Transcriptomics and Machine Learning Predict Diagnosis and Severity of Growth Hormone Deficiency

DOI:
10.1172/jci.insight.93247

Document Version
Accepted author manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):
Murray, P., Stevens, F., De Leonibus, C., Koledova, E., Chatelain, P., & Clayton, P. (2018). Transcriptomics and Machine Learning Predict Diagnosis and Severity of Growth Hormone Deficiency. Journal of Clinical Investigation Insight, 3(7), [e93247]. https://doi.org/10.1172/jci.insight.93247

Published in:
Journal of Clinical Investigation Insight

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester’s Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.
Transcriptomics and Machine Learning Predict Diagnosis and Severity of Growth Hormone Deficiency

Philip G Murray1#, Adam Stevens1#, Chiara De Leonibus1, Ekaterina Koledova2, Pierre Chatelain3, and Peter E Clayton1

1Institute of Human Development, Faculty of Medical and Human Sciences, University of Manchester and Manchester Academic Health Science Centre, Royal Manchester Children’s Hospital, Central Manchester University Hospitals NHS Foundation Trust, Manchester, UK; 2Global Medical, Safety & CMO, Merck KGaA, Darmstadt, Germany; 3Department Pediatrie, Hôpital Mère-Enfant – Université Claude Bernard, 69677 Lyon, France.

# These authors contributed equally to the work.

Address all correspondence to:

Peter Clayton, MRCP FRCPCH, Royal Manchester Children’s Hospital, 5th Floor Research, Oxford Road Manchester M13 9WL, UK
Tel: +44 161 701 6949; E-mail: peter.clayton@manchester.ac.uk

Running title: Genomic signature of Childhood GHD

Abstract Word Count: 295    Manuscript Word count: 4,712

Figures: 4 (plus two supplemental figures)

Tables: 3 (plus one supplemental table)

Key words: Gene expression, GH treatment, growth hormone (GH) deficiency.

Classifications: Genetics/genomics, growth hormone

Trial registration number: NCT00256126 & NCT00699855
Abstract

Background: The impact of gene expression data on diagnosis remains limited. Here we show how diagnosis and classification of Growth Hormone Deficiency (GHD) can be achieved from a single blood sample using a combination of transcriptomics and Random Forest analysis.

Methods: Pre-pubertal treatment naïve children with GHD (n=98) were enrolled from the PREDICT study and controls (n=26) acquired from online datasets. Whole blood gene expression (GE) was correlated with peak growth hormone (GH) using rank regression and a Random Forest algorithm tested for prediction of the presence of GHD and in classification of GHD into severe (peak GH<4 µg/L) and non-severe (peak>4 µg/L). Performance was assessed using Area under the Receiver Operating Characteristic Curve (AUC-ROC).

Results: Rank regression identified 347 probesets where gene expression correlated with peak GH concentrations: \((r = \pm 0.28, p<0.01)\). These 347 probesets gave an AUC of 0.95 for predicting GHD status versus controls and an AUC of 0.93 for prediction of GHD severity.

Conclusion: This study demonstrates highly accurate diagnosis and disease classification for GHD using a combination of transcriptomics and Random Forest analysis.

Trial registration: NCT00256126 & NCT00699855

Funding: The PREDICT study was funded by Merck KGaA. Salary funding to PGM was from the NIHR.
Introduction

High throughput technologies including next generation sequencing, microarrays, mass-spectrometry and protein chips now allow measurement of many thousands of biological variables at relatively low cost. While next generation based DNA sequencing panels are now impacting upon clinical practice (1, 2) for diagnosis of genetic disease, gene expression data are far less frequently utilised in routine clinical practice. A large number of studies have examined the utility of gene expression data in prognosis and classification of tumors (3, 4) but in other fields where affected tissue is of more limited availability, the clinical impact of transcriptomics has been minimal. One of the main challenges in utilising these technologies is identifying the useful biological signal in such complex data. Combining transcriptomics with machine learning approaches has proven useful in disease classification in autism (5). Here we show how the utility of transcriptomic data for diagnosis can be refined using a combination of machine learning and network based prioritisation.

Many diagnostic tests in endocrinology require administration of pharmacological agents, multiple blood sampling and hospital admission making them expensive and unpleasant for patients. This study, using Growth Hormone Deficiency (GHD) as an exemplar, demonstrates how a single blood test with extracted mRNA applied to a microarray could replace endocrine stimulation tests.

GHD is a rare but important cause of short stature in childhood with a prevalence of ~1:4,000 (6). Consensus guidelines recommend an approach to the diagnosis of childhood GHD integrating clinical, biochemical and auxological data (7). Biochemical investigations are key to the diagnosis, particularly pharmacological growth hormone stimulation testing where a cut off level is used below which children are diagnosed with GHD. There are, however, many problems associated with these stimulation tests – they display poor reproducibility (8) and in addition the peak growth hormone (GH) level achieved varies with body composition (9, 10), pharmacological stimuli (11) and assay (12) used. The first cut off level proposed for the diagnosis of GHD was 5 µg/L (13) in 1968 on the basis
that this seemed to best identify children with a GHD phenotype. With the increased availability of GH this cut off was subsequently increased to 7 µg/L and then 10 µg/L based on very limited evidence. Dependent upon the assay used, recent studies classifying children as GHD or not-GHD, based on auxological criteria, have suggested cut-off levels between 4 and 7 µg/L (12). Clearly there remains uncertainty as to the optimal cut-off level for the diagnosis or GHD. **Given the multiple problems associated with pharmacologic stimulation tests, there is also no clear cut-off for the differentiation between “mild” and “severe” GHD. A peak GH cut-off of 10 µg /L was used to define GHD in the PREDICT study, and, in the present analysis, we have chosen a cut-off of 4 µg /L to define “severe” GHD.**

There is therefore a need to develop new tools to aid with the diagnosis and classification of childhood GHD which are not susceptible to the many problems associated with pharmacological stimulation tests. The PREDICT study (14) was a 1-month, phase IV, open-label, prospective multicentre study in GH-treatment-naïve children with GHD (NCT00256126) that aimed to identify genetic and transcriptomic markers of response to GH therapy. Children enrolled in the study all had a peak GH level on two stimulation tests of <10 µg/L and blood samples taken for whole genome gene expression analysis and candidate single nucleotide polymorphism (SNP) genotyping prior to starting treatment.

Using this cohort along with gene expression data from healthy control children, this exploratory study aimed to:

I. Define the set of genes whose expression correlates with peak GH levels.

II. Determine the usefulness of these gene expression data in the diagnosis and classification of GHD.

III. Identify the biological function and regulators of these genes.
IV. Identify SNPs associated with peak GH levels and examine the utility of these SNPs, either alone or in combination with the gene expression data and/or demographic/biochemical data, for classification of GHD subjects.

Results

SNPs associated with peak GH

18 SNPs in 12 genes were associated with peak GH concentrations (see Table 1). 16 of the 18 SNPs are intronic with one synonymous exonic SNP and one missense exonic SNP. None of the SNPs were rare (defined as a minor allele frequency <1%) with 16/18 SNPs having a minor allele frequency >10%. The function of the genes associated with the SNPs included pituitary developmental transcription factor (POU1F1), generation of oestrogen (CYP19A1), IGF binding (IGFBP1), apoptosis (BCL2 and SHC1), cell cycle (CCND3 and CDK2), angiogenesis (CYR61) growth factors (TGFA), transcription factor (SREBF1) and signal transduction (PTPN1, RARA).

Principal Component Analysis (PCA) and Gene Expression (GE) profiling in GHD and Control Children

No differences were observed in the overall distribution of gene expression between GHD and control subjects using unsupervised PCA on the transcriptomic data from the different studies described (Supplemental Figure S1). This determined that there was no overall effect of study or associated co-variates on the data sets and therefore further direct comparison was meaningful. Rank regression identified 347 probesets (representing 271 unique genes) where expression correlated with peak GH concentrations in the 98 GHD children (188 probesets positively and 159 negatively, R = 0.28, p<0.01) and were also present in control arrays (see Supplementary Table S1). Of these 347 probesets 65 were identified as also being expressed in the human growth plate. The gene expression of the 347 probesets is displayed on a heat map for both children with GHD and normal children (assigned to a peak GH of 10 µg/L) in Figure 1A. A clear distinction was seen between the normal subjects and the subjects with GHD and, in addition, a point of inflexion is seen in the GHD subjects at a peak GH of 4.75 µg/L (Figure 1A).
These 347 probesets were then displayed on a heat map with two-way hierarchical clustering using both the control and GHD subjects (Figure 1B). On the x-axis dendrogram four clusters of GHD subjects can be seen, including 45 subjects (20 GHD, 25 controls) where it was not possible to classify them into clusters (Figure 1B). Only one control subject was classified into one of the four clusters with the remaining being GHD subjects. There were no significant differences in age, gender, height standard deviation score (SDS), weight SDS or body mass index (BMI) SDS between the 20 GHD subjects not classified into the four GHD clusters and the 78 GHD subjects classified into one of those four clusters.

A further heat map of the 347 probesets identified by rank regression using peak GH as a continuous variable in the GHD group only was generated (Figure 2A). In this heat map, 5 clusters of gene expression – 2 related to genes where there is a positive correlation with peak GH and 3 related to genes where there is a negative correlation with peak GH – were identified. In this heat map, all subjects could be classified via a Euclidian metric into one of the 5 clusters.

Supervised principal component and iso-map multidimensional scaling identified three distinct groups of GHD patients (Figure 2B). There were significant differences between groups for peak GH levels, distance to target height SDS, baseline Insulin-like Growth Factor 1 (IGF1) SDS and baseline Insulin-like Growth Factor Binding Protein 3 (IGFBP3) SDS over 1 month of GH treatment (see Table 2).

Network Analysis
Using the genes identified by the rank regression a network with 2427 nodes and 3604 links was generated. Decomposition into a hierarchical modular structure revealed 43 network modules.
Functionality was assessed on the top 15 modules as ranked by network centrality (Figure 3). Of the
15 modules, 5 were related to circadian clock, 4 related to growth factor signalling and 3 related to DNA replication.

The gene expression clusters (Figure 2) were overlapped with the network modules (See Figure 3 and Supplemental Figure S2). Overlapping simply involves comparing the list of genes present in the gene expression clusters and network modules (those with > three shared genes were considered to be linked). Gene cluster 1 linked to only one network module (HSD17B14) related to cell cycle while gene cluster 5 was also linked to only one network module – CASP2 related to apoptosis pathways. Gene cluster 2 associated with the 2nd, 3rd, 4th and 10th network modules related to Circadian Clock, chromatin organisation and growth factor signalling. Gene clusters 3 and 4 each linked to 4 network modules covering the whole spectrum of pathways identified except apoptosis. SSX2IP, STRN3 and PTGDS contained within the first and second (SSX2IP), third (STRN3) and fifth (PTGDS) clusters as determined by centrality hierarchy (Figure S2) had previously been identified in the top 10 genes with variable importance in the Random Forest model.

Causal network analysis (15) identified four causal elements within the network model which mapped to the 15 network modules (see Figure 4). This provides robust supporting evidence for the functional pathways described by the network modules. Master regulators identified included APC2 regulating the STRN3 network module related to apoptosis and gene cluster 4. SOX2, PI3KR3 and SIRT2 were identified as regulators of the ARHGAP1, TRIM54 and SUFU network modules linked to gene clusters 3 and 4 affecting Hedgehog signalling, Circadian Clock, Mitochondrial biogenesis and myogenesis pathways.

**Prediction of GHD Status and Classification of GH severity**

Random Forest analysis for predicting GHD vs Control subjects gave an AUC (Area under the Receiver Operating Characteristic Curve) of 0.95 (95%CI 0.91-0.99) with sensitivity of 96%, specificity of 100%
and an OOB-ROC AUC (Out of Box Area under the Receiver Operating Characteristic Curve) of 0.99. 53 probesets representing 40 named and 13 unnamed genes were confirmed by BORUTA as containing predictive capacity (see Supplementary Table S2). Of the 53 probesets identified by BORUTA 10 of these were also expressed in the growth plate. This represents an enrichment of growth plate genes in comparison to those selected just by the rank regression model (hypogeometric p=1.14x10^{-12}). Although the predictive capacity given by the transcriptomic data was excellent, we also assessed the ability of network biology to improve prediction by selecting probesets ranked by network centrality. Selecting the top 10 probesets ranked by network centrality gave an AUC of 0.94 (95% CI 0.91 – 0.95) while 4 different combinations of 10 probesets (randomly selected from the probesets where expression was correlated with peak GH) gave an inferior AUC of 0.84 (95% CI 0.78 – 0.90).

Demographic, biochemical, genomic and transcriptomic data were used with Random Forest analysis to assess their predictive value in determining severe GHD (defined by peak GH <4μg/L). Each of these data categories were assessed separately, and then, in combination (see Table 3). The transcriptomic data (AUC 0.93) performed better than the genomic (AUC 0.85) or biochemical/demographic data (AUC 0.88). The addition of the genomic or biochemical/demographic data (either alone or in combination) to the transcriptomic data did not improve the AUC (all 0.93). In the model using all data categories, of the top 10 variables of greatest importance (as ranked by mean decrease in accuracy) nine were gene expression probesets (NRXN1, SSX2IP, STRN3, RNF43, SUZ12P, RAB7A, PROC, GATSL3 and PTGDS) and one was a SNP (rs2715553). The functions of the encoded proteins were diverse but included several which were clearly linked to growth: STRN3 is a WD40 domain containing protein which enhances cancer cell survival and activating AKT (16), RAB7A an oncogene involved in RAS pathway, RNF43 a tumour suppressor involved in ubiquitination and SSX2IP is known to bind to a synovial sarcoma associated protein which promotes growth (17). Other genes and their encoded proteins did not have a clear role in GH
secretion or growth – PROC is a coagulation factor, SUZ12P a pseudogene, GATSL3 is associated with rheumatoid arthritis (18) but has no known function, NRXN1 is a neuroligin synapse receptor (19) and PTGDS is involved in prostaglandin production
Discussion

This exploratory study aimed to identify whether gene expression profiling could aid with the diagnosis of GHD or in our classification and understanding of the factors influencing the severity of GHD. Despite the use of GH as a therapeutic agent since 1958 (20) and the ability to measure serum GH levels since 1963 (21) the diagnosis of GHD remains challenging and there is no “gold standard” test for diagnosis. In this study we examined whether gene expression profiling could distinguish children from the PREDICT study with GHD from healthy controls. The development of a test based on gene expression would be a significant advance for patients potentially avoiding the need for hospital admission and the use of pharmacological stimulation tests. The AUC of the random forest based algorithm from our gene expression data was excellent at 0.98 and, with a specificity of 100% and a sensitivity of 96%, clearly distinguishing GHD subjects from controls thus suggesting this could be developed into a useful test for diagnosing of GHD.

In addition to assessing the predictive ability of Random Forest analysis, we also assessed whether network prioritisation of input genes improved prediction. Limiting a prediction algorithm to a small number of parameters may be helpful in developing a cost-effective test which can easily be applied to large numbers of patients. The predictive ability of 10 probesets was increased by selecting them based on their network centrality. This combination of network analysis with a machine learning approach (in this case Random Forest analysis) may be particularly effective in developing ‘omic–based’ approaches to diagnosis. Selecting probesets based on the BORUTA algorithm, which is designed to identify the probesets most likely to contain true predictive capacity, resulted in an enrichment of selected probesets for those expressed in the growth plate.

These data support the use potential use of a gene expression based test but there are significant limitations to our study. Firstly the patient and control children were assembled from different studies although extensive work was undertaken to normalise both between and within batches of
arrays. Secondly, rather than comparing children with GHD to normal healthy children, it would be better to compare GHD children to short children without GHD as this is the distinction that is required of a clinical test. In addition, as a multi-centre international study with the GHD diagnosis made at each study centre, GH stimulation test and assay were not standardised in the PREDICT study. PREDICT aimed to recruit a cohort of children with GHD diagnosed according to international practice, encompassing the variations in diagnostic criteria between centres and countries. In general, there will be reasonable correlations between peak GH levels after different stimuli, such that a low peak GH after arginine will also be low after an insulin tolerance test or a glucagon test, and higher levels in a test will be mirrored in a 2nd test. Within KIGS, which collected ‘real world’ data on GHD patients, the correlation between the first and second GH stimulation tests in >3000 patients was reasonable at $r=+0.515$ (22). A third limitation is that we did not have a peak GH level for the control subjects, due to the inaccuracy of GH stimulation tests we cannot be sure if they had been tested that they would indeed have had a peak GH level $>10 \mu g/L$, although as healthy control it is highly unlikely that any of them had GHD. Future studies should concentrate on prospective recruitment of children undergoing GH stimulation testing using standardised stimulation tests and growth hormone assays to determine in that cohort whether there is any evidence of a change in pattern of gene expression at any particular cut off level for peak GH. A small number of subjects in this study received additional hormone supplementation with hydrocortisone or thyroxine, given that this was physiological replacement we do not expect this to have significantly affected gene expression.

This study has also demonstrated the utility of gene expression profiling and SNP genotyping in identifying a cohort of children with more severe GHD. A cut-off of 4 $\mu g/L$ was selected as this allowed us to create two groups (one with more severe GHD) with sufficient numbers for prediction. The most accurate tool for identifying GH severity status were the transcriptomic data, performing better than the genomic data, clinical data or genomic and clinical data combined. This is highly
suggestive that it is possible to accurately identify on the basis of a basal gene expression a child
with severe GHD from among a cohort of subjects with short stature and a range of peak GH
concentrations classified as GHD.

Unsupervised principal component analysis did not identify clinically distinct groups of GHD patients
and we therefore undertook a supervised principal component analysis using those genes where
expression correlated with severity of GHD as defined by peak GH concentration to stimulation
testing. This supervised analysis identified three groups of GHD patients. There was a clear clinical
separation between group 1 and groups 2 and 3, with group 1 representing a less severe cohort of
patients with a higher peak GH level, higher IGF1 SDS and lower distance to target height SDS (Table
2). There was clear separation of groups 2 and 3 on the principal component analysis but no clear
auxological/biochemical differences between these groups.

In addition to identifying a GE profile associated with peak GH concentrations we also identified 18
SNPs from 12 different genes where genotype was associated with peak GH concentrations. Five of
these twelve genes (SHC1, CCND3, BCL2, CDK2 and RARA) were also present in the network. Of
those five genes, two each are involved with apoptosis and cell cycle. For many patients with GH
deficiency, anterior pituitary hypoplasia is also present and these SNPs may mediate their effects by
affecting somatotrope differentiation and survival. One of the SNPs was within POU1F1, a pituitary
transcription factor essential for differentiation of somatotropes, lactotropes and thyrotropes (23).
Five SNPs were identified within CYP19A1, the gene that encodes the enzyme aromatase responsible
for the generation of oestrogen. Sex steroids augment GH peak concentrations during stimulation
tests and mediate pituitary growth during puberty and oestrogen inhibits GH signal transduction by
stimulating expression of SOCS-2 (Suppressor of Cytokine Signalling 2) (24). Although all children
enrolled in PREDICT were pre-pubertal, it is possible that very low pre-pubertal oestrogen
concentrations can be influenced by these SNPs and hence impact on GH levels.
To explore the function of the genes whose expression was linked to peak GH levels, we generated a network model and ranked functional modules of genes according to the network centrality. Gene clusters 2, 3 and 4 all mapped to network modules involved in growth factor signalling, including WNT and hedgehog signalling, while gene cluster 5 mapped to a module involved in apoptosis. It is perhaps not surprising that a strong signature for growth factor signalling and apoptosis would be identified in the genes related to severity of growth hormone deficiency. Both growth hormone and its downstream effector hormone IGF1 are known inhibitors of apoptosis (25). Clearly, with increasing severity of GHD, we would expect reduced growth factor signalling and increased apoptosis. This study, however, defines the distinct gene expression clusters which differentially link to growth factor signalling and apoptosis. It was interesting to find a strong signature for the circadian clock. GH is secreted in pulses mainly overnight (26) and this finding may reflect either disturbances of the circadian clock leading to reduced secretion of GH or perhaps an acceleration of the circadian clock rhythm in an attempt to maximise GH pulse frequency where pulse amplitude has been limited by somatotroph hypoplasia. GH secretion has not only circadian but also ultradian rhythms (27) and disturbance of these can lead to a form of GHD termed neurosecretory dysfunction. This is a disorder where the child presents with GH deficiency with a normal pharmacological GH stimulation test but abnormal spontaneous GH secretion with reduced frequency and amplitude of GH pulses (28).

Causal network analysis allowed us to identify four master regulators – APC2, SOX2, PIK3R3 and SIRT2. Loss of function mutations in Anaphase Promoting Complex 2 (APC2) have been associated with a Sotos syndrome like phenotype of overgrowth and neurodevelopmental delay (29) and it is a negative regulator of WNT signalling by targeting β–catenin for ubiquitin mediated proteolysis (30). WNT signalling is known to be involved in pituitary development, promoting the expression of PITX2 and proliferation of pituitary precursors (31). SOX2 (Sex Determining Region Y Box 2) is a member of
the SRY-related HMG box B1 (SOXB1) subfamily of transcription factors and is expressed in the developing brain and posterior neural tube including Rathke’s pouch and hypothalamus (32). In humans, heterozygous loss of function mutations in SOX2 lead to eye abnormalities (microphthalmia and anophthalmia) and hypopituitarism (hypogonadotropic hypogonadism and variable GHD). SOX2 expression in the postnatal and adult pituitary marks a subpopulation of hormone negative cells, which are pituitary progenitor stem cells capable of differentiating into endocrine producing cells (31). SOX2 may therefore be regulating developmental processes such as pituitary stem cell proliferation in addition to myogenesis. PIK3R3 encodes a regulatory subunit of phosphoinositide-3-kinase, a component of both the GH and IGF1 signal transduction systems as well as many other cell signal transduction cascades. PI3K is involved in a diverse range of functions including proliferation, cell survival, degranulation, vesicular trafficking and cell migration. SIRT2 is one of a class of NAD(+) dependent deacetylases with anti-ageing activity in model organisms, an effect increased by caloric restriction (33). The sirtuins induce mitochondrial biogenesis (the generation of new mitochondria) to reduce the accumulation of toxic reactive oxygen species seen in caloric restriction (33). In addition, SIRT2 regulated adipocyte differentiation inhibits p53 accumulation and is regulated by Src tyrosine kinase (a component of the GH signal transduction system) (34).

This study has demonstrated the potential for gene expression profiling to aid in both the diagnosis and classification of GHD and, in addition, has identified the functions of the networks of genes related to peak GH concentrations along with their master regulators. Moving from a diagnosis requiring the use of pharmacological stimulation tests to a single blood sample would be a major advance, particularly for paediatric patients. Potentially this work could be extended from GHD to other hormone deficiencies allowing the full assessment of pituitary function with a single blood test.
Methods

Patients

The PREDICT study was a phase IV, open label, prospective pharmacogenomic study examining response to GH therapy and enrolled 125 prepubertal children (78 male, 47 female) aged 2 to 15 with a diagnosis of GH deficiency, reached following two pharmacological stimulation tests according to local protocols, with a peak GH concentration of < 10µg/L. Details of the inclusion and exclusion criteria have previously been reported (14, 35). In brief, prior to enrolment in the study, none of the children had received GH therapy; children with GHD due to central nervous system tumours or radiotherapy were excluded but children born small for gestational age were not. Of the 125 children in PREDICT baseline gene expression data were available on 98 subjects aged 2 – 15 years (34 female, 64 male). Bone age was available for 92 patients. Mean bone age delay was 2.2 ± 1.5 years. A delay in bone age of > 1 year was present in 72 patients. Birth weight was available for 82 patients of whom 12 (15%) were born small for gestational age. In addition to growth hormone therapy 5 patients received both thyroxine and hydrocortisone replacement, 3 patients thyroxine alone and 1 patient hydrocortisone alone.

This PREDICT study was conducted in compliance with ethical principles based on the Declaration of Helsinki, the International Conference on Harmonization Tripartite Guideline for Good Clinical Practice, and all applicable regulatory requirements.

Serum Samples and Assays

Blood samples were drawn in the morning, after an overnight fast, prior to and 1 month after start of treatment with recombinant human growth hormone. Samples were centrally assayed at qLAB (Livingston, Edinburgh, UK). Serum IGF1 and IGFBP-3 were measured using DPC chemiluminescent immunoassays (Immunolite 2000®, Siemens Healthcare Diagnostics Inc., Norwood, MA, USA). Plasma insulin was measured with two-site immunoenzymometric assay (AIA-PACK IRI; Tosoh,
Tokyo, Japan). Plasma glucose was determined by using the glucose oxidase method and HDL cholesterol was measured with an enzymatic-calorimetric test.

**Genotyping**

Genotyping of 1536 SNPs, located on 103 candidate genes (related to 1. the GH–IGF1 axis, 2. bone and cell growth and 3. glucose and lipid metabolism), was performed as previously described (14) using the Illumina GoldenGate assays (Illumina, San Diego, CA, USA). Before analysis, genotyping data were filtered to remove SNPs with a call rate <95%, and those showing significant deviation from the Hardy–Weinberg equilibrium using a Bonferroni correction. After data cleaning, 1171 SNPs in 97 genes remained for analysis.

**Continuous analysis**

SNPs associated with peak GH were identified using the Kruskal–Wallis rank sum test on the following models: genotype, presence of the major allele (dominant model) and presence of the minor allele (recessive model). For non-pseudoautosomal X-linked markers, boys were analysed separately from girls. As a candidate gene rather than a whole genome approach was being used, P-values were adjusted for multiple testing using a Bonferroni correction, taking into account the number of linkage disequilibrium blocks present in the gene containing the SNP of interest and considered significant where adjusted P-value <0.05.

**Categorical analysis**

Markers were also tested in a categorical analysis, with patients classified by quartiles of peak GH; comparisons were made between those with low peak GH concentration versus those with intermediate + high peak GH, and those with high peak GH concentration versus those with intermediate + low peak GH concentrations. All P-values were calculated using Fisher’s exact test.
and adjusted for multiple testing using a Bonferroni correction, taking into account the number of linkage disequilibrium blocks within each candidate gene.

Transcriptome Analysis for subjects in PREDICT study and Gene Expression Datasets from normal Childhood Control Subjects

To assess transcriptomic relationships GE profiling was carried out on whole blood RNA as previously described (14) and hybridised to Affymetrix GeneChip Human Genome U133 Plus 2.0 Arrays. This gene expression data has previously been uploaded to the NCBI Gene Expression Omnibus database - GSE72439 (36). For control subjects, gene expression analysis was conducted on a library of gene expression datasets from normal children with age annotation collated from the publically available NCBI Gene expression Omnibus (GEO) and EBI ArrayExpress databases. The original Affymetrix CEL files from GSE9006 (37), GSE26440 (38) and TABM666 (39) were downloaded and combined into one group to form a main analysis dataset following published guidelines (40). Details of the generation of this combined normative dataset have previously been published (41). As the children in PREDICT were all above the age of two, GE profiles from normal children were removed if they were aged <2 years old, which left 26 subjects (14 male, 12 female) aged 2 to 11 years. Mean age was significantly lower for control patients at 5.8 ± 2.7 years compared to PREDICT subjects at 8.8 ± 2.9 years (p<0.01) but there was no significant difference in male gender (14/26 - 54% in control subjects and 61/98, 62% of PREDICT subjects, p=0.58).

Normalisation and Quality Control of Gene Expression Data

For background correction the Robust Multichip Average (RMA) (42) was applied to the combined CEL files (derived from GEO or PREDICT). The dataset generated was subject to quality control to investigate the presence of outliers and further confounding effects. Dimensional scaling using PCA and Isomap multidimensional scaling (MDS) (43, 44) was used to demonstrate data homogeneity.
Gene expression data was normalised for batch, age and gender.

Unsupervised & Supervised PCA

Unsupervised PCA was performed on the GE profiles to identify whether the variance in the datasets (GHD and controls) was consistent, a requirement for further statistical analysis between these groups. Supervised PCA was performed after initial statistical evaluation (rank regression – see below) to determine the presence of patient sub-groups. All PCA was performed using Qlucore Omics Explorer (Qlucore, Lund, Sweden). Quality control of all PCA was undertaken using cross-validation (sequential removal of all samples) to determine group stability. Unsupervised PCA was refined using variance filtering and a projection score (45). PCA was also confirmed using Isomap multidimensional scaling.

Analysis of Network Models

Network analysis allows the identification and prioritisation of key functional elements within interactome models which this study has used to prioritise genes for prediction and also to gain insights into biological function. To derive an interactome model, genes whose expression correlated with peak GH concentrations were used as “seeds” and all known protein:protein interactions between the seeds and their inferred immediate neighbours were identified to generate a biological network using the output of the BioGRID model of the human Interactome (3.3.122) (46). Network generation and processing was performed using Cytoscape 2.8.3 (47).

Clustering and “community structure” of modules within biological networks arise from variation in connectivity within the network and are known to be associated with function (48-50). To prioritise these functional components within interactome models, we used the ModuLand plugin for Cytoscape 2.8.3 to determine overlapping modules and to identify hierarchical structure within the
model, thus enabling the identification of key network elements (S1). Network modules were prioritised for further investigation by their centrality property and the most central set of ten genes within each module was used to assess associated biological pathways using the Reactome database (S2). The network structure observed with community modelling in Moduland was confirmed by cluster analysis using the ClusterOne algorithm (S3).

Causal Network Analysis (CNA)

CNA allows the identification and prioritisation of regulatory system elements within transcriptomic models. CNA was performed within Ingenuity Pathways Analysis (IPA, Redwood City, CA, USA) using the genes whose expression was correlated with peak GH concentrations.

CNA identifies upstream molecules, up to three steps distant, that control the expression of the genes in the dataset (15). A prediction of the activation state for each regulatory factor (master regulator), based on the direction of change, was calculated (z-score) using the gene expression patterns of the transcription factor and its downstream genes. An absolute z-score of ≥ 1.4 and a corrected P-value <0.05 (Fisher’s Exact Test) was used to compare the regulators identified in each of the comparisons made using hierarchical clustering (Euclidean metric).

Regression and Random Forest Analysis

Rank Regression of probesets for association with peak GH levels was performed in Qlucore Omics Explorer (Qlucore, Lund, Sweden). Differences in demographic characteristics were assessed using SPSS version 20 (IBM, Armonk, NY) via a Kruskall Wallis Test.

A Random Forest algorithm (S4) was used to predict severity of GHD (<4µg/L) based on demographic, biochemical, genomic and transcriptomic data. A cut-off of 4 µg/L was chosen as this divided the subjects into two groups of approximately equal size, maximising accuracy of the
Random Forest classification. Biochemical and demographic data included age, gender and IGF1/BP3 before start of GH therapy. Genomic data comprised the SNPs identified as being associated with peak GH. For prediction of GHD vs Control, the data were unbalanced (98 GHD subjects and 26 controls) and with an unbalanced dataset Random Forest poorly predicts the minority class (in this case control subjects). To overcome this problem a synthetic minority over-sampling technique (55) was used to rebalance the dataset prior to Random Forest prediction using age, gender and transcriptomic data (final data used 114 subjects 57 controls and 57 patients). The predictions were assessed based on the AUC-ROC and the out-of-box ROC curve (OOB-ROC) as a validation set. Identifying those probesets most likely to contain predictive capacity was achieved with the use of BORUTA (56). All statistical analyses were performed using R 2.15.3. Random Forest analysis requires no explicit test and validation set as the OOB-ROC functions as a validation data set. In developing the random forest algorithm hundreds or thousands of decision trees (in our case 1000 trees) are created. Each tree is generated using a random selection of input variables and randomly selected 2/3 of the subjects. Each tree produced a classification vote and the majority vote across all trees determines final classification. For each tree there is a random 1/3 of subjects whose data was not used in generating that tree – this data is then used to generate the OOB-ROC which essentially functions as a validation data set.

Human Growth Plate Gene Expression and Overlap with Probesets Identified via Random Forest Analysis

Human gene expression data from growth plate derived RNA was available for two subjects (1 male, 1 female) from the NCBI GEO database (GSE9160). For each subject a sample of the distal femoral growth plate had been obtained and populations of cells from reserve, proliferative, prehypertrophic, and hypertrophic zones obtained by laser microdissection. RNA from each population of cells, corresponding to each zone of the growth plate was amplified and hybridized to Affymetrix HU-133 2.0 arrays. Frozen robust multi-array analysis (fRMA) (57) was used to define
absolute expression by comparison to publically available microarray datasets within R and an
eexpression barcode (58) was defined for each growth plate zone for each patient. Expression within
the growth plate was defined as by a gene expression barcode value of one in any zone of the
growth plate in either patient.

Statistics Overview

Study Approval: The PREDICT (NCT00256126) and PREDICT long term follow up (NCT00699855)
studies were approved by the Scotland Medical Research and Ethics Committee (reference
05/MRE10/61) and North West Research Ethics Committee (reference 08/H1010/77) respectively.
Informed consent was obtained from parents for all study participants.
Author Contributions

PEC, EK and PC conceived and designed the PREDICT project and this study. Data analysis and method development was undertaken by PM, AS and CdL. The manuscript was written by PM and AS and revised by EK, PC and PEC.

Conflict of Interest

AS, PM, CdL PC and PEC have received honoraria from Merck. EK is an employee of Merck.

Acknowledgements

The authors would like to thank all the PREDICT Investigators and the patients who took part in the study. PM thanks the National Institute of Health Research for their support (CL-2012-06-005).

PREDICT investigator group: Argentina: A Belgorosky, Buenos Aires. Australia: G Ambler, Westmead. Austria: K Kapelari, Innsbruck. Canada: C Deal, Montreal; J Hamilton, Toronto. Finland: J Jääskeläinen, Kuopio. France: Y Brusquet, Puyricard; P Chatelain, Lyon; M Colle, Bordeaux; R Coutant, Angers; Y Le Bouc, Paris; R Reynaud, Marseille; J-P Salles, Toulouse; J Weill, Lille. Germany: R Pfaffle, Leipzig; M Ranke, Tübingen; G Binder, Tübingen. Italy: M Bozzola, Pavia; F Buzi, Brescia; M Cappa, Rome; A Cicognani, Bologna; M Maghnie, Genova; L Tato, Verona; F Antoniazzi, Verona. Korea: DH Kim, Seoul; SW Yang, Seoul; HW Yoo, Seoul. Norway: E Vangsøy Hansen, Bergen; D Veimo, Bodø. Russia: E Bashnina, St Petersburg; V Peterkova, Moscow; J Skorodok, St Petersburg; L Sultanova, Kazan. Spain: A Carrascosa, Barcelona; A Ferrandez Longas, Zaragoza; R Gracia Bouthellier, Madrid; JP Lopez Siguero, Malaga; S Quinteiro, Las Palmas de Gran Canaria; MD Rodriguez-Arnao, Madrid; A Rodriguez Sanchez, Madrid. Sweden: J Dahlgren, Göteborg; L Hagenäs, Stockholm. Taiwan: JW Hou, Taoyuan; TJ Wang, Kaohsiung County. UK: P Clayton, Manchester; C Kelnar, Edinburgh.
References

1. Larsen CP, Durfee T, Wilson JD, Beggs ML. A Custom Targeted Next-Generation Sequencing Gene Panel for the Diagnosis of Genetic Nephropathies. American journal of kidney diseases : the official journal of the National Kidney Foundation. 2016.

2. Weisschuh N, Mayer AK, Strom TM, Kohl S, Glockle N, Schubach M, et al. Mutation Detection in Patients with Retinal Dystrophies Using Targeted Next Generation Sequencing. PloS one. 2016;11(1):e0145951.

3. Pinto R, De Summa S, Petriella D, Tudoran O, Danza K, Tommasi S. The value of new high-throughput technologies for diagnosis and prognosis in solid tumors. Cancer biomarkers : section A of Disease markers. 2014;14(2-3):103-17.

4. Wang K, Huang C, Nice EC. Proteomics, genomics and transcriptomics: their emerging roles in the discovery and validation of colorectal cancer biomarkers. Expert review of proteomics. 2014;11(2):179-205.

5. Latkowski T, Osowski S. Computerized system for recognition of autism on the basis of gene expression microarray data. Computers in biology and medicine. 2015;56:82-8.

6. Lindsay R, Feldkamp M, Harris D, Robertson J, Rallison M. Utah Growth Study: growth standards and the prevalence of growth hormone deficiency. The Journal of pediatrics. 1994;125(1):29-35.

7. Growth Hormone Research S. Consensus guidelines for the diagnosis and treatment of growth hormone (GH) deficiency in childhood and adolescence: summary statement of the GH Research Society. GH Research Society. The Journal of clinical endocrinology and metabolism. 2000;85(11):3990-3.

8. Hilczer M, Smyczynska J, Stawerska R, Lewinski A. Stability of IGF-I concentration despite divergent results of repeated GH stimulating tests indicates poor reproducibility of test results. Endocrine regulations. 2006;40(2):37-45.

9. Martha PM, Jr., Gorman KM, Blizzard RM, Rogol AD, Veldhuis JD. Endogenous growth hormone secretion and clearance rates in normal boys, as determined by deconvolution analysis: relationship to age, pubertal status, and body mass. The Journal of clinical endocrinology and metabolism. 1992;74(2):336-44.

10. Stanley TL, Levitsky LL, Grinspoon SK, Misra M. Effect of body mass index on peak growth hormone response to provocative testing in children with short stature. The Journal of clinical endocrinology and metabolism. 2009;94(12):4875-81.

11. Ghigo E, Bellone J, Aimaretti G, Bellone S, Loche S, Cappa M, et al. Reliability of provocative tests to assess growth hormone secretory status. Study in 472 normally growing children. The Journal of clinical endocrinology and metabolism. 1996;81(9):3323-7.

12. Wagner IV, Paetzold C, Gausche R, Vogel M, Koerner A, Thierry J, et al. Clinical evidence-based cutoff limits for GH stimulation tests in children with a backup of results with reference to mass spectrometry. European journal of endocrinology. 2014;171(3):389-97.

13. Kaplan SL, Abrams CA, Bell JJ, Conte FA, Grumbach MM. Growth and growth hormone. I. Changes in serum level of growth hormone following hypoglycemia in 134 children with growth retardation. Pediatric research. 1968;2(1):43-63.

14. Stevens A, Clayton P, Tato L, Yoo HW, Rodriguez-Arnao MD, Skorodok J, et al. Pharmacogenomics of insulin-like growth factor-I generation during GH treatment in children with GH deficiency or Turner syndrome. Pharmacogenomics J. 2014;14(1):54-62.

15. Kramer A, Green J, Pollard J, Jr., Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics (Oxford, England). 2014;30(4):523-30.

16. Tanti GK, Pandey S, Goswami SK. SG2NA enhances cancer cell survival by stabilizing DJ-1 and thus activating Akt. Biochemical and biophysical research communications. 2015;463(4):524-31.

17. de Bruijn DR, dos Santos NR, Kater-Baats E, Thijszen J, van den Berk L, Stap J, et al. The cancer-related protein SSX2 interacts with the human homologue of a Ras-like GTPase interactor, RAB3IP, and a novel nuclear protein, SSX2IP. Genes, chromosomes & cancer. 2002;34(3):285-98.
18. Orozco G, Viatte S, Bowes J, Martin P, Wilson AG, Morgan AW, et al. Novel rheumatoid arthritis susceptibility locus at 22q12 identified in an extended UK genome-wide association study. Arthritis & rheumatology. 2014;66(1):24-30.
19. Arac D, Boucard AA, Ozkan E, Strop P, Newell E, Sudhof TC, et al. Structures of neuroligin-1 and the neuroligin-1/neurexin-1 beta complex reveal specific protein-protein and protein-Ca2+ interactions. Neuron. 2007;56(6):992-1003.
20. Raben MS. Treatment of a pituitary dwarf with human growth hormone. The Journal of clinical endocrinology and metabolism. 1958;18(8):901-3.
21. Glick SM, Roth J, Yalow RS, Bersohn SA. Immunoassay of Human Growth Hormone in Plasma. Nature. 1963;199:784-7.
22. Rochiccioli PT, M.T. GHD: Establishment of the Diagnosis by means of GH secretion tests in KIGS. GH Therapy – 5 years of KIGS. Heidelberg: Verlag; 1994.
23. Kelberman D, Rizzoti K, Lovell-Badge R, Robinson IC, Dattani MT. Genetic regulation of pituitary gland development in human and mouse. Endocrine reviews. 2009;30(7):790-829.
24. Meinhardt UJ, Ho KK. Modulation of growth hormone action by sex steroids. Clinical endocrinology. 2006;65(4):413-22.
25. Kiess W, Gallaher B. Hormonal control of programmed cell death/apoptosis. Eur J Endocrinol. 1998;138(5):482-91.
26. Veldhuis JD, Keenan DM, Pincus SM. Motivations and methods for analyzing pulsatile hormone secretion. Endocrine reviews. 2008;29(7):823-64.
27. Thalange NK, Gill MS, Gill L, Whatmore AJ, Addison GM, Price DA, et al. Infradian rhythms in urinary growth hormone excretion. The Journal of clinical endocrinology and metabolism. 1996;81(1):100-6.
28. Spiliotis BE, August GP, Hung W, Sonis W, Mendelson W, Bercu BB. Growth hormone neurosecretory dysfunction. A treatable cause of short stature. Jama. 1984;251(17):2223-30.
29. Almuriekhi M, Shintani T, Fahiminiya S, Fujikawa A, Kuboyama K, Takeuchi Y, et al. Loss-of-Function Mutation in APC2 Causes Sotos Syndrome Features. Cell reports. 2015.
30. Yamulla RJ, Kane EG, Moody AE, Politi KA, Lock NE, Foley AV, et al. Testing models of the APC tumor suppressor/beta-catenin interaction reshapes our view of the destruction complex in Wnt signaling. Genetics. 2014;197(4):1285-302.
31. Kioussi C, Briata P, Baek SH, Rose DW, Hamblet NS, Herman T, et al. Identification of a Wnt/Dvl/beta-Catenin --> Pitx2 pathway mediating cell-type-specific proliferation during development. Cell. 2002;111(5):673-85.
32. Kelberman D, de Castro SC, Huang S, Crolla JA, Palmer R, Gregory JW, et al. SOX2 plays a critical role in the pituitary, forebrain, and eye during human embryonic development. The Journal of clinical endocrinology and metabolism. 2008;93(5):1865-73.
33. Guarente L. Sirtuins in aging and disease. Cold Spring Harbor symposia on quantitative biology. 2007;72:483-8.
34. Choi YH, Kim H, Lee SH, Jin YH, Lee KY. Src regulates the activity of SIRT2. Biochemical and biophysical research communications. 2014;450(2):1120-5.
35. Clayton P, Chatelain P, Tato L, Yoo HW, Ambler GR, Belgorosky A, et al. A pharmacogenomic approach to the treatment of children with GH deficiency or Turner syndrome. Eur J Endocrinol. 2013;169(3):277-89.
36. De Leonibus C, Chatelain P, Knight C, Clayton P, Stevens A. Effect of summer daylight exposure and genetic background on growth in growth hormone-deficient children. Pharmacogenomics J. 2016;16(6):540-50.
37. Kaizer EC, Glaser CL, Chaussabel D, Banchereau J, Pascual V, White PC. Gene expression in peripheral blood mononuclear cells from children with diabetes. The Journal of clinical endocrinology and metabolism. 2007;92(9):3705-11.
38. Elo LL, Mykkanen J, Nikula T, Jarvenpaa H, Simell S, Aittokallio T, et al. Early suppression of immune response pathways characterizes children with prediabetes in genome-wide gene expression profiling. Journal of autoimmunity. 2010;35(1):70-6.

39. Wong HR, Cvijanovich N, Lin R, Allen GL, Thomas NJ, Willson DF, et al. Identification of pediatric septic shock subclasses based on genome-wide expression profiling. BMC medicine. 2009;7:34.

40. Ramasamy A, Mondry A, Holmes CC, Altman DG. Key issues in conducting a meta-analysis of gene expression microarray datasets. PLoS medicine. 2008;5(9):e184.

41. Stevens A, Hanson D, Whatmore A, Destenaves B, Chatelain P, Clayton P. Human growth is associated with distinct patterns of gene expression in evolutionarily conserved networks. BMC Genomics. 2013;14.

42. 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):185-93.

43. Nilsson J, Fioretos T, Hoglund M, Fontes M. Approximate geodesic distances reveal biologically relevant structures in microarray data. Bioinformatics (Oxford, England). 2004;20(6):874-80.

44. Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science. 2000;290(5500):2319-23.

45. Fontes M, Soneson C. The projection score--an evaluation criterion for variable subset selection in PCA visualization. BMC Bioinformatics. 2011;12:307.

46. Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, et al. The BioGRID interaction database: 2013 update. Nucleic Acids Res. 2013;41(Database issue):D816-23.

47. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics (Oxford, England). 2011;27(3):431-2.

48. Ravasz E, Barabasi AL. Hierarchical organization in complex networks. Physical review E, Statistical, nonlinear, and soft matter physics. 2003;67(2 Pt 2):026112.

49. Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS computational biology. 2007;3(4):e59.

50. Sun J, Zhao Z. A comparative study of cancer proteins in the human protein-protein interaction network. BMC Genomics. 2010;11 Suppl 3:S5.

51. Kovacs IA, Palotai R, Szalay MS, Csermely P. Community landscapes: an integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics. PLoS One. 2010;5(9).

52. Haw R, Stein L. Using the reactome database. Current protocols in bioinformatics / editorial board, Andreas D Baxevanis [et al]. 2012;Chapter 8:Unit8 7.

53. Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. Nat Methods. 2012;9(5):471-2.

54. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of chemical information and computer sciences. 2003;43(6):1947-58.

55. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002;16:321-57.

56. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. J Stat Softw. 2010;36(11):1-13.

57. McCall MN, Bolstad BM, Irizarry RA. Frozen robust multiarray analysis (fRMA). Biostatistics. 2010;11(2):242-53.

58. McCall MN, Jaffee HA, Zelisko SJ, Sinha N, Hooiveld G, Irizarry RA, et al. The Gene Expression Barcode 3.0: improved data processing and mining tools. Nucleic Acids Res. 2014;42(Database issue):D938-43.
Figure 1 – Heatmap of gene expression for those probesets whose expression correlated with peak GH levels. A) Normal children (n=26) were combined with GHD patients (n=98), rank regression analysis was adjusted for gender and age as covariates, clusters of similar gene expression are identified using the Euclidean metric and marked using a dendrogram and white boxes (347 probesets, 271 unique genes). The distinction between normal subjects is marked by the break in the heatmap; GHD as defined by a cut off level of 10 mcg/L growth hormone as measured by provocation testing. The vertical white line demarcates the point of inflexion for gene expression at a peak GH level of 4.75 mcg/L while the horizontal white line demarcates those probesets positively and negatively associated with peak GH levels (< or >4.75 mcg/L).

B) Two way cluster analysis of gene expression in GHD and control subjects. Four distinct clusters of GHD subgroups can be seen from the dendrogram on horizontal axis derived via a Euclidian metric. There are, however, a large number of subjects where it was not possible to classify (right of white line). This group contained all bar one of the normal control subjects and 20 GHD subjects.
Figure 2 – Identification of clusters of variation of gene expression related to GHD severity.
A) Heat Map for the probesets identified by correlation with peak GH (347 probesets, 271 unique genes). Five distinct clusters of gene expression are identified via the dendrogram – two positively correlated (red) with peak GH and three negatively correlated (green). Pink, yellow and blue squares indicate which principal component analysis group for each patient (see Figure 1B) B) Isomap Supervised Principal Component Analysis using only those probesets whose expression correlated to peak GH identified 3 distinct groups of GHD subjects (n=98, coloured pink n=59, yellow n=37 and blue n=12)
Figure 3. Network modelling of the overlap of gene expression between clinical markers. A) Network models generated using BioGRID (version 3.2.117) were analysed to define modules of functionally related genes. The “community structure” of these modules was assessed and ranked by their “centrality” score to form a hierarchy related to the biological action of the network. B) Community structure of modules within the network was assessed using the Moduland algorithm in Cytoscape 2.8.3. Hierarchy of the first fifteen network modules in each of the network models of gene expression overlap between clinical markers. Modules are shown as octagons labelled with the most central gene in the cluster and ranked by network centrality (1st through 15th).
Figure 4. Summary of predicted activity and regulators derived via causal network analysis for the network modules. The hierarchy of clusters of gene expression shown in figure 2 were mapped onto identified causal networks. Activity of pathways and master regulators are coloured red to show a positive correlation with the GHD severity or green where activity is negatively correlated. Pathway ontology of all modules in the hierarchy is shown in Figure S2.
Supplemental Figure S1 — Unsupervised principal component analysis (PCA) of samples used in the analysis presented in Figure 1 (Normal children (n=26) combined with GHD patients (n=98)). Yellow = GHD patients <4mcg/L GH Peak test; magenta = GHD patients ≥4 - <10 mcg/L GH Peak test; white = normal controls.
### GH Severity: Hierarchy of functions and pathways ranked by centrality of module

#### Supplemental Figure S2 - Hierarchy of network modules and the functions associated with each module.

| Cluster | Module | Hierarchy of Functions and Pathways | Core nodes of Module |
|---------|--------|-------------------------------------|----------------------|
| 2       | UBC    | Circadian clock, Cellular responses to hypoxia, DNA replication & repair (q<1.0x10⁻⁵). | UBC, GOLGA2, DCUN1D1, SRRM2, RPS14, LRF1, SSX2IP, SART3, TRIM54, KLHL12 |
| 4       | POLH   | Circadian clock, Cell cycle, Cellular responses to hypoxia, DNA replication & repair (q<1.0x10⁻⁵). | POLH, UBC, PCNA, GOLGA2, SRRM2, RPS14, LRF1, REV1, RAD18, SSX2IP |
| 5       | CHD4   | Chromatin organisation (q<1.6x10⁻³). | CHD4, IKZF1, SRRM2, HDAC1, HDAC2, SUMO2, TRIM54, SAP30L, SART3, MAFK |
| 3       | GRB10  | Growth factor signalling (q<1.5x10⁻⁵). | GRB10, INSR, IGF1R, EGFR, IRS1, RAF1, APP, IQCB1, ELAVL1, AKT1 |
| 4       | STRN3  | WNT signalling (q<8.8x10⁻⁴) | STRN3, PPME1, PPP2R1A, STK24, PPP2CB, MOB4, STRN4, PPP2CA, FGFR10P2, APC |
| 5       | CASP2  | Apoptosis (q<2.2x10⁻²) | CASP2, CRADD, PIDD, LRF1, CASP3, CASP8, BCL2, TRIM54, SUMO1, GRB2 |
| 3       | ARHGAP1| Myogenesis (q<4.2x10⁻³) | ARHGAP1, CDC42, RHOA, GNA12, BNIP2, LRF1, NEK6, RAC1, CHEK2, SOX2 |
| 4       | LGALS3 | Growth factor signalling (q<4.8x10⁻²) | LGALS3, EGFR, LGALS3BP, GRB10, SGSM2, APP, PTGDS, BARD1, CSRP1, GEMIN4 |
| 1       | HSD17B1| M Phase (q<9.8x10⁻²) | HSD17B14, WDHYH1, NUDT18, TBC1D22B, PHF1, LOC541471, NEK6, SNRPC, TRIM54, PSMA1 |
| 2       | SETD1A | Chromatin organisation (q<9.2x10⁻¹¹). Circadian Clock (q<6.0x10⁻³). | SETD1A, RBBP5, WDR5, ASH2L, WDR82, HIST1H3A, HCFC1, DPY30, CREBBP, POLR2A |
| 3       | TRIM54 | Circadian clock, Cellular responses to hypoxia, DNA replication & repair (q<1.0x10⁻⁵). | TRIM54, SXX2IP, TRIM42, OTUB2, C15orf55, CRP, NELL2, ATXN7, PIK3R3, ORM1 |
| 4       | SELENBP1| Detoxification of reactive oxygen species (q<3.7x10⁻⁴) | SELENBP1, EED, MED31, LRF1, BARD1, USP33, TMED9, AR, GPX1, MLH1 |
| 3       | SUFU   | Mitochondrial biogenesis (q<5.1x10⁻³). Hedgehog signalling (q<1.6x10⁻⁴) | SUFU, GLI1, ZNF747, SIRT3, FBXL17, RCN3, SIRT2, TRIM42, DMPK, TRAF1 |
| 4       | DYSK1B| Circadian Clock (q<8.2x10⁻³). Growth factor signalling (q<6.3x10⁻³) | DYSK1B, DCAF7, CREBBP, PRKACA, STUB1, CTBP2, WDR6, USP11, HDAC5, CCNA2 |
| 3       | GRIN2A | Growth factor signalling (q<5.5x10⁻²) | GRIN2A, DLG4, FYN, DLG3, SRC, NEDD4, SPTAN1, DLGAP3, GRB10, ARHGAP1 |

The overlap between network modules and gene expression clusters is indicated. Core nodes within the module listed (in bold: the most central gene in the module). Biological pathways and functions were derived from the Reactome database, q-values are the false discovery rate modified p-values determined using the hypogeometric test.
Table 1 – SNPs associated with peak GH concentrations. Genotype categorisation indicates whether peak GH is associated with minor allele carriage, major allele carriage or whether a relationship with peak GH exists across all three genotypes (i.e. major homozygote, heterozygote, minor homozygote) and is labelled as nominal genotype. MAF = minor allele frequency. For continuous analysis a Kruskall-Wallis Test was used and a Fisher’s exact test for categorical analysis, p-values are Bonferroni corrected.

| Gene  | SNP          | MAF    | Region | Genotype categorisation       | Phenotype variable type | P-value   |
|-------|--------------|--------|--------|-------------------------------|-------------------------|-----------|
| BCL2  | rs4987786    | 0.0397 | Intron | minor allele carriage         | categorical             | 0.036     |
| CCND3 | rs3218100    | 0.0242 | Intron | nominal genotype              | categorical             | <0.001    |
| CDK2  | rs2069408    | 0.1849 | Intron | major allele carriage         | continuous              | <0.001    |
|       | rs10459592   | 0.4433 | Intron | minor allele carriage         | continuous              | <0.001    |
|       | rs4545755    | 0.2780 | Intron | major allele carriage         | continuous              | <0.001    |
|       | rs700518     | 0.3259 | Exon   | major allele carriage         | continuous              | <0.001    |
|       | rs7172156    | 0.3884 | Intron | nominal genotype              | continuous              | 0.001     |
|       | rs767199     | 0.2933 | Intron | major allele carriage         | continuous              | 0.001     |
| CYR61 | rs2297141    | 0.4675 | Intron | major allele carriage         | continuous              | 0.024     |
| IGFBP1| rs4619       | 0.3760 | Exon   | nominal genotype              | continuous              | 0.012     |
| POU1F1| rs12486159   | 0.2504 | Intron | minor allele carriage         | categorical             | <0.001    |
|       | rs17189466   | 0.1651 | Intron | minor allele carriage         | categorical             | <0.001    |
| PTPN1 | rs6126033    | 0.1272 | Intron | major allele carriage         | continuous              | <0.001    |
|       | rs941798     | 0.4251 | Intron | minor allele carriage         | continuous              | <0.001    |
| RARA  | rs2715553    | 0.4050 | Intron | major allele carriage         | categorical             | <0.001    |
| SHC1  | rs4845401    | 0.4904 | Intron | minor allele carriage         | categorical             | 0.002     |
| SREBF1| rs9899634    | 0.3351 | Intron | major allele carriage         | categorical             | <0.001    |
| TGFA  | rs6749533    | 0.1571 | Intron | minor allele carriage         | categorical             | 0.035     |
Table 2 – Baseline Auxological and Biochemical Parameters in Groups of GHD Children Identified by Supervised Principal Component Analysis. There is evidence of increasing GH severity between the groups with peak GH, pre-treatment IGF1 SDS decreasing across groups 1-3 as defined by supervised PCA. Data are presented as median (range) with differences between groups assessed using a Kruskal Wallis test.

|                        | Group 1 N=12 | Group 2 N=37 | Group 3 N=49 | p-value |
|------------------------|-------------|-------------|-------------|---------|
| Age years              | 9.7 (6.3 – 13.3) | 9.2 (2.5 – 13.4) | 8.5 (2.3 – 15.6) | NS      |
| Male (n, %)            | 8 (66)      | 25 (67)     | 31 (63)     | NS      |
| Birth Weight SDS       | 0.6 (-1.2 to 1.0) | -0.6 (-3.0 to 0.9) | -0.1 (-2.7 to 4.6) | NS      |
| Birth Length SDS       | -0.1 (-0.9 to 1.1) | -0.6 (-7.0 to 0.9) | -0.4 (-5.1 to 1.8) | NS      |
| Height SDS             | -1.9 (-2.9 to -0.2) | -2.0 (-6.5 to -1) | -2.2 (-3.4 to -0.3) | NS      |
| Weight SDS             | -1.0 (-2.0 to 0.5) | -1.5 (-4.2 to 3.9) | -1.4 (-3.2 to 2.3) | NS      |
| Distance to Target     | -0.73 (-3.2 to 2.0) | -1.4 (-5.6 to 0.9) | -1.4 (-4.3 to 3.3) | 0.037   |
| Height SDS             |             |             |             |         |
| Peak GH μg/L           | 6.5 (4.9 - 9.3)  | 3.9 (0.12 – 7.1)  | 3.3 (0.1 – 7.7)  | 0.001   |
| Pre-treatment IGF1 SDS | -0.8 (-3.3 to -0.4) | -1.3 (-5.2 to +0.3) | -2.1 (-7.8 to 0) | 0.031   |
| Pre-treatment          | 0.4 (-1.5 to 0.9)  | 0.1 (-4.1 to 1.5)  | -0.4 (-6.5 to 1.9) | NS      |
| IGFBP-3 SDS            |             |             |             |         |
Table 3 – Prediction of GH severity (peak GH <4 µg/L or >4 µg/L) via Random Forest model. Data used in the model was classified into biochemical and demographic data (baseline IGF1, IGFBP-3, age), genotype data (SNPs identified as being associated with peak GH – see Table 1) and transcriptomic data (top 50 probesets identified via rank regression model ranked by network centrality). The predictive capacity of each of the three classes of data was assessed independently and then in combination. AUC – area under the receiver operator characteristic curve.

| Data used in the prediction model | Predictive capacity |
|----------------------------------|---------------------|
| Biochemical and Demographic Data | Genotype Data       | Transcriptomic Data | AUC   | 95%CI AUC |
| Yes                             | No                  | No                  | 0.88  | 0.81-0.94 |
| No                              | Yes                 | No                  | 0.85  | 0.78-0.91 |
| No                              | No                  | Yes                 | 0.93  | 0.88-0.98 |
| Yes                             | Yes                 | No                  | 0.83  | 0.76-0.90 |
| Yes                             | No                  | Yes                 | 0.93  | 0.88-0.98 |
| No                              | Yes                 | Yes                 | 0.93  | 0.88-0.98 |
| Yes                             | Yes                 | Yes                 | 0.93  | 0.88-0.97 |
**Supplemental Table S1: Gene expression associated with GH Peak.** Rank regression (p<0.01) normalised for microarray batch.

| varID               | Gene Symbol | p-value     | R-statistic |
|---------------------|-------------|-------------|-------------|
| 231702_at           | TDO2        | 5.77E-06    | 0.477366046 |
| 1560550_at          | ---         | 1.80E-05    | -0.454363438|
| 215370_at           | ---         | 2.60E-05    | 0.446637952 |
| 225920_at           | LOC148413   | 9.84E-05    | 0.416742117 |
| 243494_at           | LOC100506926| 0.000134922 | 0.409215953 |
| 224234_at           | ---         | 0.000178275 | -0.389426118|
| 220265_at           | GPR107      | 0.000230419 | 0.396045376 |
| 231206_at           | ---         | 0.000299045 | -0.389418807|
| 242764_at           | DCHS2       | 0.00029913  | -0.389418807|
| 217145_at           | IGK@ /// IGKC /// LOC | 0.000329888 | 0.386897332 |
| 217267_s_at         | RAB7A       | 0.000363501 | 0.384377656 |
| 209915_s_at         | NRXN1       | 0.000383637 | 0.382968816 |
| 1564004_at          | ---         | 0.000466313 | -0.377815609|
| 219969_at           | TXLNG       | 0.000466313 | -0.377815609|
| 43427_at            | ACACB       | 0.000564526 | 0.369047159 |
| 222959_at           | CNGB3       | 0.000708933 | 0.36459924  |
| 205808_at           | ASPH        | 0.000735315 | 0.36549648  |
| 233528_s_at         | GATSL3 /// TBC1D10A | 0.000736399 | -0.365408821|
| 244652_at           | ---         | 0.000748581 | 0.364954017 |
| 233298_at           | C13orf38    | 0.000800587 | -0.36305247 |
| 239589_s_at         | ---         | 0.000825327 | -0.36224752 |
| 202966_at           | CAPN6       | 0.000825946 | 0.362213771 |
| 211689_s_at         | TMPRSS2     | 0.000866184 | 0.36087914  |
| 206259_at           | PROC        | 0.000892986 | 0.36022069  |
| 214602_at           | COL4A4      | 0.000955786 | 0.358100722 |
| 220288_at           | MYO15A      | 0.001017921 | -0.356309469|
| 1557652_a_at        | LOC348817   | 0.001023403 | 0.356156243 |
| 201616_s_at         | CALD1       | 0.001081548 | -0.354575272|
| 239514_at           | LOC100508951| 0.001107779 | -0.35387166 |
| 215973_at           | HCG4P6      | 0.001147992 | -0.352860446|
| 243307_at           | ---         | 0.001158518 | 0.352597084 |
| 239856_at           | ---         | 0.001211503 | 0.351207238 |
| 212187_x_at         | PTGDS       | 0.001233366 | 0.350784594 |
| 210824_at           | ---         | 0.001264386 | -0.350062505|
| 49452_at            | ACACB       | 0.00135921  | 0.347950547 |
| 228011_at           | FAM92A1     | 0.001367323 | 0.347776085 |
| 1553383_at          | ARHGAP42    | 0.001535038 | 0.344364806 |
| 1566688_at          | ---         | 0.001570263 | -0.343691422|
| 1569840_at          | ---         | 0.001606817 | 0.343006902 |
| 220703_at           | C10orf110   | 0.001670352 | -0.34189889 |
| 233032_x_at         | ---         | 0.00167036  | 0.341849756 |
| 221409_at           | OR2S2       | 0.001680778 | -0.341663839|
| 229807_s_at         | MAZ         | 0.00170203  | -0.341287788|
| 211663_x_at         | PTGDS       | 0.001724632 | 0.340892483 |
| Gene ID   | Symbol | Log2 Fold Change (Down) | Log2 Fold Change (Up) |
|----------|--------|------------------------|----------------------|
| 233669_s_at | TRIM54 | 0.001732399            | -0.340757707         |
| 236279_at  |        | 0.001755487            | 0.340360298          |
| 216665_s_at | TTTY2  | 0.001777037            | -0.33999359          |
| 236993_at  |        | 0.001810467            | 0.339432589          |
| 230642_at  |        | 0.001872657            | -0.338413341         |
| 233410_at  |        | 0.001894739            | 0.338058765          |
| 224631_at  | ZFP91  | 0.001940294            | -0.337338886         |
| 243408_at  |        | 0.001966645            | 0.336929389          |
| 216665_s_at | TTTY2  | 0.001777037            | -0.33999359          |
| 236993_at  |        | 0.001810467            | 0.339432589          |
| 230642_at  |        | 0.001872657            | -0.338413341         |
| 233410_at  |        | 0.001894739            | 0.338058765          |
| 224631_at  | ZFP91  | 0.001940294            | -0.337338886         |
| 243408_at  |        | 0.001966645            | 0.336929389          |
| 216665_s_at | TTTY2  | 0.001777037            | -0.33999359          |
| 236993_at  |        | 0.001810467            | 0.339432589          |
| 230642_at  |        | 0.001872657            | -0.338413341         |
| 233410_at  |        | 0.001894739            | 0.338058765          |
| 224631_at  | ZFP91  | 0.001940294            | -0.337338886         |
| 243408_at  |        | 0.001966645            | 0.336929389          |
| 216665_s_at | TTTY2  | 0.001777037            | -0.33999359          |
| 236993_at  |        | 0.001810467            | 0.339432589          |
| 230642_at  |        | 0.001872657            | -0.338413341         |
| 233410_at  |        | 0.001894739            | 0.338058765          |
| 224631_at  | ZFP91  | 0.001940294            | -0.337338886         |
| 243408_at  |        | 0.001966645            | 0.336929389          |
| Gene Symbol | Gene Name  | Fold Change | t-Value |
|-------------|------------|-------------|---------|
| 209409_at   | GRB10      | 0.002879741 | 0.325118483 |
| 1559620_at  | LOC441167  | 0.002913239 | -0.324753028 |
| 1564658_at  | C7orf52    | 0.002939365 | 0.323633503 |
| 210215_at   | TFR2       | 0.002967502 | -0.323199598 |
| 211748_x_at | PTGDS      | 0.003018039 | 0.323633503 |
| 236941_at   | C22orf30   | 0.003059556 | -0.323199598 |
| 209812_x_at | CASP2      | 0.003099271 | 0.322789423 |
| 241898_at   | LIPIH      | 0.003135148 | 0.322422907 |
| 244269_at   | WISP1      | 0.00314508  | 0.322322109 |
| 224203_at   | SUFU       | 0.003187697 | 0.321783384 |
| 211312_s_at | WISP1      | 0.003211597 | 0.321654277 |
| 213671_s_at | MARS       | 0.003222639 | 0.321544614 |
| 240447_at   | ---        | 0.003244208 | -0.321331363 |
| 215118_s_at | IGHJ1      | 0.003298012 | 0.320804882 |
| 243630_at   | NDUFB1     | 0.003321597 | 0.321654277 |
| 1566984_at  | ---        | 0.003322639 | 0.321544614 |
| 207459_x_at | GYPB       | 0.003435421 | -0.319494364 |
| 229653_at   | VPS53      | 0.00344044  | 0.319447392 |
| 206529_x_at | SLC26A4    | 0.003485287 | 0.319030341 |
| 238216_at   | ---        | 0.003493972 | -0.318950122 |
| 215250_at   | TMEM111    | 0.003499698 | 0.318897335 |
| 201183_s_at | CHD4       | 0.003505389 | 0.318844954 |
| 214566_at   | SMR3A      | 0.003520477 | 0.318706428 |
| 240526_at   | ---        | 0.003545076 | -0.31848172 |
| 202358_s_at | SNX19      | 0.003553465 | 0.318405405 |
| 209720_s_at | SERPINB3   | 0.003581973 | 0.318147261 |
| 1560912_at  | LOC389043  | 0.003612722 | -0.31787087 |
| 240428_at   | LOC285000  | 0.003630577 | 0.317711341 |
| 205502_at   | CYP17A1    | 0.003649385 | -0.317544057 |
| 210745_at   | ONECUT1    | 0.003699964 | -0.317097967 |
| 221631_at   | CACNA1I    | 0.003700836 | 0.317090327 |
| 203238_s_at | NOTCH3     | 0.003757657 | 0.316595924 |
| 234450_at   | PROKR2     | 0.003827304 | -0.315998496 |
| 200771_at   | LAMC1      | 0.003907935 | -0.315319065 |
| 216235_s_at | EDNRA      | 0.003920295 | -0.315216021 |
| 203623_at   | PLXNA3     | 0.003956944 | 0.314912177 |
| 239918_at   | ---        | 0.00400513  | -0.314516469 |
| 224681_at   | GNA12      | 0.00402771  | -0.314332499 |
| 242094_at   | ---        | 0.004035577 | -0.314268621 |
| 228022_at   | CCDC18     | 0.004060271 | 0.314068818 |
| 1566851_at  | TRIM42     | 0.004091225 | -0.313819896 |
| 205753_at   | CRP        | 0.004132416 | 0.313491231 |
| Gene ID     | Gene Symbol | E-values | Log2 Fold Changes |
|------------|-------------|----------|------------------|
| 1554643_at | RGS11       | 0.004145465 | 0.313387715       |
| 232517_s_at| PRIC285     | 0.004213604 | -0.312851881      |
| 222094_at  | SULT1A3 /// SULT1A4 | 0.004255152 | 0.312151766       |
| 203631_s_at| GPRC5B      | 0.004271779 | 0.312528916       |
| 220622_at  | LRRC31      | 0.004295221 | -0.311737837      |
| 220506_at  | GUClB2      | 0.004304133 | 0.312151766       |
| 206750_at  | MAFK        | 0.004358467 | -0.311290010      |
| 240098_at  | RIF1        | 0.004378442 | 0.311284176       |
| 1568902_at | ---         | 0.004383403 | 0.311549400       |
| 240091_at  | PSMA8       | 0.004417936 | -0.311284176      |
| 1555839_a_at| C3orf79    | 0.004418716 | -0.311284176      |
| 224937_at  | PTGFRN      | 0.004450415 | 0.311047690       |
| 210842_at  | NRP2        | 0.004455853 | -0.310972586      |
| 239244_at  | ---         | 0.004460525 | -0.310972586      |
| 219472_at  | CENPO       | 0.004468481 | 0.310913588       |
| 236787_at  | LOC100507286| 0.004497116 | -0.310702009      |
| 241078_at  | SLC35E4     | 0.004501386 | 0.310670556       |
| 219092_s_at| IPPK        | 0.004520148 | 0.310532693       |
| 1563086_at | ---         | 0.004527986 | 0.310475248       |
| 221348_at  | NPPC        | 0.004590122 | -0.310022966      |
| 235512_at  | CDKL1       | 0.004654325 | -0.309561307      |
| 221137_at  | ---         | 0.004655871 | -0.30955026       |
| 205595_at  | DSG3        | 0.004672477 | -0.309431813      |
| 201967_at  | RBM6        | 0.004680829 | 0.309372377       |
| 232009_at  | EMR2        | 0.004694908 | 0.309272409       |
| 227643_at  | TPPP        | 0.004696136 | -0.309263697      |
| 219380_x_at| POLH        | 0.004812862 | 0.308445125       |
| 236012_at  | PSMF1       | 0.004839326 | -0.308261992      |
| 222506_at  | LMBR1       | 0.004856587 | -0.308143021      |
| 224771_at  | NAV1        | 0.004859647 | 0.308121975       |
| 204507_s_at| PPP3R1      | 0.004970425 | -0.30736768       |
| 1561678_at | ---         | 0.004971689 | 0.307359159       |
| 241733_at  | C18orf54    | 0.005027006 | -0.306988156      |
| 214944_at  | PHLPP2      | 0.005052297 | -0.306819735      |
| 211912_at  | MERTK       | 0.005093075 | 0.306549752       |
| 219113_x_at| HSD17B14    | 0.005138008 | 0.306254469       |
| 243616_at  | ---         | 0.005154913 | 0.30614397        |
| 200851_s_at| KIAA0174    | 0.00519766  | 0.30586598        |
| 1564083_at | ---         | 0.00520798  | -0.305799174      |
| 213393_at  | MFS9D       | 0.005245121 | 0.3055597         |
| 233830_at  | LOC90246    | 0.00528199  | -0.30532346       |
| 227842_at  | RAB30       | 0.005304515 | -0.305799174      |
| 223631_s_at| C19orf33    | 0.005325114 | -0.305048992      |
| 1556891_at | SORCS1      | 0.005330656 | -0.305013858      |
| 243555_at  | ---         | 0.005338319 | 0.304993827       |
| 217066_s_at| DMPK        | 0.005338525 | -0.304964033      |
| 210558_at  | AKR1C4      | 0.00535823  | 0.304893551       |
| 1570285_at | ---         | 0.005366561 | 0.304787045       |
| Gene Symbol | Gene Name | Log2 Ratio (Case) | Log2 Ratio (Control) |
|-------------|-----------|------------------|---------------------|
| 219931_s_at | KLHL12    | 0.005373958      | -0.304740479        |
| 234592_at  | ---       | 0.00544307       | -0.304308173        |
| 217292_at  | MTMR7     | 0.005464988      | -0.304172087        |
| 235454_at  | ---       | 0.005495865      | 0.303981198         |
| 1559840_s_at | TBX18 | 0.005544523      | -0.30368229         |
| 203568_s_at | TRIM38    | 0.00561207       | 0.303271182         |
| 217021_at  | CYB5A     | 0.005679086      | 0.302867611         |
| 238980_x_at | C17orf56 | 0.005720634      | 0.302619514         |
| 1569990_at | NUDT3     | 0.005770473      | -0.302324007        |
| 216910_at  | XPNPEP2   | 0.005835199      | -0.301943588        |
| 241352_at  | ---       | 0.005869679      | 0.300483081         |
| 1555127_a_at | ITGA11 | 0.005897061      | -0.301180214        |
| 212705_x_at | PNPLA2    | 0.005998634      | -0.30099364         |
| 208610_s_at | SRRM2     | 0.006048353      | 0.30071663          |
| 215434_x_at | LOC100288142 | 0.006060751 | 0.300646448 |
| 241352_at  | ---       | 0.006089697      | 0.300483081         |
| 1557953_at | ZKSCAN1   | 0.006102547      | 0.300410778         |
| 225172_at  | CRAMP1L   | 0.006129486      | 0.300259634         |
| 209127_s_at | SART3    | 0.006155092      | 0.300116519         |
| 239855_at  | PPM1L     | 0.006162523      | -0.30011004         |
| 207600_at  | KCNC3     | 0.006163972      | -0.300067004        |
| 223597_at  | ITLN1     | 0.006165239      | -0.300059947        |
| 211691_x_at | ---       | 0.006176034      | -0.299998855        |
| 223365_at  | DHX37     | 0.006180478      | 0.299975142         |
| 277327_at  | MEGF8     | 0.006204056      | 0.299844298         |
| 1555505_a_at | TYR    | 0.006226321      | -0.299721146        |
| 203502_at  | BPGM      | 0.006280699      | -0.299421985        |
| 220082_at  | PPP1R14D  | 0.006407819      | 0.29873147          |
| 204991_s_at | NF2       | 0.006462013      | -0.298440754        |
| 241288_at  | ---       | 0.006462759      | 0.298436768         |
| 244069_at  | ---       | 0.006480128      | 0.298344063         |
| 231441_at  | C7orf62   | 0.006497121      | -0.298253574        |
| 239625_at  | ---       | 0.006500608      | 0.29823503          |
| 208997_s_at | UCP2      | 0.006539199      | -0.2980304          |
| 234638_at  | ---       | 0.006645096      | 0.297474304         |
| 201053_s_at | PSMF1     | 0.00666174       | -0.297387613        |
| 219233_s_at | GSDMB     | 0.006754643      | 0.296907206         |
| 1557170_at | NEK8      | 0.006838457      | 0.296478774         |
| 209376_x_at | SRSF2IP   | 0.006854349      | 0.296398059         |
| 239457_at  | ATP8B3    | 0.006885973      | 0.296237933         |
| 213202_at  | SETD1A    | 0.006940895      | 0.295961375         |
| 215366_at  | SNX13     | 0.006943012      | 0.29590758          |
| ProbeID   | GeneSymbol | CValue        | PValue       |
|-----------|------------|---------------|--------------|
| 213850_s_at | SRSF2IP    | 0.006987884   | 0.295726292  |
| 207167_at  | CD101      | 0.007017457   | 0.295579053  |
| 232545_at  | LRRC29     | 0.007066379   | 0.295366758  |
| 1557164_a_at | ---       | 0.00704421    | 0.29526971   |
| 213279_at  | DHRS1      | 0.007105015   | 0.295146299  |
| 224350_at  | ---        | 0.007108426   | 0.295129536  |
| 200759_x_at | NFE2L1     | 0.007116777   | 0.295088522  |
| 238167_at  | ACOT12     | 0.007123123   | 0.295057388  |
| 224157_at  | KAAP1      | 0.007155818   | 0.294897355  |
| 239546_at  | LOC100131053 | 0.007174306 | 0.294807143  |
| 1555829_at | ESYT2      | 0.007176335   | 0.294797259  |
| 216689_x_at | ARHGAP1    | 0.007197632   | 0.294641888  |
| 1559630_at | DLGAP3     | 0.007237858   | 0.294498609  |
| 1563204_at | ZNF627     | 0.007282322   | -0.294039452 |
| 1552499_a_at | ZSCAN20    | 0.007337467   | -0.29401973  |
| 213908_at  | WHAMML1###WHAML | 0.00737521 | -0.294019527 |
| 230904_at  | FSD1L      | 0.007350103   | -0.29395452  |
| 238135_at  | ---        | 0.007373325   | -0.29388812  |
| 1555866_a_at | HEXDC      | 0.007431993   | -0.29376059  |
| 214433_s_at | SELENBP1   | 0.007453158   | -0.29347083  |
| 209952_s_at | MAP2K7     | 0.007481754   | -0.29336242  |
| 1556885_at | LAYN       | 0.007490202   | -0.29329657  |
| 240677_at  | ---        | 0.007494095   | -0.29327316  |
| 235376_at  | ---        | 0.007497189   | -0.29326380  |
| 241717_at  | LOC285281  | 0.007514247   | -0.29318391  |
| 1557841_at | ---        | 0.007519695   | -0.29315842  |
| 229178_at  | PRTG       | 0.007605286   | -0.29276017  |
| 224907_s_at | SH3GLB2    | 0.007611731   | -0.29270346  |
| 227392_at  | NISCH      | 0.00762835    | -0.29265352  |
| 210203_at  | CNOT4      | 0.007655655   | -0.29252764  |
| 244626_at  | ---        | 0.007669754   | -0.29246280  |
| 229599_at  | LOC440335  | 0.007686839   | -0.29238436  |
| 237238_at  | WWC1       | 0.007689966   | -0.29237002  |
| 223998_at  | TTL2       | 0.007692117   | -0.29236016  |
| 202236_s_at | SLC16A1    | 0.007742197   | -0.29213125  |
| 222796_at  | PTCD1      | 0.007761398   | -0.29203834  |
| 219265_at  | MOBKL2B    | 0.00782682    | -0.2917474  |
| 241295_at  | ---        | 0.007981817   | -0.29105368  |
| 208949_s_at | LGALS3     | 0.008016107   | -0.29090181  |
| 223561_at  | NEK6       | 0.008016266   | -0.29090111  |
| 235856_at  | ---        | 0.008027194   | -0.29085283  |
| 229151_at  | SLC14A1    | 0.008036478   | -0.29081185  |
| 230556_at  | IMMP1L     | 0.008053687   | -0.29073602  |
| 1555913_at | GON4L      | 0.00809016    | 0.29057577   |
| 1561232_at | LOC100270680 | 0.008093244 | 0.290562249  |
| 1570224_at | ---        | 0.008116976   | 0.290458348  |
| 1555993_at | CACNA1D    | 0.008163429   | 0.290255748  |
| 207109_at  | POU2F3     | 0.008181455   | 0.290177398  |
| probe_id | gene_symbol | p_value | fold_change |
|----------|-------------|---------|-------------|
| 1561969_at | ZPLD1 | 0.008253129 | 0.289867373 |
| 213751_at | LRRC68 | 0.008265399 | -0.28987102 |
| 1561394_s_at | KIAA1755 | 0.008287217 | -0.289721554 |
| 223663_at | CCDC88B | 0.008342221 | 0.289485284 |
| 1559060_a_at | FNIP1 | 0.008347294 | 0.289463637 |
| 223663_at | MPHOSPH8 | 0.008350905 | 0.289448231 |
| 242741_x_at | --- | 0.008350905 | 0.289448231 |
| 1552950_at | C15orf26 | 0.008388762 | -0.289287102 |
| 212055_at | C18orf10 | 0.008392423 | -0.289271554 |
| 1564603_at | C15orf55 | 0.008410083 | -0.289196638 |
| 208666_s_at | ST13 | 0.008410116 | -0.289196498 |
| 1553936_a_at | MGC2848 | 0.008448528 | -0.289034021 |
| 228451_at | TSSK3 | 0.008541514 | 0.288643371 |
| 228922_at | SHF | 0.008581023 | 0.288478516 |
| 1553354_a_at | FLJ31958 | 0.008603126 | -0.288386578 |
| 223316_at | CCDC3 | 0.008629499 | 0.288413559 |
| 237319_at | C2orf53 | 0.008677952 | 0.288076871 |
| 1557029_at | --- | 0.008700647 | -0.2879834 |
| 234245_at | --- | 0.008733359 | 0.287849043 |
| 211059_s_at | GOLGA2 | 0.008754149 | 0.287763888 |
| 207892_at | CD40LG | 0.008770052 | 0.287698864 |
| 214979_at | ABCC3 | 0.008830446 | -0.287452874 |
| 1566834_at | --- | 0.008924204 | -0.287073895 |
| 237319_at | C2orf53 | 0.009000029 | -0.286799945 |
| 234263_at | --- | 0.00909193 | -0.28673336 |
| 208844_at | VDAC3 | 0.009052955 | -0.286559114 |
| 1565817_at | IKZF1 | 0.009058697 | -0.286536306 |
| 212931_at | TCF20 | 0.009071077 | 0.286487171 |
| 207865_s_at | BMP8B | 0.009082172 | 0.286443192 |
| 216660_at | MYO7B | 0.009100098 | 0.286372226 |
| 207481_at | --- | 0.009216338 | 0.285915018 |
| 207883_s_at | TFR2 | 0.009272096 | -0.285697496 |
| 208193_at | IL9 | 0.00932119 | 0.285506914 |
| 224418_x_at | PMCHL1 | 0.009352317 | 0.285386537 |
| 1554710_at | KCNMB1 | 0.00939482 | 0.285222727 |
| 205940_at | MYH3 | 0.009396291 | 0.28521707 |
| 1562853_x_at | --- | 0.009406353 | -0.28517839 |
| 210822_at | RPL13P5 | 0.009424842 | 0.285107411 |
| 230526_at | LOC100131096 | 0.009439944 | 0.285049525 |
| 49077_at | PPME1 | 0.009456222 | -0.284987223 |
| 241241_at | RPS14 | 0.009487364 | 0.28486829 |
| 241418_at | LOC344887 | 0.009499548 | 0.284821849 |
| 208417_at | FGF6 | 0.009527 | -0.284717409 |
| 239656_at | LOC723809 | 0.00953748 | 0.284677608 |
| 233831_at | LOC100289465 | 0.00954032 | -0.284666829 |
| 235683_at | SESN3 | 0.009688623 | -0.284107802 |
| Probe ID | Gene Symbol | Expression Value 1 | Expression Value 2 |
|----------|-------------|--------------------|--------------------|
| 204954_s_at | DYRK1B | 0.009710181 | -0.284027165 |
| 1564985_a_at | SLC8A1 | 0.009722071 | 0.283982756 |
| 1560475_at | LOC100129455 | 0.009741407 | 0.283910641 |
| 1554398_at | LYG2 | 0.009772762 | 0.283793962 |
| 205040_at | ORM1 | 0.009775476 | -0.28378388 |
| 219245_s_at | OGFOD2 | 0.00978062 | 0.283764773 |
| 1558444_at | --- | 0.00978697 | 0.2837412 |
| 201613_s_at | AP1G2 | 0.009810077 | 0.283655529 |
| 222678_s_at | DCUN1D1 | 0.009840647 | -0.283542467 |
| 1560971_a_at | --- | 0.009852319 | -0.283499378 |
| 237202_at | PGPEP1 | 0.009858424 | 0.283476857 |
| 230213_at | C19orf43 | 0.00986748 | 0.283443475 |
| 1561593_at | --- | 0.009910214 | -0.283286305 |
| 242033_at | RNF180 | 0.009931975 | 0.283206501 |
| 206249_at | MAP3K13 | 0.00993722 | -0.283187289 |
Supplementary Table S2 – List of the probesets (n=53) confirmed by Boruta as containing the predictive capacity for the classification of patients into GHD and control subjects. Where available the associated gene name for the probeset is given. The probesets highlighted in grey are those which are also expressed in the human growth plate.

| Affymetrix HG U133-PLUS-2 probeset | Associated Gene Name |
|-----------------------------------|----------------------|
| 214602_at                        | COL4A4               |
| 202358_s_at                      | SNX19                |
| 1554348_s_at                     | CDKN2AIPNL           |
| 219380_x_at                      | POLH                 |
| 1564658_at                       | NAT16                |
| 210203_at                        | CNOT4                |
| 1557953_at                       | ZKSCAN1              |
| 43427_at                         | ACACB                |
| 211462_s_at                      | TBL1Y                |
| 226545_at                        | CD109                |
| 209812_x_at                      | CASP2                |
| 214433_s_at                      | SELENBP1             |
| 239244_at                        | RP11-693J15.5        |
| 224418_x_at                      | PMCHL1               |
| 224203_at                        | SUFU                 |
| 208193_at                        | IL9                  |
| 201616_s_at                      | CALD1                |
| 203502_at                        | BPGM                 |
| 224937_at                        | PTGFRN               |
| 232698_at                        | BPIFB2               |
| 219182_at                        | TMEM231              |
| 217066_s_at                      | DMPK                 |
| 201967_at                        | RBM6                 |
| 219931_s_at                      | KLHL12               |
| 208417_at                        | FG6                  |
| 217445_s_at                      | GART                 |
| 202236_s_at                      | SLC16A1              |
| 208949_s_at                      | LGALS3               |
| 1557029_at                       | HMMR-AS1             |
| 209127_s_at                      | SART3                |
| 215505_s_at                      | STRN3                |
| 230556_at                        | IMMP1L               |
| 206534_at                        | GRIN2A               |
| 222005_s_at                      | GNG3                 |
| 228011_at                        | FAM92A1              |
| 1561593_at                       | RP11-400D2.2         |
| 211748_x_at                      | PTGDS                |
| 243408_at                        | RP11-77403.2         |
| 239625_at                        | RP11-1109F11.3       |
| 225981_at                        | HID1                 |
| 230526_at                        | -                    |
| 237238_at                        | -                    |
| 1566602_at                       | -                    |
| 212187_x_at                      | -                    |
| 15699990_at                      | -                    |
| 1559630_at                       | -                    |
| 1560171_at                       | -                    |
| 1562853_x_at                     | -                    |
| 1565875_at                       | -                    |
| 221137_at                        | -                    |
| 233032_x_at                      | -                    |
| 236993_at                        | -                    |
| 237539_at                        | -                    |