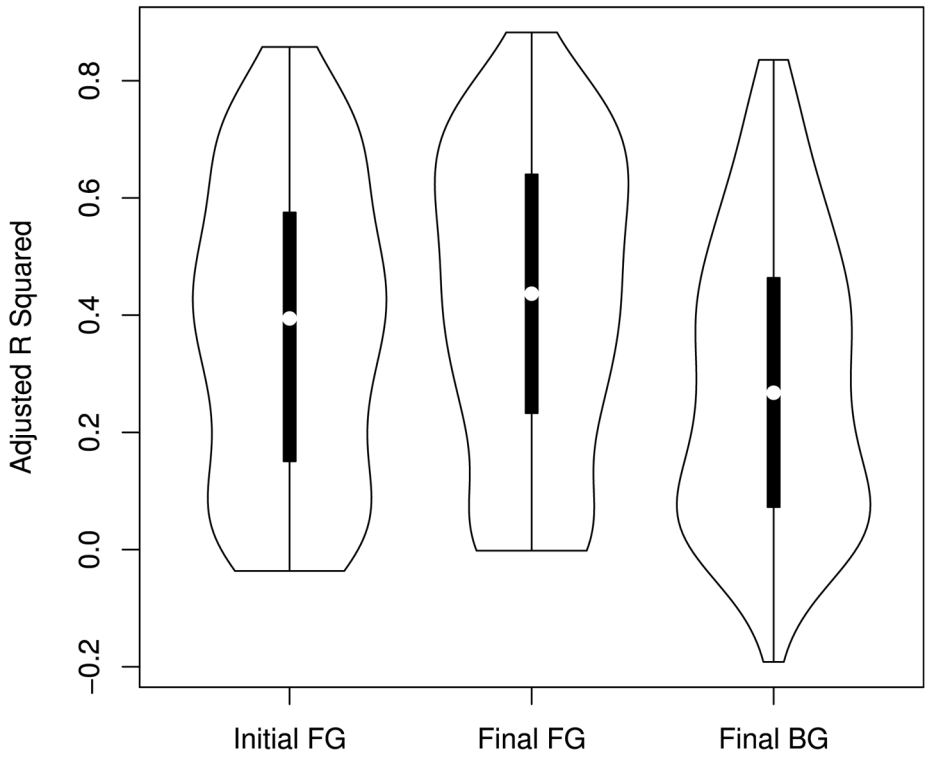



**Fig. 3.**

Goodness of fit (Adj-$R^2$) distributions for gene clusters. Initial FG plot: initial clustering, Final FG plot: Final refined clustering, Final BG plot: Refined clustering for randomly permuted gene expression data.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

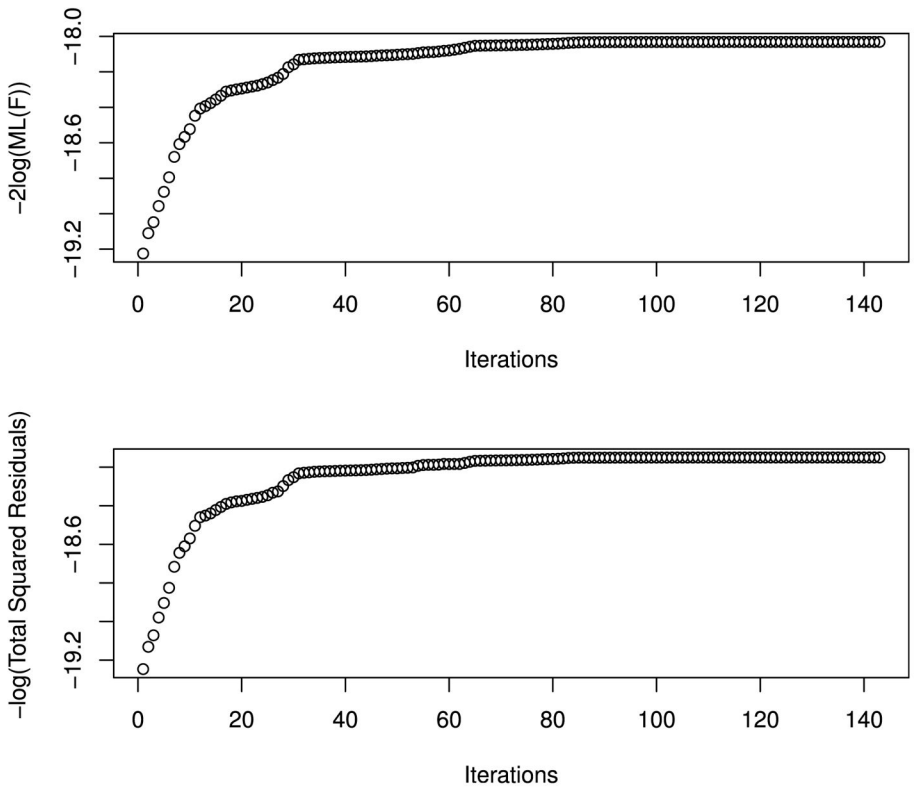



**Fig. 4.**
Convergence of $\tilde{F}^2$ and total squared residuals of linear regression.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

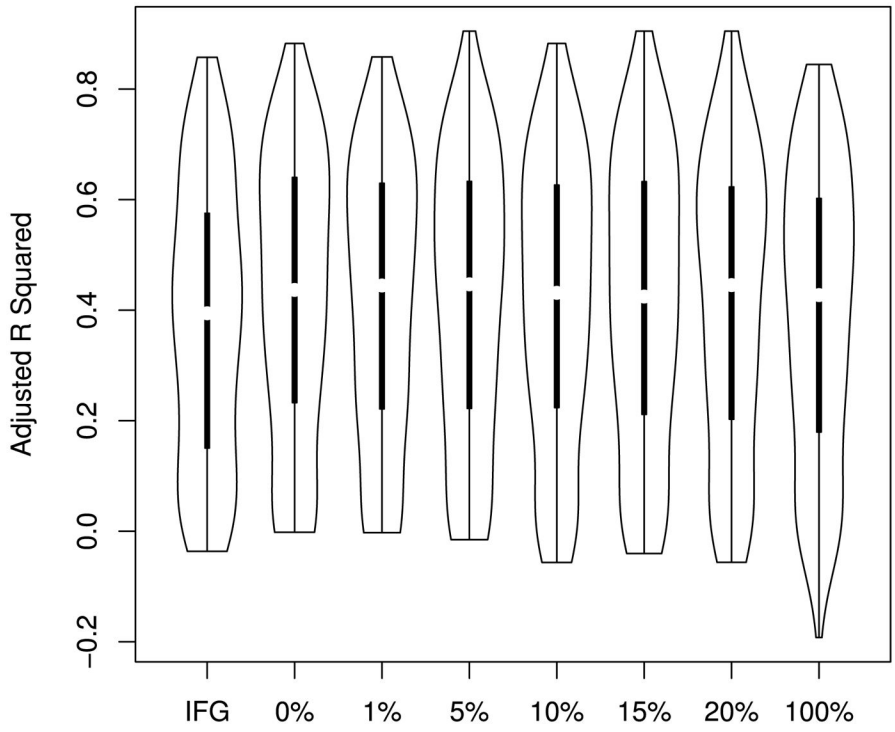



**Fig. 5.**

Robustness of the iterative refinement. For varying degree ($a$) of perturbation of the initial clustering the plots show the Adj-$R^2$ distribution of the refined clustering. IFG represents initial clustering.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

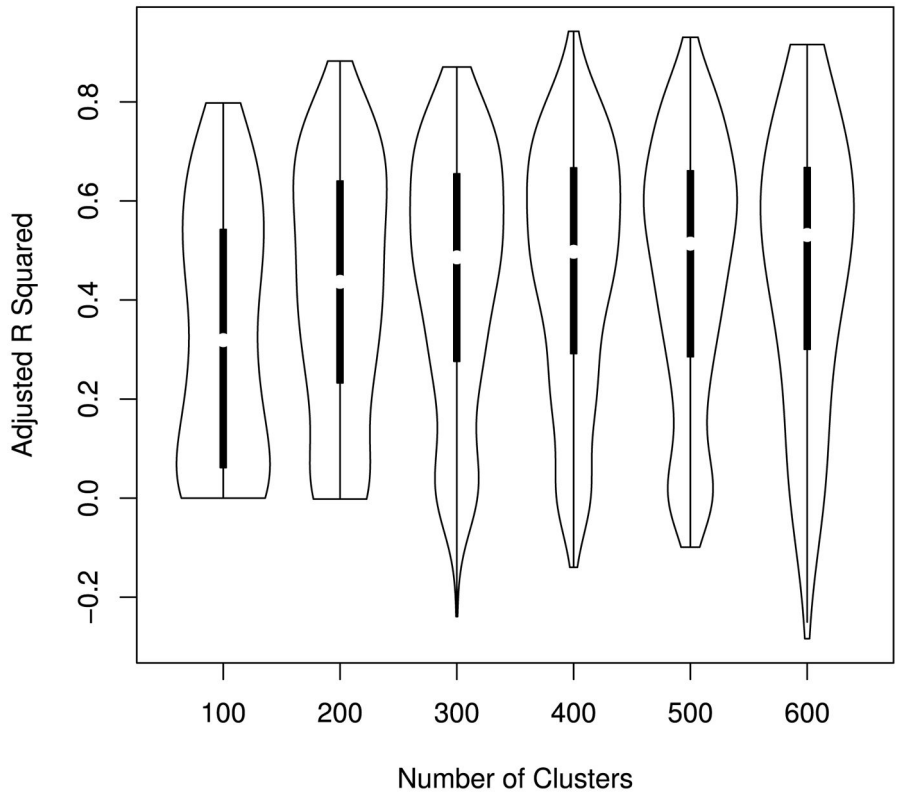



**Fig. 6.**

Effect of cluster size on cluster quality. For varying number of clusters the plots show the Adj-$R^2$ distribution of the refined clustering.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

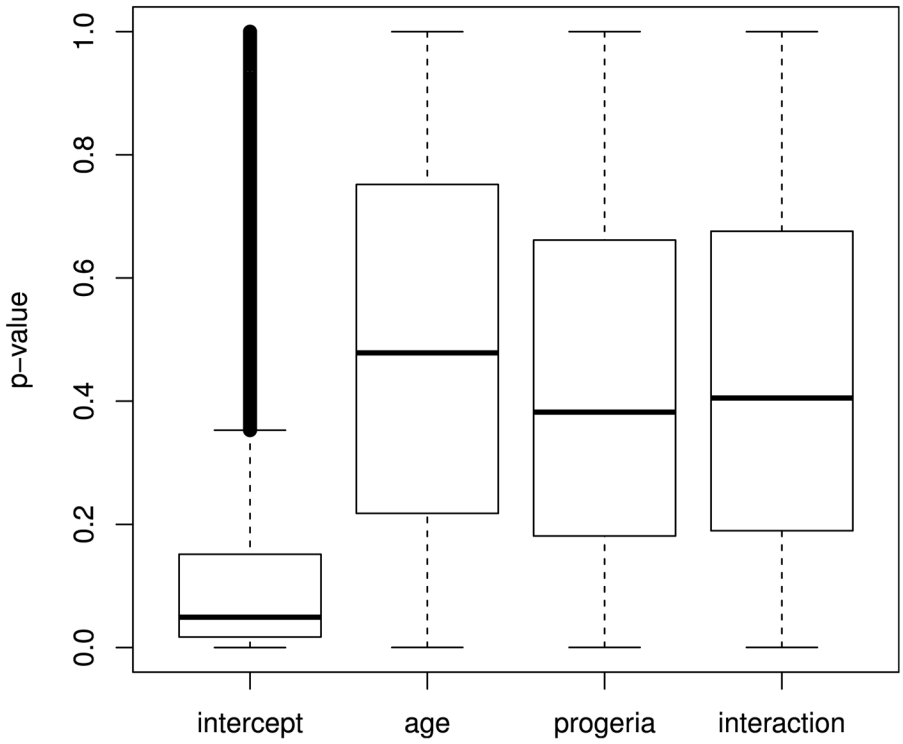



**Fig. 7.**
Distribution of p-values of the four parameters in the single gene fitting model.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

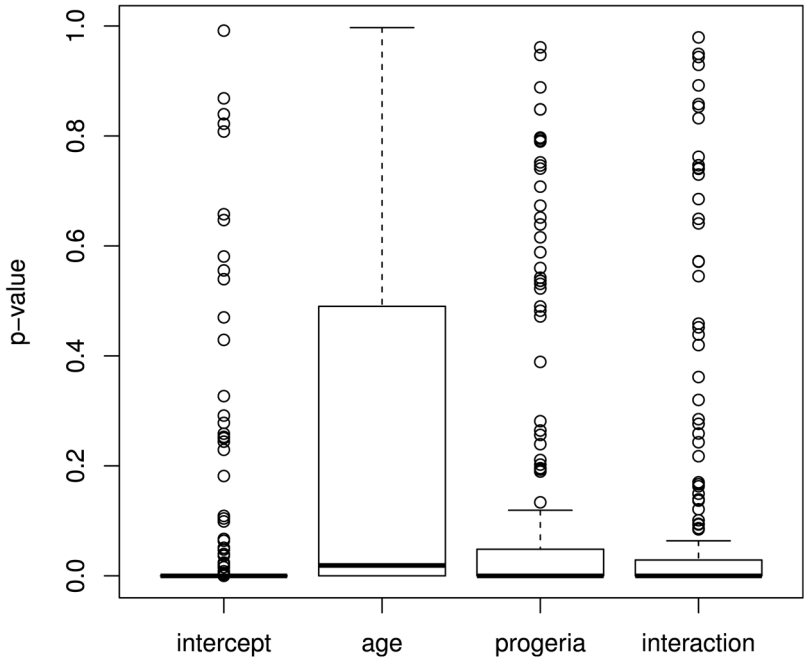



**Fig. 8.**
Distribution of p-values of the four parameters in the gene cluster fitting model.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

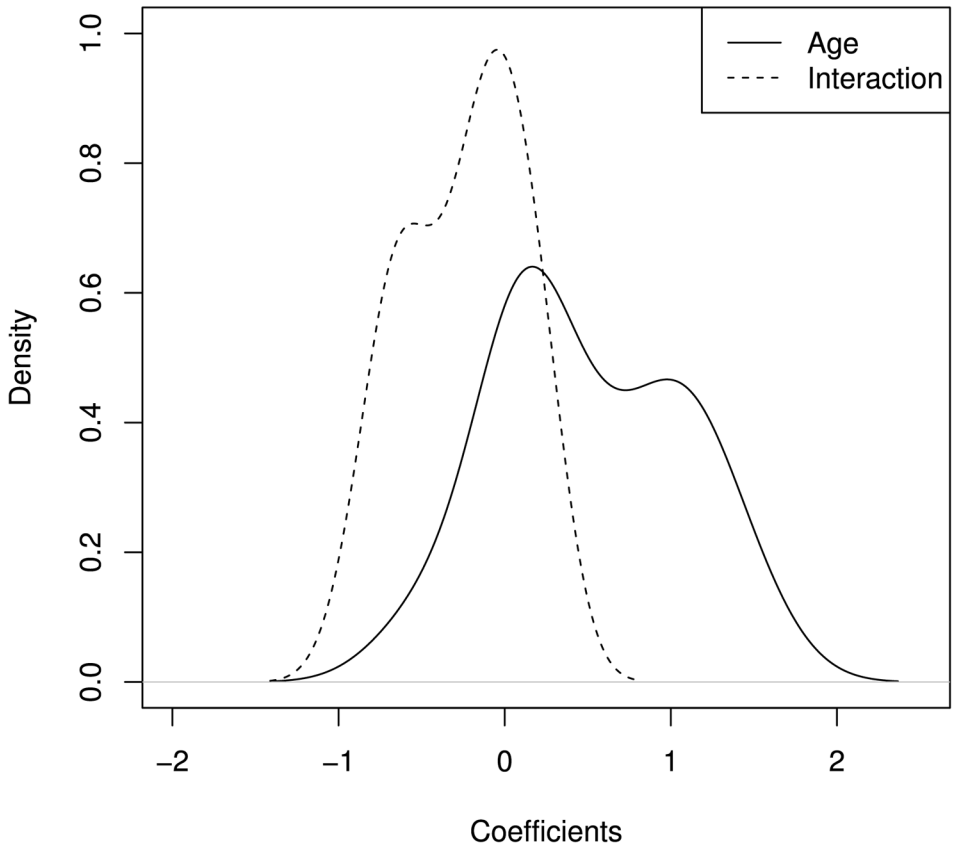



**Fig. 9.**
Distribution of normalized coefficients $\beta_1$ and $\beta_3$ specifically for the clusters for which only these two coefficients were significant.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

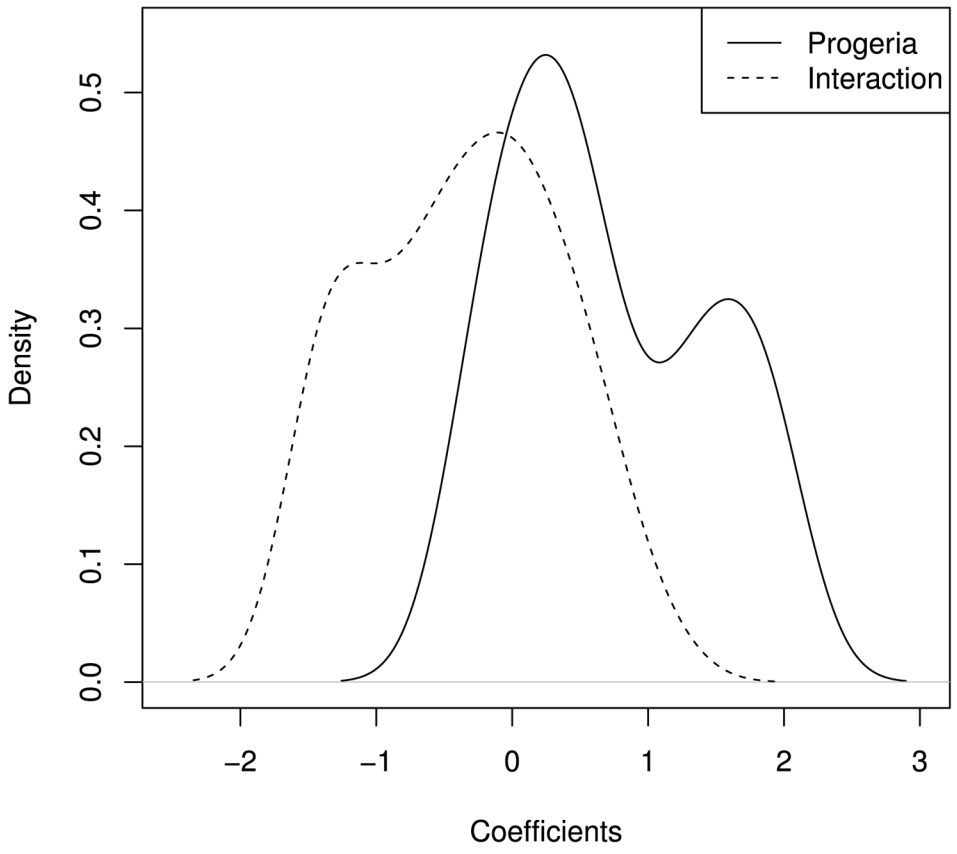



**Fig. 10.**
Distribution of normalized coefficients $\beta_2$ and $\beta_3$ specifically for the clusters for which only these two coefficients were significant.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

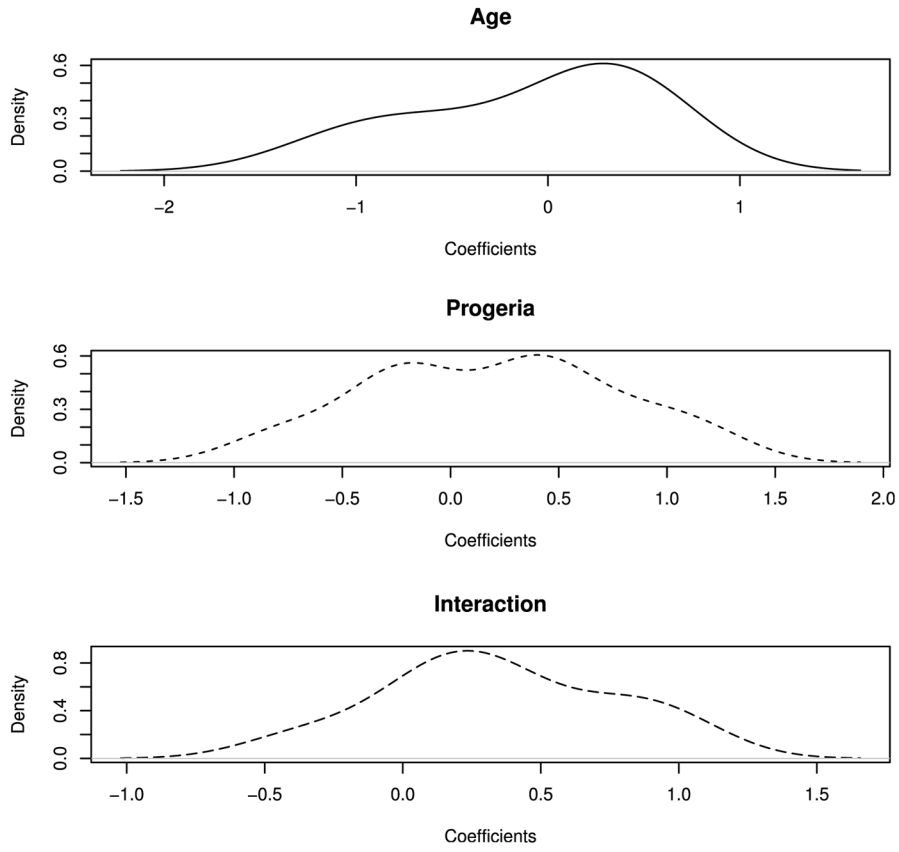



**Fig. 11.**

Distribution of normalized coefficients $\beta_1$, $\beta_2$, and $\beta_3$ specifically for the clusters for which only one of the coefficients was significant.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

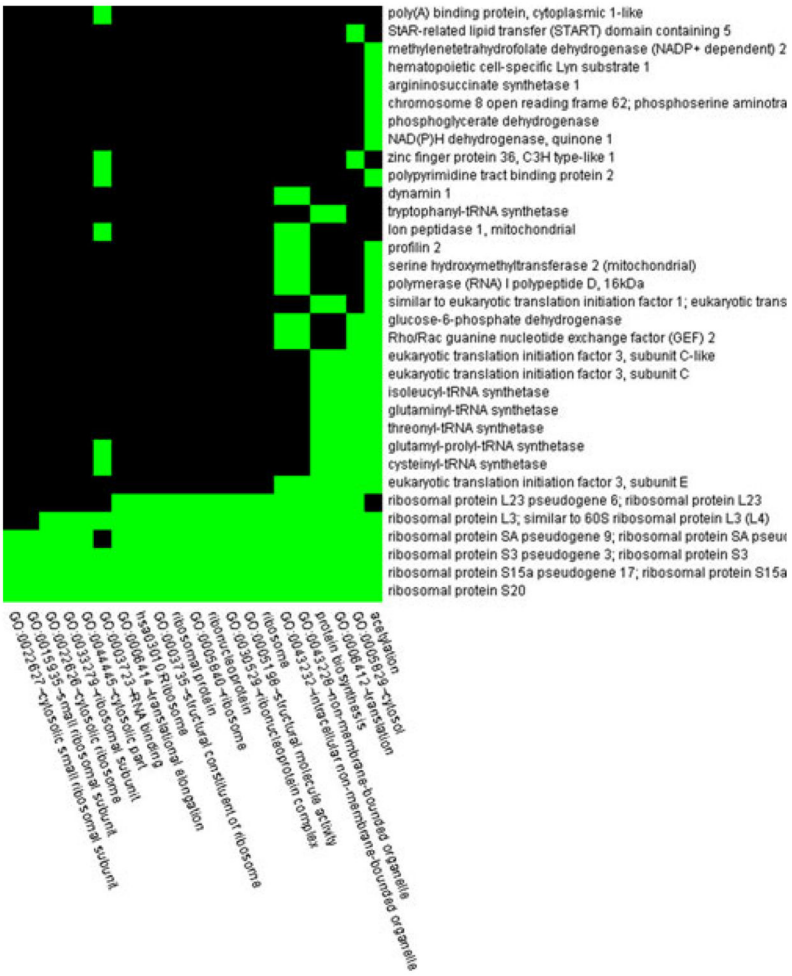



**Fig. 12.**

Cluster of enriched terms relating to translations are enriched among gene clusters in two-category.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**TABLE 1**

Number of Clusters in Various Categories Based on Which Combination of Coefficients Were Significant and with Large Effect Size (Absolute Value ≥ 0.5)

| Category | Number of Clusters |
|---|---|
| 1+ | 3 |
| 1− | 3 |
| 2+ | 6 |
| 2− | 6 |
| 3+ | 11 |
| 1−2− | 2 |
| 1+3− | 9 |
| 2+3− | 22 |
| 2−3+ | 1 |
| 1+2+3− | 4 |

The numbers in the first column represent significant coefficients where 1: $\beta_1$, 2: $\beta_2$, 3: $\beta_3$, and the sign represents direction of influence. For instance category "1−2−" represents the clusters for which $\beta_1$ and $\beta_2$ were significant and ≤ −0.5 and $\beta_3$ was not significant.