|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | # of peaks | # of predicted motifs | # of predicted motif modules | Percentage motifs similar to known motifs (Evalue<1E-10)  | Percentage motifs similar to known motifs (Evalue<1E-5) | Percentage motifs similar to known motifs (Evalue<1E-4) | # of motifs predicted in random data sets |
| Sox2 | 7761 | 68 | 227 | 8/68=11.8% | 54/68=79.4% | 66/68=97.1% | 0 |
| E2f1 | 20670 | 63 | 427 | 12/63=19% | 51/63=81% | 61/63=96.8% | 0 |
| Stat3 | 5347 | 97 | 1058 | 28/97=28.9% | 83/97=85.6% | 94/97=96.9% | 0 |
| Nanog | 17834 | 81 | 190 | 17/81=20.1% | 64/81=79% | 79/81=97.5% | 0 |
| Oct4 | 6915 | 58 | 322 | 9/58=15.5% | 51/58=87.9% | 56/58=96.55% | 0 |
| c-Myc | 6492 | 79 | 1214  | 17/79=21.6% | 66/79=83.5% | 77/79=97.5% | 0 |
| Klf4 | 18144 | 94 | 433 | 19/94=20.2% | 75/94=79.8% | 91/94=96.8% | 0 |
| Ctcf | 49114 | 88 | 123 | 16/88=18.1% | 70/88=79.5% | 85/88=96.6 | 0 |
| Zfx | 17201 | 95 | 811 | 17/95=17.9% | 76/95=80% | 92/95=96.8% | 0 |
| Tcfcp2l1 | 45885 | 71 | 133 | 11/71=15.5% | 56/71=78.9% | 68/71=95.8% | 0 |
| Esrrb | 49127 | 45 | 43 | 6/45=13.3% | 37/45=82.2 | 43/45=95.5% | 0 |
| n-myc | 10987 | 87 | 767 | 22/87=26.3% | 70/87=80.5% | 87/87=100% | 0 |
| Smad1 | 2185 | 20 | 29 | 2/20=10% | 20/20=100% | 20/20=100% | 0 |

Table 1. Predicted motifs by SIOMICS\_Extension in 13 Chip-seq data sets and 13 random data sets

Table 2. Predicted motifs by SIOMICS in 13 Chip-seq data sets and 13 random data sets (SIOMICS Paper)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | # of peaks | # of predicted motifs | # of predicted motif modules | Percentage motifs similar to known motifs (Evalue<1E-10)  | Percentage motifs similar to known motifs (Evalue<1E-5) | Percentage motifs similar to known motifs (Evalue<1E-4) | # of motifs predicted in random data sets |
| Sox2 | 7761 | 99 | 89 | 8/99=8.1% | 78/99=78.8% | 96/99=97.0% | 0 |
| E2f1 | 20670 | 99 | 2510 | 8/99=8.1% | 79/99=79.8% | 94/99=94.9% | 0 |
| Stat3 | 5347 | 91 | 1256 | 9/91=9.9% | 72/91=79.1% | 85/91=93.4% | 0 |
| Nanog | 17834 | 99 | 1131 | 7/99=7.1% | 76/99=76.8% | 96/99=97.0% | 0 |
| Oct4 | 6915 | 73 | 719 | 2/73=2.7% | 64/73=87.7% | 69/73=94.5% | 0 |
| c-Myc | 6492 | 96 | 1901 | 5/96=5.2% | 74/96=77.1% | 94/96=97.9% | 0 |
| Klf4 | 18144 | 99 | 2052 | 5/99=5.1% | 83/99=83.8% | 96/99=97.0% | 0 |
| Ctcf | 49114 | 99 | 784 | 5/99=5.1% | 78/99=78.8% | 94/99=94.9% | 0 |
| Zfx | 17201 | 98 | 1945 | 6/98=6.1% | 75/98=76.5% | 93/98=94.9% | 0 |
| Tcfcp2l1 | 45885 | 71 | 782 | 2/71=2.8% | 55/71=77.5% | 68/71=95.8% | 0 |
| Esrrb | 49127 | 43 | 308 | 2/43=4.7% | 35/43=81.4% | 41/43=95.3% | 0 |
| n-myc | 10987 | 94 | 1766 | 5/94=5.3% | 72/94=76.6% | 91/94=96.8% | 0 |
| Smad1 | 2185 | 21 | 33 | 2/21=9.5% | 21/21=100% | 21/21=100% | 0 |

From the comparison of the above 2 tables, we found out that SIOMICS Extension is able to find more accurate motifs compared with original SIOMICS. Reasons:

1. Under E-value cutoff 1E-5 and 1E-4, the SIOMICS Extension is just slightly better than SIOMICS
2. However, if we look at the E-value cutoff 1E-10 (which means extremely similar to known motifs), the original SIOMICS able to predict about 5% of motifs, similar to known motifs. In contrast, SIOMICS Extension is able to predict average 19% motifs, similar to known motifs. The performance is around 3 times better.

**Note: For the column: Percentage motifs not in original top 100.**

**Since our SIOMICS Extension has made no changes on the SIOMICS procedure, we only extend or shorten after we get predicted motifs. Therefore, this column was not included in the table since it should be the same for SIOMICS and SIOMICS\_Extension.**

Table 3. Predicted co-factors

|  |  |  |
| --- | --- | --- |
| Data set | SIOMICS  | SIOMIC Extension |
| Sox2 | 8/9(Sox2,Klf4,Stat3,Zic3,Hoxa5,Tcf3,Tead1,Oct4) | 8/9(Sox2,Klf4,Stat3,Tead1,Oct4,Zic3,Tcf3,Hoxa5) |
| E2f1 | 7/10(E2f1,Stat3,Klf4,Fox,Sp1,Nfkb1,Tbp) | 8/10(E2f1, Stat3,Klf,Fox,Sp1,Nfkb1, Creb1,Tbp) |
| Stat3 | 6/8(Stat3,Klf4,Sox2,Myc,Sp1,Irf) | 6/8(Stat3,Klf4,Sox2,Myc,Sp1,Irf) |
| Nanog | 7/8(Nanog,Sox2,Oct4,Zic3,Klf4,Elf5,Tead1) | 8/8(Nanog,Sox2,Oct4,Zic3,Klf4,Esrrb,Elf5,Tead1) |
| Oct4 | 8/10(Oct4,Sox2,Klf4,Sox10,Ewsr1,Nanog,Zic,Esrrb) | 8/10(Oct4,Sox2,Klf4,Sox10,Ewsr1,Nanog,Esrrb,Tead1) |
| c-Myc | 3/4(Stat3,Egr1,Sp1) | 3/4(Stat3,Egr1,Sp1) |
| Klf4 | 4/10(Klf4,Stat3,Sox2,Sp1) | 5/10(Klf4,Stat3,Sp1,Myc,Sox2) |
| Ctcf | 5/6(Ctcf,Stat3,Gabpa,Yy1,Smad3) | 5/6(Ctcf,Stat3,Gabpa,Smad3,Myc) |
| Zfx | 2/4(Zfx,Stat3) | 2/4(Zfx,Stat3) |
| Tcfcp2l1 | 7/12(Tcfcp2l1,Stat3,Klf4,Sox2,Esrrb,Fox,Sp1) | 10/12(Tcfcp2l1,Stat3,Klf4,Sox2,Fox,Sp1,Oct4,Creb,Myc,Tead1) |
| Esrrb | 4/10(Esrrb,Klf4,Rxra,Sp1) | 8/10(Esrrb, Klf4,Rxra,Sp1,Ewsr1,Creb,Sox2,Stat3) |
| n-Myc | 2/5(Stat3,Creb) | 2/5(Stat3,Creb) |
| Smad1 | 5/9(Sox2,Oct4,Esrrb,Klf4,Stat3) | 5/9(Sox2,Oct4,Klf4,Stat3,Esrrb) |

Based on the comparison on the above table, SIOMICS Extension is slightly better than SIOMICS.

One reason why can’t show that SIOMICS Extension is significantly better than SIOMICS is: The limitation of known co-factors.

For example, for Sox2(8/9) ,Stat3 (6/8) ,Nanog (7/8), Oct4 (8/10),c-Myc (3/4),Ctcf (5/6) datasets, SIOMICS already predicted most of known co-factors. Therefore, there is no much space for SIOMICS Extension to show its advantage.

However, in other data sets, e.g. Tcfcp2l1, Esrrb, in which SIOMICS failed to predict around half know co-factors.

SIOMICS Extension is able to find more.

Table 4. Software running time

|  |  |  |
| --- | --- | --- |
| Data set | SIOMICS (secs) | SIOMIC Extension (secs) |
| Sox2 | 7761 | 8578 |
| E2f1 | 20670 | 21280 |
| Stat3 | 5347 | 8775 |
| Nanog | 17834 | 18039 |
| Oct4 | 6915 | 9516 |
| c-Myc | 6462 | 9570 |
| Klf4 | 18114 | 19282 |
| Ctcf | 49114 | 57830 |
| Zfx | 17201 | 18362 |
| Tcfcp2l1 | 45885 | 50413 |
| Esrrb | 49127 | 51412 |
| n-Myc | 10987 | 12321 |
| Smad1 | 2185 | 6264 |

The following table shows the consensus of predicted co-factors

Table 5. Consensus of predicted co-factors for SIOMICS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data set | Known co-factors consensus | SIOMICS Predicted consensus | Known co-factors consensus | SIOMICS Extension Predicted consensus |
| Sox2 | Sox2:TNTGCATNACAAWGGKlf4: CCTTYYYTTNStat3: TTCCNGGAAG-Zic3: GANCMCCC--Hoxa5:NGRNNNNNNNNANTAATGRNGNNNNNNGCTcf3:NNAMCATCTGKTTead1: CNSWGGAATGTROct4:ATTTGCATNACAAWG | Sox2: ------- AACAAAAG(4.8E-6)Klf4: CTTTTCTT--(2.5E-5)Stat3: ---CAGGAAAG(1.4E-5)Zic3: --CCACCCAC(4.3E-6)Hoxa5: ----------------TGGGGGTG-----(5.6E-5)Tcf3:--ACCAGCTG--(9.0E-07)Tead1:CAGAGGAA----(2.2E-06)Oct4: -----CAAAACAA--(9.0735E-06) | Sox2: CCWTTGTNATGCANAKlf4: NAARRRAAGG---Stat3: TTCCNGGAAG-Zic3: GANCMCCC--Hoxa5:NGRNNNNNNNNANTAATGRNGNNNNNNGCTcf3:NNAMCATCTGKTTead1: CNSWGGAATGTROct4: ATTTGCATNACAAWG | Sox2: -CTTTGTTCT-----(8.9E-07)Klf4: --AAGGAAGGAAG (6.6E-7)Stat3: ---CAGGAAAG(1.4E-5)Zic3: --CCACCCAC(4.3E-6)Hoxa5: ----------------TGGGGGTG-----(5.6E-5)Tcf3:--ACCAGCTG--(9.0E-07)Tead1:CAGAGGAA----(2.2E-06)Oct4: ------AGAACAAAG(4.0E-06) |
| E2f1 | E2F: NTTTCCCGCNStat3: CTTCCNGGAAKlf4: CCTTYYYTTNFox: ANTGTTTRTTTWSp1: GGGGGNGGGGNfkb1: GGGGATTCCCCTbp: WATTTAT--| | E2f: -TTTCCCAG-(5.4575E-05)Stat3:--GCCTGGAG (2.9E-05)Klf4: CCTTTCCT--(2.1E-07)Fox: ---ATTTATTT-(6.9E-07)Sp1: GAGGGAGG--(8.5E-06)Nfkb1: ---CCTTCCCC(8.7E-05)Tbp:-ATTTATTT (1.7E-07) | E2F1: NTTTCCCGCNStat3: -CTTCCNGGAAKlf4: CCTTYYYTTNFox: ANTGTTTRTTTWSp1: CCCCNCCCCCNfkb1: GGGGATTCCCCTbp: WATTTAT--Creb: NACGTCANCMNN | E2f: -TTTCCCAG-(5.4575E-05)Stat3: CCTTCCTGC-- (1.2E-6)Klf4: CCTTTCCT--(2.1E-07)Fox: ---ATTTATTT-(6.9E-07)Sp1: CCCCTCCTCC(1.1E-11)Nfkb1: ---CCTTCCCC(8.7E-05)Tbp: -ATTTATTT (1.7E-07)Creb: TCACCAG (5.6E-05) |
| Stat3 | Stat3: CTTCCNGGAAKlf4: CCTTYYYTTNSox2: TNTGCATNACAAWGGMyc:-GAGCACGTGGTSp1: CCCCNCCCCC-Irf: NTTTCWNTTT | Stat3: -TTCCTGGA-(2.5E-12)Klf4: --TTTCTTTC (3.1E-05 )Sox2: ----CAAAACAA---(4.7E-06)Myc: AGAGCAGG---- (8.9E-05)Sp1:---CTCCCCCT(6.6.E-07)Irf: CTTTCTCT--(3.3E-05) | Stat3: CTTCCNGGAAKlf4: CCTTYYYTTNSox2: TNTGCATNACAAWGGMyc:-GAGCACGTGGTSp1: GGGGGNGGGG- Irf: NTTTCWNTTT | Stat3: -TTCCTGGAA-(7.3E-15)Klf4: --TTTCTTTC (3.1E-05 )Sox2: ----CAAAACAA---(4.7E-06)Myc: AGAGCAGG---- (8.9E-05)Sp1: GGGGGAGGGGM (6.5E-15)Irf: CTTTCTCT--(3.3E-05) |
| Nanog | Nanog: -GGMAATGGNCCSox2: -CCWTTGTNATGCANAOct4: CWTTGTNATGCAAATZic3: GANCMCCC—Klf4: GCCMCRCCCNElf5: -NMGGAARTNTead1: CNSWGGAATGTR | Nanog:AGGAAATG----(1.6E-07)Sox2: TCCTTTGT--------(7.2E-6)Oct4: --TTGTTTTG-----(9.5E-06)Zic3: --CCACCCCA (3.6E-06)Klf4: -CCACACCC-(2.5E-09)Elf5: CCAGGAAG--(9.6057e-05)Tead1: CAGAGGAA----(1.5253e-06) | Nanog: -GGMAATGGNCCSox2: -CCWTTGTNATGCANAOct4: CWTTGTNATGCAAATZic3: GANCMCCC—Klf4: GCCMCRCCCNElf5: -NMGGAARTNTead1: CNSWGGAATGTREsrrb: NNNYCAAGGTCA-- | Nanog: AGGAAATG----(1.6E-07)Sox2: TCCTTTGT--------(7.2E-6)Oct4: --TTGTTTTG-----(9.5E-06)Zic3: --CCACCCCA (3.6E-06)Klf4: -CCACACCC-(2.5E-09)Elf5: CCAGGAAG--(9.6057e-05)Tead1: CAGAGGAA----(1.5253e-06)Esrrb:------AGGTCAGG(5.1E-05) |
| Oct4 | Oct4: ATTTGCATNACAAWGSox2: CCWTTGTNATGCANA-Klf4: GCCMCRCCCSox10: ACAAWG--Ewsr1: GGAAGGAAGGAAGGAAGGNanog: GGMAATGGNCCZic2: NACCACCC-Esrrb: NNNYCAAGGTCA-- | Oct4: ATTTGCAT-------(8.0E-11)Sox2: --------ATGCAAAT(5.3923e-07|)Klf4: -CCACACCC-(1.1E-09)Sox10: ACAAAGCC(8.1E-08)Ewsr1: : ---------GAAGGAAG-(1.1E-10)Nanog: AGGAAAGG----(9.4E-05)Zic2: -TCCACCCC(1.7E-06)Esrrb: ------AGGTCAGG(5.1E-05) | Oct4: ATTTGCATNACAAWGSox2: CCWTTGTNATGCANA-Klf4: GCCMCRCCCN-Sox10: ACAAWG--Ewsr1: GGAAGGAAGGAAGGAAGGNanog: GGMAATGGNCCZic2: NACCACCC-Esrrb: NNNYCAAGGTCA-- | Oct4: ATTTGCAT-------(8.0E-11)Sox2: --------ATGCAAAT(5.3923e-07|)Klf4: GCCCCGCCCCC (7.1E-10)Sox10: ACAAAGCC(8.1E-08)Ewsr1: : ---------GAAGGAAG-(1.1E-10)Nanog: AGGAAAGG----(9.4E-05)Zic2: -TCCACCCC(1.7E-06)Esrrb: ------AGGTCAGG(5.1E-05) |
| c-Myc | Stat3: -CTTCCNGGAAEgr1: NGCGTGGGCGKSp1: CCCCNCCCCC | Stat3: CCTTCCTG---(3.7E-08)Egr1: TGGGTGGG---(8.4E-05)Sp1: CCCCCCCC--(4.8E-07) | Stat3: -CTTCCNGGAAEgr1: NGCGTGGGCGKSp1: CCCCNCCCCC- | Stat3: CCTTCCTG---(3.7E-08)Egr1: TGGGTGGG---(8.4E-05)Sp1: CCCCGCCCCKC (1.3E-11) |
| Klf4 | Klf4: -CCTTYYYTTNStat3: TTCCNGGAAGSox2:TNTGCATNACAAWGGSp1: GGGGGNGGGG | Klf4: TCCTTCCT---(4.7E-6)Stat3: --CCAGGAAG(6.4E-12)Sox2: -------GACAAAGG(5.4E-5)Sp1: -GGGGCAGG- (1.3E-6) | Klf4: -GCCMCRCCCN;Stat3: TTCCNGGAAGSox2:TNTGCATNACAAWGGSp1: GGGGGNGGGGMyc:-GCCACGTGSN | Klf4: GGCCACGCCC-(2.1E-10)Stat3: --CCAGGAAG(6.4E-12)Sox2: -------GACAAAGG(5.4E-5)Sp1: -GGGGCAGG- (1.3E-6)Myc: GGCCACGTCC(4.7E-05) |
| Ctcf | Ctcf: NNSYGCCMCCTRSTGGNNRStat3: TTCCNGGAAGGabpa: SNCTTCCGGTYy1: WNSANNCAAGATGGCNGNNSmad3: AGNCAGAC | Ctcf: --------CCTGCTGG---(9.4E--07)Stat3: CTCCAGGA—(3.8E-08)Gabpa: -TCTTCCTG-(1.4E-07)Yy1: --CAGCCCAG---------(3.1E-05)Smad3: AGAGAGAG(2.3E-05) | Ctcf: YNNCCASYAGGKGGCRSNNStat3: TTCCNGGAAGGabpa: SNCTTCCGGTYy1: WNSANNCAAGATGGCNGNNSmad3: ----AGNCAGAC | Ctcf: ---CCAGCAGGGGGCGC--(0)Stat3: CTCCAGGA—(3.8E-08)Gabpa: -TCTTCCTG-(1.4E-07)Yy1: --CAGCCCAG---------(3.1E-05)Smad3: ACACAGACAGA- (8.1E-07) |
| Zfx | Zfx: NAGGCCNNGGCNNStat3: CTTCCNGGAA | Zfx: CAGGCCTGGG----(6.4E-07)Stat3: --TCCTGGGA(4.6E-07) | Zfx: NNNGCCNNGGCCTNStat3: CTTCCNGGAA | Zfx: ---GCCCAGGC---(4.2E-06)Stat3: --TCCTGGGA(4.6E-07) |
| Tcfcp2l1 | Tcfcp2l1: CYGGNTNNRNCYGG-Stat3: CTTCCNGGAAKlf4: NAARRRAAGG-Sox2: --TNTGCATNACAAWGGEsrrb: --TGACCTTGRNNNFoxp1: TATTTGTGTTGTTTTTTATSp1: GGGCGGGGN | Tcfcp2l1: -------GAACTGGA(6.0E-6)Stat3: CTTCCTCT—(5.0E-5)Klf4: ---AGGAAGGA(7.4E-6)Sox2: CCTTTGCA---------(9.0E-5)Esrrb: CCTGACCT------(5.0E-5)FoxP1: ---------TTTTTTTT—(8.3e-06)Sp1: GGGCTGGG-(6.7E-8) | Tcfcp2l1: CYGGNTNNRNCYGG-Stat3: CTTCCNGGAAKlf4: NAARRRAAGG-Sox2: --TNTGCATNACAAWGGEsrrb: --TGACCTTGRNNNFoxp1: TATTTGTGTTGTTTTTTATSp1: GGGCGGGGNCreb: TGACGTCA---Myc: ACCACGTGSTNTead1:----CNSWGGAATGTR- | Tcfcp2l1: -------GAACTGGA(6.0E-6)Stat3: CTTCCTCT—(5.0E-5)Klf4: ---AGGAAGGA(7.4E-6)Sox2: CCTTTGCA---------(9.0E-5)Esrrb: CCTGACCT------(5.0E-5)FoxP1: ---------TTTTTTTT—(8.3e-06)Sp1: GGGCTGGG-(6.7E-8)Creb:-GACGTCAGGG(8.1E-05)Myc: CNCACCTGCT-(2.2E-05)Tead1:GGGGGGGAGGAATGTGG (1.1E-05) |
| Esrrb | Esrrb: TGACCTTGRNNNKlf4: CCTTYYYTTNRxra: STTGACCTTTGACCTTTSp1: GGGGGNGGGG | Esrrb: TGACCTTG----(7.6E-12)Klf4: -TTTCCTTT- (6.9E-06)Rxra: --TGACCTTG-------(3.6E-07)Sp1: GGAGGAGG--(1.7E-06) | Esrrb: TGACCTTGRNNNKlf4: CCTTYYYTTNRxra: STTGACCTTTGACCTTTSp1: GGGGGNGGGGEwsr1: -CCTTCCTTCCTTCCTTCCCreb: --NACGTCANCMNNSox2: TNTGCATNACAAWGGStat3: CTTCCNGGAA | Esrrb: TGACCTTG----(7.6E-12)Klf4: -TTTCCTTT- (6.9E-06)Rxra: --TGACCTTG-------(3.6E-07)Sp1: GGAGGAGG--(1.7E-06)Ewsr1: TTTTTCTTTCCTTTCT---(2.8E-10)Creb: CCCAGGTCTCCAGG(5.8E-05)Sox2: TCTGCATNACG---- (4.5E-5)Stat3: CTTTCTGGG-(4.3E-6) |
| n-Myc | Stat3: CTTCCNGGAACreb: NGNTGACGTNN | Stat3: CTTCCTCT--(6.3E-05)Creb: AGGTGAGG---(6.2E-05) | Stat3: CTTCCNGGAACreb: NGNTGACGTNN | Stat3: CTTCCTCT--(6.3E-05)Creb: AGGTGAGG---(6.2E-05) |
| Smad1 | Sox2: TNTGCATNACAAWGGOct4:CWTTGTNATGCAAATEsrrb: NNNYCAAGGTCAKlf4: GCCMCRCCCNStat3: --CTTCCNGGAA | Sox2: ----CAAAACAA---(4.4E-06)Oct4: --TTGTTTTG-----(8.6E-06)Esrrb: ----AAAGGGCA(8.3E-05)Klf4: -CCCCACCC-(1.7E-09)Stat3: TCCTTCCT----(5.7e-05) | Sox2: TNTGCATNACAAWGGOct4:CWTTGTNATGCAAATEsrrb: NNNYCAAGGTCAKlf4: GCCMCRCCCNStat3: --CTTCCNGGAA | Sox2: ----CAAAACAA---(4.4E-06)Oct4: --TTGTTTTG-----(8.6E-06)Esrrb: ----AAAGGGCA(8.3E-05)Klf4: -CCCCACCC-(1.7E-09)Stat3: TCCTTCCT----(5.7e-05) |

Note: known motifs consensus can be known consensus or reverse complement of known consensus. E.g. Sp1: GGGGGNGGGG/ CCCCNCCCCC