# Using Gene Sets To Analyze Genomic Compression

Eric You Mentor: Dr. Gil Alterovitz 7th Annual Primes Conference May 21 2017

| 21111 ganged contrained graphing and targenergy part of the constrained graphing and the constrained gr   |       | 40421551    | 40421561  | 4042       | 1571   | 40421581  | 48421         | 591 4042  | 21601 4  | 84216  | 11 404   | 21621 4   | 0421631   | 4842164 | 4842    | 1651 4042  | 1661 4042   | 671 4   | 8421681  | 404216    | 91 4042176  | 404217     | 11 40421                                       |
|--|-------|-------------|-----------|------------|--------|-----------|---------------|-----------|----------|--------|----------|-----------|-----------|---------|---------|------------|-------------|---------|----------|-----------|-------------|------------|--|
| L  | 721tt | tgagcagacct | atataagat | ggttat     | gaagat | tcacacago | cggctca       | tgcctgtga | atcccage | acttt  | gggaggc  | tgaggcaa  | gtggagcac | ctgagat | catgagt | tcaagaccag | cctggccaaca | tggtga  | aaccccat | tctctacta | aaagatacaaa | aattatcca  | ggtgtggtg                                      |
| A generging of a far ange get a generging of a generging a gene   |       | A           |           |            |        |           | · · · · · · · |           |          |        |          |           |           |         |         |            |             |         | ******   |           |             |            |  |
| etcl         copiect all allanging getting and the copiect of get construction of get constructions of get construntions of get construntions of get constructions of  | A     | gaacagacc   | atataagat | 99         | gaagat | cacacag   | ggctca        | tgcctgtga | atcccage | ac t   | 99939990 | tgagtcaa  | gtggagcac | ctgagat | ca gag  | ACCAG      | CCTGGCCAAC/ | GGTGA   | AACCCCA  | CICIACI   | AAA ATACAAA | MATTATCCA  | GG G GG G                                      |
| Bit  | aca   | cagacci     | atataagat | ggtt       | aagat  | acacacag  | ggclica       | tgcctgtg  | atcccage | actt   | GGGAGGC  | GAGGCAA   | GTGGAGCAC | C GAGA  | CA GAG  | C cag      | cctggccaaca | 19919a  | aacccca  | c c acta  | aaaga ACAA  | AATTATCCA  | GG G GG G                                      |
| ALT IT GULGE         ALT ALAG         ALT ALAG         GULGE CLUCK         CONSCREDUNCT         GULGE CLUCK         CONSCREDUNCT         GULGE CLUCK   | acatt | GACC        | A A AAGA  | GGTTAT     | GAAGA  | CACACAG   | GGC           | CC G G    | CCCAG    | ACTU   | GEGNEEC  | GAGGCAA   | GTGGAG AC | C GAGA  | CA GAG  | TCAAGACCAG | CCTGGACAAC  | TGG     | AACCCCA  | CCACI     | AAAGATACAA  | MATTA CCA  | GGTGT G  |
| LATT GMAGACCA A A MAGA GT A GAVAA TI CALCAG GGC CA A CCCCACACT GACCAGAC TO GACACCACT CALCAGAC CALCAGACAC CALCAGAC CALCAGAC CALCAGAC CALCAGAC CALCAGAC CALCAGAC CALCAG   | ACATH | GAACAG      | A A AAGA  | GGITAT     | GAAGA  | CACACAG   | GGC CA        | TGCC tga  | atcccage | acttt  | 999899   | TGAGGCAA  | GTGGAGCAC | C GAGA  | CA GAG  | TCAAGACCA  | GCCAAC      | GG GA   | AACCCCA  | CCACI     | AAAGATACAA  | A ATCCA    | GG G GG G                                      |
| Heist generation         Beild generation         File Charlen Charlen Control         Generation         File Charlen Charlen Control         Generation         File Charlen Charlen Charlen Control         Generation         File Charlen   | ACATH | GAACAGAC    | TCAGA     | GGITA      | GAAGAT | CACACAG   | GGC CA        | GCC G     | ATCCCAG  | ACTO   | GGGAGGC  | GAGGCAA   | GGGGAGCAC | CG      | AGAG    | TCAAGACCAG | CCTGGCCAAC  | GG GA   | AACCCCA  | CTCTACT   | AAAGATACAA  | WATTA a    | <u>56 5 6 5 6 5 6 6 6 6 6 6 6 6 6 6 6 6 6 </u> |
| Bit Mind State         Bit Min   | acatt | gaacagccc   | ata aagar | 99 a.      | gaagat | cacacag   | ggc ca        | tgcc g g  | CCCAG    | ACTIN  | GGGAGCC  | GAGGCAA   | GTGGAGCAC | C GA    | AIGAG   | CAAGACCAG  | CC GGCCAACA | GG GA   | AACCCCA  | TATACT    | AAAGAT CAAA | AATTATCCA  | GG G GG G                                      |
| ACV       gale ta as again againaginaginaginaginagin centricing again (************************************  | acatt | gaacagacc   | atata gal | ggilai     | gaagat | cacacag   | agetca        | tgcctgtg  | AG       | ACTU   | GGGAGGC  | TGAGGCAA  | GGGGAGCAC | GGA     | GAG     | TCAAGACCAG | CC GGCCAAC/ | GG GA   | AACCCCA  | C CTACT   | AAAGATACAA  | MATTA CCA  | GGIGIGGIG                                      |
| RCATT GAUCADECT A.A. AM. GETTA ELVARITORIZAGIO GET C. E.G. GALCE C. E.C. E. Bage E.B. C. BARCELACC GETALLA GE GALACE C. C. C. A. AMAR ALCAMANT TACCELAGE GETALLA GE GALCE GETALLA GENERAL C. BARCELACC GETALLA GENERAL C. C. C. C. AMAR ALCAMANT TACCELAGE GETALLA GENERAL C. C. C. C. C. AMARCELACC GETALLA GENERAL C. C. C. C. C. AMARCELACC C. C. C. C. ALCAMAR TACCELAGE GETALLA GENERAL C. C. C. C. C. AMARCELACC GETALLA GENERAL C. C. C. C. C. AMARCELACC GETALLA GENERAL C. C. C. C. C. AMARCELACC C. C. C. C. ALCAMAR TACCELAGE GETALLA GENERAL C. C. C. C. C. AMARCELACC C. C. C. C. ALCAMAR TACCELAGE GETALLA GENERAL C. C. C. C. C. AMARCELACC C. C. C. C. ALCAMAR TACCELAGE GETALLA GENERAL C. C. C. C. C. AMARCELACC C. C. C. C. ALCAMAR TACCELAGE GETALLA GENERAL C. C. C. C. C. AMARCELACC C. C. C. C. ALCAMAR TALLA C. C. G. G. C. C. C. C. C. AMARCELACC C. C. C. C. ALCAMAR TALLA C. C. G. G. C. C. C. C. C. C. MARCELACC C. C. C. C. ALCAMAR TALLA C. C. G. G. C. C. C. C. C. C. MARCELACC C. C. C. C. ALCAMAR TALLA C. C. G. G. G. G. G. C. C. C. C. C. C. C. MARCELACC C. C. C. C. C. AMARCELACC C. C. C. C. C. C. MARCELACC C. G. C. C. C. C. C. MARCELACC MARCELACC MARCELACC C. C. C. MARCELACC MARCELACC C.   | ACAT  | gacc        | atataagat | 99 tat     | gaaga  | cacacag   | ggclc         | CTGTA     | ATCCCATO | ACTI   | GGGAGGC  | TGAGGCAA  | GIGGAGC C | CIGAGA  | CA GAG  | CAAGA AG   | CC GGCCAAC/ | CGTGA   | AACCCCA  | ACACI     | AAAGAT caaa | aattatcca  | ggtgtggtg                                      |
| NCH TG         ACC A TA ANGH TG         A GUART TO CARLENDER OF CARL  | ACATT | GAACAGACC   | ATATAA    | GGILAT     | GAAGA  | CACACAG   | GGC CA        | GCC G G   | ATCC (   | actt   | 999a gc  | gaggcaa   | giggagcac | c gaga  | cat     | CAAGACCAG  | CC GGCCAAC/ | GG GA   | AACCCCA  | CCAC      | AGAAATACAA  | MATTATCCA  | G G GG G                                       |
| NATT GAACABACC 1 AT AAAA GO TACEAACA (CALCAGE) GO TACE IS A LOCK OF CALL CC CALCAGE OF CA   | ACATT | G ACC       | A A AAGA  | GGA        | GAAGA  | CACACAG   | GGC CA        | GG        | ATCCCAG  | AC 111 | GGGAGGC  | TGAGGCAA  | GIGGAGCA  | C GAGA  | CACGAG  | CAAGACCAG  | CC GCCCAAC/ | GGC     | AACCCCA  | CCACI     | AAAGATACAAA | AAT ACCCA  | GG G GG  |
| BERK Spaces  | ACAT  | GAACAGACC   | ATATAAGA  | GGILAC     | GAAGA  | CACACAG   | GGC CA        | GCC G G   |          | acall  | 999-99C  | gaggcaa   | grggagcac | c gaga  | ca      | AAGACCAG   | CC GGCCAAC/ | GG GA   | AACCCCA  | CCAC      | AAGATACAA   | MATTATCCA  | GG G GG GG                                     |
| BCAT: GRANCHARCE: A 13 ABBA 593         BCAT: GRANCHARCE: A 13 ABBA 503         BCAT: GRANCHARCE: A 13 ABBA 503         BCAT: GRANCHARCE: A 13 ABBA 503         BCAT: GRANCHARCE: A 13 ABBA 503        BCAT: GRANCHARC   | acatt | gaacagacc   | alalaaga  | Ltat       | gaaga  | cacacag   | ggc ca        | gcc g g   | atcccag  | CIT    | GGGAGGC  | GAGGCAA   | GTGGAGCAC | GAGA    | CA GA   | 29         | cclggccaaca | .99 .90 | aacccca  | C C aC    | aagat A     | MATTATCCA  | GG G GG GG                                     |
| BCAT: GALAGANCC, AL ALAMA, GGI YA, BOG YA, GUE, C. Y. C. MARA, M. C. T. ALCAR, M. C. T. C. C. G. C. G. C. T. C. C. C. G. G. C. T. C. C. C. G. G. C. C. C. T. C. C. C. G. G. C. C. C. T. C. C. C. G. G. C. C. C. T. C. C. C. G. G. C.   | aca   | gaacagacc   | atataaga  | 99         | aagar  | cacacag   | ggc ca        | gccag g   | a ceeage | acti   | GGGAGGC  | IGAGGCAA  | GIGGAGCAC | GAGA    | AA GAG  | IC G       | CC GGCCAACA | GG GA   | AA CCCA  | C C AC    | MAGATACAA   | MATTA CCA  | GG G GG GG                                     |
| BCATT GAMAGANGC A A A MAGA GGT A'       CAGA GGT CAGA GGT A'       CAGA   | ACA   | GAACAGACC   | A A AAGA  | GGTA       | aga    | cacacaga  | igge ca       | dcc d d   |          | actu   | AGGC     | TGAGGCAA  | GIGGAGCAC | GAGA    | CA GAG  | TCAAG (    | CCTGGCCAAC  | GG GA   | AACCCCA  | CCACI     | AAAGATAC    | TATCCA     | GG G GG GG                                     |
| Rein Tewardenics An Anden Gen A Guide<br>Text Text Text Text Anden Anden A Guide<br>Text Text Text Text Anden Anden A Guide<br>Text Text Text Text Anden Anden A Guide<br>Text Text Text Anden Anden Anden A Guide<br>Text T   | acatt | gaacagacc   | atataaga  | 9 <b>9</b> |        | CACACAG   | GGC CA        | GCC G G   | ATCCCAG  | ACCIT  | GGG GC   | GAGGCAA   | GEGAGCAC  | GAGA    | CA GAG  | CAAGAC     | CCAAC       | GG GA   | AACCCCA  | CICIACI   | AAAGATACAA  | WA atcca   | <u>aa a</u> aa ac                              |
| Incl. Tegewidelic () A.A.Men (Gr. )         ALCA Ge (EWACCCCA )         ALCA Ge (EWACCCCA )         ALCA Ge (EWACCCCA )         ALCA (Ge (EWACCCA )         ALCA (G   | ACAT  | GAACAGACC   | ATA AAGA  | GG         |        | CAG       | GGC CA        | GCC G G   |          | ACIT   | GGGAGGC  | GAGGCAA   | GGAGCAC   | GAGA    | CA G    |            | CAAC        | GG GA   | AACCCCA  | CCACI     | AAAGATACAA  | WAT TCCA   | GG G GG GG                                     |
| NCH TIGMUCHARCH A A ADACK GO TA RUMART       C C C C C C C C C C C C C C C C C C C   | ACATT | GAACAGACC   | A A AAGA  | GG         | GAAG   | CAG       | GGC CA        | GCC G G   |          | ACTIC  | GGGAGG   | IGAGGCAA  | GEGAGCAC  | GAGA    | CAG     |            | AACI        | GG GA   | AACCCCA  | CICIACI   | NAAGATACAG  | AATT a     | 99 9 99 96 G                                   |
| Inclumination         Get Al (SALA)         CCCC (C) (C) (C) (C) (C) (C) (C) (C) (C)   | ACAT  | GAACAGACC   | ATA AAGA  | GGTAT      | GAAGA  | CAG       | GGCTCA        | GCC G G   | ATCC     | CC C   | GGGAGGC  | IGAGGCAA  | GGAGCAC   | GAGA    | CAG     |            | AC          | GG GA   | AACCCCA  | CCACI     | AAAGATACAAA | AATTA      | G G GG GG                                      |
| LEAT IDAUCADEL I AL ALADA GOLLA GOLLA I LOGANGEL CALLE I GAN LOL COLLA. I DAGANGEL CALLE GAN LOL GOLL CALLE GAN LOL GOLLA I LOGANGEL CALLE GANGEL GANGEL GANGEL GAN LOL GAN LOL GAN LOL GANAL I LOC GOLLA I LOGANGEL GANGEL I LOGANGEL GANGEL GANGEL GANGEL GANGEL GANGEL CALLE GAN LOL GAN LOL GANAL I LOC GAN LOL GAN LOL GAN LOL GAN LOL GANAL I LOC GAN LOL GAN LOL GAN LOL GAN LOL GANAL I LOC GANGEL GANGEL GANGEL GANGEL GANGEL GANGEL GANGEL CALLE GANAL I LOC GANGEL GANGEL GANGEL GANGEL GANGEL GANGEL GANGEL CALLE GANAL I LOC GANGEL  | ACATT | GAACAGACC   | ACTAAGA   | GG         | GAAGA  | GC        | GGCTCA        | GCC G     | ATC      | CIT    | GGGAGGC  | GAGGCAA   | GTGGAGCAC | GAGA    | CAGA    |            | ACA         | GG GA   | AACCCCA  | CTATACT/  | AAAGATACAA  | AATTA      | G GG GG  |
| ACA T TGAUCAGUC TA A AMAGA GGA TA GAUGATTA COLOR TO FOR TO CONSIDER TO BUT CONSIDER.           ALCART TO ACCART TO   | ACA   | GAACAGACC   | A A AAGA  | GG         | GAAGA  | 5         | CC            | GCC G G   | ATCCCAG  | AC     | GGGAGGC  | TGACGCAA  | GGAGCAC   | GAGA    | CA GAG  | CAAGACCAG  | CC GGCCA    | GGIGA   | AACCCCA  | CCACI     | AAAGATACAA  | MATTATEC   | 99 90  |
| ALCA TEMACAGUE (A A MADA TGOT) TA GARGA CA CARCAG GARGALLE GARGALLE GARGALLE (A CARGA) CANARALE CAR SIG GARGALLE (C TAC ANARA) A CANARALT A TCCAGG G G G CARCELLE (C TAC ANARA) A CANARALT A TCCAGG G G G CARCELLE (C TAC ANARA) A CANARALT A TCCAGG G G G CARCELLE (C TAC ANARA) A CANARALT A TCCAGG G G G CARCELLE (C TAC ANARA) A CANARALT A TCCAGG G G G CARCELLE (C TAC ANARA) A CANARALT A TCCAGG G G G CARCELLE (C G G C CARCELLE) (C G G G G G G C CARCELLE) (C G G G G G G C CARCELLE) (C G G G G G G C CARCELLE) (C G G G G G G C CARCELLE) (C G G G G G C CARCELLE) (C G C CARCELLE) (C G G G G G G C CARCELLE) (C G C CARCELLE) (C G C CARCELLE) (C C C CARCELLE) (C C C CARCELLE) (C C C C CARCELLE) (C C C C C C C C C C C C C C C C C C C   | ACATT | GAACAGACC   | A A AAGA  | GG         | GAAGA  | E.        | CTCA          | GCC G G   | ATCCCAG  | ACTU   | GGGAGGC  | IGAGGCAA  | GGAGCAC   | GAGA    | CA GAG  | CAAGACCAG  | CCTGGCCA    | GGIGA   | AACCCCCA | CGCIACI   | AAAGATACAA  | MATTATECA  | Ge   |
| AKA GETA GAAAT TOCACAG GGC CA TOCC GA       CONSCICT GAAT CAAAGA TACAAAGA TACAAAAATTA       GGTA GAAAGA TOCACAG GGC CA TOCC GAAGA CCC GAAGA CAC GAAGA CACAAAATTA       GGTA GAAGA TOCACAG GGC CA TOCC GAAGA CAC GAAGA CACAAAATTA       GGTA GAAGA TOCACAG GGC CA TOCC GAAGA CAC GAAGA CACAAAATTA       GGTA GAAGA TOCACAG GGC CA TOCC GAAGA CAC GAAGA CACAAAATTA       GGTA GAAGA TOCACAG GGC CA TOCC GAAGA CAC GAGA CA TOGA GAAGA CACAAAAATTA       GGTA GAAGA TOCACAG GGC CA TOCC GAAGA CAC GAGA CA TOGA GAAGA CACAAAATTA       GGTA GAAGA TOCACAG GGC CA TOCC GAAGA CAC GGC GAGGA CA TOGA GAAGA CACAAAATTA       GGTA GAAGA TOCACAG GGC CA TOCC GAACAA TOCAGAG CAC GAAGA CA TOGA GAAGA CACAAAATTA       GGTA GAAGA TOCACAG GGC CACGC GAAGA CA TOGA GAAGA CACAAAATTA       GGTA GAAGA CACAAGA GAACAAATTA       TOCAGAGAGC GAAGACCC GAAGA CA TOGA GAACACCC GAACAAATTA       GGTA GAAGA CACAAGA GAACAAATTA       TOCAGAGAGC GAAGACCC GAAGA CA TOGA GAACACCC GAACAAATTA       GGTA GAAGA CACAAAATTA       TOCAGAGAGC GAAGACCC GAAGA CA TOGA GAACACCC GAACAAATTA       GGTA GAAGA CACAAAATTA       GGAAGCCCC TOC TOCAC AAAGA AAAAATTA       TOCAGAGGAGC GAAGACCC GAACACC GAACACC GAACACCC TOCACAAAATTA       TOCAGAGGAGC GAAGACCC GAACACC GAACACCC GAACACCC TOCACAAGA AAAAATTA       TOCAGAGGAGCACC GAAGACCC GAACACC GAAGA CACAAAATTA       TOCAGGAGCACC GAAGACCC GAAGA CACGAAAATTA       TOCAGGAGCACC GAAGACCC GAAGA CAC GAACACCC TOCACAC TOCACAC CACAGA CACAAAATTA       TOCAGG GAAGCCCC GAAGA CAC GAAGCCCC GAACACCC TOCACAC CACAC COCAC COCACAC COCAC COCACAC COCAC CO  | ACATT | GAACAGACC   | A A AAGA  | GG         | GAAGA  | CA        |               | GG        | ATCCCAG  | AC     | GGGAGGC  | TGAGGCAA  | GIGGAGCAC | GA      | CA GAG  | CAAGACCCG  | CCTGGCCAAC  | GG GA   | AAC CCa  | C C aC    | aagatacaa   | aalacca    | 99 9 99 90                                     |
| ALC GGT A GUNGATT CALCAGE CG CG AT CCC       CTC CGGGAGCC CGGGCCG CG GGCC CGGCC GGGCCC CGGCCGGCCCGGGCCCGGCCGGCCGGCCGGCCGGCCGCGC  |       |             | AGAT      | GG         | GAAGA  | CACACAG   | GGCTCA        | GCC G G   | CCAG     | ACTI   | GGGAGGC  | TGAGGCAA  | GTGGAGTAC | GAGA    | GAG     | CAAGACCAG  | CC GGCCAACA | GG GA   | AACCCCA  | C AC I    | AAAGATACAA  | AATTATECA  | 66 6 66 6                                      |
| Gold National Calculation used Calculation of the Sector (2002) (2012  |       |             | ACAT      | GGTA       | GAAGAT | CACACAG   | GGCTCA        | GCC G G   | Tece     | -      | GGGAGGC  | IGAGGCAA  | GIGGAGCAC | GAGA    | CA GA   |            | U           | GG G    | AACCCCA  | CCACI     | AAAGATACAA  | MATTAT     | 90   |
| A I GAMAA TI CACACAG GOC L'A I GOC IGA GOC LCC I GAGCACCI GAGACAL CGAGA CAL GAGCAC I GAGCACL CGAGACACI GAGCACCI GAGACACI CGAGCACCI GAGCACI CGAGCACCI GAGCACCI GAGCACI CGAGCACCI GAGCACI CGAGCACCI GAGCACCI GAGACCACGAGCAGCI GAGCACCI GAGACCCAGCI GAGCACCI GAGACCACGAGCI GAGCACCI GAGACCCAGCI GAGCACCI GAGACCACGAGCAGCI GAGCACCI GAGACCACGAGCI GAGCACCI GAGACCACGACGI GAGCACCI GAGACCACGAGCI GAGCACCI GAGACCACGAGCI GAGACCCCGI GAGACCACGACGI GAGCACCI GAGACCACGAGCI GAGACCCAGCI GAGACCACGACGI GAGCACCI GAGACCACGAGCI GAGACCACGACGI GAGCACCI GAGACCACGAGCI GAGACCACGACGI GAGCACCI GAGACCACGAGCI GAGACCCCGACI CC CACCI CC CACCI CC CACAAAAAT CACCAAGGI GAGCAGCAGCI GAGCACCI GAGACCAGCAGCI GGCC   |       |             |           | GG         | GAAGA  | CACACAG   | GGC CA        | GCC G G   | ALCCC    | CIC    | 6664666  | GAGGCAA   | G G agcac | c gaga  | ca gag  | caagaccag  | cc g**caaca | ga      | aacccca  | C C AC    | aagatacaa   | aattatcca  | 99 6   |
| A GMDAT COLOR GOL CAT GUE WIND CONDUCT GAGAT CATAGO       CATAGO GAMACCCCAT CT CT ALAGAA ALAAAAAT A T COLAGG G G GOL         g at coccago t att ggagg gg cago gg gg cago cag       CATAGO GAAACCCCAT CT CT ALAGAA ALAAAAAT A T COLAGG G G G G G         a t conduct to double to moto cago t att gg gg gg cago cag       CATAGO GAAACCCAT CT CT ALAGAA ALAAAAAT A T COLAGG G G G G G         a t conduct to double to moto cago cago cag gg cago cag   |       |             |           |            | GAAGA  | CACACAG   | GGCTCA        | 9         |          | ac     | dad addr | g ggc a   | g ggagcac | CICLCL  | ag .    | caagaccag  | cerggecaaca | 99 ga   | aacccca  | C AC I    | AAAGATALAA  | MATTAICCA  | 66 6 6 66 66                                   |
| gat cccag cast trygglaget agg gas cast       ccc agg cast trygglaget agg cast       ccc agg cast trygglaget agg cast         a) cccag cast trygglaget agg cast       ccc agg cast trygglaget agg cast       ccc agg cast trygglaget agg cast         a) cccag cast trygglaget agg cast       ccc agg cast trygglaget agg cast       ccc agg cast trygglaget agg cast         cccag cast trygglaget agg cast       cccag cast trygglaget agg cast cast       cccag cast trygglaget agg cast cast       cccag cast trygglaget agg cast cast         cccag cast trygglaget agg cast cast cast trygglaget agg cast cast cast cast cast cast cast cast  |       |             |           | ~          | GHAGAT | CACACAG   | OGCICA        | 10001010  | I LCCAG  | A      | 0004000  | 10-00C/V  | GTOGAGCAC | CIGHGA  | CALIGHE |            | 0           | 00.04   | AACCCCA  | C C AC I  | CALL CALL   | AATTATCC.  | CC C CC CC                                     |
| Critical Control Contr   |       |             |           |            |        |           |               | 9         | cccage   |        | gggagge  | gaggaaa   | grggageac | C.      |         |            | 0           | GG GA   | AACCCCA  |           | AAAGA ACAAA | MATTA ICCA | 66 6 66 66                                     |
| GGANGCC (THAG CANTON TOURDOUL ORDAD CANGANCCAN CANGANAL CANAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA  |       |             |           |            |        |           |               |           | steecage | action | ggg ggc  | TGAGGC AN | g gg gc c | CIGACA  | CALCAC  |            | <u>v</u>    | 6 64    | AACCCCA  |           | AAGA ACAA   | AATTATCCA  | c (  |
| Odder GL     Odder  |       |             |           |            |        |           |               |           |          |        | CCCA TCC | TAG CAA   | G GGAGCAC | CIGAGA  | CA GAG  | C          |             | GIGA    | AACCCCA  | CCAC      | A A GATACAA | AATTATCCA  |  |
| Import in yosteness       Import in yosteness<   |       |             |           |            |        |           |               |           |          |        | 0004100  | T AND CAV |           | CIGHGA  | 04 040  | Lenne.     |             | 0 04    | AALLULA  | C C AC    |             | WAT TATECO | 9  |
| ggggcasg igggscacc igggr catagasca       GAAACCCCATC ICTACTAAAGATACAAAAATTA ICCAGGT ICTACTAAAGATACAAAAATTAAITACCAGGT ICTACTAAAGATACAAAAATTAAITACAAAGT ICCAGGT ICTACTAAAGATACAAAAATTAAITACAAGT ICCAGGT ICTACTAAAGATACAAAAATTAAITACCAGGT ICTACTAAAGATACAAAAATTAAITACCAGGT ICTACT   |       |             |           |            |        |           |               |           |          |        | ggu      | Lagage as | g ggogcac | C gaga  |         | LCaagacca  |             | 9.90    | ancoala  | C C DC    |             |            | g at at an al                                  |
| GAGGCAACI GAGA CCA GAGA CA TGAGTI CAAGACCAGC<br>GAGCCAACI GAGA CA TGAGTI CAAGACCAGC<br>GAAACCCCATCI CTACTAAAGATACAAAAATTA TCCAGGT<br>GCAAGT GGAGCACC GAGA TCA<br>GCAAGT GGAGCACC GAGA TCA<br>GCAAGT GGAGCACC GAGA TCA<br>GGAGCACC GAGA TCA<br>GGCAGCACC GAGA TCA<br>GGAGCACC G |       |             |           |            |        |           |               |           |          |        |          | 9999000   | giggogcac | chaga   | ca gag  | tcaagacca  |             | 9.90    | AA CCCA  | CCAC      | AAGATACAA   | AATTATCCA  | 99 9 99 90 GC                                  |
| AGGCAAG GGAGCACC GAGA CA GAGT CAAGACCAGC       GAAACCCCA C C CAACAAAAATTA CCAGG G G         aggcaal tigagcic cigaga caigagi caagaccagc       gaaacccaic cic cigcigaaga gcaagaata       g         GCAAG GGAGCACC GAGA CA       AACCCCAIC C CAACAAAAATTA CCAGG G       AACCCCAIC CIC AC TAAGAAAAATTA TCCAGG G         CAAG GGAGCACC GAGA CA       GAAACCCCAIC CIC AC TAAGAAAAATTA TCCAGG G       CAACCCCAIC CIC AC TAAGAAAAAATTA TCCAGG G         CAAG GGAGCACC GAGA CA GAGT CAAGACCAGCC G       AA CCCCAIC CIC AC TAAGAAAAATTA TCCAGG G       CCCGG G         CAAG GGAGCACC GAGA CA GAGT CAAGACCAGCC G       AA CCCCAIC CIC AC TAAGAAAAATTA TCCAGG G       CCCGG G         AAG GGAGCACC GAGA CA GAGT CAAGACCAGCC GC       AACCCCAIC CIC AC TAAGA TAAAAAAATTA TCCAGG G       CCCGG G         AAG GGAGCACC GAGA CA GAGT CAAGACCAGCC GC       GAACCCCAIC CIC AC TAAGA TAACAAAAATTA TCCAGG G G       CCCGAIC CIC AC TAAGA TAACAAAAATTA TCCAGG G G         AG GGAGCACC GAGA CA GAGT CAAGACCAGCC GGC       ACCCCCAIC CIC AC TAAGA TAACAAAAATTA TCCAGG G G G       CCCCAIC CIC AC TAAGA TAACAAAAATTA TCCAGG G G G G         AG GGAGCACC GAGA CA GAGT CAAGACCAGCC GGCC       GCCCCAIC CIC AC TAAGA TAACAAAAATTA TCCAGG G G G G       GCCCCAIC CIC AC TAAGA TAACAAAAATTA TCCAGG G G G G G         GGAGCACC GAGA CA GAGT CAAGACCAGCC GGCCAA       CAIC CIC AC TAAGA TAACAAAAATTA TCCAGG G G G G G       GGAGCACC GAGA CA GAGT CAAGACCAGCC GCC GCCAA       CAIC CIC AC TAAGA TAACAAAAATTA TCCAGG G G G G G G       GGAGCACC GAGA CA GAGT CAAGACCAGCC GCCC GC   |       |             |           |            |        |           |               |           |          |        |          | GLGGCAL   | GTGGAGCAC | CIGLGA  | CA GAG  | TCAAGACCAG |             | G       | AACCCCA  | CCAC      | AAGATACAA   | AATTATCCA  | GGT C  |
| aggcaattigagciccigagaica:gagaica:gagticaagaccagc       gaabccccatcictgcigagaigcaagaatta       g         aggcaattigagciccigagaica:gagaica:gaggicaagaccagccig       gaabccccatcictgcigagaigaccaggggg       aAccccatcictaciaagaatta       g         cAAGIGGAGCACCIGAGAICA       AACCCCATCICTACIAAAATTACAAAAATTAICCAGGIGT       AACCCCATCICTACIAAAATATACAAAAATTAICCAGGIGT         caagigggagcaccigagaica:gaggicaagagcaccigg       gabccccatciccicacciccicacagaattaicccaggggg       g         caagigggagcaccigagaica:gaggicaagaccigg       gabccccatcicicacagaattaicccaggggg       g         caagigggagcaccigagaica:gaggicaagaccigg       gabccccatcicicacagaattaicccagggg       g         caagigggagcaccigagaica:gaggicaagaccigg       gabccccatcicicacagaattaiccagggg       g         caagigggagcaccigagaica:gaggicaagaccigg       gabccccatcicicacciccicacagaattaiccagggg       g         AAGIGGAGCACCIGAGAICA TAGAGITCAAGACCAGCCIGGC       AACCCCATCICTACIAAAAATATAICCAGGIG       g         AAGIGGAGCACCIGAGAICA TGAGITCAAGACCAGCCIGGC       ACCCCATCICTACIAAAGATACAAAAATTAICCAGGIGIGG       g         GGAGCACCIGAGAICA TGAGITCAAGACCAGCCIGGCCAA       CATCICTACIAAAGATACAAAAATTAICCAGGIGIGGIGG       g         ggagcaccigagaigaigaigaigagticaagaaccagggigggccaa       CATCICTACIAAAGATACAAAAATTAICCAGGIGIGGIGG       g         ggagcaccigagaigaigaigaigaigaigaigaigaagaigaigagcagagaigaigaigaigaigaigagcacgagacgiggccaa       CATCICTACIAAAGATACAAAAATTAICCAGGIGIGGG       G         ggagc   |       |             |           |            |        |           |               |           |          |        |          | AGGCAA    | GTGGAGCAC | GAGA    | CA GAG  | TCAAGACCAG | ÷           | GA      | AACCCCA  | CCAC      | AATAAACA    | ateca      | ontotonto                                      |
| GCAAG IGAGCACC IGAGA CA       AACCCCAI CI CI ACTAAAGA IACAAAAA IT ATCCAGG IGT         CAAG IGGAGCACC IGAGA CA TGAGT CAAGACCAGCC IG       AACCCCAI CI CI ACTAAATA IACAAAAATT ATCCAGG IGT         Caag Iggagcacc Igaga ca Igagt caagaccagcc Ig       aacccca ci ci actaaaga ccaagagg caagagg I         Caag Iggagcacc Igaga ca Igagt caagaccagcc Ig       aacccca ci ci actaaaga ccaagagg I         Caag Iggagcacc Igaga ca Igagt caagaccagcc Ig       aacccca ci ci actaaaga ccaagagg I         Caag Iggagcacc Igaga ca Igagt caagaccagcc Ig       aacccca ci ci actaaaga ccaagagg I         AAG IGGAGCACC IGAGA CA IGAGT CAAGACCAGCC IGG       AACCCCAI CI CI ACTAAAGA IACAAAAATTA TCCAGGIG IG         AAG IGGAGCACC IGAGA CA IGAGT CAAGACCAGCC IGGC       ACCCCCGTTC TACTAAAGA IACAAAAATTA TCCAGGIG IG         AG IGCAGCACC IGAGA CA IGAGT CAAGACCAGCC IGGC       acccca i ci ci actaaaga iacaaaaatt a ccaggg Ig g         G IGGAGCACC IGAGA CA IGAGT CAAGACCAGCC IGGC       acccca i ci ci actaaaga iacaaaaatt a ccagg Ig g         G IGGAGCACC IGAGA CA IGAGT CAAGACCAGCC IGGC       acccca i ci ci actaaaga iacaaaaatt a ccagg Ig g         G IGGAGCACC IGAGA CA IGAGT CAAGACCAGCC IGGC       acccca i ci ci actaaaga iacaaaaatt a ccagg Ig g         G IGGAGCACC IGAGA CA IGAGT CAAGACCAGCC IGGC       acccca i ci ci actaaaga ccagg Ig Ig g         G IGGAGCACC IGAGA CA IGAGT CAAGACCAGCC IGGCCA       CA I CI CI ACTAAAGA IACAAAAATAT ATCCAGGI IG G         G IGGAGCACC IGAGA CA IGAGT CAAGACCAGCC IGGCCA       CA I CI CI ACTAAAGA IACAAAAAAT TA TCCAGGI IG G   |       |             |           |            |        |           |               |           |          |        |          | anneaa    | tttaaacto | c nana  | Ca 030  | tcaagaccag | -           |         | aaccoca  | e e ac    | naanatocaa  | aatta      | AA A AA A                                      |
| CAAG IGGAGCACC IGAGA ICATGAGTICAAGACCAGCCIG<br>Caag Iggggcacc Iggg icalgog I caagaccogccig<br>AAG IGGAGCACC IGAGA ICA IGAGTICAAGACCAGCCIGG<br>AAG IGGAGCACC IGAGA ICA IGAGTICAAGACCAGCCIGG<br>AG IGCAGCACC IGAGA ICA IGAGTICAAGACCAGCCIGGC<br>AG IGCAGCACC IGAGA ICA IGAGTICAAGACCAGCCIGGC<br>AG IGCAGCACC IGAGA ICA IGAGTICAAGACCAGCCIGGC<br>AG IGCAGCACC IGAGA ICA IGAGTICAAGACCAGCCIGGC<br>AG IGCAGCACC IGAGA ICA IGAGTICAAGACCAGCCIGGC<br>GIGGAGCACC IGAGA ICA IGAGTICAAGACCAGCCIGGCCA<br>CA ICIC I ACTAAAGA IACAAAAATTA ICCAGGIG IGGIGG<br>GIGGAGCACC IGAGA ICA IGAGTICAAGACCAGCCIGGCCA<br>CA ICIC I ACTAAAGA IACAAAAATTA ICCAGGIG IGGIGG<br>GIGGAGCACC IGAGA ICA IGAGTICAAGACCAGCCIGGCCA<br>CA ICIC I ACTAAAGA IACAAAAATTA ICCAGGIG IGGIGG<br>GIGGAGCACC IGAGA ICA IGAGTICAAGACCAGCCIGGCCA<br>CA ICIC I ACTAAAGA IACAAAAATTA ICCAGGIG IGGIGG<br>GIGGACCCIGAGA ICA IGAGTICAAGACCAGCCIGGCCA<br>CA ICIC I ACTAAAGA IACAAAAATTA ICCAGGIG IGGIGG<br>GIGGACCIGAGA ICA IGAGTICAAGACCAGCCIGCCIGCCAA<br>CA ICIC I ACTAAAGA IACAAAAATTA ICCAGGIG IGGIGG<br>GAGCACCIGAGA ICA IGAGTICAAGACCAGCCIGCCCACC<br>CG ICIC I ACTAAAGA IACAAAAATTA ICCAGGIG IGGIGG<br>GAGCACCIGAGA ICA IGAGTICAAGACCAGCCIGCCAACC<br>CG ICIC I ACTAAAGA IACAAAAATTA ICCAGGIG IGGIGG  |       |             |           |            |        |           |               |           |          |        |          | GCAA      | GTOGAGCAC | CIGAGA  | CA      | condoceod. | <b>.</b>    | A.,     | AACCCCA  | CCAC      | AAGATACAA   | AATTATCCA  | GGTGT  |
| Caag lggggcacc lgggg i calggg i caaggaccagcc lg       Baccca i ci ci ac Baagg i ccaagaa i a i ccagggg i caaggac i gg gg i caagga ccagca i ga gg i ccaagaa i a i ccagggg i caagga ccagca i ga gg i caagga ccaaca i ci ci ac Baagg i ccaagaa i a i ccagggg i caagga ccagca i ga gg i caagga ccaaca i ga caaga i caagaa i a i ccagg gg   |       |             |           |            |        |           |               |           |          |        |          | CAA       | GTGGAGCAC | GIGIGI  | CA GAG  | TCAAGACCAG | CCIG        |         | AATCCCA  | CCAC      | AATATACAA   | MATTATCCA  | GG G   |
| AAG IGGAGCACC IGAGA ICA IGAG I CAAGACCAGCC IGG<br>AG IGGAGCACC IGAGA ICA IGAG I CAAGACCAGCC IGG<br>AG IGGAGCACC IGAGA ICA IGAG I CAAGACCAGCC IGGC<br>AG IGGAGCACC IGAGA ICA IGAG I CAAGACCAGCC IGGC<br>AG IGGAGCACC IGAGA ICA IGAG I CAAGACCAGCC IGGC<br>G IGGAGCACC IGAGA ICA IGAG I CAAGACCAGCA IGGCC<br>G IGGAGCACC IGAGA ICA IGAG I CAAGACCAGCC IGGCCA<br>G IGCACC IGAGA ICA IGAG I CAAGACCAGCC IGCCAA<br>CAI CIT I A IAAAGA I A CAAAAA I TA ICCAGG I G IGG IG<br>G IGCACC IGAGA ICA IGAG I CAAGACCAGCC IGCCAA<br>CAI CIT I A IAAAGA I A CAAAAA I TA ICCAGG I G I G I G I<br>G IGCACC IGAGA ICA IGAG I CAAGACCAGCC I G I G I G I G I G I G I G I G I G  |       |             |           |            |        |           |               |           |          |        |          | Cas       | a agaacac | 0.000   | CA 030  | caagaccag  | cela        |         | aarccca  | C C ACL   | aaaaatccaaa | aattateca  | 0000 0   |
| AGIGGAGCACCIGAGAICAIGAGITCAAGACCAGCCIGGC<br>AGIGCAGCACCIGAGAICAIGAGITCAAGACCAGCCIGGC<br>GTGGAGCACCIGAGAICAIGAGITCAAGACCAGCCIGGC<br>GGAGCACCIGAGAICAIGAGITCAAGACCAGCCIGGCC<br>GGAGCACCIGAGAICAIGAGITCAAGACCAGCCIGGCCA<br>GGAGCACCIGAGAICAIGAGITCAAGACCAGCCIGGCCA<br>Ggagcaccigagaigatgatgatgatgagtcaaggaccagggiggccaa<br>CAICICIACAAAAAAITAITCCAGGIGIGGGG<br>GGAGCACCIGAGAICAIGAGITCAAGACCAGCCIGGCCAA<br>CAICICIACAAAAAAITAITCCAGGIGIGGGGG<br>GGAGCACCIGAGAICAIGAGITCAAGACCAGCCIGGCCAA<br>CAICICIACAAAAAAITAITCCAGGIGIGGGGG<br>GGAGCACCIGAGAICAIGAGITCAAGACCAGCCIGGCCAA<br>CAICICIACAAAAAAITAITCCAGGIGIGGGGG<br>GAGCACCIGAGAICAIGAGITCAAGACCAGCCIGGCCAA<br>CAICICIACAAAAAAITAITCCAGGIGIGGGGG  |       |             |           |            |        |           |               |           |          |        |          | 44        | GIGGAGCAC | GAGA    | CA GAG  | CAAGACCAG  | CC GG       |         | AACCCCA  | C C AC    | AAGATACAA   | AATTAICCA  | GGIGI  |
| AG IGCAGCACC IGAGA ICA IGAGT ICAAGACCAGCC IGGC<br>GIGGAGCACC IGAGA ICA IGAGT ICAAGACCAGCA IGGCC<br>GGAGCACC IGAGA ICA IGAGT ICAAGACCAGCA IGGCC<br>GGAGCACC IGAGA ICA IGAGT ICAAGACCAGCC IGGCCAA<br>Ggagt GGC Iggagt G Iggagt I GabgacCaggg I ggCCaa<br>CA ICTCI ACTAAAGA IACAAAAA TI A TCCAGG IG IGG IG<br>Ggagt GGC Iggagt G Iggagt I GabgaCCaggg I ggCCaa<br>CG ICTCI ACTAAAGA IACAAAAA TI A TCCAGG IG IGG IG<br>GGAGCACC IGAGA ICA IGAGT ICAAGACCAGCC IGGCCAA<br>CG ICTCI ACTAAAGA IACAAAAA TI A TCCAGG IG IGG IG<br>GGAGCACC IGAGA ICA IGAGT ICAAGACCAGCC IGGCCAA<br>CG ICTCI ACTAAAGA IACAAAAA TI A TCCAGG IG IGG IG<br>GAGCACC IGAGA ICA IGAGT ICAAGACCAGCC IGGCCAAC<br>CG ICTCI ACTAAAGA IACAAAAA TI A TCCAGG IG IGG IG   |       |             |           |            |        |           |               |           |          |        |          |           | GTGGAGCAC | GAGAT   | CA GAG  | TCAAGACCAG | CCIGGC      |         | ACCCCG   | CAC       | AAGATACAA   | AATTATCCA  | GGIGIG C                                       |
| GIGGAGCACCIGAGAICAIGAGITCAAGACCAGCAIGGCC CCCCAICICIACTAAAGAIAC alccaggigigiggigg<br>GGAGCACCIGAGAICAIGAGITCAAGACCAGCCIGGCCAA CAICTCIAAAAGAIACAAAAAITAICCAGGIGIGGIGG<br>ggagcaccigagaigaigagitcaagaaccagggiggccaa CAICICIACTAAAGAIACAAAAAITAICCAGGIGIGGIGC<br>ggagcaccigagaigaigaigagitcaagaaccagggiggccaa CGICICIACTAAAGAIACAAAAAITAICCAGGIGIGGIG<br>ggagcaccigagaigaigaigagitcaagaaccaggciggccaa CGICICIACTAAAGAIACAAAAAITAICCAGGIGIGGIG<br>GAGCACCIGAGAICAIGAGIICAAGACCAGCCIGGCCAAC CAICICIACTAAAGAIACAAAAAITAICCCAGGIGIGGIG   |       |             |           |            |        |           |               |           |          |        |          |           | GIGCAGCAC | GAGA    | CA GAG  | CAAGACCAG  | CC GGC      |         | accesa   | c c acu   | aadatacaa   | aattateca  | aatata c                                       |
| GGAGCACC IGAGA I CA IGAGT I CAAGACCAGCC I GGCCAA<br>ggagcacc Igaga iga Igagt i caagaccaggg I ggccaa<br>Ggagcacc Igaga i ca Igagt i caagaccaggc I ggccaa<br>GAGCACC IGAGA I CA IGAGT I CAAGACCAGCC I GGCCAAC<br>CA I C T CI ACTAAAGA I ACAAAAA T A TCCAGG IG I GG I GG<br>GAGCACC IGAGA I CA IGAGT I CAAGACCAGCC I GGCCAAC<br>CA I C T CI ACTAAAGA I ACAAAAA T A TCCAGG I G I GG I GG   |       |             |           |            |        |           |               |           |          |        |          |           | GTGGAGCAC | TGLG    | CA GAG  | TCAAGACCAG | CATGGCC     |         | CCCCA    | CCAC      | AAGATAC     | atcca      | optotoptop                                     |
| ggagcacctgagatgatgagttcaagaccagggtggccaa CAICTCIACTAAAGAIACAAAAAITATCCAGGTGTGCIGC<br>ggagcacctgagatcatgagttcaagaccagcctggccaa CGICTCIACTAAAGAIACAAAAAITATCCAGGTGTGGIG<br>GAGCACCTGAGAICATGAGTTCAAGACCAGCCTGGCCAAC CAICTCIACTAAAGAIACAAAAAITATCCAGGTGTGGIG  |       |             |           |            |        |           |               |           |          |        |          |           | GGAGCAC   | CTG.G.  | CA GAG  | CAAGACCAG  | CELOGCCAA   |         | CAT      | CCAA      | AAGATACAA   | AATTATCCA  |  |
|  |       |             |           |            |        |           |               |           |          |        |          |           | 0030030   | claada  | 08 080  | tcaagaccag | ogtogccaa   |         | CA       | CCACI     | AAGATACAA   | MATTATCCA  | GG G GC GC                                     |
| GAGCACC GAGA CA GAG CAAGACCAGCC GGCCAAC CA C CA C C AC AAAGA ACAAAAA I A CCAGG G G G   |       |             |           |            |        |           |               |           |          |        |          |           | 0030030   | ctoadat | ca gag  | tcaagaccag | cctogccaa   |         | CG       | CICIACT   | AAGATACAA   | WATTATCCA  | GG G GG GG                                     |
|  |       |             |           |            |        |           |               |           |          |        |          |           | GAGCAC    | C GAGA  | CA GAG  | CAAGACCAG  | CC GGCCAAC  |         | CAT      | C C ACT   | AAAGATACAA  | MATTATCCA  | GGTGTGGTG                                      |

Next Generation Sequencing (NGS) readout from autism's genetics

# An Introduction to Genomic Data

- Next Generation Sequencing (NGS)
  - Easy human genomic data
- NGS for genetics research
  - Precise detection of variants
  - Personalized drug and medicines
- NGS data
  - Large, difficult to store and process



## An Introduction to Genomic Compression

- Genomic compression would
   greatly improve handling
- Conventional compression can be improved through intrinsic patterns in variant data (SNPs)
- Group's ongoing focus on autoencoders and convolutional networks



#### **Current Work in Compression**



Fully connected convolutional neural network



autoencoder with smaller hidden layer

#### **Current Research**

- Correlate compression results to properties of genomic data
  - Gene sets organize data
- Quantify differences in genomic data
- Similar genomic data should be similar after compression!

...ATCGTGTACTTCGTGTGAGGG... ...ATCGTGTACTTCGTGTGAGGG... ...TACTCGGTAGCTATGCAGTGT...

## Analysis of Gene Sets

- Gene Ontology (GO) consortium classification
  - Assigns biological significance
- Gene Set Enrichment Analysis (GSEA) finds hidden patterns



## Gene Set Enrichment Analysis

- In gene expression profiles
   no individual gene may be statistically significant
  - significant genes with no biological theme
  - effects on pathways
     improperly described



#### Gene Set Enrichment Analysis

- Rank expression data set
- Running sum down gene list
- Enrichment Score (ES) calculated for each gene set





- •Phenotype labels permuted and ES calculated vs null
- Nominal p-value estimation

## VCF Enrichment Analysis

- VCF files require preparation
  Annotate with GO labels
- Running sum down VCF labels
   to calculate ES
- Similarity score calculated using Kolmogorov-Smirnov test



### Future Work

- Analyze more genomic data, including FASTA, to be able to find more potential biological patterns
- Improve the metric for comparing VCFs to be more robust and include various genomic data types
- Experiment with convolutional network layers to be able to take advantage of biological patterns
- Investigate this process to test whether biological patterns exist in other genomic datasets

### Conclusions

- Genomic data is difficult to use without compression
- Optimal compression methods have to use intrinsic patterns
- Gene sets allow for quantification of biological significance and patterns
- Labeling VCF files allow for us to analyze intrinsic patterns
- Current algorithms are not able to show much compression towards VCF, but FASTA

## Acknowledgements

- MIT PRIMES for allowing us to take on this exciting and challenging research opportunity
- Dr. Gil Alterovitz and Maksym for their guidance and support
- Adithya, Kalyan, and other members of our group for their previous work and assistance
- My parents for their support



## ALL THESE GENESPIP