Subspace Clustering of CFS Data

NYU Bioinformatics
Oct 5, 2009
The Epidemiology of CFS

• Clinic populations
  – Women
  – Primarily white
  – Middle upper socioeconomic status
  – Sudden onset (recovery more likely)
  – Illness duration 5 yrs
  – 22 physician visits/year

• General population
  – Women
  – All race/ethnicities
  – Low socioeconomic
  – Gradual onset (lower recovery rates)
  – Illness duration 5 yrs
  – 15% have been diagnosed/treated

Economic impact - $9 billion/year, $20,000 per household
CFS Heterogeneity

- Regarding CFS heterogeneity.
- Here are some of Q’s related to CFS heterogeneity…
  - Since we observe differing insults (viral or infection or stress, … or a combination of the these) preceding the CFS state, is there a difference in what’s going wrong ‘under the hood’ physiologically, in people with CFS?
  - Since there are natural genetic variations between people, does this lead to variations in susceptibility to the disease from these insults?
  - What is an exhaustive list of disturbed biological pathways in chronic fatigued persons – does this list vary by person?
  - What are the biological sequence of events (in terms of biological pathway disruptions) leading to a CFS state, and does this sequence of events vary from person to person?
  - Are there common themes in the answer to the above questions? Can we define subgroups of people with CFS from these common themes?
- Can we use these subgroups to develop optimal, targeted treatments - custom tailored to address each subgroups specific characteristic problems?
Study Subjects

This in-hospital study enrolled people who were identified with CFS according to the 1994 CFS case definition as described by Fukuda and colleagues during the 4-year longitudinal study of CFS in Wichita, KS, USA.

Never fatigued and no exclusionary conditions
- NF (n = 55)

Everyone ever clinically evaluated between 1997 and 2000
- CFS (n = 58)
- CFS-MDDm (n = 27)
- ISF (n = 59)
- ISF-MDDm (n = 28)

People with ISF were also enrolled, in addition to those with CFS and ISF with MDDm. NF controls were selected from people surveyed in the longitudinal study that did not report fatigue, medical and psychiatric exclusionary conditions and were similar in age, race and BMI to people with CFS and ISF.

BMI: Body mass index; CFS: Chronic fatigue syndrome; ISF: Insufficient symptoms and fatigue; MDDm: Major depressive disorder with melancholy; NF: Nonfatigued.
Data Integration to Identify Biomarkers

**Descriptive Data**
- Body – physical and clinical
- Instruments to describe symptom domains (SF-36, MFI, CDCSI)
- Psychology/psychiatric

**Neuroendocrine/Immune**
- HPA, HPG, HPT
- ANS
- Immune system

**Sleep**
- Brain and muscle

**Cognition**
- CANTAB

**Targeted Genetics**
- Single nucleotide polymorphisms of genes involved in HPA axis function
Initial Dataset (Table List)

- Blood Pressure.csv
- CANTAB.csv
- Catecholamines.csv
- Class and Demo.csv
- Complete Blood Evaluation.csv
- Endocrine.csv
- Gynecologic History.csv
- Medications.csv
- MEDID.csv
- Menstrual WASI WRAT.csv
- MFI.csv
- PharmaActionTable.csv
- Physical Exam.csv
- SF36 Summary Scores.csv
- Sleep Evaluation.csv
- Symptom Inventory.csv
- Urine Profile.csv
- Zung.csv

plus cytokine data

- abtinfo
- MF1
- zung
- sf36
- tilt_info
- tilt_bp
- symptom_inventory
- catecholamines
- cytokines
- urine_profile
- pregnancy
- med_id
- med_usage
- pharma_action
- blood_eval
- endocrine
- salivary_cortisol
- gynecologic_history
- gynecologic_surgery
- wasi_wrat
- sleep_gen
- sleep_nightly
- sleep_hr_variability
- cantab
- oxygen_saturation

- urine_profile3a
- urine_profile_bacteria
- urine_profile_epithelial_cells
- urine_profile_fix
- urine_profile_mucus
- urine_profile_red_blood_cells
- urine_profile_vol_tcr_microexam
- urine_profile_white_blood_cells
- urine_summary
- urine_summary2
- wasi_wrat2
- zung2
- cytokine
- blood_eval
- cantab
- catecholamines
- endocrine
- gynecologic_history
- gynecologic_surgery
- med_usage
- mfi
- microarray_normalized
- microarray_raw
- microarray_annotation
- oxygen_saturation
- pharma_action
- med_id
- pregnancy
- salivary_cortisol
- sf36
- sleep_gen
- sleep_hr_variability
- sleep_nightly
- symptom_inventory
- tilt_bp
- tilt_info
- urine_profile
- wasi_wrat
- zung
- abtinfo
### Initial Dataset

<table>
<thead>
<tr>
<th>abtid</th>
<th>intake_index</th>
<th>onset_index</th>
<th>prc_class</th>
<th>eric_inde</th>
<th>cluster_ind</th>
<th>years_ill</th>
<th>sex_bin</th>
<th>age</th>
<th>race_ol</th>
</tr>
</thead>
<tbody>
<tr>
<td>10043905</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>2</td>
<td>26.3</td>
<td>0</td>
<td>40</td>
<td>3</td>
</tr>
<tr>
<td>10081101</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>18.2</td>
<td>0</td>
<td>58</td>
<td>6</td>
</tr>
<tr>
<td>10103103</td>
<td>3</td>
<td>2</td>
<td>12</td>
<td>8</td>
<td>3</td>
<td>12.1</td>
<td>0</td>
<td>50</td>
<td>6</td>
</tr>
<tr>
<td>10193601</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>3</td>
<td>8</td>
<td>0</td>
<td>57</td>
<td>6</td>
</tr>
<tr>
<td>10203401</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>9.8</td>
<td>0</td>
<td>54</td>
<td>6</td>
</tr>
<tr>
<td>10215301</td>
<td>5</td>
<td>1</td>
<td>11</td>
<td>13</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>35</td>
<td>6</td>
</tr>
<tr>
<td>10215901</td>
<td>5</td>
<td>1</td>
<td>15</td>
<td>15</td>
<td>1</td>
<td>19.7</td>
<td>0</td>
<td>57</td>
<td>6</td>
</tr>
<tr>
<td>10240402</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>8</td>
<td>3</td>
<td>8.7</td>
<td>1</td>
<td>43</td>
<td>6</td>
</tr>
<tr>
<td>10243501</td>
<td>5</td>
<td>1</td>
<td>11</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>41</td>
<td>6</td>
</tr>
<tr>
<td>10261501</td>
<td>5</td>
<td>1</td>
<td>11</td>
<td>13</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>53</td>
<td>6</td>
</tr>
<tr>
<td>10268605</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>9.3</td>
<td>0</td>
<td>54</td>
<td>6</td>
</tr>
<tr>
<td>10323401</td>
<td>5</td>
<td>1</td>
<td>11</td>
<td>13</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>58</td>
<td>6</td>
</tr>
<tr>
<td>10372601</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>10.7</td>
<td>0</td>
<td>54</td>
<td>6</td>
</tr>
<tr>
<td>10528403</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10.7</td>
<td>0</td>
<td>45</td>
<td>6</td>
</tr>
<tr>
<td>10689003</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>7.6</td>
<td>0</td>
<td>47</td>
<td>6</td>
</tr>
<tr>
<td>10803801</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>3</td>
<td>8.6</td>
<td>1</td>
<td>59</td>
<td>6</td>
</tr>
<tr>
<td>10860201</td>
<td>5</td>
<td>1</td>
<td>11</td>
<td>13</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>39</td>
<td>6</td>
</tr>
<tr>
<td>20052705</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>11</td>
<td>1</td>
<td>12.5</td>
<td>0</td>
<td>43</td>
<td>6</td>
</tr>
<tr>
<td>20077904</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>15</td>
<td>1</td>
<td>31.3</td>
<td>0</td>
<td>37</td>
<td>6</td>
</tr>
<tr>
<td>20082302</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>3</td>
<td>5.3</td>
<td>0</td>
<td>53</td>
<td>6</td>
</tr>
<tr>
<td>20086501</td>
<td>2</td>
<td>2</td>
<td>9</td>
<td>3</td>
<td>1</td>
<td>12.3</td>
<td>0</td>
<td>61</td>
<td>6</td>
</tr>
<tr>
<td>20129103</td>
<td>5</td>
<td>1</td>
<td>15</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>65</td>
<td>3</td>
</tr>
<tr>
<td>20366001</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>8.8</td>
<td>0</td>
<td>48</td>
<td>6</td>
</tr>
<tr>
<td>20416901</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>11</td>
<td>1</td>
<td>14.3</td>
<td>0</td>
<td>62</td>
<td>6</td>
</tr>
</tbody>
</table>
2 Approaches

• PCA/LCA
  – Top Down

• Subspace Clustering
  – Bottoms Up
PCA

- Utilized correlation coefficient based PCA
- Reduces dimensionality while preserving full variation
- Helps to establish a variable set which spans variations in the dataset
## PCA Summary

<table>
<thead>
<tr>
<th>PV1</th>
<th>PV1</th>
<th>PV2</th>
<th>PV2</th>
<th>PV3</th>
<th>PV3</th>
<th>PV3</th>
<th>PV3</th>
<th>PV4</th>
<th>PV4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent</td>
<td>6.475078</td>
<td>Percent</td>
<td>3.396883</td>
<td>Percent</td>
<td>3.225891</td>
<td>Percent</td>
<td>2.950088</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cum Percen</td>
<td>6.475078</td>
<td>Cum Percen</td>
<td>9.871961</td>
<td>Cum Percen</td>
<td>13.09785</td>
<td>Cum Percen</td>
<td>16.04794</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Variables and Loadings

<table>
<thead>
<tr>
<th>Variable</th>
<th>PV1</th>
<th>PV2</th>
<th>PV3</th>
<th>PV4</th>
</tr>
</thead>
<tbody>
<tr>
<td>mfi.generalfat</td>
<td>-0.12066</td>
<td>wasi_wrat2</td>
<td>0.240773</td>
<td>gynecologic</td>
</tr>
<tr>
<td>sf36.vitality</td>
<td>0.123514</td>
<td>wasi_wrat2</td>
<td>0.240769</td>
<td>gynecologic</td>
</tr>
<tr>
<td>sf36.social functioning</td>
<td>-0.10374</td>
<td>wasi_wrat2</td>
<td>0.235238</td>
<td>gynecologic</td>
</tr>
<tr>
<td>abtinfo3.emploi</td>
<td>-0.1406</td>
<td>wasi_wrat2</td>
<td>0.212085</td>
<td>gynecologic</td>
</tr>
<tr>
<td>symptom_in</td>
<td>0.039662</td>
<td>wasi_wrat2</td>
<td>0.21114</td>
<td>gynecologic</td>
</tr>
<tr>
<td>zung2.sdsint</td>
<td>0.012426</td>
<td>wasi_wrat2</td>
<td>0.209433</td>
<td>gynecologic</td>
</tr>
<tr>
<td>symptom_in</td>
<td>-0.02687</td>
<td>wasi_wrat2</td>
<td>0.209421</td>
<td>gynecologic</td>
</tr>
<tr>
<td>mfi.physicalfunction</td>
<td>0.022401</td>
<td>wasi_wrat2</td>
<td>0.208413</td>
<td>gynecologic</td>
</tr>
<tr>
<td>symptom_in</td>
<td>0.002692</td>
<td>wasi_wrat2</td>
<td>0.205281</td>
<td>pharma_clas</td>
</tr>
<tr>
<td>sf36.bodilypain</td>
<td>-0.01031</td>
<td>wasi_wrat2</td>
<td>0.197316</td>
<td>gynecologic</td>
</tr>
<tr>
<td>sf36.rolephys</td>
<td>0.015727</td>
<td>wasi_wrat2</td>
<td>0.196524</td>
<td>gynecologic</td>
</tr>
<tr>
<td>mfi.activityrate</td>
<td>0.006969</td>
<td>wasi_wrat2</td>
<td>0.194707</td>
<td>pharma_clas</td>
</tr>
<tr>
<td>sf36.gnrlhlth</td>
<td>0.005082</td>
<td>wasi_wrat2</td>
<td>0.19301</td>
<td>gynecologic</td>
</tr>
<tr>
<td>symptom_in</td>
<td>-0.00025</td>
<td>wasi_wrat2</td>
<td>0.185406</td>
<td>gynecologic</td>
</tr>
<tr>
<td>mfi.motivation</td>
<td>-0.01941</td>
<td>wasi_wrat2</td>
<td>0.184849</td>
<td>gynecologic</td>
</tr>
<tr>
<td>mfi.mentalfatigue</td>
<td>0.001438</td>
<td>wasi_wrat2</td>
<td>0.171769</td>
<td>gynecologic</td>
</tr>
<tr>
<td>abtinfo3.onset</td>
<td>0.018855</td>
<td>wasi_wrat2</td>
<td>0.161124</td>
<td>gynecologic</td>
</tr>
<tr>
<td>sf36.physfunc</td>
<td>-0.01398</td>
<td>sleep_hr_var</td>
<td>-0.09508</td>
<td>gynecologic</td>
</tr>
<tr>
<td>oxygen_saturation2.o2lowp</td>
<td>0.016793</td>
<td>oxygen_saturation2.o2avp</td>
<td>-0.1184</td>
<td>oxygen_saturation2.o2avp</td>
</tr>
<tr>
<td>Dataset Name</td>
<td>Notes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>----------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all_data_outer</td>
<td>all data - with null data elements (outer join)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all_data_inner</td>
<td>all data - no null data elements (inner join)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bio_data_outer</td>
<td>contains only the non-survey (biological) data, remove sf36, mfi, gynecologic_history, symptom_inventory</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>select_data_outer</td>
<td>remove blood_eval2 cols, cantab cols, endocrine2A cols, gynacoloic_history3 cols, mfi cols, sf36 cols, sleep_hrバリエーション性質cols, symptom_inventorycols, tilt_bp cols, tilt_info3 cols, uring_profile9A cols, wasi_wrat2 cols</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>select2_data_outer</td>
<td>remove pharma_class_usage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>select3_data_outer</td>
<td>remove sleep related variables</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>select5_data_outer</td>
<td>remove redundant measures such as 3 measures of testo</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>select9_data_outer</td>
<td>remove waist_circum, cytokine cols, some endocrine cols (androstenedione), mfi cols, sf36 cols, sleep_nightly2 cols</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>select10_data_outer</td>
<td>remove med/psych exclusions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>select11_data_outer</td>
<td>remove men</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>select12_data_outer</td>
<td>remove DHEA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>select13_data_outer</td>
<td>remove gabdpn, years_ill_index, potassium. Wbc, rti_five_choice_reaction_time_index, soc_num_problems_solved_in_min_num_moves_index, swm_num_total_errors_index, metanephrine, epinephrine, thyroxin_t4_index, lm_index, specific_gravity</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Class attributes in order of significance to the solution

<table>
<thead>
<tr>
<th>Class attributes</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Class 5</th>
<th>Class 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>% unrefreshing_sleep</td>
<td>98</td>
<td>13</td>
<td>79</td>
<td>78</td>
<td>95</td>
<td>100</td>
</tr>
<tr>
<td>high_sensitivity_crp</td>
<td>5.3</td>
<td>4.2</td>
<td>2.1</td>
<td>0.9</td>
<td>1</td>
<td>5.1</td>
</tr>
<tr>
<td>% sleep_problems</td>
<td>100</td>
<td>29</td>
<td>75</td>
<td>96</td>
<td>91</td>
<td>91</td>
</tr>
<tr>
<td>O₂lowp</td>
<td>85</td>
<td>88</td>
<td>86</td>
<td>90</td>
<td>89</td>
<td>89</td>
</tr>
<tr>
<td>%sore_throat</td>
<td>29</td>
<td>8</td>
<td>17</td>
<td>13</td>
<td>59</td>
<td>73</td>
</tr>
<tr>
<td>% photophobia</td>
<td>56</td>
<td>11</td>
<td>38</td>
<td>22</td>
<td>64</td>
<td>73</td>
</tr>
<tr>
<td>% abdominal_pain</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>35</td>
<td>68</td>
<td>45</td>
</tr>
<tr>
<td>% concentration</td>
<td>27</td>
<td>0</td>
<td>21</td>
<td>4</td>
<td>86</td>
<td>45</td>
</tr>
<tr>
<td>% joint_pain</td>
<td>63</td>
<td>32</td>
<td>54</td>
<td>30</td>
<td>91</td>
<td>82</td>
</tr>
<tr>
<td>rdi_index</td>
<td>3.8</td>
<td>1</td>
<td>6</td>
<td>0.4</td>
<td>0.8</td>
<td>1.1</td>
</tr>
<tr>
<td>insulin_serum</td>
<td>10</td>
<td>7</td>
<td>10</td>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>epworth</td>
<td>11</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>% short_breath</td>
<td>32</td>
<td>5</td>
<td>8</td>
<td>4</td>
<td>50</td>
<td>45</td>
</tr>
<tr>
<td>progesterone</td>
<td>14</td>
<td>12</td>
<td>17</td>
<td>35</td>
<td>37</td>
<td>0</td>
</tr>
<tr>
<td>% post_exertional_fatigue</td>
<td>78</td>
<td>0</td>
<td>33</td>
<td>43</td>
<td>82</td>
<td>91</td>
</tr>
<tr>
<td>% fever</td>
<td>10</td>
<td>5</td>
<td>4</td>
<td>13</td>
<td>50</td>
<td>18</td>
</tr>
<tr>
<td>% muscle_pain</td>
<td>93</td>
<td>47</td>
<td>54</td>
<td>70</td>
<td>100</td>
<td>91</td>
</tr>
<tr>
<td>BMI</td>
<td>32</td>
<td>30</td>
<td>30</td>
<td>24</td>
<td>27</td>
<td>26</td>
</tr>
<tr>
<td>free_t3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>sleep_period_time_spt</td>
<td>471</td>
<td>452</td>
<td>467</td>
<td>456</td>
<td>448</td>
<td>439</td>
</tr>
<tr>
<td>il_6</td>
<td>68</td>
<td>56</td>
<td>45</td>
<td>32</td>
<td>66</td>
<td>50</td>
</tr>
<tr>
<td>arousals_total_index_psg</td>
<td>18.5</td>
<td>11.2</td>
<td>13.4</td>
<td>13.4</td>
<td>9.9</td>
<td>17.1</td>
</tr>
<tr>
<td>neck_circum</td>
<td>37</td>
<td>35</td>
<td>35</td>
<td>33</td>
<td>35</td>
<td>33</td>
</tr>
<tr>
<td>hgb</td>
<td>14.2</td>
<td>13.8</td>
<td>13.6</td>
<td>13.5</td>
<td>13.6</td>
<td>12.8</td>
</tr>
<tr>
<td>sleep_HR variability(sdn)</td>
<td>60</td>
<td>62</td>
<td>38</td>
<td>50</td>
<td>54</td>
<td>37</td>
</tr>
<tr>
<td>pct_free_testo_ratio</td>
<td>2.6</td>
<td>1.6</td>
<td>2</td>
<td>2.5</td>
<td>2.6</td>
<td>0.8</td>
</tr>
<tr>
<td>% headache</td>
<td>66</td>
<td>42</td>
<td>50</td>
<td>48</td>
<td>91</td>
<td>64</td>
</tr>
<tr>
<td>latency_to_sleep_onset</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>5</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>urine_free_cortisol_24h</td>
<td>17</td>
<td>19</td>
<td>10</td>
<td>21</td>
<td>20</td>
<td>14</td>
</tr>
<tr>
<td>ZUNG</td>
<td>50</td>
<td>35</td>
<td>51</td>
<td>48</td>
<td>54</td>
<td>55</td>
</tr>
</tbody>
</table>

**Class 1:** OBESE, HYPNOEIC, FATIGUED & PAINED

**Class 2:** WELL

**Class 3:** HYPNOEIC, OBESE, STRESSED - LOW CORTISOL

**Class 4:** YOUNG, normal BMI, fewer symptoms besides SLEEP

**Class 5:** INTEROCEPTIVE, depressed & symptomatic

**Class 6:** SYMPTOMATIC, OLDer, THINer, DEPRESSED, STRESSED, poor sleep, NO SEX,
CFS is Heterogeneous

Depressed, menopausal, stressed, low cort, low HRV

Depressed, interoception

Interoception

Obese, hypopnea

Well, obese

Obese, hypopnea

Stressed, low cort, low HRV
The Genomics of CFS

- Depressed, menopausal, stressed, low cort, low HRV
  MAOA and TPH 2

- Interoception

- Obese, hypopnea, Stressed, low cortisol, low HRV
  MAOA and B, TPH2

- Obese, hypopnea
  POMC, GR

- Well, obese
Gene Expression Differentiates Fatigue

1. **Vac14**: regulates phosphatidylinositol kinases (stress response and membrane trafficking)
2. **SLC1A6**: an excitatory aa transporter (glutamate/aspartate)
3. **Fbxo7**: Fbxo7 has been characterised as a selective enhancer of cdk6 activity (regulate major cell cycle transitions)
4. **ZNF350**: crucial roles in ubiquitination events involved in diverse cellular processes including signal transduction (MAPK), differentiation and apoptosis

1. **PTCH2**: receptor for shh signaling which is active in T cell growth and differentiation and proliferation
2. **TCL1A**: TCL1A regulates the growth and survival of peripheral T cells
CFS Heterogeneity

• Problems with PCA/LCA approach.
  – Are we studying illness end-stage symptomatology rather than more helpful subgroups?
• More helpful subgroups may answer these Q’s related to CFS heterogeneity…
  – Since we observe differing insults (viral or infection or stress, … or a combination of the these) preceding the CFS state, is there a difference in what’s going wrong ‘under the hood’ physiologically, in people with CFS?
  – Since there are natural genetic variations between people, does this lead to variations in susceptibility to the disease from these insults?
  – What is an exhaustive list of disturbed biological pathways in chronic fatigued persons – does this list vary by person?
  – What are the biological sequence of events (in terms of biological pathway disruptions) leading to a CFS state, and does this sequence of events vary from person to person?
  – Are there common themes in the answer to the above questions? Can we define subgroups of people with CFS from these common themes?
• Can we use these subgroups to develop optimal, targeted treatments - custom tailored to address each subgroups specific characteristic problems?
Clustering

• Clustering algorithms find associations between sets of variables and subjects by identifying regions or clusters of closely spaced values.
Why Clustering?

• Differs from variance based calculations
• Can elucidates trends obscured through heterogeneous data

Why Subspace Clustering?

Relative high dimensionality of dataset (3686 variables) compared to the number of participants (227) favors subspace clustering algorithms
Subspace Clustering

- Subspace clustering considers all combinations of subsets of the original space.
- We utilize portions of IBM’s Genes@Work tool (clique) which identifies clusters under a user specified p-value.
- The algorithm identifies maximally sized clusters.

Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications, Agrawal et. al. (1998)
Genes@Work: an efficient algorithm for pattern discovery and multivariate feature selection in gene expression data, Lepre, J. et. al. Bioinformatics 2004
Example

No Cluster

Variable A

Variable B

Variable C

Cluster

Variable A

Variable B

Variable C
Subspace Clustering Interpretation

- We utilize VENN diagrams to visualize the information in the clusters:
Results

• Stringent selection of significant clusters
  – 183 of 1,042 (18%) significant clusters with (p-value = 0.005) were free of null values and analyzed further
  – Three major “themes” were identified
    • Cognition (29 clusters)
    • Sleep (31 clusters)
    • Allostatic load (7 clusters)
The CANTAB was used to assess function in the basal ganglia (especially psychomotor retardation) and/or frontal cortex (attention shift and problem solving/planning abilities).
Sleep Polysomnographic monitoring and multiple sleep latency tests were used to diagnose sleep pathology.

35% CFS, 60% ISF

35% CFS, 60% ISF

15 REM and total IL1B

28 Respiratory mechanics Paradoxical breathing DBH and COMT

17 Sleep latency TH

19 REM latency Stage 1 sleep

19 REM latency Stage 1 sleep

58% CFS, 33% ISF

70% CFS, 29% ISF

35% CFS, 52% ISF
Allostatic Load

1. Aldosterone and DHEA (NR3C1)
   - 35% CFS, 44% ISF, 14% NFMDDm, NFMed
   - 57% CFS, 36% ISF

2. Aldosterone and Hemoglobin
   - 19% CFS, 62% ISF, 13% NFMEd
   - 45% CFS, 29% ISF, 22% NFMDDm, NFMed

3. Aldosterone and psych
   - 53
   - 13%

4. Cortisol and hysterectomy
   - 19
   - 48% CFS, 48% ISF

5. DBH
   - 15
   - 33
   - 48% CFS, 58% ISF

6. BMI and CRP
   - 19
   - 33
   - 48% CFS, 48% ISF

7. Empiric and fatigue
   - 7
   - 100% NFMDDm, NFMed

8. Aldosterone and psych
   - 19
   - 32% CFS, 58% ISF

9. Aldosterone and DHEA
   - 19
   - 48% CFS, 48% ISF

10. HTR7
    - 13
    - 29% ISF, 22% NFMDDm, NFMed
Important Clusters

- Allostatic load and aldosterone
  - 48% CFS, 48% ISF

- Aldosterone, DHEA and HTR7
  - 52% CFS, 26% ISF

- Cognition and HTR2A
  - 40% CFS, 47% ISF

- Fatigue and NR3C1 (n=7)
  - 100% NFMed, NFMDm

- Sleep and DBH (n=14)
  - 45% CFS, 29% ISF, 22% NFMed, NFMDm

- 18 Cognition NR3C1
  - 27% CFS, 60% ISF

- 13 Aldosterone
  - 45% CFS, 29% ISF, 22% NFMed, NFMDm

- 19 Cognition DBH
  - 31% CFS, 46% ISF, 16% NFMed

- 33

- 17

- 19
Summary

• Heterogeneity at a biological level has been demonstrated through PCA/LCA and subspace clustering techniques
• Subspace clustering technique validated existing results and revealed hidden associations
• Predominant themes are cognition, sleep and allostatic load
Heterogeneity Revealed