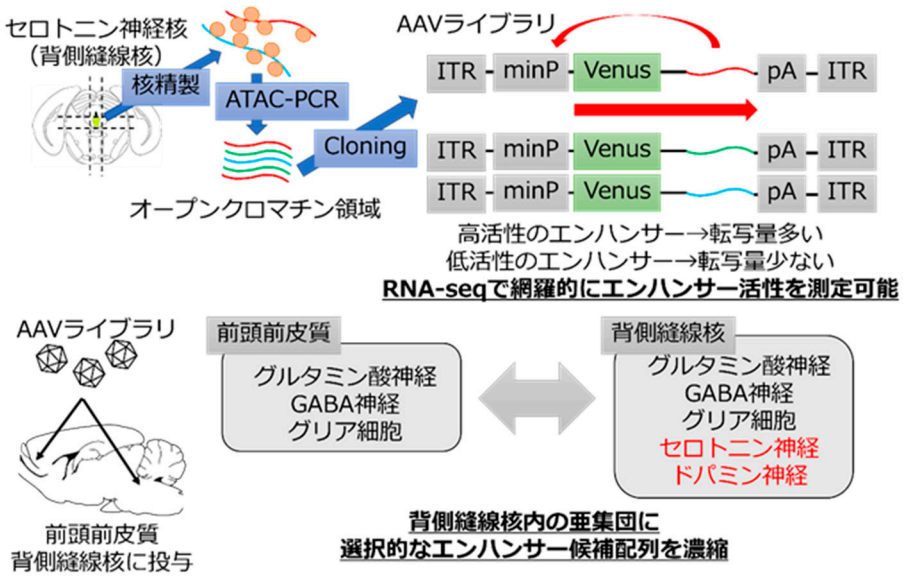
206　AIを活用した細胞亜集団選択的制御手法の開発	永安　一樹

【目的】本研究では、細胞亜集団の制御手法の開発を目指し、新たな制御因子活性網羅的測定手法の開発を情動制御において重要なセロトニン神経核である背側縫線核を一例として実施した。さらに情報科学的アプローチから任意の核酸配列の制御因子活性を予測する手法の開発を実施した。

【方法】エンハンサー活性が特に高いゲノム配列を濃縮するため、マウス背側縫線核から ATAC 法によりオープンクロマチン領域配列を単離した。単離した配列のエンハンサー活性は STARR-seq 法を用いて測定した。AAV のパッケージング後、マウス背側縫線核および前頭前皮質に投与した。4 週間の回復後、背側縫線核および前頭前皮質を単離し total RNA を回収し、逆転写反応および PCR 増幅を行ったのち次世代シーケンサーで解析した。制御因子活性の *in silico* 予測のため、まず核酸配列の前処理を行い 2,002 次元×201 ビン（200 bp／ビン）のクロマチン特徴量を算出した。このクロマチン特徴量を入力として線条体の神経細胞亜集団の遺伝子発現量を出力するような予測モデルを、WaveNet、DenseNet あるいは画像認識で一般的に用いられる 2 次元の畳み込みを多層に配置したニューラルネットワークを用いて構築した。ハイパーパラメーター探索は GPyOpt によるガウス過程最適化により行った。

【結果】縫線核に由来する AAV ライブラリであることから、まず縫線核が含まれる中脳領域の細胞種決定に重要な転写因子群の結合配列が濃縮されているか否かを調べたところ、神経細胞の細胞種決定に重要である CTCF、Nfil3 および Rfx1 について AAV ライブラリ中で高度に濃縮されていることが明らかになった。次に背側縫線核および前頭前皮質における結果を解析したところ、両神経核でのピークは高度に類似しており、単独の神経核の解析では細胞集団選択的エンハンサー候補の単離が困難であることが示唆された。そこで、両神経核のデータを定量比較した結果、背側縫線核のみで高度に発現するエンハンサー領域を 6 か所同定することに成功した。本結果は、今回開発した手法の高い有用性を示すものと考えられた。次に、核酸配列のみからその制御因子活性を予測する手法を開発した。その結果、画像同様の 2 次元の畳み込みフィルターを多層に重ねた手法が、その他の手法よりも高い精度で予測を行えることが明らかになった。また学習に使わなかった評価データセットについて、構築した予測モデルを用いた遺伝子発現予測値と実測値の相関を調べたところ、そのスピアマンの相関係数は 0.607（$P=2.26\times10^{90}$）と有意な相関を示すことが明らかになった。これらの結果は、本予測手法の高い性能と妥当性を示すものと考えられた。今後、前項で得られた制御因子活性の網羅的測定手法と組み合わせてさらなる学習を行わせることで、あらゆる核酸配列の制御因子活性の予測が可能になると考えられる。

網羅的エンハンサー活性測定法による亜集団選択的エンハンサー配列の濃縮



206. AI を活用した細胞亜集団選択的制御手法の開発

永安　一樹

京都大学 薬学研究科 生体機能解析学分野

Key words：エンハンサー，ATAC-Seq，縫線核，畳み込みニューラルネットワーク

緒　言

生体は、推定 37 兆個もの細胞が密接に連携しあうことで、個体としての機能を保っている。この機能の欠損は、個体の生命機能に大きな障害をもたらし、疾患の発症につながる。この膨大な数の細胞の持つ個々の機能とそれらの相互作用の詳細を明らかにするため、形態、発現遺伝子／タンパク質マーカー等による「細胞集団」への分類と、その集団に対する分子生物学的・遺伝学的介入が行われてきた。例えば脳においては、その構成細胞は神経細胞とグリア細胞に大別することができる。神経細胞はさらに、パルバルブミン発現神経やソマトスタチン発現神経など、単一遺伝子マーカーの発現を基に分類され、それぞれの生理的役割が解明されてきている。我々はこれまでに、不安などの情動制御に関わるセロトニン神経を標的として、選択的に外来遺伝子を導入できるウイルスベクターを開発し、セロトニン神経活動を高めることで抗うつ効果が得られることを見出し報告してきた［1，2］。しかし近年、均一とみなされてきた神経集団の中に、複数の亜集団が存在することが、単一細胞の網羅的遺伝子発現解析（scRNA-seq）およびオープンクロマチン領域の網羅的解析（scATAC-seq）の結果から明らかになりつつある［3，4］。この急速に同定が進む神経亜集団の機能とその相互作用の詳細を完全に解明するためには、その集団に選択的に遺伝子を発現させるための手法が必須となる。しかし、この様なラベリング手法は脳に限らずどの生体組織においても確立されておらず、scRNA-seq および scATAC-seq により存在が示されつつある細胞亜集団の機能は、特異的遺伝子マーカーが発見された一部を除いて全く明らかになっていない。

近年、深層学習に代表される情報学的手法の長足の進歩を背景として、核酸配列の機能予測手法が開発されている。Kelley らは 1-ホット表現へと変換した核酸配列に対して、畳み込み層と指数関数的膨張畳み込み層（Exponential dilating convolve filters）を連続的に適用することで、オープンクロマチン領域やクロマチン免疫沈降-シーケンス（ChIP-seq）の結果を予測可能な手法を開発し Basenji の名称で報告している［5］。また、Zhou らは 2,002 種類のヒストン修飾、クロマチンアクセシビリティおよび転写因子結合の実アッセイデータを学習させた制御因子予測モデルと、組織における遺伝子発現プロファイルとを組み合わせることで、組織特異的な制御因子活性予測アルゴリズムが構築可能であることを報告している［6］。これらの報告は、十分な学習データが存在すれば核酸配列のみからの制御因子活性予測が可能であることを強く示唆している。

そこで本研究では、細胞亜集団の制御手法の開発を目指し、新たな制御因子活性網羅的測定手法の開発を情動制御において重要なセロトニン神経核である背側縫線核を一例として実施した。さらに情報科学的アプローチから任意の核酸配列の制御因子活性を予測する手法の開発を実施した。

方　法

1. ATAC-STARR-seq

エンハンサー活性が特に高いゲノム配列を濃縮するため、マウス背側縫線核から ATAC 法［7］によりオープンクロマチン領域配列を単離した。単離した配列は制限酵素サイトを付加した PCR プライマーにより増幅したのち精製しクローニングに用いた。単離した配列のエンハンサー活性を網羅的に測定するため、最小プロモーター下に蛍光タンパク

質、エンハンサー配列およびポリ A 付加配列をこの順に配置したプラスミドライブラリを構築した。これにより、エンハンサーの活性を mRNA 中の比率として測定可能とした。本研究ではこれをマウス脳内で実施するため、上記のカセットをアデノ随伴ウイルスベクター（AAV）の ITR 配列間に挿入したプラスミドライブラリを作製したのち、AAV のパッケージングおよび精製を行った［2］。精製した AAV 1 μL をマウス背側縫線核（AP −4.4 mm、ML +0 mm、DV +3.4 mm from the Bregma）および前頭前皮質（AP +1.8 mm、ML −0.3 mm、DV +2.5 mm from the Bregma）に投与した。4 週間の回復後、背側縫線核および前頭前皮質を単離し total RNA を回収した。オリゴ dT プライマーを用いた逆転写反応後、エンハンサー配列前後に付加したアダプター配列に特異的なプライマーで PCR 増幅を行ったのち精製し、次世代シーケンサーでの解析に供した（PE150、Novaseq 6000、Illumina）。得られたリードはアダプター配列の除去ののち、Bowtie2 を用いてマウスゲノム DNA 配列（mm10）へとマッピングした。ピークコーリングおよび背側縫線核および前頭前皮質サンプル間のピークの定量比較には HOMER を用いた。

2. 制御因子活性予測

マウスゲノム配列（mm10）から、全遺伝子の転写開始点上流 20 kb および下流 20 kb の配列を取得したのち、既報［6］と同様の前処理を行い、2,002 次元×201 ビン（200 bp／ビン）のクロマチン特徴量を算出した。線条体の単一細胞 RNA-seq 解析を行った結果（GSE82187）を、Seurat を用いて解析し、神経細胞と推定される細胞群をさらに 7 つのクラスターに分離した。各クラスターに含まれる細胞の平均の遺伝子発現量を算出し、上記のクロマチン特徴量と並べたテーブルを作成した。深層学習を用いたクロマチン特徴量からの遺伝子発現予測には、Tensorflow およびそのラッパーである Keras を用いた。最適なハイパーパラメーターを得るため、GPyOpt によるガウス過程最適化を行い 1,000 イテレーションの最適化を実施しホールドアウトセットに対する予測精度が最も良い条件を探索した。比較のため既報［6］同様に XGBoost を用いた遺伝子発現予測を行った。

結果および考察

1. ATAC-STARR-seq 法の組織間比較による細胞集団選択的エンハンサー候補配列の網羅的探索

ATAC-STARR-seq 法は［8］、オープンクロマチン領域を濃縮しその活性を評価することが可能であるが、ATAC 法の過程での増幅効率の差が、見かけ上の活性の差につながってしまうという欠点がある。また解析対象とする組織 1 種類のみの解析では、ユビキタスに低活性なエンハンサーと少数の細胞で高活性なエンハンサーの区別がつかないという点は、本研究の目的である細胞集団選択的エンハンサーの単離には適さないと推定された。

そこで我々は、2 つの脳神経核の ATAC-STARR-seq 法の結果の差分に注目する Diff-ATAC-STARR-seq 法を考案した（図 1a）。2 つの神経核を、大部分の細胞集団組成に重なりをもたせながら、一部の細胞集団については 1 種類の脳神経核にしか存在しないように選択することで、1 種類の神経核にしか存在しない細胞集団に選択的なエンハンサー配列を単離できると考えられた。まず背側縫線核から細胞核を精製したのち方法に記載の手法を用いて、背側縫線核のオープンクロマチン領域を含むエンハンサーAAV ライブラリを作製した。AAV ライブラリを背側縫線核および対照領域として前頭前皮質に投与したのち、神経核を回収しエンハンサー活性の定量を行った。

縫線核に由来する AAV ライブラリであることから、まず縫線核が含まれる中脳領域の細胞種決定に重要な転写因子群［9］の結合配列が濃縮されているか否かを、HOMER を用いて検討したところ、神経細胞の細胞種決定に重要である CTCF、Nfil3 および Rfx1 について AAV ライブラリ中で高度に濃縮されていることが明らかになった（図 1b）。次に背側縫線核および前頭前皮質から RNA を回収し、逆転写反応ののちエンハンサー配列前後に付加したアダプター配列に特異的なプライマーで PCR 増幅を行い得られた産物を次世代シーケンサーで解析した。得られたリード配列をマウスゲノム配列にマッピングし両神経核で比較した（図 1c）。予想通り、両神経核でのピークは高度に類似しており、単独の神経核の解析では細胞集団選択的エンハンサー候補の単離が困難であることが示唆される。そこで、両神経核の結果を定量比較し、背側縫線核のみで高度に発現するエンハンサーの同定を試みた。その結果、背側縫線核サンプルで 30 倍以上の濃縮が見られ、全ての生物学的反復において再現されたエンハンサー領域を 6 か所同定することに成功した（図 2）。本結果は、今回開発した手法の高い有用性を示すものと考えられた。

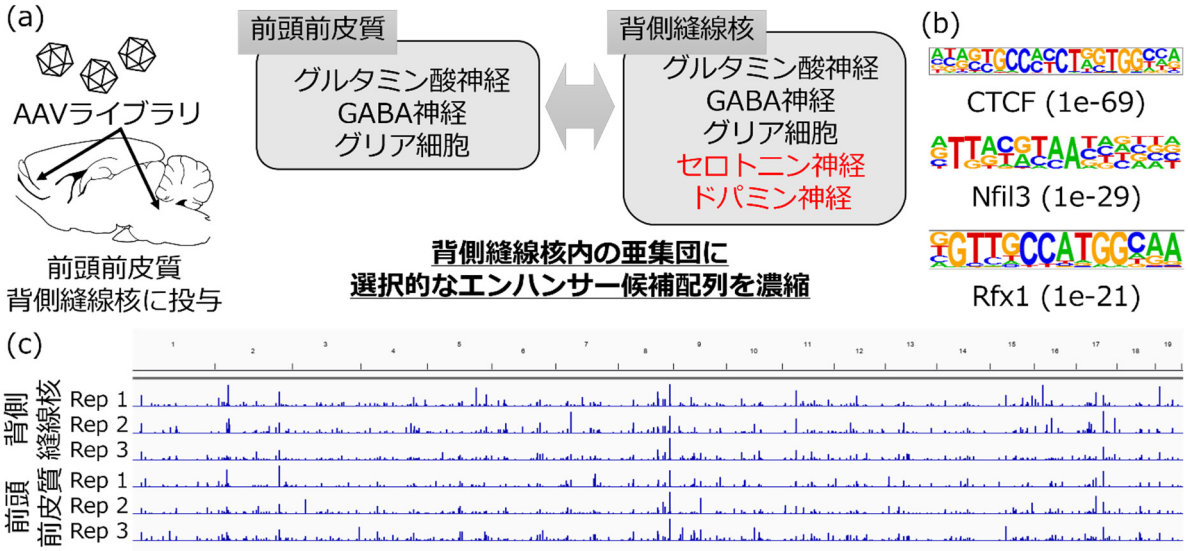


図 1. Diff-ATAC-STARR-seq 法による縫線核亜集団選択的エンハンサー濃縮

a) Diff-ATAC-STARR-seq 法の概略図。

b) AAV ライブラリ中エンハンサーのモチーフ解析。転写因子名および濃縮度の P 値。

c) 前頭前皮質および背側縫線核における ATAC-STARR-seq の結果。それぞれの脳領域について 3 生物学的反復を行いその結果を示した。染色体番号の小さい順に左から右に記載した。

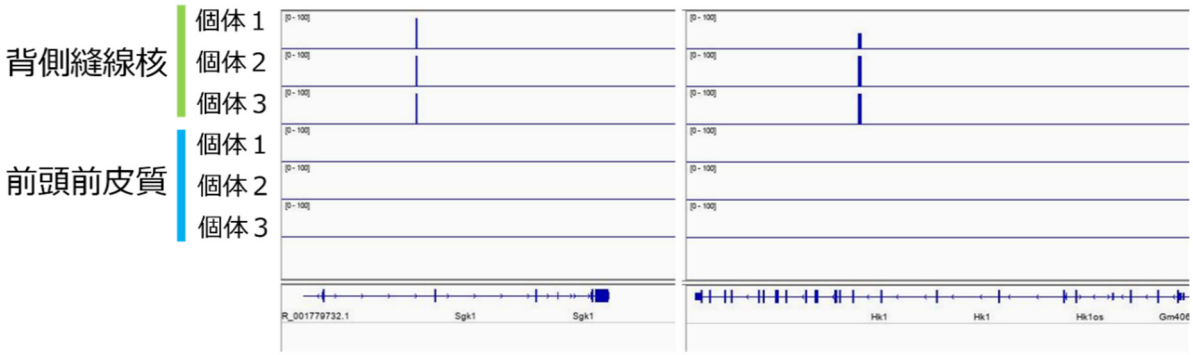


図 2. Diff-ATAC-STARR-seq 法で同定された背側縫線核選択的エンハンサー

背側縫線核および前頭前皮質サンプルから得られたリード配列をマウスゲノム配列にマッピングした。左には Sgk1 遺伝子座におけるエンハンサーを右には Hk1 遺伝子座におけるエンハンサーを示した。いずれも背側縫線核のみで高度な発現が検出された。

2. 制御因子活性網羅的予測手法の開発

網羅的な制御因子活性測定によっても、測定可能な制御因子配列の種類には限界が存在する。AAV のウイルスタイターは通常 10^{10} gc/μL 程度であり、同一配列を 100 細胞に感染させ評価するためには、制御因子を 10^7種類以下に抑える必要がある。一方、わずか 1 kb の核酸配列であってもその取り得る配列の種類は $4^{1,000}$（≒10^{602}）にも達することから、制御因子の網羅的実アッセイは現実的ではない。そこで本研究では、深層学習などの情報科学的手法を用いることで、核酸配列のみからその制御因子活性予測が可能なのではないかと考え検討を行った（図 3a）。できる限り均一な細胞集団からの遺伝子発現量を得るために、単一細胞 RNA-seq 法による神経細胞のクラスタリングと得られたクラスターの平均遺伝子発現量の算出を行った。既報に基づき［6］、各遺伝子の転写開始点上流および下流それぞれ

20 kb の配列からクロマチン特徴量を計算した。遺伝子ごとのクロマチン特徴量を入力として、前述の遺伝子発現量を出力するように深層学習モデルの学習を行った。手法ごとに最適なハイパーパラメーターは異なると考えられる一方で、グリッドサーチによる網羅的探索は計算時間の観点から現実的ではない。そこで限られた計算時間で予測精度の高いハイパーパラメーターに到達し得るガウス過程最適化を実施した。その結果、画像同様の 2 次元の畳み込みフィルターを多層に重ねた手法が、その他の手法よりも高い精度で予測を行えることが明らかになった（図 3b）。また学習に使わなかった評価データセットについて、構築した予測モデルを用いた遺伝子発現予測値と実測値の相関を調べたところ、そのスピアマンの相関係数は 0.607（$P=2.26\times10^{-90}$）と有意な相関を示すことが明らかになった。これらの結果は、本予測手法の高い性能と妥当性を示すものと考えられた。今後、前項で得られた制御因子活性の網羅的測定手法と組み合わせてさらなる学習を行わせることで、あらゆる核酸配列の制御因子活性の予測が可能になると考えられる。

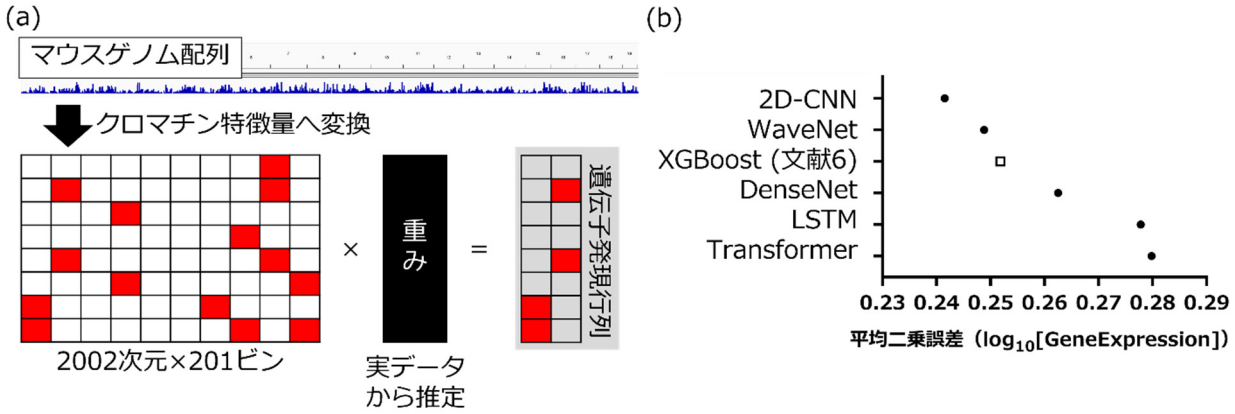


図 3. 制御因子活性予測手法の開発とその評価

a) 開発手法の概略。各遺伝子につき転写開始点の上流および下流それぞれ 20 kb ずつの配列を、既報［6］に基づきクロマチン特徴量へと変換した。変換後のクロマチン特徴量行列を入力とし、線条体神経細胞の各クラスターにおける遺伝子発現量を出力するような重み行列を学習により得る。

b) 異なる重み行列を用いた際の評価データセットに対する予測精度。平均二乗誤差が小さいほど予測精度が高いことを示す。ガウス過程最適化を行った際の各手法の最良値をプロットしている。

共同研究者・謝辞

本研究の共同研究者は、京都大学大学院薬学研究科生体機能解析学分野の金子周司教授である。本研究は、上原記念生命科学財団研究奨励金の支援を受け実施された。この場を借りて厚く御礼申し上げる。

文　献

1） Nishitani N, Nagayasu K, Asaoka N, Yamashiro M, Andoh C, Nagai Y, Kinoshita H, Kawai H, Shibui N, Liu B, Hewinson J, Shirakawa H, Nakagawa T, Hashimoto H, Kasparov S, Kaneko S. Manipulation of dorsal raphe serotonergic neurons modulates active coping to inescapable stress and anxiety-related behaviors in mice and rats. Neuropsychopharmacology. 2019 Mar;44(4):721-732. Epub 2018 Oct 30. PubMed PMID: 30377380 DOI: 10.1038/s41386-018-0254-y.

2) Nagai Y, Takayama K, Nishitani N, Andoh C, Koda M, Shirakawa H, Nakagawa T, Nagayasu K, Yamanaka A, Kaneko S. The Role of Dorsal Raphe Serotonin Neurons in the Balance between Reward and Aversion. Int J Mol Sci. 2020 Mar 21;21(6). PubMed PMID: 32245184 DOI: 10.3390/ijms21062160.
3) Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, Filippova GN, Huang X, Christiansen L, DeWitt WS, Lee C, Regalado SG, Read DF, Steemers FJ, Disteche CM, Trapnell C, Shendure J. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell. 2018 Aug 23;174(5):1309-1324.e18. Epub 2018 Aug 2. PMID: 30078704 DOI: 10.1016/j.cell.2018.06.052.
4) Rosenberg AB, Roco CM, Muscat RA, Kuchina A, Sample P, Yao Z, Graybuck LT, Peeler DJ, Mukherjee S, Chen W, Pun SH, Sellers DL, Tasic B, Seelig G. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science. 2018 Apr 13;360(6385):176-182. Epub 2018 Mar 15. PMID: 29545511 DOI: 10.1126/science.aam8999.
5) Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Res. 2018 May;28(5):739-750. Epub 2018 Mar 27. PMID: 29588361 DOI: 10.1101/gr.227819.117.
6) Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat Genet. 2018 Aug;50(8):1171-1179. Epub 2018 Jul 16. PMID: 30013180 DOI: 10.1038/s41588-018-0160-6.
7) Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013 Dec;10(12):1213-8. Epub 2013 Oct 6. PMID: 24097267 DOI: 10.1038/nmeth.2688.
8) Wang X, He L, Goggin SM, Saadat A, Wang L, Sinnott-Armstrong N, Claussnitzer M, Kellis M. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. Nat Commun. 2018 Dec 19;9(1):5380. PMID: 30568279 DOI: 10.1038/s41467-018-07746-1.
9) Gui Y, Grzyb K, Thomas MH, Ohnmacht J, Garcia P, Buttini M, Skupin A, Sauter T, Sinkkonen L. Single-nuclei chromatin profiling of ventral midbrain reveals cell identity transcription factors and cell-type-specific gene regulatory variation. Epigenetics Chromatin. 2021 Sep 9;14(1):43. PMID: 34503558 DOI: 10.1186/s13072-021-00418-3.