**T.C**

**İNÖNÜ ÜNİVERSİTESİ**

**FEN BİLİMLERİ ENSTİTÜSÜ**

# UZAMSAL TRANSKRİPTOMİKS ANALİZLERİ İÇİN TEK HÜCRE RNA SEKANSLAMA VERİLERİNDEN HAREKETLE İŞARETÇİ GEN SEÇİMİ YAPAN YENİ BİR YÖNTEMİN GELİŞTİRİLMESİ

**YÜKSEK LİSANS TEZİ**

**Yusuf BARAN**

**Biyomedikal Mühendisliği Anabilim Dalı**

**Tez Danışmanı: Doç. Dr. Berat DOĞAN**

**ARALIK 2022**

**T.C**

**İNÖNÜ ÜNİVERSİTESİ**

**FEN BİLİMLERİ ENSTİTÜSÜ**

# UZAMSAL TRANSKRİPTOMİKS ANALİZLERİ İÇİN TEK HÜCRE RNA SEKANSLAMA VERİLERİNDEN HAREKETLE İŞARETÇİ GEN SEÇİMİ YAPAN YENİ BİR YÖNTEMİN GELİŞTİRİLMESİ

**YÜKSEK LİSANS TEZİ**

**Yusuf BARAN**

(36203630006)

**Biyomedikal Mühendisliği Anabilim Dalı**

**Tez Danışmanı: Doç. Dr. Berat DOĞAN**

**ARALIK 2022**

# TEŞEKKÜR VE ÖNSÖZ

Bu çalışmada in situ uzamsal transkriptomik yöntemlerinde kullanılmak üzere işaretçi gen seçimi yapan yeni bir algoritma önerilmiş ve yazılım paketi haline getirilerek araştırmacıların kullanımına sunulmuştur. Çalışmamı tamamlamamda desteklerini ve yardımlarını esirgemeyen saygıdeğer hocalarım, ailem ve arkadaşlarıma sonsuz şükranlarımı sunarım.

Bu tezin gerçekleştirilmesinde, kıymetli bilgilerini ve zamanını benimle her zaman paylaşan, ne zaman olursa olsun sorularıma cevap veren, tezim haricinde bildiğim birçok şeyi kendisinden öğrendiğim, sadece bilimsel anlamda değil her anlamda bana destek olan ve teşvik eden, hoca kelimesinin hakkını her anlamda veren saygıdeğer hocam Doç. Dr. Berat Doğan'a; derslerini her zaman merakla ve zevkle dinlediğim hocam Doç. Dr. Olcay Kızılaslan'a ve bizlere öğretmek için her zaman elinden gelenin fazlasını yapan çaba gösteren Dr. Öğr. Üyesi Reyhan Zengin'e sonsuz şükranlarımı sunar ve teşekkürü borç bilirim. Ayrıca bu süreçte her derdimde yanımda olan sevgili arkadaşlarım Gökhan Kesin ve Uygar Aslan'a da teşekkür ederim.

Hayatım boyunca her zaman yanımda olan, hiçbir zaman desteklerini esirgemeyen sevgili anneme, babaanneme ve halama; gece gündüz yollarda bizim için çalışan sevgili babama ve hayatımın neşesi olan sevgili kardeşlerime sonsuz sevgilerimi sunarım.

Kasım 2022 Yusuf BARAN

## ONUR SÖZÜ

Yüksek lisans tezi olarak sunduğum “Uzamsal Transkriptomik Analizleri İçin Tek Hücre RNA Sekanslama Verilerinden Hareketle İşaretçi Gen Seçimi Yapan Yeni Bir Yöntemin Geliştirilmesi” başlıklı bu çalışmanın bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurmaksızın tarafımdan yazıldığına ve yararlandığım büyün kaynakların hem metin içinde hem de kaynakçada yöntemine uygun biçimde gösterilenlerden oluştuğunu belirtir, bunu onurumla doğrularım.

Yusuf BARAN

# İÇİNDEKİLER

# ÇİZELGE LİSTESİ

# ŞEKİL LİSTESİ

## SEMBOLLER VE KISALTMALAR

**$\mu m$** **:** Mikro Metre

**RNA** **:** Ribonucleic Acid

**DNA** **:** Deoxyribonucleic Acid

**mRNA** **:** Messenger RNA

**tRNA** **:** Transfer RNA

**ncRNA** **:** non-coding RNA

**RNA-seq** **:** RNA Sequencing

**NGS** **:** Next Generation Sequencing

**scRNA-seq** **:** Single Cell RNA Sequencing

**FACS** **:** Fluorescence-activated Cell Sorting

**LCM** **:** Laser Capture Microdissection

**UMI** **:** Unique Molecular Identifier

**cDNA** **:** Complementary DNA

**IVT** **:** In Vitro Transcription

**FISH** **:** Fluorescence In Situ Hybridization

**smFISH** **:** single-molecule FISH

**seqFISH** **:** sequential FISH

**MERFISH** **:** Multiplexed Error-Robust FISH

**osmFISH** **:** cyclic-ourobos smFISH

**RF** **:** Random Forest

**BRF** **:** Balanced Random Forest

**DNN** **:** Deep Neural Network

**PCA** **:** Principal Component Analysis

**CLI** **:** Command Line Interface

**IDE** **:** Integrated Development Environment

**M** **:** Milyon

**Gb** **:** Giga Byte

**Mb** **:** Mega Byte

# ÖZET

Yüksek Lisans Tezi

UZAMSAL TRANSKRİPTOMİKS ANALİZLERİ İÇİN TEK HÜCRE RNA SEKANSLAMA VERİLERİNDEN HAREKETLE İŞARETÇİ GEN SEÇİMİ YAPAN YENİ BİR YÖNTEMİN GELİŞTİRİLMESİ

Yusuf BARAN

İnönü Üniversitesi

Fen Bilimleri Enstitüsü

Biyomedikal Mühendisliği Anabilim Dalı

91+X sayfa

2022

Danışman: Doç. Dr. Berat DOĞAN

Son birkaç yılda geliştirilen teknolojiler ile birlikte artık herhangi bir anda hücre içerisindeki transkriptomun uzamsal olarak analiz edilebilmesi mümkün hale gelmiştir. Hücrelerin gen ifadesine ilişkin konum bilgisinin artık elde edilebiliyor olması hücrelerin uzamsal olarak birbirleriyle olan ilişkilerini ve dolayısıyla hücresel düzeyde vücutta meydana gelen birçok biyolojik süreci daha iyi anlamamıza imkân tanımaktadır. Ancak gelişmekte olan uzamsal teknolojilerin deneysel maliyetleri hala yüksektir ve bu teknolojilerin bazıları deney öncesinde hücre tiplerini birbirlerinden ayırt edebilen kısıtlı sayıdaki işaretçi genin önceden belirlenmesine ihtiyaç duymaktadır. Bu tez kapsamında literatürde in situ uzamsal transkriptomik deneyleri için gen seçimi işlemini gerçekleştiren yöntemler incelenmiş, eksiklikleri belirlenmiş ve eksiklikleri gidermek için bir yöntem geliştirilmiştir. Geliştirilen yöntem literatürde önerilen diğer yöntemler ile birçok açıdan karşılaştırılmış, üstünlükleri ve eksiklikleri tartışılmıştır. Sonuçlar, önerilen yöntemin değerlendirilen birçok parametre için mevcut yöntemlerden üstün olduğunu açıkça göstermektedir. Ayrıca önerilen yöntem açık erişimli bir yazılım paketi haline getirilerek araştırmacıların kullanımına sunulmuştur.

**Anahtar Kelimeler:** Uzamsal Transkriptomik, İşaretçi Gen Seçimi, Prob Seti Seçimi

## ABSTRACT

Master Thesis

A NEW MARKER SELECTION METHOD FROM SINGLE CELL RNA SEQUENCING DATA FOR SPATIAL TRANSCRIPTOMICS ANALYSES

Yusuf BARAN

Inonu University

Graduate School of Nature and Applied Sciences

Department of Biomedical Engineering

91+X pages

2022

Supervisor: Assoc. Prof. Dr. Berat DOĞAN

Recent studies showed that cell-cell communication regulated by biochemical signals is an important aspect of tissue structure and function and regulates individual cellular processes. Therefore, knowing the spatial positions of cells in the tissue is critical to understanding normal development and disease pathology. With the help of the recent technological developments, it has now become possible to spatially analyze the transcriptome of a cell at any time. However, the experimental costs of emerging spatial transcriptomics technologies are high, and some of these technologies require predetermination of the limited number of marker genes that can distinguish cell types from each other prior to experimentation. Within the scope of this thesis, the methods that perform the gene selection process for in situ spatial transcriptomics experiments in the literature were examined, the deficiencies were determined, and a method was developed to eliminate the deficiencies. The developed method was compared with other methods suggested in the literature in many respects, and its advantages and disadvantages were discussed. The results clearly show that the proposed method is superior to existing methods for many parameters evaluated. In addition, the proposed method has been turned into an open access software package and made available to researchers.

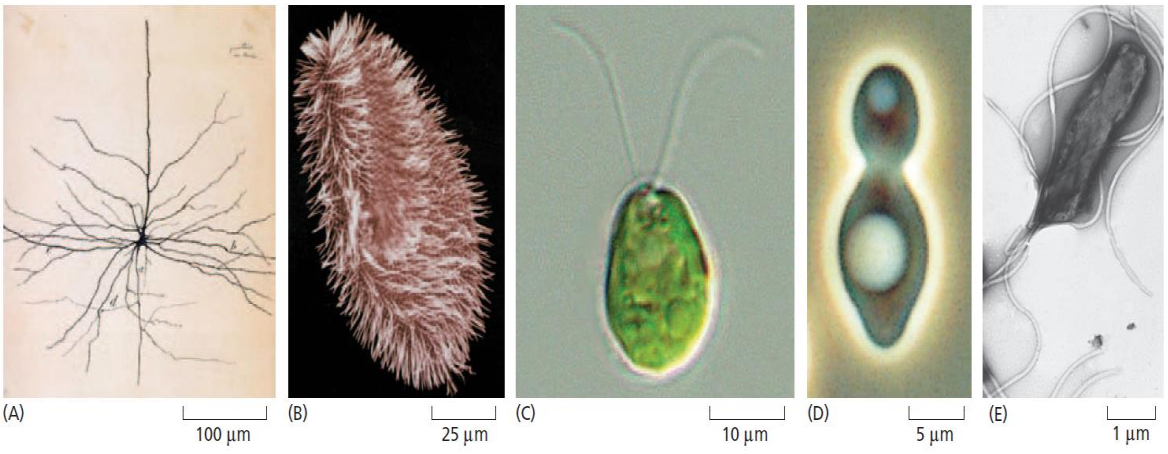


**Keywords:** Spatial Transcriptomics, Marker Gene Selection, Prob Set Selection

# 1. GİRİŞ

Dünya'da yaklaşık 100 milyon çeşit kompleks veya basit canlının yaşadığı tahmin edilmektedir ve bu organizmaların temel yapı taşı birçok farklı çeşidi bulunan hücrelerdir [1]. Şekil 1.1'de ilkel ve gelişmiş canlılardan elde edilen farklı hücre türlerinin sadece yapısal anlamda bile ne kadar farklı oldukları açıkça görülmektedir. Yaşayan canlılar arasındaki en kompleks tür olan insan ise birçok farklı hücre tipini içeren yaklaşık 40 trilyon hücreden meydana gelmektedir [2].

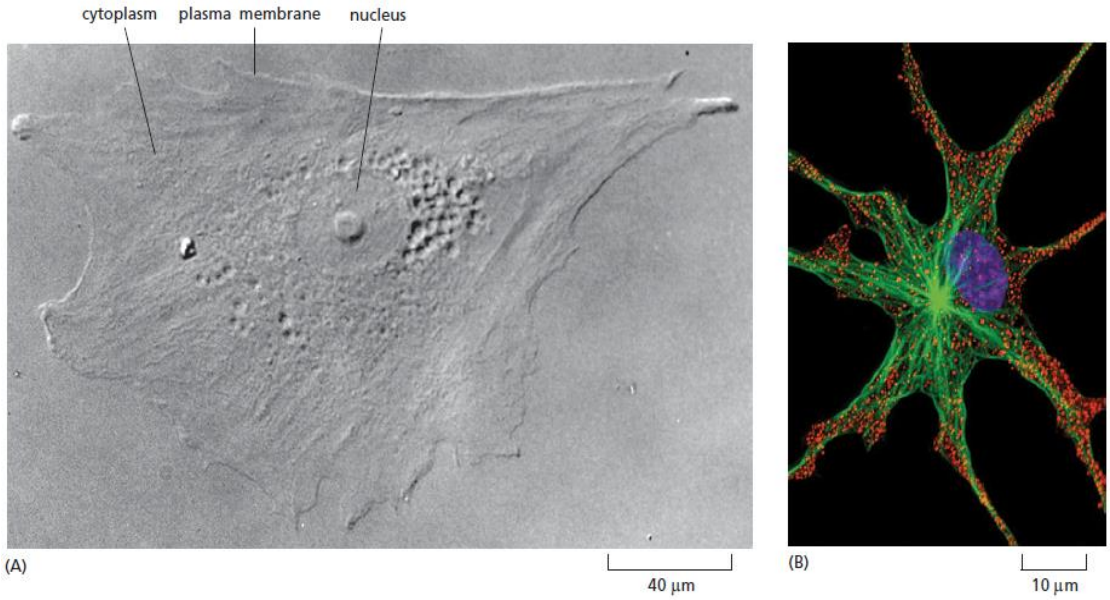
Farklı yaşam türlerinin nasıl oluştuğunu, çalıştığını, adapte olduğunu veya geliştiğini ortaya çıkarmak, diğer yandan bu türleri tehdit eden hastalıkların bu sistemleri nasıl etkilediklerini anlamak ve etkilenen mekanizmalara yönelik tedaviler geliştirebilmek gibi birçok konunun araştırılması, yaşamın temel yapı birimi olan hücrenin incelenmesinden geçmektedir. Hücrelerin en temelden yani moleküler düzeyde sistemlerinin, organizasyonlarının ve dolayısıyla birbirleriyle olan ilişkilerinin incelenmesi çözülememiş birçok soruya ışık tutmakta ve insanoğlunun kompleks mekanizmasını anlamamıza yardımcı olmaktadır.




**Şekil 1.1:** Farklı boyut ve biçimlerde hücre tipleri a) *Memeli Sinir Hücresi* b) *Paramecium* c) *Chlamydomonas* d) *Saccharomyces Cerevisiae* e) *Helicobacter Pylori* [1].

## 1.1 RNA, RNA Üretimi ve Ölçüm Yöntemleri

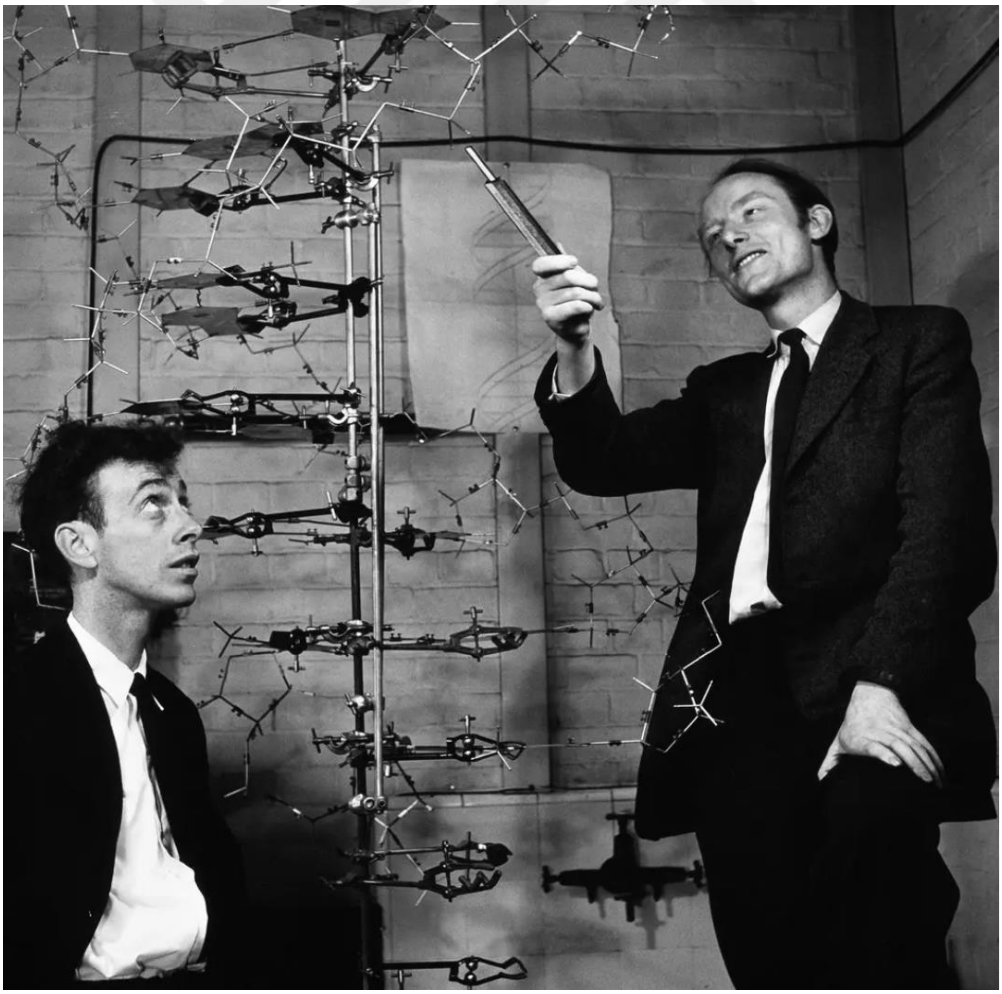
İnsan gibi karmaşık ve kompleks organizmalar, sadece spesifik görevleri olan veya genel görevler yapmak üzere modifiye edilmiş birçok hücre tipinden meydana gelmektedir. Örneğin sperm ve yumurta hücreleri organizmanın devamı için gereken genetik bilgiyi bir sonraki nesle aktarma görevini üstlenmişlerdir. Bu hücreler organizmadan çıkarıldığında tek başına bir anlam ifade etmese de çok hücreli bir organizmanın içerisinde canlının soyunun devam ettirilmesi gibi çok önemli bir görevi ifa etmektedirler. Sistemin işlemesi için, üretmek üzere modifiye olmuş bazı hücre tipleri ise hormonlar, yağlar, pigmentler veya nişasta gibi sadece belirlenmiş spesifik maddeleri oluşturmak üzere tasarlanmış fabrikalar gibidir. Veya başka bir hücre tipi olan kaslar, adeta motorlar gibi fiziksel işler yapmak için diğer hücreler tarafından üretilen yakıtları yakar ve kinetik enerjiye çevirirler. Diğer bir tür olan yılan balıklarında ise kaslar üretilen yakıtları sadece kinetik enerjiye değil, aynı zamanda elektriksel enerjiye de çeviren kas hücrelerine sahiptir ve bu hücreler adeta elektrik jeneratörleri gibi işlev görmektedirler [1]. Nihai olarak birçok farklı hücre tipi büyük bir yaşam döngüsünde farklı görevler üstlenirler ve hepsi döngünün devamı için birbirlerine bağlıdır.




**Şekil 1.2:** Yaşayan hücrelerin mikroskop görüntüleri a) İnsan cildinden alınan bir hücre b) Floresan boyalarla boyanmış bir hücre [1].

Farklı hücre tipleri organizmalar içerisinde farklı görevler üstlense de aslında her ökaryotik hücre Şekil 1.2 (a)' da görülen çekirdek olarak isimlendirilen, zarla kaplı bir organelin içinde

DNA (Deoxyribonucleic Acid) adı verilen aynı genetik materyali içermektedir [3]. DNA ilk olarak İsviçreli bir bilim adamı olan Johann Friedrich Miescher tarafından 'nuclein' olarak tanımlanmıştır. Daha sonra Şekil 1.3'te görülen Watson ve Crick DNA'nın sarmal yapısını keşfetmiş ve Francis Crick transkripsiyon ve translasyon mekanizmaları hakkında çok önemli keşiflerde bulunmuştur [4]. DNA aslında her hücre içerisinde bulunan bir kitaplık olarak değerlendirilebilir. Ancak bu kitaplıktan her hücre kendisi ile ilgili olan bölümleri okumakta ve verilen görevleri uygulamaktadır. Bu bölümlerin okunması ve hayata geçirilmesi ise transkripsiyon, translasyon yani protein üretimi gibi başlıkları içermektedir. RNA (Ribonucleic Acid); mRNA (messenger RNA) olarak protein üretimi, prokaryotik ncRNA (non-coding RNA) olarak transkripsiyon ve translasyon, ökaryotik ncRNA olarak ise gen susturma, epigenetik düzenleme gibi çeşitli görevlerle ilgilenmektedir [5]. Farklı hücre tipleri görevleri sebebiyle farklı genleri eksprese ederler ve hücrelerin RNA ekspresyon karakterlerini incelemek bu hücre tiplerini karakterize etmemize ve yukarıda bahsedilen birçok mekanizmayı anlamamıza olanak sağlar.

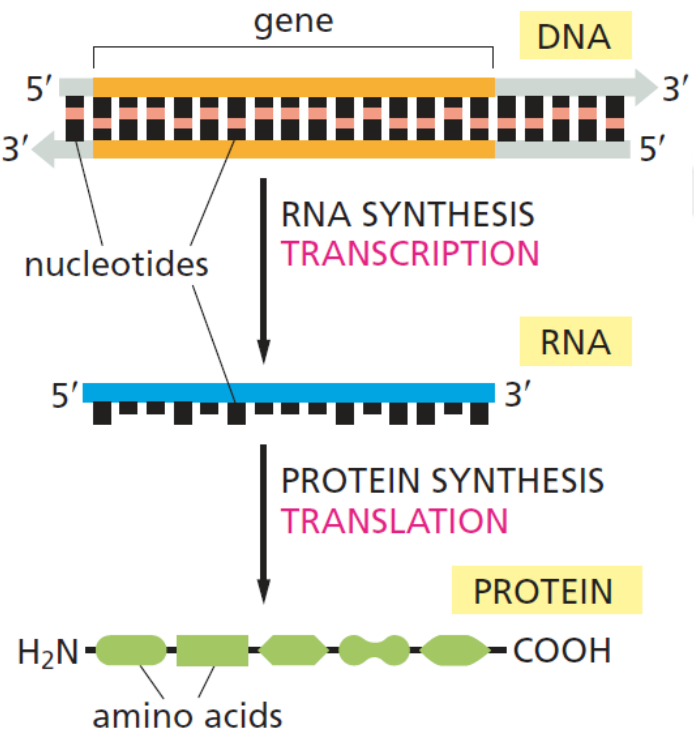


**Şekil 1.3:** James Watson, Francis Crick ve önerdikleri "Double Helix" DNA modeli [6].

### 1.1.1 RNA üretimi

Hücreler anlık ihtiyaçlarına bağlı olarak DNA'yı baz alarak hücrenin temel bileşeni olan proteinleri üretmektedirler. 20 temel amino asit dizisi farklı biçimlerde kombine edilerek farklı proteinleri oluşturmakta ve bu diziler DNA'nın içinde bulunan yönergeler sonucu üretilmektedir. DNA'nın kendisi protein sentezlememektedir ancak hücrenin ihtiyacı olduğu durumlarda protein sentezi için bir kılavuz görevi görmektedir. Proteinleri üretecek olan moleküller yani RNA'lar ise DNA'nın ilgili kısımlarından belirli yönergelere göre kopyalanmaktadır.

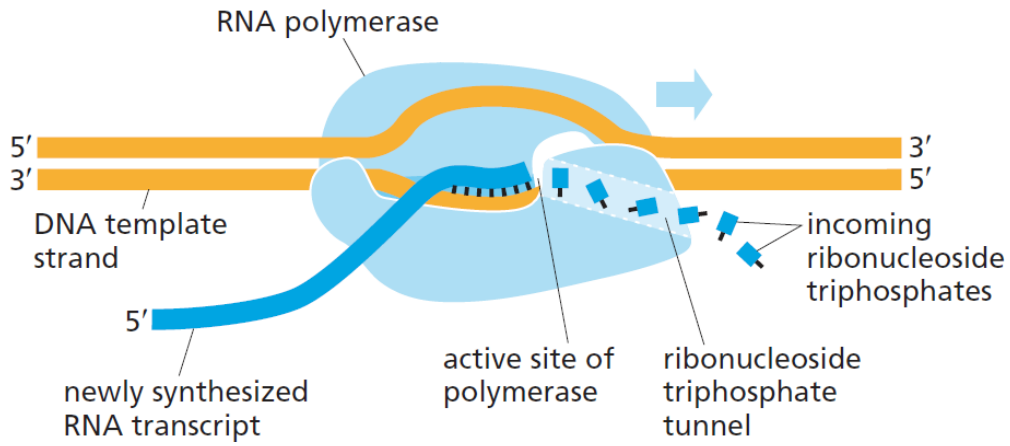
Şekil 1.4'te bu işlemlerin kısa bir özeti görülmektedir. DNA'nın ilgili segmentinin, ki bu kısma gen denir, sarmal yapısı açılarak RNA polimeraz enzimi tarafından kopyalanmaktadır. RNA polimeraz, nükleotidleri karşısındaki DNA'yı rehber alarak teker teker ekleyip birleştirmekte ve RNA transkriptlerini oluşturmaktadır.




**Şekil 1.4:** Protein üretiminin ana aşamaları [1].

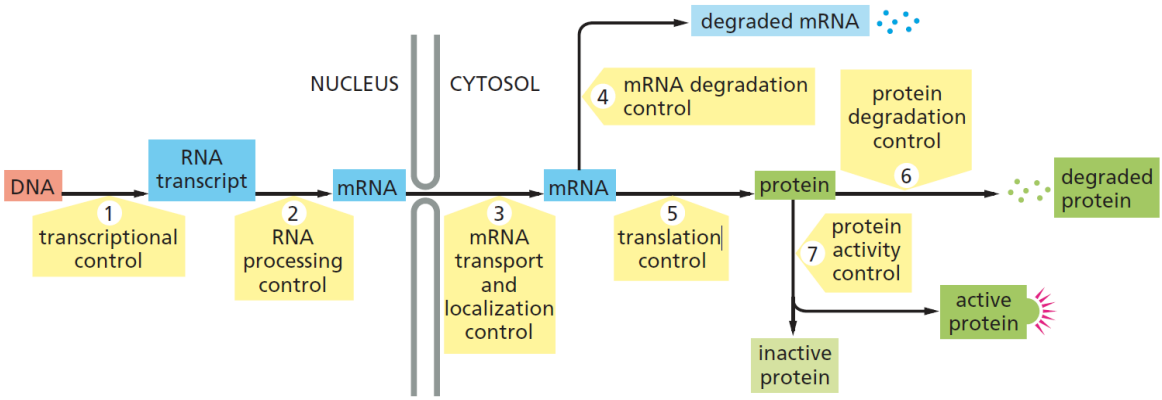
Şekil 1.5'te de görüldüğü gibi RNA polimeraz DNA üzerinde ilerlerken adeta bir fermuarı açar gibi DNA'nın çift sarmal yapısını ayırmakta ve RNA transkriptini sentezlemektedir. Daha sonra DNA yeniden eski haline dönmektedir. Bu olay transkripsiyon olarak isimlendirilmektedir [1]. Burada belirtilen tüm süreç yani transkripsiyon, gen ekspresyonu olarak da adlandırılmakta ve gen ekspresyonu protein üretimi sürecinde bulunan translasyon

adımını içermek zorunda değildir. Özet olarak, gen ekspresyonu DNA dizisindeki direktiflerin organizma veya hücre üzerinde etkisi olan bir çıktıya çevrildiği süreçtir [1].




**Şekil 1.5:** RNA polimeraz tarafından DNA'nın okunması [1].

Hücreleri birbirlerinden farklı kılan gen ekspresyonu Şekil 1.6'te görüldüğü üzere birçok farklı adımda denetlenmektedir. Kısaca 1. adımda genlerin hangi zamanlarda ve hangi sıklıklarda kopyalandığı, 2. adımda üretilen RNA transkriptlerinin nasıl splice edildiği, 3. adımda hangi mRNA'ların çekirdek zarından geçeceği, 4. adımda degredasyon, 5. adımda hangi mRNA'ların proteine dönüştürüleceği, 6. adımda proteinlerin ne kadar sürede yok edileceği, 7. adımda ise proteinlerin aktivite durumları kontrol edilmektedir [1]. Bahsedilen kontrol süreçleri hassas bir işlem olan gen ekspresyonu sürecinin düzgün bir şekilde işlemesi için önemli bir görev üstlenmektedir.



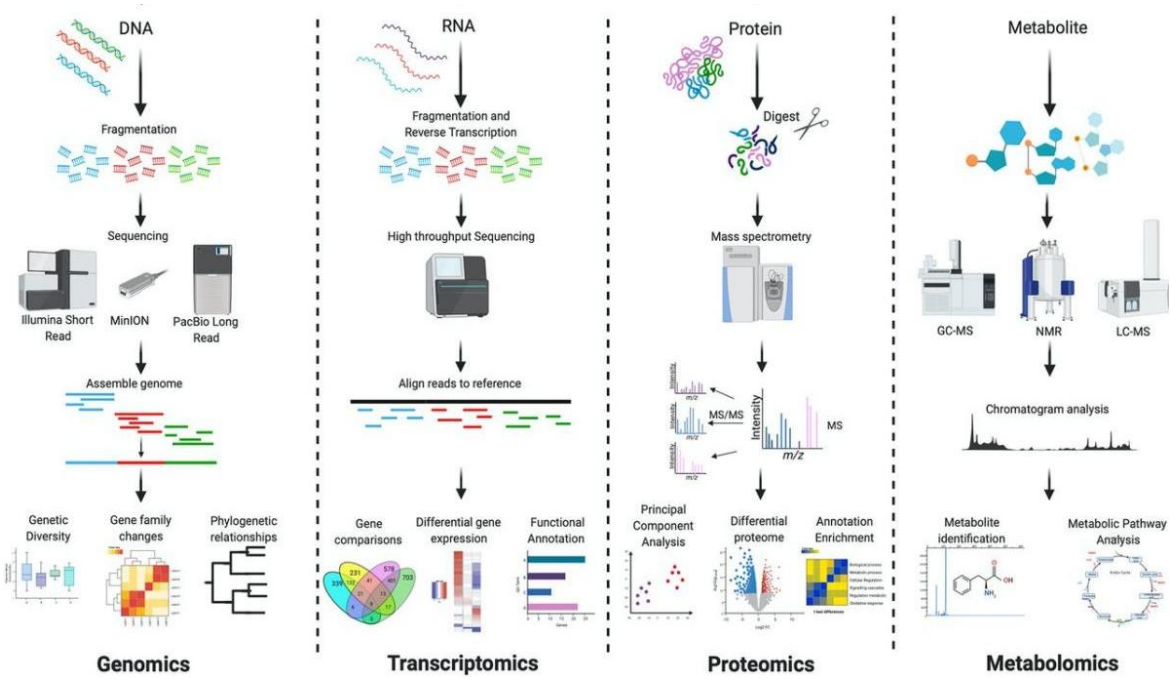

**Şekil 1.6:** Ökaryotik hücrelerde gen ekspresyonu ve protein üretimi adımları [1].

Bir hücrenin DNA'sından taşınan genlerin çoğunluğu proteinlerin yapısını belirlemekte ve bu genler daha önce de bahsedildiği üzere mRNA ismini almaktadır. Ancak transkripsiyon

sonrası proteine çevrilmeyen ncRNA veya small-RNA olarak adlandırılan RNA molekülleri de bulunmaktadır. Bunların hücre içerisinde düzenleyici, yapısal veya katalitik komponent olarak çeşitli rolleri vardır; örneğin rRNA (ribosomal RNA) ribozomların yapısını meydana getirirken, tRNA'lar (transfer RNA) protein üretim sürecinde gerekli amino asit dizilerini taşımaktadırlar. Diğer kodlamayan RNA'lar ise RNA splicing, (uçbirleştirme, transkripsiyon sonrasında RNA'daki kodlamayan bölümlerin çıkartılması) gen regülasyonu (genlerdeki bilginin ürüne dönüşüm süreci), DNA replikasyonu, telomer bakımı ve daha keşfedilmeye devam edilen birçok hücre içi proseslerde görevler almaktadırlar. Sonuç olarak gen ekspresyonu kavramı eksprese edilen, kodlayan veya kodlamayan tüm fonksiyonel gen ürünlerinin üretimini kapsamaktadır [1].

### 1.1.2 Transkriptomiks, RNA-seq ve NGS teknolojileri

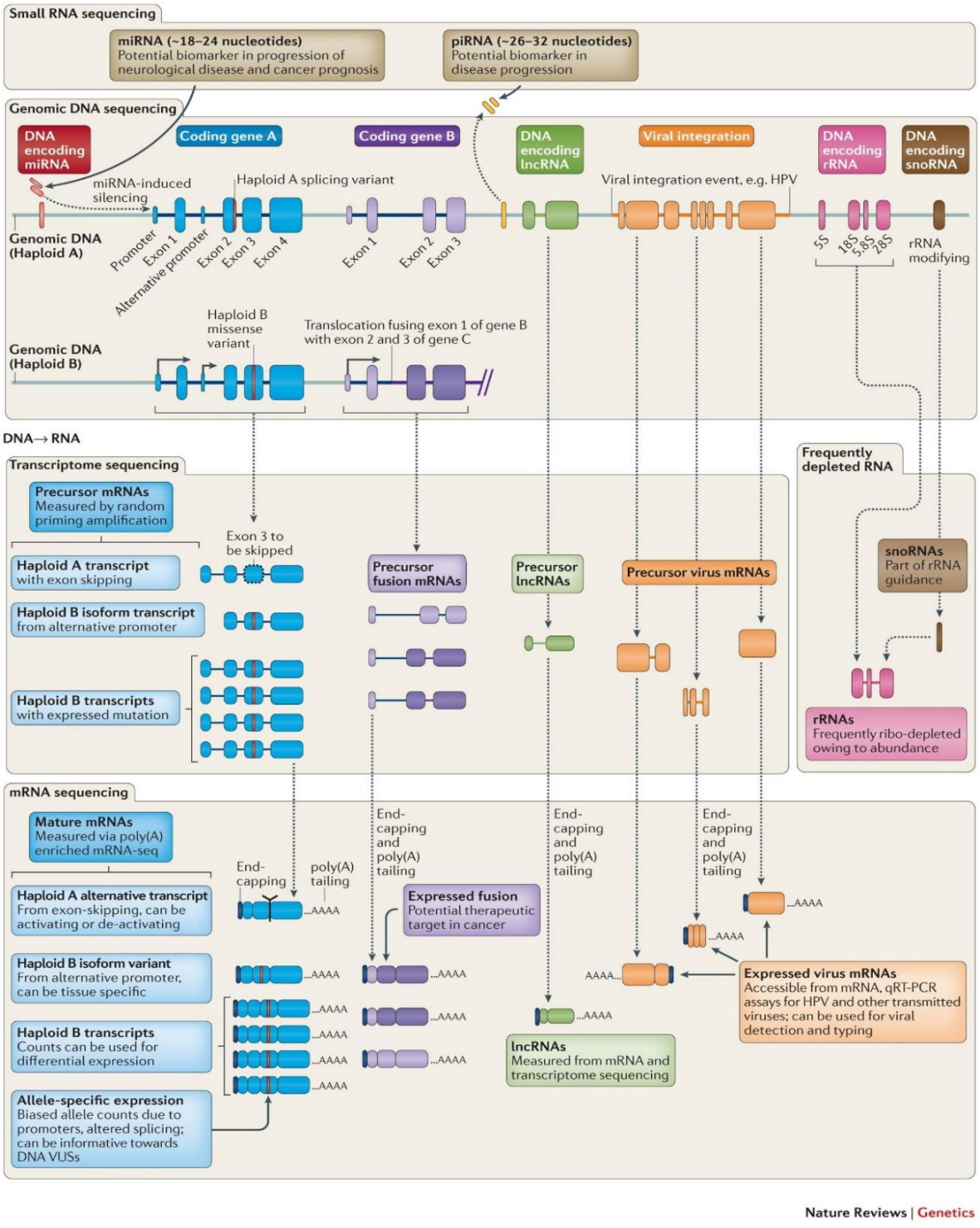
DNA'nın keşfinin ardı sıra teknolojide ve dolayısıyla enstrümanlarda meydana gelen gelişmeler İnsan Genom Projesi'nin (Human Genom Project) önünü açmıştır. Ancak yüksek maliyetler her zaman sekanslama teknolojilerine büyük engeller oluşturmaktaydı [7]. Yüksek verimli (High-Throughput) sekanslama teknolojileri, sekanslama maliyetlerini düşürmüştür. Yeni Nesil Sekanslama (Next Generation Sequencing, NGS) teknolojileri, daha uzun okuma olanakları sağladıkları ve daha ulaşılabilir oldukları için artık kliniklerde bile vazgeçilmez araçlar haline gelmiştir [7].




**Şekil 1.7:** Omik teknolojilerine genel bakış [8].

Yeni nesil sekanslama teknolojileri ile artık gen ve transkriptom düzeyinde hücre içinde meydana gelen biyolojik süreçleri daha iyi anlayabilmekteyiz. Buna ek olarak, gelişen teknoloji ile birlikte protein ve metabolit düzeyinde vücut içerisinde meydana gelen biyolojik süreçler hakkında da artık eskiye nazaran daha çok bilgi sahibiyiz. Bu teknolojilerin bütününe günümüzde omik teknolojileri adı verilmektedir. Bu teknolojilerin ürettiği veriler ise omik veriler olarak isimlendirilmektedir. Şekil 1.7'de günümüzde kullanılan omik teknolojileri kısaca özetlenmiştir. Omik teknolojiler genomik, transkriptomik, proteomik ve metabolomik olarak sıralanabilir. Bu teknolojiler sistem biyolojisi yaklaşımı ile karmaşık biyolojik süreçleri anlamamız açısından büyük önem arz etmektedir [9]. Bu tezin kapsamını ilgilendiren ve yukarıda sözü edilen teknolojilerden biri olan transkriptomiks, bir hücrede belirli bir andaki tüm RNA moleküllerini tespit etmemizi sağlayan teknolojidir. Transkriptomik verisinin anlaşılması, DNA'nın meydana getirdiği eylemlerin anlaşılması anlamına gelmektedir [10]. Çünkü transkriptom DNA kitaplığından hangi kitapların hangi sayfalarının okunduğunun veya eyleme geçirildiğinin bir raporu olarak değerlendirilebilir. Transkriptomik çalışmalarının ana amacı mRNA'lar ve kodlamayan RNA'ları da içermek üzere tüm transkript türlerini ölçmek, kategorize etmek, splicing örüntülerini keşfetmek ve farklı koşullar altında hücrelerin değişen transkript seviyelerini değerlendirmektir [9, 10]. RNA-seq (RNA sequencing) ise transkriptomu profillemek için son zamanlarda geliştirilmiş bir yeni nesil sekanslama teknolojisidir.

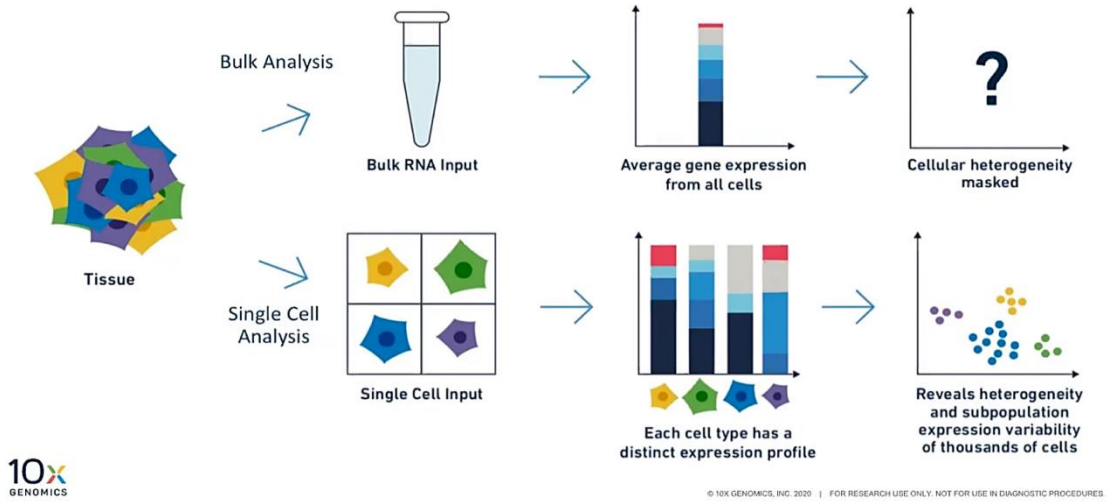
RNA tabanlı ölçümler Şekil 1.8'de de görüldüğü gibi klinik olarak ilgili RNA türlerini ölçmek için kullanılmaktadır ve dolayısıyla klinik uygulamalarda hastalık teşhisi, bulaşıcı hastalıklar veya kanser çalışmaları gibi çeşitli alanlarda geniş uygulama potansiyeline sahiptir. RNA sekanslama insan hastalıklarında; mRNA ekspresyonu profilleme, gen füzyonlarını inceleme, splicing veya yapısal alternatif transkript varyantlarını (Alternative transcripts) tanımlama, alel spesifik ekspresyonlarını (Allele-specific expression ASE) değerlendirme, hücre dışı (Extracellular RNA) veya ncRNA'ları incelenme ve değerlendirme gibi birçok önemli süreçte rol oynar [11]. Ayrıca bulaşıcı hastalık teşhisinde; RNA bazlı patojen diyagnozlar (RNA-based pathogen diagnostics), mikrobiyal eksojen küçük RNA (microbial exogenous small RNA) analizleri, patojen mRNA ve host RNA analizleri ile RNA sekanslama klinik mikrobiyolojinin genomik tabanlı teşhislerinde giderek vazgeçilmez bir araç haline gelmektedir [11].




**Şekil 1.8:** Çeşitli RNA-seq metodolojileri [11].

Bulk RNA-seq (toplu RNA sekanslama) yukarıda bahsedildiği üzere geniş bir uygulama alanı bulsa da NGS (Yeni Nesil Sekanslama) yöntemlerindeki hızlı gelişmeler, scRNA-seq (Single Cell RNA sequencing Tek Hücre RNA Sekanslama) ile hücrelerin teker teker karakterize edilmesine olanak sağlamış ve bulk RNA-seq'nın en büyük eksikliklerinden birine çözüm getirmiştir [12]. Vücudumuzdaki hücreler veya farklı hücre tipleri aynı

genotipleri paylaşsa da her hücrenin transkripsiyon karakteri kendine özel ve benzersizdir. Klasik bulk RNA-seq yöntemlerinin büyük bir kısmı dokulardan alınan hücrelerin homojen olduğunu varsaymakta ve elde edilen bir hücre topluluğunun Şekil 1.9'da da görüldüğü gibi ortalama ekspresyon değerlerini ölçmektedir [12]. Ancak yapılan birçok çalışma gen ekspresyon paternlerinin aynı hücre tiplerinde bile heterojeniteye sahip olduğunu göstermektedir [13]. Bu durum bulk RNA-seq'in hücreye veya hücre tipine özgü transkriptomik karakterlerini elde edemeyeceği ve hücreler arası heterojenitenin kaybolmasına neden olacağı anlamına gelir.



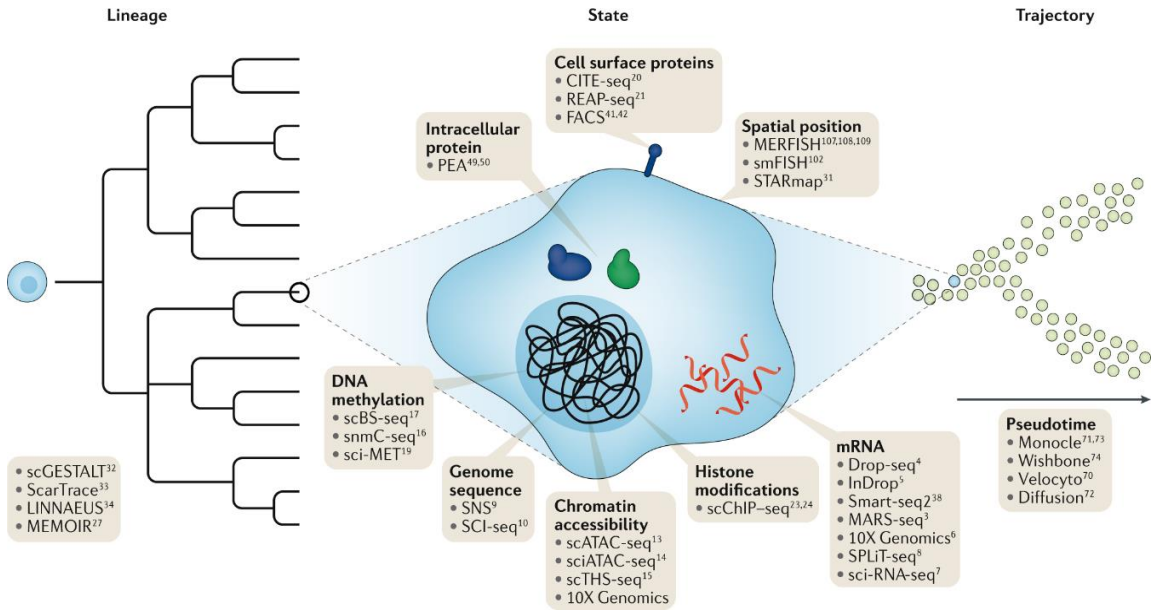

**Şekil 1.9:** Bulk RNA-seq ve scRNA-seq [14].

### 1.1.3 scRNA-seq

NGS'ye dayalı ilk tek hücre transkriptom analizi 2009'da önerilmiştir [15]. Olgunlaşmış scRNA-seq yaklaşımları, geleneksel bulk RNA-seq yöntemlerine kıyasla milyonlarca hücreyi yüksek verimle ve hücre bazında çözünürlükle analiz etmemize olanak tanır ve karmaşık hücresel fonksiyonları daha anlaşılır hale getirebilir. Örneğin tümör dokuları, vasküler hücreler, fibroblastlar, bağışıklık hücreleri (immune cells) ve kanser hücreleri gibi birçok çeşitte farklı hücre popülasyonlarını içerir. Kanser hücrelerinin karakterizasyonunun hedeflendiği bir çalışmada bulk RNA-seq Şekil 1.9'da görüldüğü gibi bu heterojen topluluğun profillerinin ortalamasını sağlar ve kanser hücrelerinin farklı olarak hangi genleri eksprese ettiğini sağladığı ortalama değerler sonucu kaybetmektedir [16].

scRNA-seq ilk önerildiği zamanlarda yüksek maliyetleri sebebiyle geniş kullanım alanları bulamamıştır. Ancak çalışmaların çok hızlı bir biçimde artması ve maliyetlerin düşürülmesi sonucunda tek-hücre (single-cell) sekanslama teknolojileri mikrobiyoloji, üreme, sindirim,

üriner sistemler, nöroloji ve tümörler gibi birçok alanda vazgeçilmez bir araç haline gelmektedir [17]. scRNA-seq, bulk RNA-seq'in aksine hücrelerin tek tek transkriptomlarının incelenmesine olanak sağlamaktadır. Buradan yola çıkarak araştırmacılar herhangi bir hücre popülasyonu içerisindeki transkriptomun hücre tipleri arasındaki veya hücreler arasındaki farklılıklarını ve benzerliklerini analiz edebilmektedir. scRNA-seq, hücresel heterojenitenin açığa çıkarılmasının haricinde monoallelic gen ekspresyonu, transkripsiyonel yanıtlar arası gürültüleri ve gen düzenleyici ağların ortaya çıkarılması gibi gen ekspresyonunun temel mekanizmaları hakkında önemli bilgiler ortaya çıkarabilme potansiyeline sahiptir [18].




**Şekil 1.10:** Tek hücreli yöntemlere genel bir bakış [19].

Ayrıca tek hücreli yöntemler sadece scRNA-seq ile kısıtlı değildir. Şekil 1.10'da görüldüğü üzere tek hücreli yöntemlerin uygulamaları DNA metilasyonundan histon modifikasyonlarına, hücre yüzeyinde bulunan proteinlerin incelenmesinden uzamsal konumların elde edilmesine kadar uzanmaktadır. Sonuç olarak hücreleri tek hücre çözünürlüğünde inceleyebilmek; hücrelerin komşu hücrelerle ilişkisinin anlaşılmasında, aykırı hücrelerin tespit edilmesinde, ilaç veya antibiyotik direncinin incelenmesinde ve daha birçok karmaşık sürecin anlaşılmasında araştırmacılara büyük bir kapı açmaktadır [20].

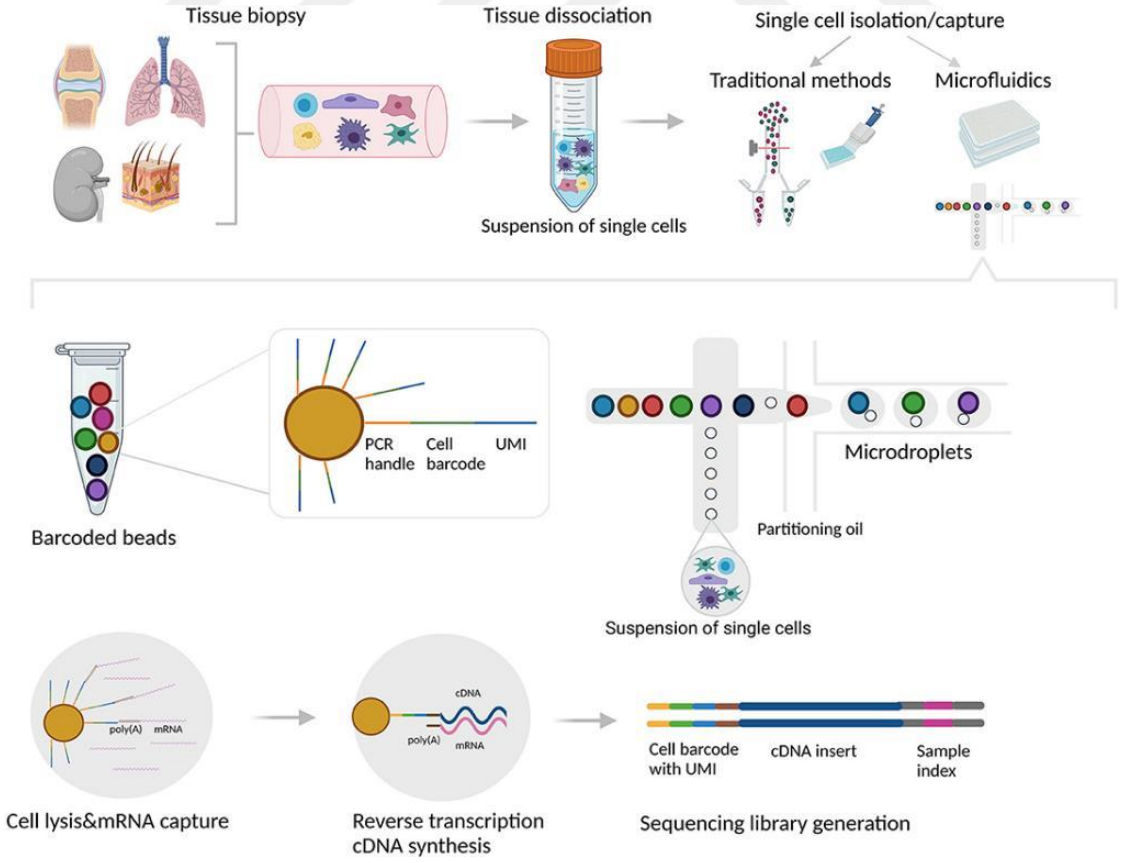
Bu tez kapsamında da scRNA-seq verilerinden hareketle işaretçi gen seçimi yapıldığı için, scRNA-seq verisinin nasıl elde edildiği, getirdiği avantajların neler olduğu ve kullanım alanları gibi konuların daha iyi anlaşılması adına, sonraki bölümde scRNA-seq teknolojisi daha ayrıntılı bir biçimde ele alınacaktır.

### 1.1.3.1 scRNA-seq genel iş akışı

scRNA-seq deneyleri genel anlamda birkaç ana adımdan oluşur ve bu adımlar sırasıyla şu şekilde ifade edilebilir [18]:

1. Numune hazırlama
2. Tek hücre izolasyonu
3. mRNA moleküllerinin yakalanması ve ters transkripsiyon
4. cDNA moleküllerinin amplifikasyonu ve kitaplık hazırlanması
5. Sekanslama
6. Veri işleme ve kalite kontrolü
7. Veri analizi ve görselleştirme

Deneysel olarak birçok sekanslama ve kütüphane hazırlama protokolü son birkaç yıl içerisinde önerilmiştir. Önerilen yöntemler; CEL-seq, Drop-seq, Smart-seq2, InDrop-seq, MARS-seq, Quartz seq ve STRT-seq şeklinde örneklendirilebilir [21–27]. Bu yöntemler izolasyon, indeksleme, cDNA çoğaltma, transkript kapsamı ve sekanslama kısımlarında bazı farklılıklar göstermektedir ancak genel olarak adımlar Şekil 1.11'daki gibi ilerlemektedir.




**Şekil 1.11:** Genel scRNA-seq analizi adımları [28].

### 1.1.3.2 Numune hazırlama

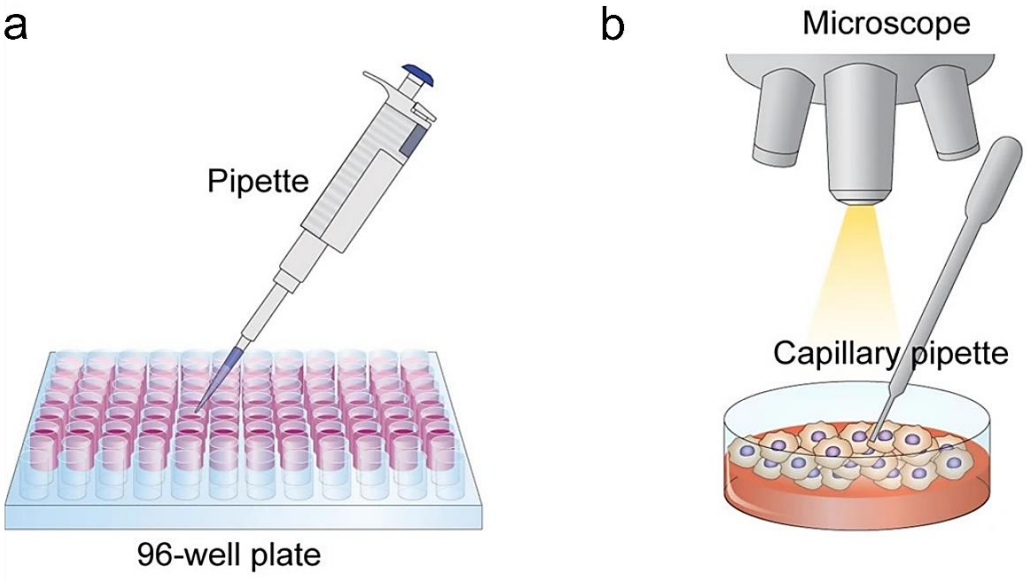
Numune hazırlama genel olarak sekanslama protokollerinde olduğu gibi scRNA-seq'in de ilk adımıdır. Hücrelerin durumu, literatürde önerilen scRNA-seq protokollerinin başarısı için en önemli kriterlerin başında gelmektedir. Numune hazırlama esnasında kullanılan pipetleme ve santrifüjlemenin yoğun kullanılması hücre hasarlarıyla sonuçlanabilmektedir. Dolayısıyla belirtilen iki proses minimum düzeyde tutulmalı ve santrifüjleme şartları optimize edilmelidir [29]. Ek olarak agregasyon ve kümeleme ile sonuçlanabilecek yüksek konsantrasyonlara sahip süspansiyonlardan kaçınılmalıdır. Kalıntı ve kümelemeleri gidermek için ise hücreler hücre boyutundan büyük filtreler ile filtrelenmeli ve süspansiyon hazırlandıktan sonra degredasyondan kaçınmak için mümkünse 30 dk içerisinde işlenmelidir [29].

### 1.1.3.3 Tek hücre izolasyon yöntemleri

Dokulardaki hücreler genellikle heterojendir, yani çoğunlukla farklı miktarlarda birbirlerine karışmış birçok farklı hücre tipini içerirler [30]. Hücrelerin karmaşık popülasyonlardan izole edilmesine dair çeşitli yöntemler literatürde önerilmiş olmasına rağmen düşük popülasyona sahip tek hücrelerin izole edilmesi hala teknik bir zorluk olmaya devam etmektedir [31]. Literatürde önerilen izolasyon yöntemleri ana başlıklar olarak şu şekilde ifade edilebilir:

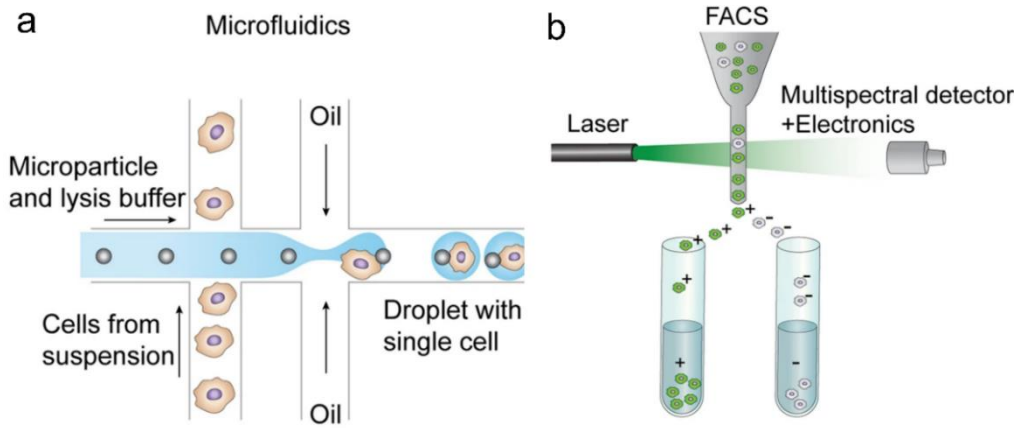
1. Sınırlayıcı Seyreltme (Limiting dilution)
2. Mikroakışkan sistemler (Microfluidic systems)
3. Mikro manipülasyon (Micromanipulation)
4. Floresanla aktive edilmiş hücre sıralama (Fluorescence-activated cell sorting FACS)
5. Lazer destekli mikrodisseksiyon (Laser capture microdissection LCM)

Sınırlayıcı seyreltme yönteminde pipetler yardımıyla hücreler bir topluluktan seyreltme yoluyla izole edilmektedir. Şekil 1.12 (a)'da görüldüğü üzere basit ve ulaşılabilir bir yaklaşımdır. Ancak bu yaklaşımın bir seferde birden çok hücreyi izole edebilme olasılığı yüksek olduğundan pek verimli değildir [12]. Şekil 1.12 (b)'de görseli görülen mikro manipülasyon tekniği robotize edilmiş mikropipetler ile hücreleri tek tek izole etmektedir [32]. Mikro manipülasyon, kolay ve ucuz olduğu, ayrıca düzgün sıvı kontrolü sağladığı için en çok kullanılan izolasyon yöntemidir. Çok kullanılmasına rağmen, yalnızca süspansiyon halindeki hücrelere uygulanabiliyor olması ve hücrelerin mikroskop altında yanlış tanımlanabilme olasılığı yöntemin dezavantajlarıdır [30].




**Şekil 1.12: a)** Sınırlayıcı Seyreltme ve **b)** Mikro manipülasyon yöntemlerinin illüstrasyonu [12].

Mikroakışkan sistemler ise tek hücre izolasyonu için düşük analiz maliyetleri ve numune tüketimi ve hassas sıvı kontrolü gibi olanaklar sağlaması nedeniyle popülerlik kazanmıştır [33]. Şekil 1.13 (a)'da görseli görülen sistem düşük maliyetlere ek olarak uygulama sırasında nano litre hacminde sıvı gerektirdiği için önemli bir problem olan dış kontaminasyon riskini de azaltmaktadır [12].




**Şekil 1.13: a)** Mikroakışkan Sistemler ve **b)** FACS metotlarının illüstrasyonu [12].

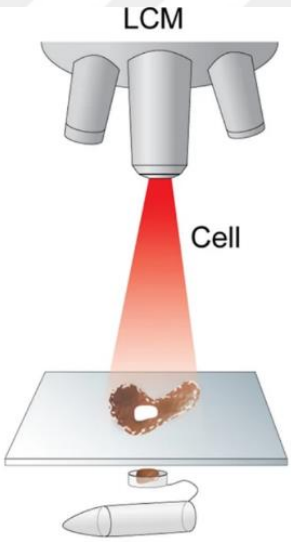
Mikroakışkanların aksine FACS her süspansiyonda 10.000'den fazla hücre gerektirmektedir. Şekil 1.13 (b)'de görselleştirilen FACS sıralama yapılmadan önce hücrelerin floresan etiketlerle etiketlenmesi prensibine dayanır ve hücre süspansiyonunun sitometriden geçmesiyle hücreleri izole etmektedir. Hücreler damlacıklar içerisinde elektrik

ile yüklendiğinden elektrostatik sapma ile istenilen hücreler izole edilebilir. Ancak ayrım sırasındaki hızlı akış hücrelere zarar verebilmekte ve izolasyon başarısız olabilmektedir [32].

Şekil 1.14'te görülen LCM ile izolasyon genel olarak 3 adımdan oluşur

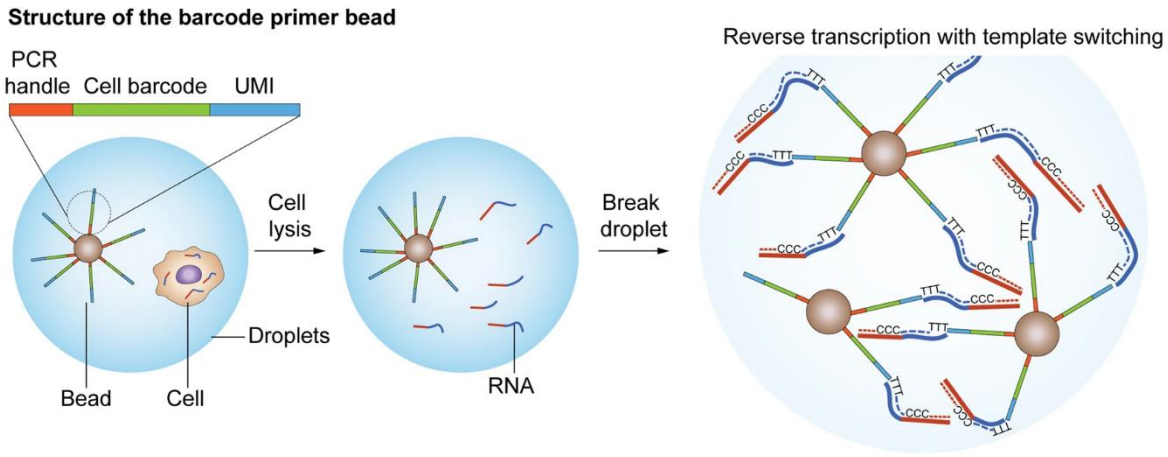
1. Mikroskop altında bir doku gözlemlenir ve tanımlanır.
2. İstenen bölüm etrafına çizgi çizilerek işaretlenir.
3. Lazer ışını istenen çizgi boyunca dokuyu keser ve hücreleri izole eder.

LCM hem fikse edilmiş (kimyasal olarak korunmuş), hem de canlı hücrelerde kullanılabilir [34]. Ayrıca diğer izolasyon yaklaşımlarının çoğu süspansiyon halinde bulunan hücre veya çekirdek gerektirir. Ek olarak süspansiyonları oluşturmak için hücrelerin bulundukları konumlardan ayrıştırılması gerektiği için hücreler dokulardaki uzamsal bağlamlarını koruyamamaktadırlar. Fakat LCM kullanılarak bu problem çözülebilmektedir [31].



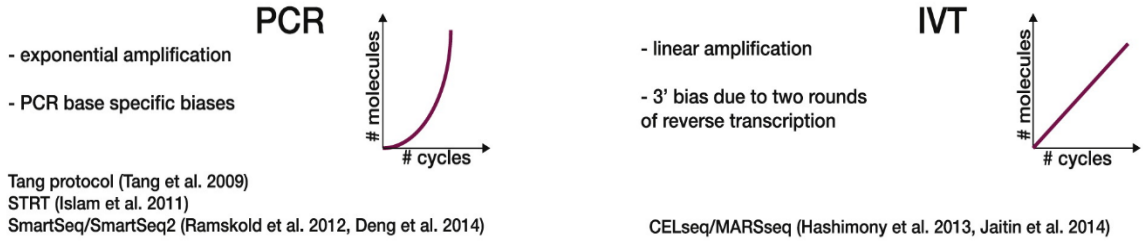

**Şekil 1.14:** LCM ile tek hücre izolasyonu illustrasyonu [12].

İzolasyon sonrasında kuyu ve damlacıklar ile yakalanan hücreler RNA molekülleri korunacak şekilde parçalanır. poli(A)-kuyruklu RNA'lar hücrelerin parçalanmasından sonra, poli(T) kuyruklu RNA'lar tarafından tutulur ve bu sayede rRNA ve tRNA gibi sekanslanması istenmeyen RNA molekülleri uzaklaştırılır. Daha sonra ters transkripsiyon vasıtasıyla RNA moleküllerinden, daha kararlı moleküller olan cDNA molekülleri elde edilmektedir. poli(T) kuyruklu oligonükleotidler Şekil 1.15'te de görüldüğü üzere içerdikleri rastgele nükleotid dizili uzantılar sayesinde, araştırmacının amplifikasyondan kaynaklanan biasları düzeltmesine ve teknik gürültüleri azaltmasına olanak sağlayan, UMI'ler (Unique Molecular Identifier) olarak görev yapmaktadırlar [29].




**Şekil 1.15:** UMI etiketleme ve ters transkripsiyon [12].

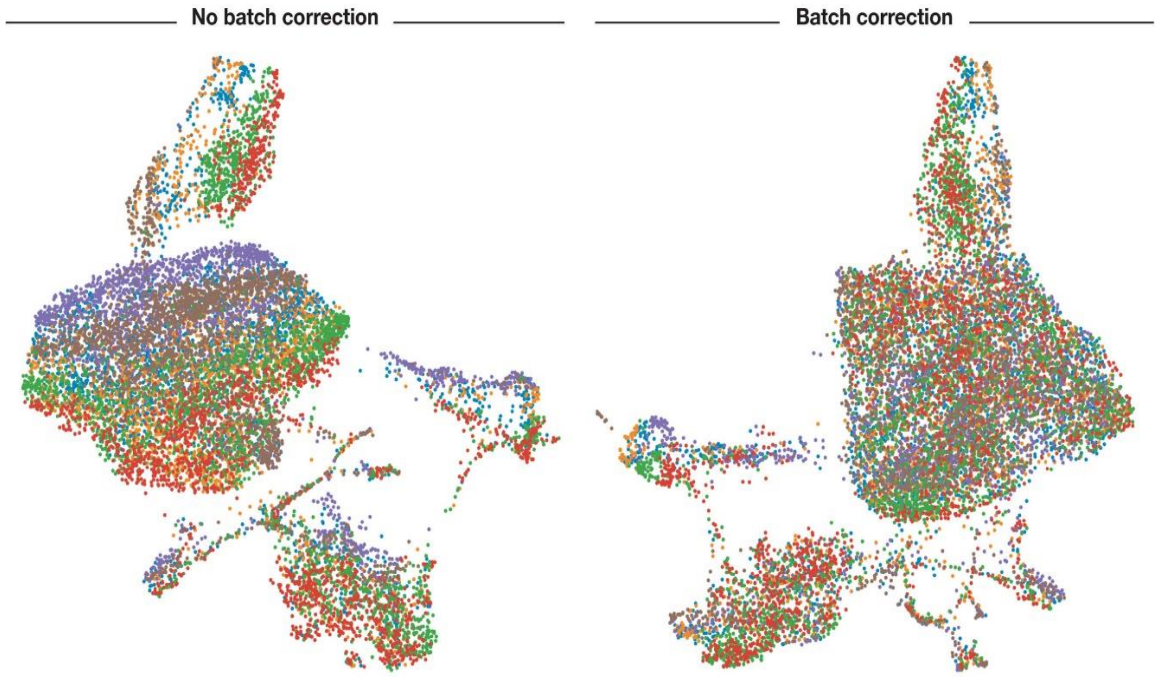
Ters transkripsiyon ile elde edilen cDNA molekülleri çok düşük miktarlardadır ve sekanslama için PCR (Polymerase Chain Reaction) veya IVT (in vitro transcription) gibi yöntemler kullanılarak çoğaltılması gerekir. Bu yöntemler arasından IVT sonuçlara bias getirmediği için PCR'dan bu anlamda daha başarılıdır çünkü Şekil 1.16'te de görüldüğü gibi lineer amplifikasyon sağlamaktadır. Ancak IVT'de amplifiye edilmiş RNA ek bir ters transkripsiyon turunu gerektirmektedir [35]. PCR tabanlı protokoller ise daha az uygulama süresi gerektirmesine rağmen üssel biaslara yol açabilmektedir [29, 36, 37].




**Şekil 1.16** PCR ve IVT yöntem çıkışlarının grafiksel karşılaştırması [35].

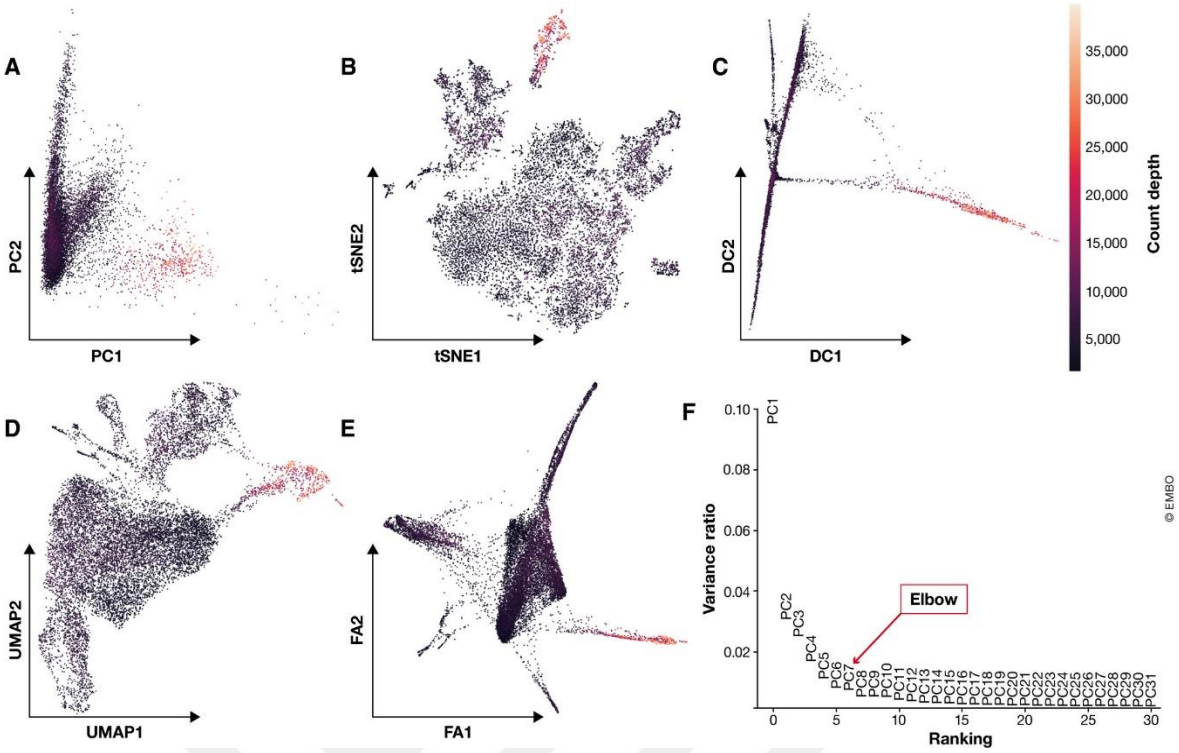
Hazırlanan kütüphaneler NGS yöntemleri kullanılarak sekanslanmakta ve ham veriler genel olarak $fastQ$ formatında depolanmaktadır Depolanan veriler için yapılması gereken ilk işlem sekanslanan okumaların kalite kontrolüdür. Kalite kontrolündeki amaç okunan hücresel barkod sayılarının canlı hücrelere karşılık gelip gelmediğini kontrol etmektir. Üç kriter kalite kontrolü için önem arz etmektedir bunlar; her barkoda karşılık gelen mitokondriyal gen sayısı, her gene karşılık gelen barkod sayısı ve her barkoda karşılık gelen gen sayısıdır (counth depth). Bu üç değişkenin dağılımları içerisindeki pikler üzerinden değerlendirmeler yapılır ve kalitesiz olduğu düşünülen kısımlar elenir [38]. Kalite kontrolü

sonrası diğer aşama normalizasyondur. Normalizasyon ile sayımlardan ekspresyon seviyelerine geçiş amaçlanmaktadır [39]. Ortalama-varyans ilişkisini azaltmak ve verideki çarpıklığı azaltmak için genel olarak veri matrislerine $log(x + 1)$ dönüşümü uygulanır [38].




**Şekil 1.17:** Batch Effect düzeltmesi öncesi ve sonrasında verinin dağılımları [38].

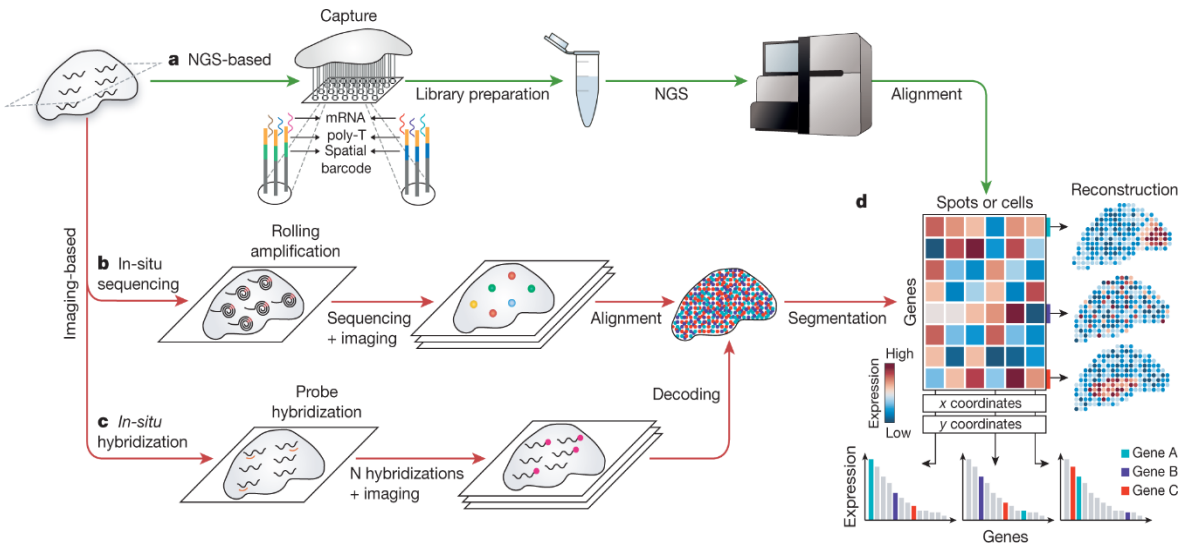
Normalizasyon sonrası deney şartlarından ve örneklerin farklı gruplar halinde sekanslanmasından kaynaklanan biyolojik olmayan etkilerin (Batch Effects) düzeltilmesi gerekmektedir. Yakalama süreleri, deney personeli, ekipman ve sekanslama platformlarındaki farklılıklar bu etkilere örnek olarak gösterilebilir. Bu etkiler veride büyük farklılıklara yol açabilmekte ve veri entegrasyonu esnasında biyolojik varyasyonları karıştırabilmektedir [40]. Şekil 1.17'de batch effect düzeltmelerinin etkileri açıkça görülmektedir. Düzeltme yapılmadan önce gruplar açıkça belli olurken düzeltme sonrasında katmanlarda azalma görülmüştür. Düzeltme aşamaları bittikten sonra veri kullanıcın amacına uygun biçimde görselleştirilebilir, yörünge analizleri yapılabilir, hücre çeşitleri veya hücreler arası diferansiyel eksprese edilmiş genler analiz edilebilir [41]. Şekil 1.18'de veri içerisindeki kümeleri ve dağılımları görebilmek için kullanılan çeşitli boyut azaltma yöntemleri görülmektedir.




**Şekil 1.18:** Çeşitli scRNA-seq görselleştirme metotları [38].

## 1.2 Uzamsal Transkriptomik (Spatial Transcriptomics)

Hücrelerin komşu hücrelerine veya hücresel olmayan diğer yapılara göre lokasyonları; hücre durumunu, hücresel fenotipi, hücre ve doku fonksiyonunu ve hücre-hücre ilişkilerini (Cell-Cell Communication) tanımlamak için faydalı olabilecek birçok bilgiyi ortaya çıkarabilir [42]. Örneğin hücre-hücre etkileşimleri yoluyla çevreye etki eden veya hareket eden çözünür sinyalleri dedekte edebilen hücre yüzeyine bağlı reseptörler ve ligand çiftlerinin mRNA'ları bulunur ve mRNA'lar transkriptomik analizleri ile incelenebilmektedirler [43]. Transkriptomik analizlerinde mRNA'ların miktarlarının haricinde konum bilgilerinin elde edilmesi, hücreden hücreye ekspresyon profillerinin nasıl değiştiğini, dolayısıyla hücre altı organizasyonları, gen düzenleyici ağları (gene regulatory networks) veya başka bir açıdan tümör içi ekspresyonel heterojenitenin incelenmesini ya da beyin hücrelerinin algı ve bilinç gibi işlevleri üretmek için moleküler düzeyde nasıl işbirliği yaptığının anlaşılabilmesi için araştırmacılara büyük bir görüş açısı kazandıracaktır [44, 45]. scRNA-seq son yıllarda inanılmaz bir biçimde popüler hale gelse de hücre izolasyonu aşaması nedeniyle bu önemli bakış açısını kaybeder.

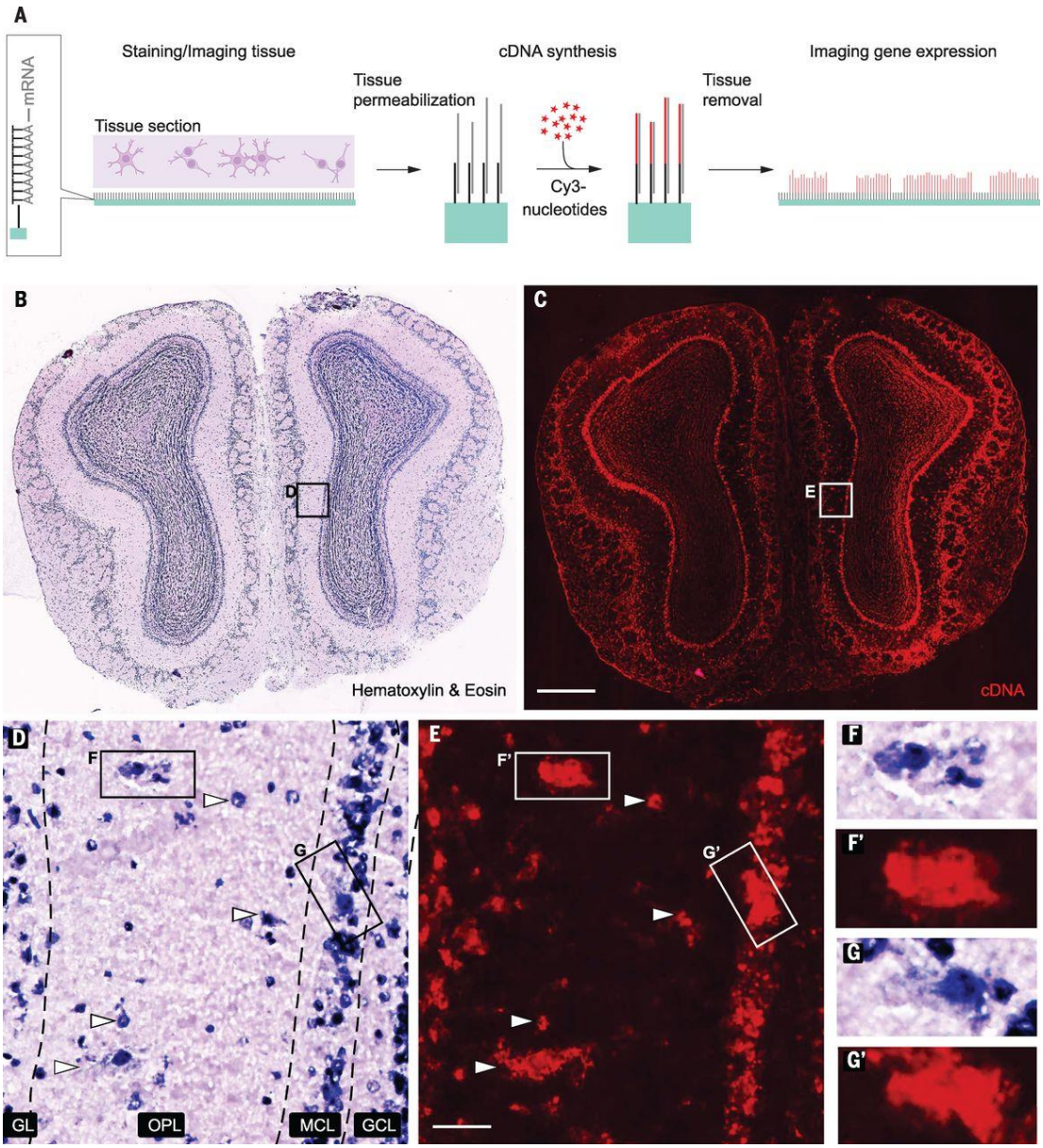



**Şekil 1.19:** Uzamsal Transkriptomik yöntemlerine genel bakış [46].

Son yıllarda popülerleşen ve olgunlaşan uzamsal transkriptomik teknolojileri ise bu sorulara hücrelerin konumlarını kaybetmeden analiz edilmesine olanak sağlayarak çözüm getirmektedir [47]. Uzamsal transkriptomik yöntemleri Şekil 1.19'da görüldüğü üzere sekanslama bazlı ve görüntüleme bazlı olmak üzere ana iki başlık altında toplanabilir. Sekanslama bazlı yöntemlerde hücre içi mRNA'ların konumları belli hedeflerle eşleştirilip sekanslanırken, in situ yöntemlerde genel olarak transkriptlerin hücre içerisinde floresan problarla görselleştirilmesi sağlanmaktadır.

Belirtilen iki ana başlık arasında bir trade-off mevcuttur [48]. Sekanslama bazlı uzamsal transkriptomik teknolojileri yüksek sayıda geni sekanslayabilir ancak düşük çözünürlüklü ve hassasiyeti sınırlıdır. Bunun yanında in situ teknolojiler ise her iki sorunu çözmekte, ancak gen verimi açısından sınırlı kalmaktadır [46].

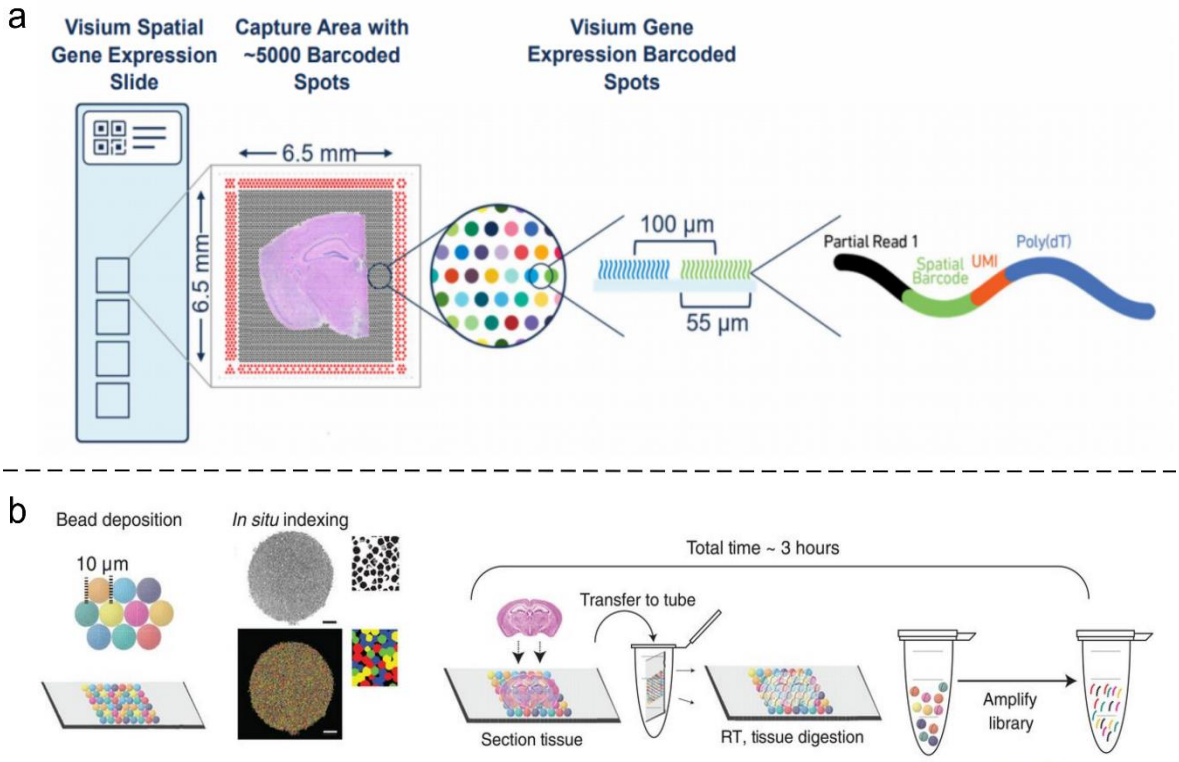
Tez kapsamında geliştirilen yazılım paketi gen kısıtlaması olan in situ yöntemlerin bu problemini çözmeyi hedeflemektedir. In situ yöntemler sınırlı sayıda gen görüntüleyebildiği için amaç bu yöntemler için deney öncesinde karar verilmesi gereken uygun gen kümesi seçimi problemini çözmektir. Bu nedenle sekanslama bazlı teknolojilerden kısaca, görüntüleme bazlı teknolojilerden ise daha ayrıntılı biçimde bahsedilecektir. Ancak, tez kapsamında geliştirilen yazılım paketi hücre tipine özgü gen seçimi gerçekleştirebildiğinden ölçüm yönteminden bağımsız olarak tüm çalışmalar için kullanılabilir.




**Şekil 1.20:** Uzamsal olarak lokalize edilmiş cDNA sentezi [49].

Sekanslama bazlı teknolojilerin ilk örneği, oligo(dT) probların cam slaytlar üzerine sabitlenmesi ve kriyoseksiyonlu doku dilimlerinin enzimatik geçirgenleştirme ile mRNA'larını salıp sabitlenen ve konumları belirli problarla hibritleşmesi ilkesine dayanmaktadır [49]. Şekil 1.20'de bu uygulamanın ilk örneği görülmektedir. Öncelikle doku kesilip oligo(dT) primerleri üzerine yerleştirilmekte daha sonra geçirgenleştirilmekte ve cDNA sentezi sonrasında floresan etiketli nükleotitler vasıtasıyla görselleştirilmektedir [49]. Şekil 1.20 (b)'de koku soğanlarının boyanmış hali (c)'de ise doku çıkarıldıktan sonraki durumda floresan içeren cDNA'lar görülmektedir.

100 $\mu m$ çapa sahip ilk nesil mikro diziler yaklaşık 1000 nokta içermekte ve her hücreyi karşılayan bir nokta olmadığı için ortalama gen ekspresyon değerlerini sağlamaktadır [50]. Uzamsal barkodlar sayesinde konumları belli olan oligo problar, üzerine eklenen doku kesitindeki hücrelerin mRNA'ları ile hibridize olmakta ve daha sonra genel sekanslama prosedürleri ile süreç devam etmektedir. Oligo(dT) probların konumları, barkodlandıkları için belli olduğundan sekanslama bilgisine ek olarak konum bilgisi de bu yolla elde edilmiş olunur. Ancak her hücreyi karşılayan bir spot mevcut olmadığından in situ yöntemler kadar yüksek çözünürlük mevcut değildir.




**Şekil 1.21: a)** 10X Genomics Visium ve **b)** Slide-seq teknolojileri [51, 52].

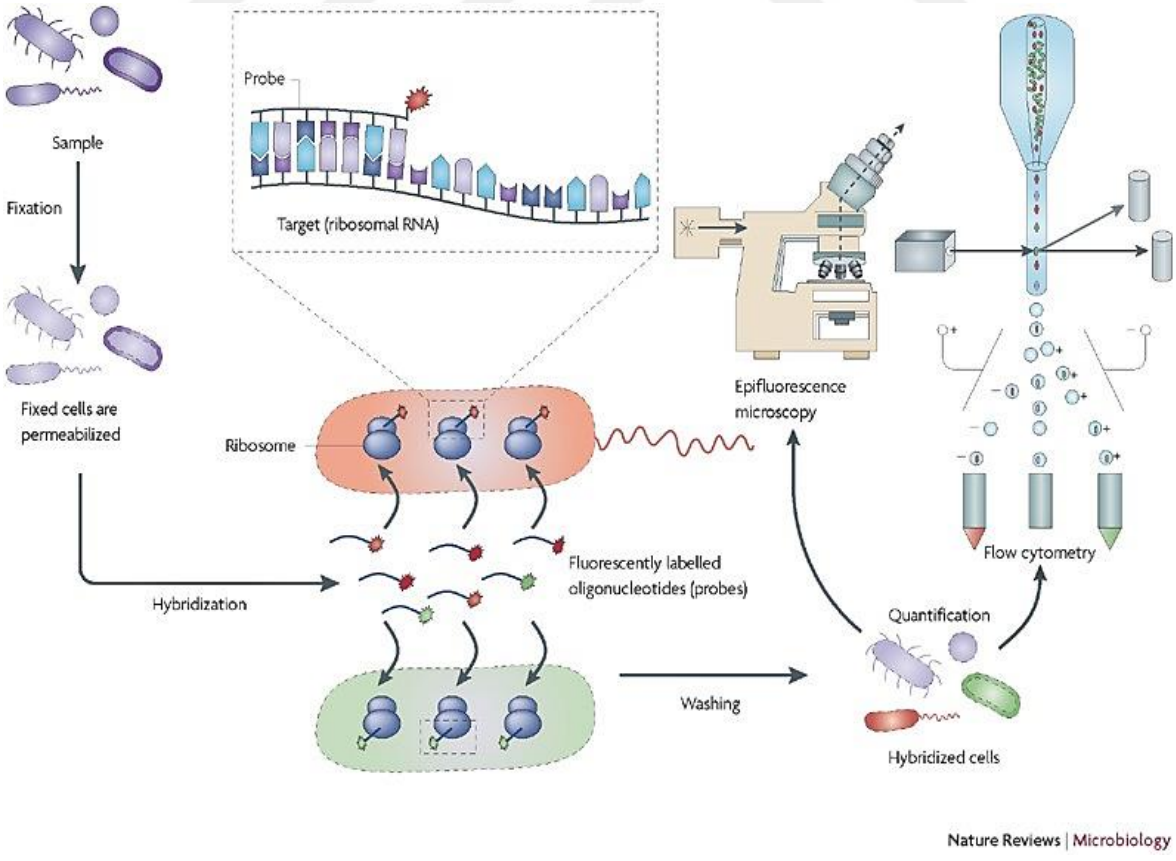
Şekil 1.21 (a)'da görülen 10X Genomics'in geliştirdiği Visium teknolojisinde ise çap 55 $\mu m$'ye küçültülürken aynı zamanda nokta sayısı da 5000'e çıkartılarak tek hücre çözünürlüğüne yakın bir şekilde yaklaşık 1-10 arası hücrenin ekspresyonu ölçülebilmektedir [50]. Şekil 1.21 (b)'de görülen Slide-seq ise cam lamel üzerine dizilmiş 10 $\mu m$ çapında, Drop-seq yönteminde kullanılanlara benzer biçimde "beads" ismi verilen boncuklar ve UMI'ler ile yaklaşık 1 ila 2 hücre çözünürlüğünde ekspresyonları ölçebilen bir yöntemdir [51].

### 1.2.1 Görüntüleme bazlı uzamsal transkriptomik yöntemleri

Görüntüleme bazlı yaklaşımların getirdiği yüksek çözünürlük hücre içi transkriptomların mekânsal mekanizmalarının anlaşılmasına olanak sağlamaktadır. Örneğin RNA'ların hücre içerisindeki konumsal organizasyonları ve ilişkileri transkripsiyon sonrası (post-transcriptional) mekanizmalar için önemli bir rol üstlenmektedir. Dolayısıyla bu mekanizmaların konumlarının hassas bir biçimde incelenebilmesi, işleyişin anlaşılabilmesi için kritik bir adımdır. Görüntüleme bazlı yöntemler ise bu hassasiyeti mRNA'ları in situ inceleyerek kullanıcıya sunmaktadır. Bu yöntemler, genel olarak in situ sekanslama ve multiplexed (çoğullanmış) yaklaşımlar olarak iki başlık altında incelenebilirler [53].

#### 1.2.1.1 FISH (Fluorescence In Situ Hybridization)

FISH tekniği bir DNA probunun kromozomal dizilere hibridize edilmesine dayanan basit ve etkili bir görüntüleme tekniğidir. Kullanılan problar; doğrudan floresan nükleotitlerin dahil edilmesiyle veya afinite moleküllerin dahil edilmesiyle olmak üzere iki şekilde etiketlenebilmektedir [54]. Daha sonra problar, dolayısıyla hedeflenen diziler veya moleküller in situ olarak mikroskopi analizleriyle görselleştirilebilmektedir.

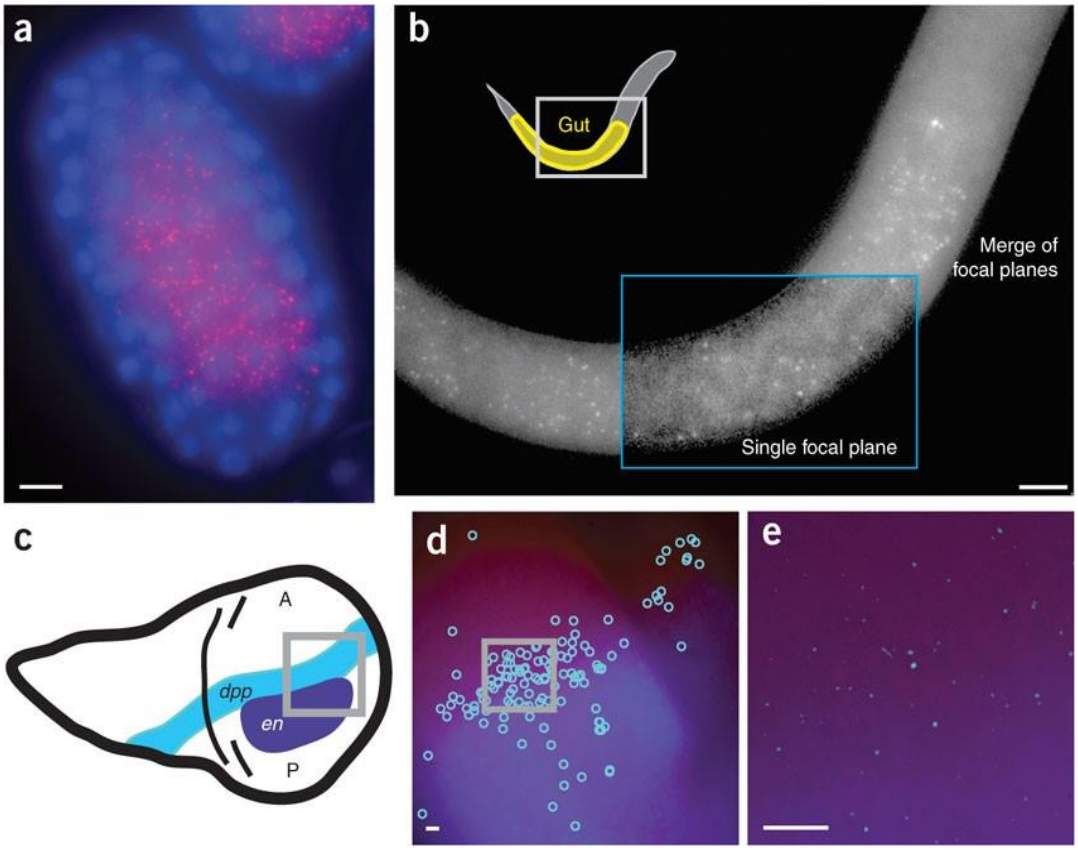



**Şekil 1.22:** Fluorescence In Situ Hybridization [55].

Şekil 1.22'de görülen FISH yöntemi klinik genetik, hücresel genomik, üreme tıbbı, karşılaştırmalı genomik, sinirbilim, toksikoloji gibi birçok alanda, getirdiği tek hücre bazında inceleme kabiliyeti DNA ve kromozomal analizler için sağladığı orta derecede çözünürlük sayesinde çok geniş bir kullanım alanı bulmuş ve şu an mevcut olan birçok yöntem bu yöntem baz alınarak geliştirilmiştir [54].

### 1.2.1.2 smFISH (single molecule Fluorescence In Situ Hybridization)

Oudenaarden ve arkadaşları radyo etiketli probların kullanıldığı yöntemlere alternatif olarak geliştirilen FISH yönteminin ve bunun üzerine geliştirilen FISH tabanlı yöntemlerin eksikliklerini gidermek amacıyla 2008 yılında smFISH yöntemini öne sürmüşlerdir [56–58].




**Şekil 1.23:** *C.elegans* ve *D.melanogaster* canlılarından elde edilen smFISH görüntüleri [56].

smFISH'ten önce geliştirilen yöntemlerin uygulanabilirliğindeki problemler ve tam olarak hassas sonuçlar vermemeleri nedeniyle, smFISH'i geliştiren araştırmacılar yoğun şekilde etiketlenmiş probların aksine 96-position DNA sentezleyicilerin yüksek veriminden yararlanarak çok sayıda prob sentezlemiş ve 3' uçlarında bir florofor parçası ile etiketlenmiş probların mRNA'ları güvenilir biçimde etiketleyebileceğini göstermişlerdir [56]. Şekil 1.23'te çıktıları görülen smFISH önceki yöntemlerin aksine çok sayıda tek etiketli prob

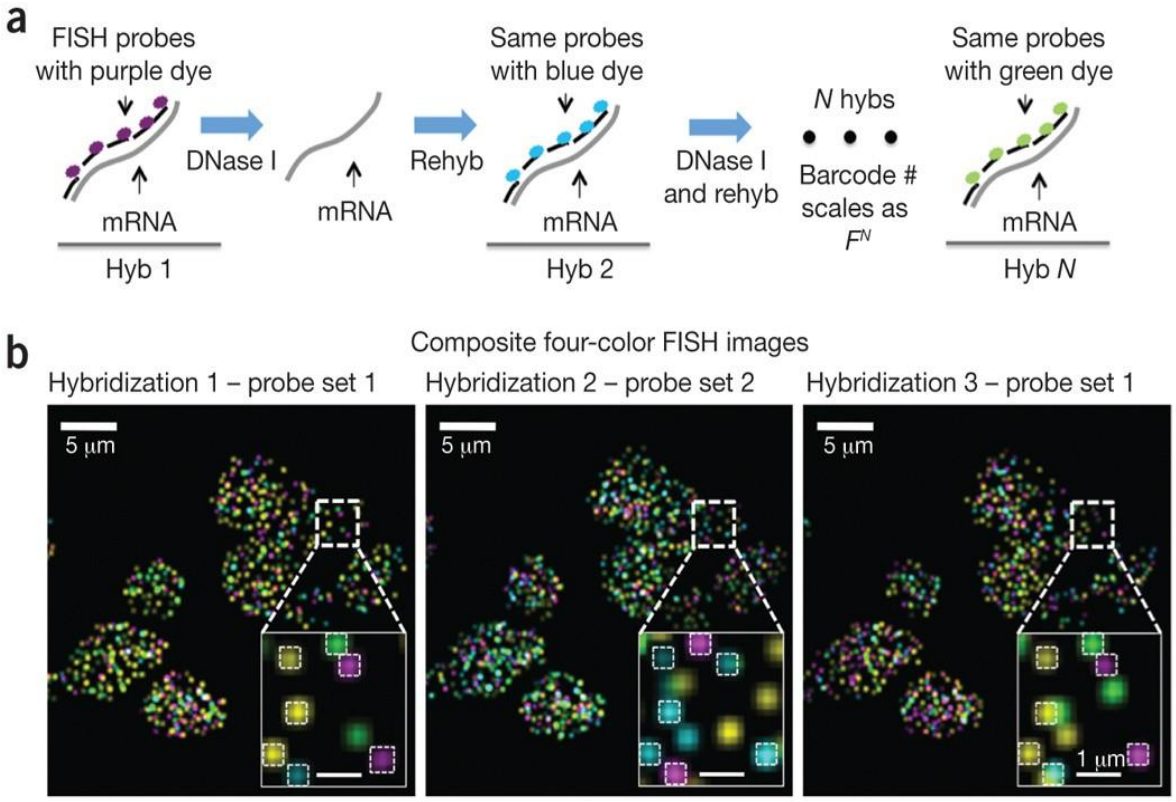
kullanmakta ve çıktı olarak tek tip sinyal üretmektedir. smFISH tek etiketli prob kullanarak, çok sayıda prob kullanan yöntemlerin ayrı prob bağlanma veya yanlış bağlanma problemleri sonucu oluşan yanlış pozitif ve yanlış negatif üretme sorununun üstesinden gelmektedir. Ayrıca prob oluşturma ve saflaştırma prosedürleri de önceki yöntemlere göre daha basittir [56].

#### 1.2.1.3 seqFISH (sequential Fluorescence In Situ Hybridization)

Cai ve Lubeck, genomik yaklaşımların uzamsal bilgiyi kaybetmesi ve buna ek olarak hücresel popülasyonlardaki heterojenliği elde edememesi, diğer yandan tek hücreli yöntemlerin (scRNA-seq) ise tek seferde ancak birkaç geni inceleyecek kapasitede olması sebebiyle bu iki yaklaşımı birleştirmek için Süper Çözünürlüklü Mikroskopi (Super Resolution Microscopy SRM) kullanmayı önermişlerdir [59]. Geleneksel floresan mikroskopisi ile barkodlama, ancak transkript seviyelerinin düşük olduğu durumlarda faydalı olabilir, çünkü transkriptlerin yoğun olduğu durumda floresan barkodlar üst üste gelmekte (overlapped) ve okunması zorlaşmaktadır. SRM'lerin yüksek çözünürlüğü ise moleküler yapılara bağlanmış florofor bazlı barkodları algılamak için yeterlidir ve bu sayede barkodun gözlemlenme sayısına göre analizler gerçekleştirilebilmektedir [59].

seqFISH yöntemi, smFISH tekniğinin getirdiği avantajlar baz alınarak geliştirilmiştir; yani oligonükleotid problarının yüksek etiketleyebilme kabiliyetinden yararlanarak, transkriptleri kombinasyonel olarak barkodlamak için aslında smFISH kullanmaktadır. Şekil 1.24'te seqFISH yöntemine dair görseller görülmektedir. seqFISH, mRNA moleküllerini farklı renklere sahip problarla birden fazla kez hibridize eder ve daha sonra görüntü işleme prosedürleriyle analizleri gerçekleştirmektedir. Ayrıca yazarlar barkodlama için uzamsal ve spektral olmak üzere iki farklı yaklaşım benimsemişlerdir. Ancak spektral barkodlama daha sağlam olduğu için spektral barkodlama ile *S.cerevisiae* hücrelerinde yöntemlerini test etmiş ve mRNA'ları teker teker barkodlayabildiğini %100 uzamsal yeniden yapılandırma (reconstruction) ile göstermişlerdir [59].

mRNA profilleme için ise *Crz1* tarafından düzenlenen 14 gen, 5 adet strese tepki veren gen ve ek olarak 13 yaşlanma ve stres işaretçileri belirlemiş ve 3 adet foto-aktive edilebilir 7 boya çiftinden üçünün kombinasyonlarını kullanmışlardır. SRM ile yapılan barkodlamanın kontrol edilebilmesi için elde edilen ölçümleri (q)PCR ve smFISH sonuçlarıyla karşılaştırmış ve her iki yöntemde de 0.95 $R\hat{}2$ değerleri elde etmişlerdir.

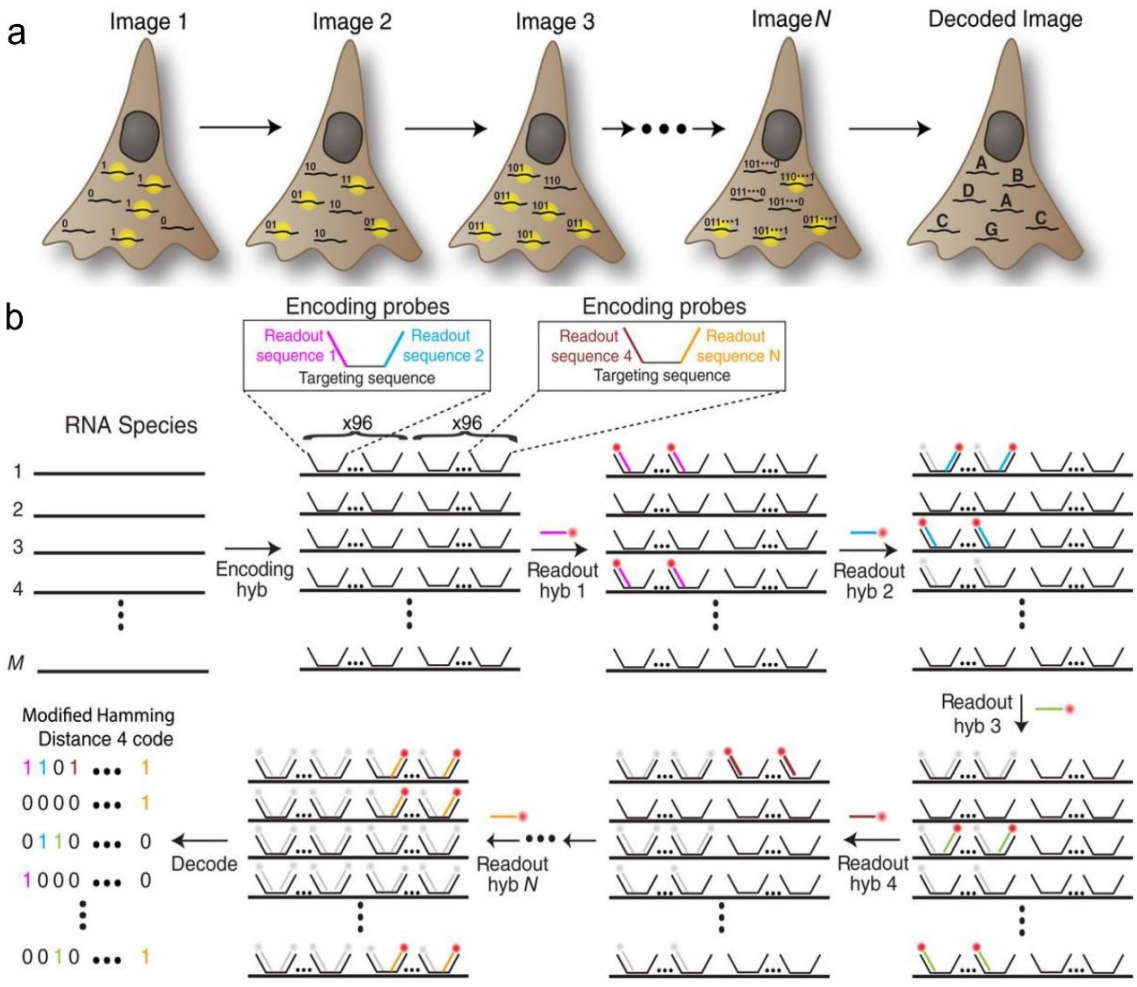



**Şekil 1.24:** seqFISH yönteminde hibridizasyon aşamaları [60].

Nihai olarak seqFISH sıralı barkodlama, hızlı bir şekilde ölçeklenebilmesi yani sadece iki adet floresan boya ile bile sınırsız kodlama kapasitesine sahip olması, hibridizasyon sırasında her bir transkripte karşı mevcut diğer FISH problarının tamamen kullanılabilir olması ve en önemlisi barkod okumanın sağlam ve güvenilir olması gibi sağladığı avantajlar nedeniyle nörobilim [61–63], biyogenez araştırmaları [64] ve kanser [65] gibi çok önemli alanlarda yaygın bir şekilde kullanılmaktadır.

### 1.2.1.4 MERFISH (Multiplexed Error-Robust FISH)

RNA transkriptlerini görüntülemek için uzun süredir kullanılmakta olan FISH yöntemleri moleküllerin konumlarını yüksek doğrulukta tespit etmeye olanak tanırlar ancak sınırlı sayıda bulunan renk kanalları bu yöntemlerin en büyük kısıtlamalarından birisi olarak bilinmektedir. Bu problemin çözümü için genleri tek renkle tanımlamak yerine, birçok hibridizasyon turu ve binary kodlama sistemi içeren bir mekanizma ile genlerin farklı renk kombinasyonları sayesinde tespit edilmesine olanak tanıyan MERFISH, (multiplexed error-robust FISH) 2015'te Zhuang ve arkadaşları tarafından önerilmiştir [44].

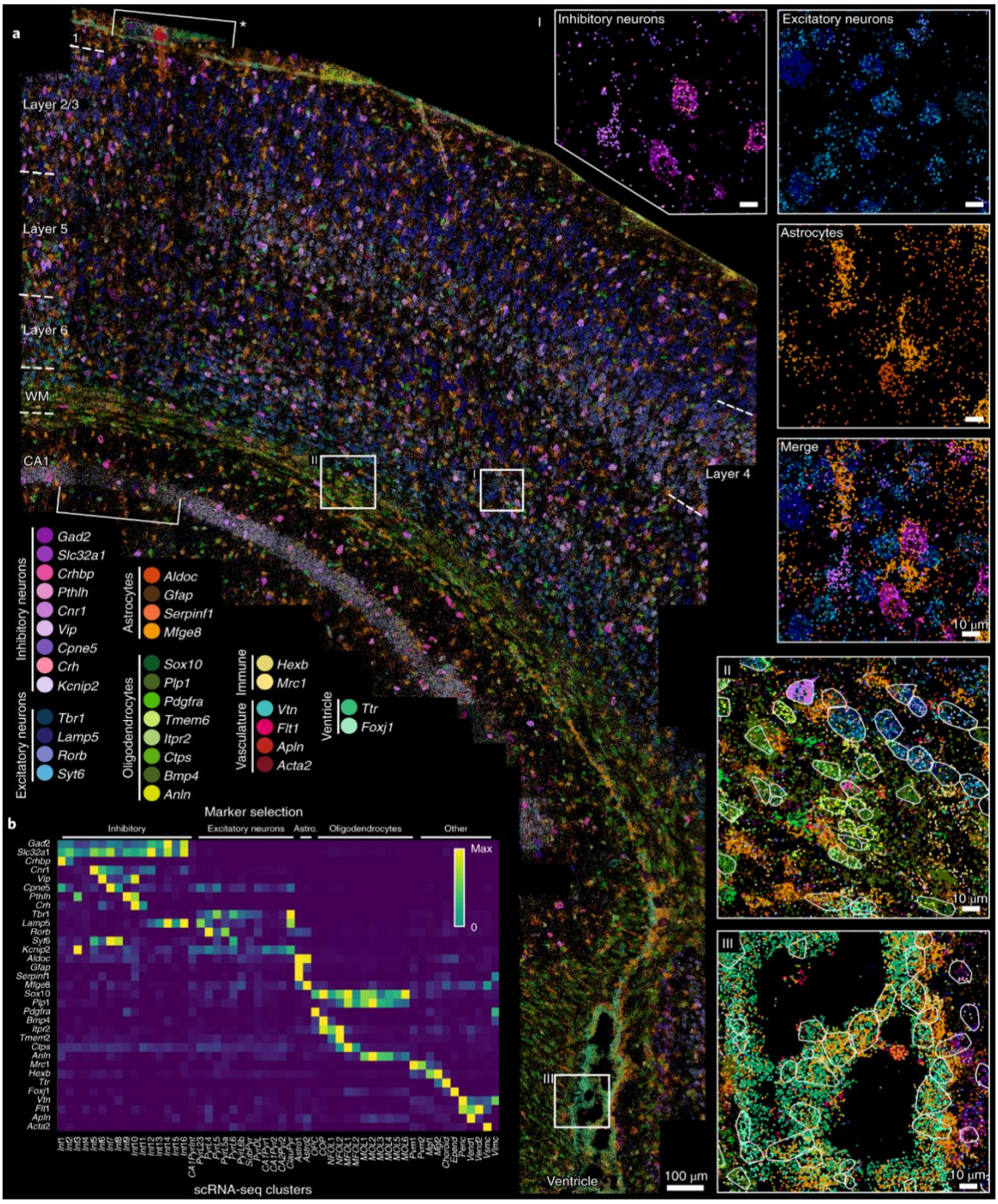



**Şekil 1.25:** MERFISH yönteminde hamming kodlama metodu ile hibridizasyon [44].

Spesifik RNA'ların amplifikasyona dayalı biaslar olmadan görüntülenebilmesine olanak sağlayan smFISH yöntemini baz alan MERFISH, her RNA molekülü için *N* bitlik ikili barkod oluşturmakta ve daha sonra hibridizasyon turları ile belirlenen kodun elde edilip edilmediğini kontrol etmektedir. Şekil 1.25 (b)'de prob mekanizmasına dair görsel görülmektedir. Her hibridizasyon turunda tasarlanmış problar ile yapılan görüntülemeler sonrasında diğer tura geçilirken parlayan problar kısa bir foto ağartma işlemi ile söndürülmektedir. Daha sonra elde edilen görüntüler analiz edilmekte ve konumları belli her RNA molekülü ışıma durumlarına göre bir ikili diziye sahip olmaktadır. Her dizi 1 ve 0'lar dan oluşmakta ve farklı bir mRNA molekülüne karşılık gelmektedir [44, 53].

### 1.2.1.5 osmFISH (cyclic-ourobos smFISH)

Linnarson ve arkadaşları smFISH yönteminin floresan sinyaller üst üste geldiğinde karşılaştığı optik kalabalık ve transkript uzunluğu ile sınırlı olması sebebiyle bazı hücre tiplerini haritalayamayacağını öne sürerek bu problemlerin üstesinden gelmek için barkodsuz ve amplifiye edilmemiş bir yöntem olan osmFISH'i önermişlerdir [66].




**Şekil 1.26:** Fare somatosensör korteksi, hipokampüs ve ventrikülünden bir parçanın osmFISH analizi sonuçları [66].

Hedef gen sayısı ve hibridizasyon turu sayısı osmFISH için lineer biçimde ölçeklenmiş ve her RNA molekülü floresan etiketli 20 nükleotid uzunluğunda DNA problarının bağlanması ile görselleştirilmektedir. Görüntüler elde edildikten sonra problar diğer hibridizasyon turu için eritilerek çıkartılmakta ve barkod kullanılmadığı için her görüntü birbirinden bağımsız biçimde analiz edilebilmektedir. Ayrıca osmFISH yöntemi ile yüksek oranda eksprese edilen genler düşük oranda eksprese edilmiş genlerin tespit edilmesine engel oluşturmamaktadır [66].

Linnarson ve arkadaşları yöntemlerini fare beyin hücreleri üzerinde test etmiş ve yöntemlerinin scRNA-seq'e kıyasla büyük oranda daha az sıfır sayım gösterdiğini, ayrıca hücre başına ortalama 4 kat daha fazla molekül tespit ettiğini dolayısıyla scRNA-seq'e kıyasla daha duyarlı olduğunu ve daha düşük dropout (yanlış negatif) oranlarına sahip olduğunu göstermişlerdir [66]. Fare beyin hücreleri için yöntemlerinin sonuçları Şekil 1.26'da görülmektedir. Şekil 1.26 (a) her noktanın bir RNA molekülüne karşılık geldiği büyük anatomik yapıları 33 adet gen ile görselleştirmektedir. Şekil 1.26 (b) ise elde edilen gen ekspresyonu ölçümünün ısı haritasını ve karşılık gelen hücreleri içermektedir.

### 1.3 Tezin Amacı

Yukarıda bahsedilen; smFISH (single-molecule FISH) [56], seqFISH (sequential FISH)[60], MERFISH (multiplexed error-robust fluorescence in situ hybridization) [67], osmFISH (cyclic-ouroboros smFISH) [68] ve bahsedilmeyen; in situ sekanslama [69], seqFISH+ [70], STARmap (spatially resolved transcript amplicon readout mapping) [71], targeted expansion sequencing [72], gibi hedefli in situ teknolojiler genellikle önceden seçilmiş sınırlı sayıda gene bağımlıdır [73]. Bu yöntemler çoğullama (multiplexing) ile gen sınırlamasını çözmeyi amaçlar ancak tekrarlayan görüntüleme turları da numunelerin stabilitesini etkileyebilmektedir [46]. Bu nedenlerden dolayı yeni bulguları engelleyebilecek ve uygulamayı sınırlayabilecek yüksek maliyetlere sahip problar için deney öncesi gen seçimi büyük bir önem arz etmektedir [74].

Bu tez kapsamında, hücre içerisinde ifade edilen RNA moleküllerinin miktarına ek olarak konum bilgisini de sağlayan, in situ bazlı uzamsal transkriptomiks yöntemlerinin ihtiyaç duyduğu kümeye özgü işaretçi gen seçimini gerçekleştiren bir yöntem önerilmiştir. Önerilen yöntem bir yazılım paketi olarak geliştirilmiş ve açık erişimli olarak literatüre sunulmuştur. Ayrıca yazılım paketinin dokümantasyonu için bir web sayfası da oluşturulmuştur.

# 2. HÜCRE TİPİNE ÖZGÜ İŞARETÇİ GEN SEÇİMİ

## 2.1 Amaç

Bu bölümde uzamsal transkriptomik deneyleri için gereken, herhangi bir hücre tipini diğer hücre tiplerinin tamamından ayırabilecek sınırlı sayıda işaretçi geni seçmek için, tez kapsamında geliştirilen scMAGS (**s**ingle-**c**ell **ma**rker **g**ene **s**election) yazılım paketi ve literatürde önerilen diğer yöntemler sunulacaktır. Öncelikle işaretçi genlerde aranan kriterlerden daha sonra ise literatürde önerilen yöntemlerden bahsedilecektir.

## 2.2 Hücre Tipine Özgü İşaretçilerde Aranan Kriterler

Uzamsal transkriptomik çalışmaları hücre-hücre arası ilişkilerin uzamsal olarak anlaşılmasında büyük rol oynasa da bu çalışmaların bazı kısıtlamaları bulunmaktadır. Bunlardan birisi çok sayıda floresan oligonükleotid prob kullanıldığında ortaya çıkan yüksek deney maliyetleridir. Diğer kısıtlamalar ise kullanılabilecek prob sayısının in situ yöntemler ile gerçekleştirilen deneylerde sınırlı (10-50) olması ve probların önceden seçilmesine gerek olmasıdır [75]. Bu iki kısıtlamanın getirdiği problemin çözümü olarak hücre tiplerinin ayırt edilebilmesi için işaretçilerin dikkatli ve olabildiğince az miktarda seçilmesi gerekmektedir. Herhangi $X$ hücre tipi için seçilecek hücre tipine özgü işaretçilerin uzamsal transkriptomik çalışmalarında $X$ hücrelerini diğer tüm hücre tiplerinden ayırt edebilmesi için, aşağıda sıralanan kriterlere sahip olmaları gerekir;

1. Hücre tipi içerisindeki hücrelerde yani $X$ hücrelerinde, yüksek ekspresyon oranı (Yani hücrelerin çoğunda eksprese ediliyor olmalı)
2. Hücre tipi dışındaki hücrelerde düşük ekspresyon oranı (Mümkünse sıfır, yani hücre tipi dışında az veya hiç eksprese edilmemeli)
3. Hücre tipi içerisindeki ekspresyon değerlerinin yüksek ortalamaya sahip olması
4. Eğer hücre tipi dışındaki hücrelerde ekspresyon varsa düşük ortalamaya sahip olması

Bu kriterler uzamsal transkriptomik deneylerinde problara bağlanacak genlerin herhangi bir hücre tipini diğer tüm hücre tiplerinden ayırt edilebilmesi açısından büyük önem arz etmektedir. Bu nedenle seçilecek işaretçiler bu kriterlere sahip olmalıdır.

## 2.3 Literatür Özeti ve Mevcut Yöntemlerin Eksiklikleri

Bu bölümde literatürde önerilen yöntemler ayrıntılı biçimde incelenecek ve eksikliklerinden bahsedilecektir.

### 2.3.1 SMaSH

SMaSH (Scalable Marker (Gene) Signal Hunter) Nelson ve arkadaşları tarafından önerilen ve uzamsal transkriptomik çalışmaları için gerekli sınırlı sayıda işaretçi geni seçmek için geliştirilmiş bir yöntemdir [76]. Yöntem Python dilinde implemente edilmiş ve Scanpy paketi içerisindeki AnnData sınıfına bağımlı olarak geliştirilmiştir [77]. SMaSH giriş verisi olarak sayım matrisini ve küme etiketlerini içermesi zorunlu olan AnnData nesnesini almaktadır. SMaSH işaretçi seçimi gerçekleştirmeden önce genleri iki filtreleme adımından geçirmektedir. Birinci filtreleme adımında Harmony kullanılarak batch-effect (Deneysel Etkiler) düzeltmesi uygulanmakta daha sonra; housekeeping genlerini, mitokondriyal aktivite ile ilişkili genleri, ribozomal biyogenez genlerini ve son olarak düşük ve yüksek düzeyde eksprese edilen genleri filtrelemektedir [78]. İkinci filtreleme adımında sayım matrisini PCA (Principal Component Analysis) kullanılarak boyutsal olarak indirgemekte ve verideki toplam varyansın %80'ini açıklayan her bir ana bileşendeki ilk 20 geni bir sonraki adım için saklamaktadır. Kalan genleri kullanarak; Rastgele Orman (Random Forest, RF) Dengelenmiş Rasgele Orman (Balanced Random Forest), XGBoost ve Derin Sinir Ağları (Deep Neural Network DNN) algoritmalarından herhangi biriyle sınıflama gerçekleştirmektedir. Sınıflama sonrasında genleri sınıflama için ne kadar yararlı olduklarını değerlendiren; RF, BRF XGBoost için Gini Importance, DNN için ise Shapley değerini kullanarak sıralamaktadır.

### 2.3.2 scGeneFit

scGeneFit Dumitrascu ve arkadaşları tarafından önerilen küme etiketlerine duyarlı sıkıştırma ve sınıflandırma yöntemlerini kullanılarak işaretçi seçimini gerçekleştiren, Python dilinde implemente edilmiş bir algoritmadır [79]. scGeneFit giriş verisi olarak sayım matrisini ve hücre etiketlerini almaktadır. scGeneFit'in amacı yüksek boyutlu bir öznitelik uzayını etikete duyarlı sıkıştırma yöntemleri kullanarak boyutu belirlenmiş ve hücre kümelerinin en iyi şekilde ayrıldığı bir uzaya yansıtmaktır (projection). Burada aynı etikete sahip hücreler farklı etikete sahip hücrelere göre düşük uzayda birbirine daha yakın konumlanmaktadır. Ayrıca scGeneFit düşük boyuttaki uzayda işaretçi seçiminin mümkün olduğundan emin olmak için

projeksiyonu sınırlandırmaktadır. Böylece düşük boyutlu uzayda her boyut tek bir işaretçiyi yakalayarak tek bir gene karşılık gelirken, genlerin lineer kombinasyonuna karşılık gelmemektedir.

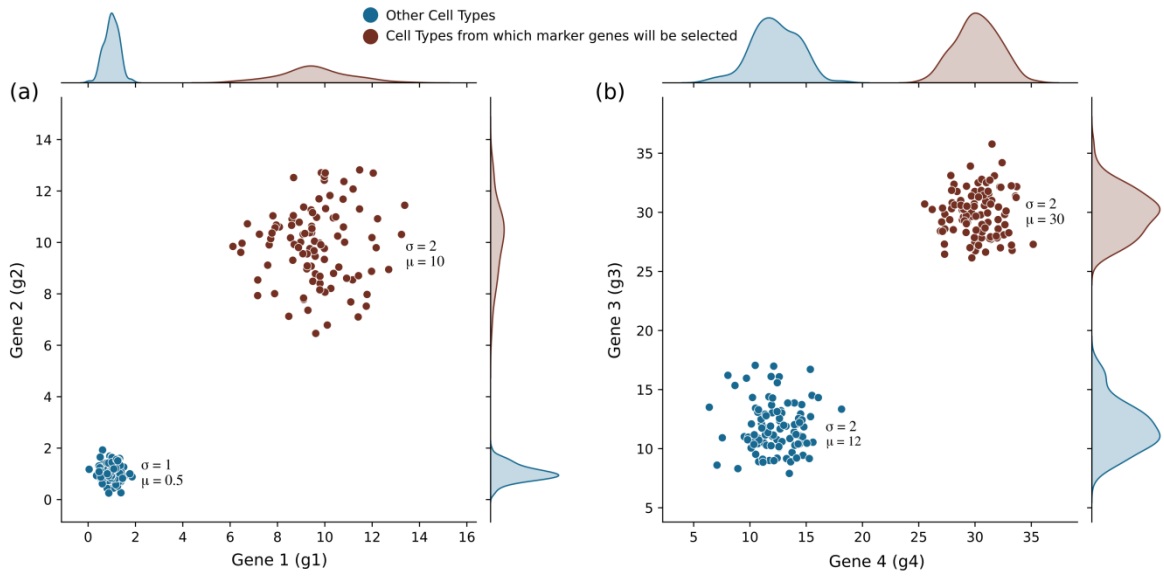
### 2.3.3 COSG

COSG Dai ve arkadaşları tarafından önerilen, hücreye özgü işaretçi genleri seçmek için kosinüs benzerliğini temel alan, Python ve R dilinde implemente edilmiş bir yöntemdir [80]. COSG hücre tipleri belirli olan scRNA-seq veya scATAC-seq verisinden işaretçi genleri seçmek için, öncelikle ideal ekspresyona sahip yapay genler oluşturmaktadır. Yapay genler her hücre tipi için ayrı ayrı oluşturulmaktadır ve sadece oluşturuldukları hedef hücre tipinde ekspresyona sahiptirler. Daha sonra oluşturulan yapay genlerin veri matrisindeki tüm genlere olan kosinüs benzerliği hesaplanmaktadır. Hesaplanan $kosin\ddot{u}s$ benzerliği kullanılarak $COSGscore$ olarak tanımlanan hedef hücre tipi dışındaki hücrelerin ekspresyonlarını bir penaltı terimi ile denetleyen skor hesaplanmaktadır. Bu skor içerisinde bulunan $\mu$ penaltı terimi kullanıcı tarafından değiştirilebilir ve penaltı teriminin varsayılanı 1 olarak ayarlanmıştır. $\mu$'nün sıfırdan küçük olması hedef hücre tipi dışındaki hücrelerdeki ekspresyonlara daha az ceza (penaltı) verilmesi anlamına gelmektedir.

### 2.3.4 Önerilen yöntemlerin eksiklikleri

Uzamsal transkriptomik deneylerinde hücre tiplerinin başarılı bir şekilde ayrılabilmesi için seçilen işaretçilerin Bölüm 2.2'de belirtilen kriterlere sahip olması gerekir. Ancak literatürde önerilen yöntemler bazı eksiklikleri nedeniyle bu kriterleri tam olarak sağlayamamaktadır. Bu bölümde bu eksikliklere değinilecektir.

SMaSH birkaç gen filtreleme adımı sonrasında kalan genleri çeşitli sınıflandırma algoritmalarında hücreleri sınıflandırmak için kullanır ve genleri sınıflandırma performasına olan etkiyi değerlendiren indekslere (Gini ve Shapley) bağlı olarak sıralar. SMaSH'in ilk problemi örnek sayısının düşük olduğu bir hücre tipinde sınıflama algoritmaları, eğitim aşamasında yeterince örneğe sahip olamayacağı için genelleme yapamayacak olmasıdır. Ve eğitim setinin boyutu sınıflama yöntemlerinin performansını etkileyen en büyük etkenlerden birisidir [81]. İkinci problem ise genleri bir sınıflama algoritmasına sağladığı katkıya göre sıralamaktır. Bu problem şu şekilde açıklanabilir; herhangi $k$ kümesi için işaretçi seçilecek olsun ve bu işaretçilerin seçilebileceği farklı tiplerdeki genler Şekil 2.1'deki gibi olsun. Şekil 2.1 (a) seçilmesi gereken işaretçi genleri, Şekil 2.1 (b) ise seçilmemesi gereken genleri görselleştirmektedir.




**Şekil 2.1:** Farklı dağılımlara sahip genler

Şekil 2.1'deki kahverengi noktalar $k$ kümesindeki hücrelerin, mavi noktalar ise $k$ kümesi haricindeki hücre tiplerinin ekspresyon değerlerini temsil etmektedir. Şekil 2.1 (a)'da görülen $g1$ ve $g2$ genleri aslında neredeyse aynı ekspresyon paternine sahiptir ancak 2 boyutta ekspresyon paternleri daha anlaşılır olacağından bu şekilde görselleştirilmiştir. Şekil 2.1 (a)'da görülen $g1$ ve $g2$ geni $k$ hücre tipinde ortalaması 10 ve standart sapması 2 olan bir ekspresyon seviyesine sahipken $k$ hücresi haricindeki hücrelerde ise eksprese edilmemekte, eksprese edildiyse bile çok düşük seviyelerde kalmaktadır. Şekil 2.1 (b)' de görülen $g3$ ve $g4$ genleri ise $k$ hücre tipinde ortalaması 30 ve standart sapması 2 olan bir ekspresyon seviyesine sahip, ek olarak $k$ hücre tipi haricindeki hücrelerde de ortalaması 12 ve standart sapması 2 olan bir ekspresyon karakterine sahiptir. Şekil 2.1 (b)'de görülen $g3$ ve $g4$ genleri $k$ hücre tipi ile geriye kalan diğer hücreleri uzayda Şekil 2.1 (a)'da görülen $g1$ ve $g2$ genlerine göre daha iyi ayırır çünkü hücreler arası mesafeler daha yüksek olacaktır. Dolayısıyla $g3$ ve $g4$ geni sınıflama performansını $g1$ ve $g2$'ye göre daha fazla artıracaktır ve bunun sonucu olarak SMaSH $g3$ ve $g4$ genlerini tercih edecektir. Ancak $k$ hücre tipi için işaretçi seçimi yapılıyorsa $k$ hücreleri haricindeki hücrelerde ekspresyon eğer mümkünse olmamalıdır yani $g1$ ve $g2$ genleri tercih edilmelidir.

Yazılımsal olarak ise SMaSH seyrek (sparse) matrisler ile çalışamamaktadır. Ancak scRNA-seq verilerinin giderek büyüyen yapısı nedeniyle veriler sparse vb. gibi daha verimli formatlarda depolanmaktadır. Bu durum kullanılabilirlik ve uygulanabilirlik açısından bir eksik olarak değerlendirilebilir.

COSG ise SMaSH'ta belirtilen probleme benzer eksiklikler gösterebilir. COSG'nin hesapladığı kosinüs mesafesi iki vektör arasındaki açıları, diğer bir deyişle iki vektörün yönelimlerini değerlendirdiğinden yönelimi doğru, ama hatalı olan genleri seçebilir. Ayrıca mesafe hesapları RAM kullanımı açısından da maliyetli işlemlerdir.

scGeneFit optimizasyonla optimal bir gen setini bulur. Ancak bu optimal gen seti bir bütün olarak değerlendirildiğinde ayırıcıdır. Yani bu gen setindeki herhangi bir gen bir hücre tipi için ayırıcı olmayabilir veya ayırıcı olsa bile Bölüm 2.2'de belirtilen kriterleri karşılayamayabilir. Çünkü bir genin diferansiyel eksprese edilmiş olması da önerilen algoritmada hücre tiplerini ayırt etmesine neden olabilir. Ayrıca işaretçi seçiminde kullanılan optimizasyon yüksek sayıda gen içeren veri setlerinde uzun hesaplama sürelerine ve yüksek hesaplama maliyetlerine neden olabilir. Yazılımsal olarak ise scGeneFit SMaSH''te olduğu gibi seyrek matrislerle çalışmaz. Bu durum daha önce belirtildiği gibi kullanılabilirlik açısından büyük bir eksiktir.

### 2.4 Tez kapsamında önerilen yöntem: scMAGS (single cell MArker Gene Selection)

scMAGS uzamsal transkriptomik çalışmaları için gereken işaretçi seçimi işlemini gerçekleştirebilen bir Python paketidir. Python 3 ve üzerindeki sürümlerde işletim sisteminden bağımsız olarak çalışabilir. Paket içerisinde dahili olarak bir CLI (Command Line Interface) oluşturulmuştur ve bu sayede herhangi bir IDE (Integrated Development Environment) ihtiyacı duymadan terminal ortamından da çalıştırılabilir. İşaretçi seçimi yapılırken süreci hızlandırmak amacıyla bir paralel arayüz ile çalışır ve kullanıcı isteğine bağlı olarak kullanılacak çekirdek sayısı ayarlanabilir. scMAGS giriş verisi olarak satırları hücrelere sütunları genlere karşılık gelen scRNA-seq sayım (count) matrisini alır. Giriş verisi, scRNA-seq verilerindeki yüksek sıfır oranları sebebiyle algoritma verimliliği açısından seyrek (sparse) veya yoğun (dense veya normal) matris şeklinde olabilir. Ayrıca kümeye özgü işaretçi seçimi yapıldığından küme etiketleri de zorunludur. Paket içerisinde işaretçi seçimine ek olarak seçilen işaretçiler görselleştirilebilir. scMAGS paketinin kaynak kodu *https://github.com/doganlab/scmags* adresinden temin edilebilir ve Python Paket Dizini (PyPı) yazılım deposundan ***pip install scmags*** komutu ile indirilebilir. Paketin kullanımına ve içeriğine dair oluşturulan dökümantasyon web sayfasına *https://scmags.readthedocs.io/en/latest/* adresinden ulaşılabilir.

### 2.4.1 Ön işleme ve normalizasyon

Klasik bir scRNA-seq deneyi genellikle 20.000'den fazla gen içerir. Tipik bir damlacık tabanlı (Drop-Seq) scRNA-seq verisi, ifade matrisinde %90'a kadar sıfır değerleri bulunabilmektedir [22, 25, 82, 83]. Yüksek miktardaki gen sayısı birkaç hücreden fazlasında ifade edilmeyen veya tüm hücrelerde aynı düzeylerde ifade edilen, dolayısıyla hücresel heterojenite hakkında bilgi vermeyen genlerin filtrelenmesi sonucu büyük oranda azaltılabilmektedir [22, 25, 82, 83]. Hücre tipleri arasında ayrım yapmayan genlerin elenmesi algoritmanın verimliliği ve ölçeklenebilirliği açısından çok önemlidir.

scMAGS ilk aşama olarak tüm genlerin tüm hücre tiplerindeki ekspresyon oranlarını hesaplamaktadır. Daha sonra tüm hücre tiplerinde %20'den az eksprese edilmiş genleri filtrelemektedir. Bu aşamada scRNA-seq'in seyrek verisinin büyük bir kısmı filtrelenmiş olur ve bu durum RAM kullanımına olumlu katkıda bulunmaktadır.

Hücresel heterojenite açısından anlamsız genlerin elenmesi sonrasında diğer adım normalizasyondur. Normalizasyon sayımlardan ekspresyon seviyelerine geçiş aşamasıdır ve sayım matrisi deneysel etkilerin getireceği biasların azaltılması amacıyla normalize edilmelidir [38]. Her hücredeki bir gen için okuma sayısının, gene özgü ifade düzeyi ve hücreye özgü ölçekleme faktörleriyle (rastgele) orantılı olması beklenmektedir [84]. Ancak hücreler özdeş bile olsa sayım derinlikleri her hücre için farklı olabilmektedir. Bu probleme neden olarak; minimum mRNA materyali içeren hücrelerden kütüphane hazırlamanın zorluğu, cDNA yakalama veya PCR amplifikasyonundaki teknik bias örnek olarak gösterilebilir [38].

Normalizasyon aşaması hücreler arasında doğru gen ekspresyon değerlerini elde etmek başka bir deyişle tüm hücreleri aynı şartlar altına getirmek için sayım verilerini ölçeklemekte ve bu sorunu halletmektedir [84]. Eğer normalizasyon yapılmadan işaretçi seçimi yapılırsa, teknik biaslar nedeniyle yanlış işaretçi seçimi yapılabilir. Bu nedenle gen filtreleme öncesinde tüm sayım (count) matrisine $log(1 + x)$ dönüşümü uygulanmaktadır. $log(1 + x)$ dönüşümünün getireceği sonsuz değerlerden kaçınmak için ise eşitliğe bir sözde (pseudo) sayı $(+1)$ eklenmektedir.

### 2.4.2 Kümeye özgü gen filtreleme

Hücre tipleri için işaretçi seçimi yaparken her hücre tipi için genlerin tamamının taranmasına gerek yoktur. Gen filtreleme scMAGS paketini diğer yöntemlerden ayıran en önemli özelliklerden biridir. Giderek büyüyen scRNA-seq veri setlerinde işaretçi seçimi işleminin

tüm genler üzerinden yürütülmesi hesapsal zorluklara ve yüksek RAM ihtiyacına neden olabilmektedir. scMAGS kümeye özgü gen filtreleme adımıyla tüm hücre tipleri için aday işaretçi genleri belirlemekte ve işaretçi seçimi aday işaretçi genler üzerinden yürütülmektedir. Bu durum RAM kullanımını ve hesaplama maliyetlerini düşürmektedir.

Uzamsal transkriptomik probları için işaretçi gen seçerken dikkat edilmesi gereken kriterlerden Bölüm 2.2'de bahsedilmiştir. Ancak bu kriterler arasında bir ödünleşim (trade-off) bulunmaktadır. Şöyle ki bazı durumlarda bir gen küme içi yüksek ekspresyona ve küme dışı düşük ekspresyona sahip olabilir, ancak bu ekspresyon değerlerinin ortalaması düşük seviyelerde olabilmektedir. Ancak ekspresyon seviyesi de yüksek olmalıdır çünkü ekspresyon seviyesinin düşük olması, deneylerde hücre tiplerinin yani ayırt edilememesine yol açabilir. Veya başka bir durumda gen, küme içi yüksek ekspresyon ve ortalamaya sahip olup aynı zamanda küme dışında da yüksek ekspresyon oranına sahip olabilir. Sonuç olarak tüm kriterler aynı anda sağlanamayabilir. Bu ve bunun gibi birçok nedenden dolayı bu kriterlerden bazılarına öncelik verilmiştir. Kriterler uzamsal transkriptomik deneyleri baz alınarak en önemliden en önemsize doğru sıralanmış ve filtre algoritması bu sıraya önem verecek şekilde geliştirilmiştir. Bu kriterler şu şekilde sıralanabilir;

1. Küme içi yüksek ekspresyon oranı
2. Küme içi ortalamanın diğer küme içi ortalamalara göre büyük olması
3. Küme dışı düşük ekspresyon oranı
4. Küme içi ortalama ile diğer kümeler arasındaki en yüksek ortalama arasındaki farkın yüksek olması

Sıralanan kriterlere göre filtreleme algoritmasının adımları herhangi bir $k$ kümesi için şu şekilde sıralanabilir;

1. Küme içi ekspresyon oranı eşiğinin otomatik olarak Eşitlik 2.1'e göre veya kullanıcı tarafından belirlenmesi (Eşikler tüm kümeler için kendi küme içi ekspresyon oranı dağılımlarına göre tanımlanır.)
2. Küme içi ekspresyon oranı eşiğinin altında kalan genlerin $k$ kümesi için filtrelenmesi (1. kriterin sağlanması için)
3. Kalan genlerden $k$ kümesinde maksimum ortalamaya sahip olanlarının haricindekilerinin filtrelenmesi (2. kriterin sağlanması için)
4. Maksimum ortalamaya ve yüksek ekspresyona sahip genlerin Eşitlik 2.2'de verilen dağılım metriğinin hesaplanması

5. Dağılım metriği değerlerinin sıralanması ve daha sonra en yüksek ilk $t$ tanesinin (varsayılan 10) $k$ kümesi için aday işaretçi gen olarak belirlenmesi (3. ve 4. kriterin sağlanması için)

5. adım sonrasında $t$ adet gen bulunup bulunmadığı kontrol edilir. Eğer $t$ adet aday gen yoksa öncelikle ekspresyon oranı eşiği, eşiğin %1'i kadar düşürülüp adımlar tekrarlanır ve tekrar aday işaretçi gen sayısı kontrol edilir. Bu işlem en fazla 10 kez olmak üzere $t$ adet gen elde edilene kadar tekrarlanır. Eğer eşik 10 kez düşürüldükten sonra yine $t$ adet gen elde edilemezse öncelikle önceki denemelerden elde edilen genler varsa kaydedilir, daha sonra sıralanan 2. kriter değiştirilir. $k$ kümesinin küme içi ekspresyon ortalaması diğer küme içi ekspresyon ortalamaları arasında 2. sırada olan genler için yukarıda belirtilen 5 adım tekrarlanır. Tekrar aday işaretçi gen sayısı kontrol edilir ve bu işlemde $k$ kümesinin ekspresyon ortalamasının sıralaması küme sayısına eşit olana dek tekrarlanır. Bu adımlara göre tüm kümeler için filtreleme işlemi tekrarlanır ve filtreleme sonrası belirlenen aday işaretçi genler üzerinden asıl işaretçi genler seçilir. Eğer bu adım sonrasında da yeterli gen sayısına ulaşılamazsa algoritma işaretçi gen seçimine elde edilen genlerle devam eder.

### 2.4.3 Kümeye özgü işaretçi gen seçimi

scMAGS'ın işaretçi gen seçmek için gerçekleştirdiği son adım kümeye özgü işaretçi gen seçimidir. Gen seçimi her küme için filtreleme sonrası kalan genlerin ortalama Silhouette indeksine veya büyük veri setleri (100.000 hücreden fazla hücre içerenler) için Calinski-Harabasz indeksine bağlı olarak gerçekleştirilir. Silhouette indeksi, bir örneğin diğer kümelere kıyasla kendi kümesine ne kadar benzediğinin bir ölçüsüdür. $-1$ ile $+1$ arasında değişmektedir. Burada yüksek bir değer nesnenin kendi kümesiyle iyi eşleştiği ve komşu kümelerle zayıf şekilde eşleştiğini göstermektedir. Varyans oranı kriteri olarak bilinen Calinski-Harabasz indeksi ise kümeler arası ve küme içi dağılımının yani mesafelerin karesinin oranıdır ve kümeler yoğun olduğunda ve ek olarak birbirlerinden iyi ayrıldığında skor daha yüksek bir değer almaktadır. Büyük veri setlerinde kullanılmasının sebebi ise hesaplama süresinin daha hızlı ve verimli olmasıdır. Ancak küçük veri setlerinde Silhouette indeksini kullanmak, Calinski-Harabasz indeksine göre tercih edilmiştir çünkü 2013 yılında yapılan ve 30 adet kümeleme değerlendirme indeksini kapsayan bir çalışmada yapılan değerlendirmeler sonucunda birçok alanda Silhouette indeksi en başarılı indeks olarak gösterilmiştir [85]. Ek olarak Calinski-Harabasz indeksi yapılan çalışmada Silhouette indeksinden sonra en başarılı indeks olarak değerlendirilmiştir.

Silhouette veya Calinski-Harabasz indeksini kullanmadaki amaç ise şu şekilde açıklanabilir; herhangi bir $k$ kümesi için filtreleme sonrası kalan 10 adet aday işaretçi gen içerisinden 3 adet asıl işaretçi gen seçilecek olsun. Kalan 10 adet gen filtrelemenin sonucu olarak, küme içi yüksek ve küme dışında düşük ekspresyon oranına sahiptir ancak küme içi yüksek ekspresyon oranı eksprese edilen değerlerin birbirine yakın olacağı dolayısıyla hücre tiplerini iyi eşleştireceği anlamına gelmez. Veya küme dışı düşük ekspresyon, düşükte olsa eksprese edilen değerlerin birbirine yakın olacağı anlamına gelmez. Ancak seçilecek genlerin küme içinde yüksek ekspresyona sahipken ekspresyon değerlerinin de birbirine yakın ve uyumlu olması yani hücre tiplerini kendi içlerinde iyi eşleştirmesi gerekmektedir. Değerlerin birbirine yakın olmaması, hücre içerisinde yüksek değişkenlik gösteren ekspresyon profilleri veya aykırı değerler hücrelerin yanlış konumlandırılmasına sebep olabilir.

Eğer problem bir kümeleme problemi olarak ele alınırsa, kalan genlerin Silhouette indeksini hesaplamak ve Silhouette indeksini maksimize eden genleri seçmek; kısacası Silhouette indeksini bu problem için bir "cost" fonksiyonu olarak ele almak bu problemi çözebilir. Aynı durum Calinski-Harabasz indeksi için de geçerlidir ve buradan sonra yapılacak olan açıklamalar ve mantıksal yaklaşım Calinski-Harabasz indeksinde de geçerlidir. İki indekste aynı amaca hizmet ettiğinden kelime karmaşası olmaması için bu kısımdan sonra sadece Silhouette indeksine değinilerek açıklamalar yapılacaktır. Şöyle ki bir veri seti eğer iyi kümeleniyorsa yani küme içerisindeki elemanlar birbirleri ile iyi eşleşiyor ve komşu kümelerle olan mesafeleri yüksekse Silhouette değeri de yüksek çıkacaktır. Bu durumdan yola çıkarak, aday işaretçi genler arasından verinin optimum şekilde kümelenmesini sağlayan genlerinde Silhouette indeksi yüksek çıkacaktır. Ancak silhouette indeksini hesaplamadan önce etiketlerin düzenlenmesi gerekir, çünkü amaç her bir $k$ kümesini diğer kümelerin tamamından ayıran işaretçi genleri bulmaktır. Bu nedenle herhangi bir küme için işaretçi seçimi yapılırken Silhouette indeksinin hesaplanması gereken etiketler, binary olacak şekilde düzenlenmektedir. Bu durumda $k$ adet hücre tipi yani küme içeren bir veri setinde $k$ kümesi için gen seçimi yapılıyorsa; etiketler $k$ kümesini bir küme, geriye kalan tüm kümeleri aynı kümeye atayacak şekilde yeniden düzenlenir. Etiketler bu şekilde düzenlenirse Silhouette indeksinin yüksek olduğu genler $k$ kümesini diğer tüm kümelerden daha iyi ayıran genler olacaktır.

Örneğin herhangi bir $k$ kümesi için işaretçi gen seçiliyor olsun; aday işaretçi genlerin filtreleme işlemi sonucunda $k$ kümesi dışındaki kümelerde ekspresyon oranları düşüktür

ancak eksprese edilen düşük miktarda hücre eğer $k$ kümesi içerisindeki profile benzer bir profil sergilerse küme dışı hücreler $k$ kümesi içerisine dahil edilebilir. Silhouette indeksi bu kısımda devreye girer ve $k$ kümesi ile diğer tüm kümelerin optimum şekilde ayrılmasını sağlayan işaretçilerin veya işaretçi kombinasyonlarının seçilmesine yardımcı olur.

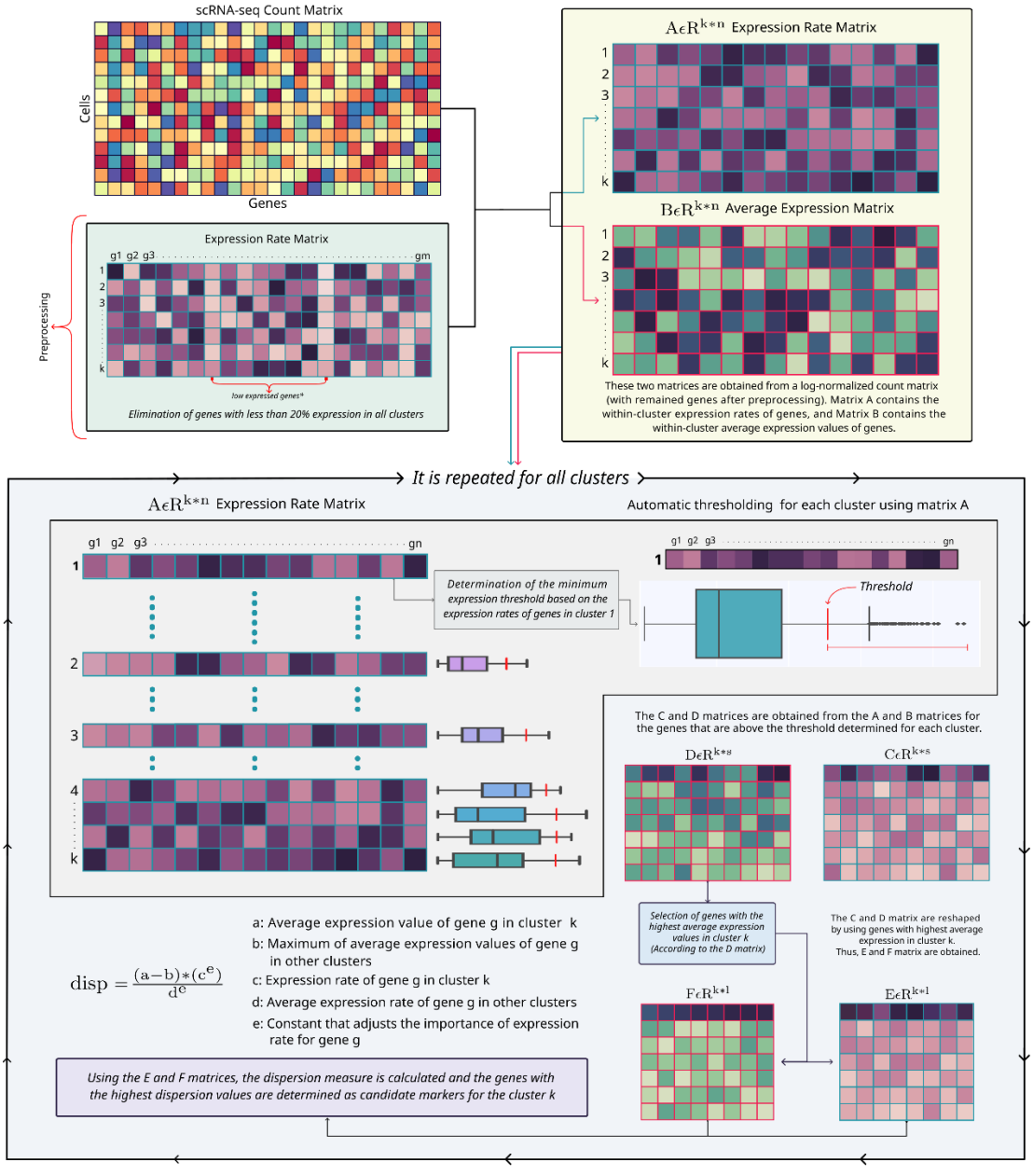
Etiketler düzenlendikten sonra Silhouette indeksi hesaplanırken farklı yollar izlenebilir bunlar, dinamik (kombinasyonel olarak indeks değerlerine bağlı) ve normal (indeks değerlerine bağlı) olmak üzere iki şekildedir. Normal programlamada herhangi $k$ kümesi için oluşturulan etiketlere bağlı olarak hesaplanır ve sıralanır. Sıralama sonrası en yüksek Silhouette indeksine sahip $n$ adet gen $k$ kümesi için işaretçi olarak nitelendirilir. Normal programlama hesapsal açıdan daha verimlidir ancak kombinasyonel ilişkileri ele almaz.

Eğer genleri kombine etmek ve bu kombinasyonlar arasından en iyisi seçilmek isteniyorsa dinamik programlama seçeneği kullanılabilir. Dinamik programlamada ise şu şekilde bir yol izlenir; herhangi $k$ kümesi için 10 adet işaretçi gen arasından 5 adet işaretçi gen seçilecek olsun. Öncelikle $k$ kümesi için oluşturulan etiketlere bağlı olarak tüm genlerin Silhouette değeri hesaplanır. Daha sonra Silhouette değeri maksimum olan gen $k$ kümesi için 1. işaretçi gen olarak nitelendirilir. Daha sonra kalan 9 gen 1. işaretçi gen ile teker teker kombine edilerek her 2'li kombinasyonun silhouette değeri hesaplanır. 2'li kombinasyonlarda maksimum silhouette değerini üreten kombinasyondaki 2. gen 2. işaretçi gen olur. Bu işlemler 5 adet işaretçi gen seçilene kadar tekrarlanır ve işlem sonunda kombinasyonel olarak en yüksek silhouette değerine sahip 5 adet gen seçilmiş olur. Bu seçim kullanıcıya bağlıdır. Kullanıcı isteklerine göre dinamik veya normal programlama seçeneği kullanılır. Bu kısma kadar anlatılan iş akışı büyük veri setlerinde farklı olarak sadece silhouette indeksi yerine, hesapsal verimlilik için Calisnki-Harabasz indeksi kullanılarak gerçekleştirilir.

### 2.4.4 Algoritmanın genel iş akışı ve matematiksel alt yapısı

scMAGS paketi Bölüm 2.3'te bahsedildiği gibi giriş verisi olarak satırları hücreler sütunları genlere karşılık gelen scRNA-seq verisini alır. Giriş verisinin ön İşlemden geçirilmesi sonrasında $X \in R^{m \times n}$, $n$ hücre ve $m$ gen ile elde edilen $log(1+x)$ dönüştürülmüş ekspresyon matrisi olsun. Ayrıca $X$, $k$ adet hücre türü, tipi veya durumu içersin. Bu durumda her hücre tipi veya durumu ayrı bir küme olarak kabul edilebilir. Buradan $X = \{x_1, x_2, \dots, x_k\}$ şeklinde yazılabilir ve $x_1, x_2, \dots, x_k$ toplamda $m$ adet gen içeren ancak her biri farklı hücre tiplerini içeren küme matrisleri olur. Gen filtreleme için öncelikle $A \in R^{k \times m}$ matrisinin hesaplanması gerekir. Burada $A$, $k$ adet satır, $m$ adet sütundan oluşur ve

tüm genlerin küme içi ekspresyon oranlarını içerir. Örneğin $A$ matrisinin 3. Satırının 5. sütunundaki elemanı, 3. kümedeki hücrelerin 5 numaralı geni hangi oranda eksprese ettiğine dair $0-1$ arası bir değer içerir. $A$ matrisi hesaplandıktan sonra $A$ matrisinde tüm satırları $0.2'$den küçük olan yani tüm hücre tiplerinde %20'den az eksprese edilen genler elenir. Elenen genler $A$ matrisinden çıkarılır ve $B$ matrisi de kalan genler için genlerin küme içi ortalama ekspresyon değerlerini içerecek şekilde hesaplanır. Buna göre $A$ ve $B$ matrisleri $k$ adet satır $n$ adet sütun içerirler. $A$ ve $B$ matrisleri üzerinden gen filtreleme işlemi her bir $k$ kümesi için aşağıdaki adımlar takip edilerek gerçekleştirilir. Bu adımlar Şekil 2.2'de ayrıntılı olarak gösterilmiştir.




**Şekil 2.2:** Kümeye özgü gen filtreleme iş akışı

1) $A$ matrisinin $k$ kümesine karşılık gelen satırı kullanılarak öncelikle $k$ kümesi için aday işaretçi genlerin sahip olması gereken minimum küme içi ekspresyon oranı eşiği belirlenir. Her $k$ kümesi için Eşitlik 2.1'de verilen denkleme göre hesaplanır. Denklemde $Q_3$ %75'lik dilime karşılık gelen değeri yani 3. çeyreği, IQR (Interquartile Range) ise çeyrekler arası ağırlığı temsil eder.

$$Eşik = 0.5 \times (2 \times Q_3 + 1.5\, IQR) \qquad \text{Eşitlik 2.1}$$

2) Eşik belirlendikten sonra eşiğin üzerindeki genler ile $A$ matrisinden $D \in R^{k \times s}$ matrisi oluşturulur. Burada $s$ eşiğin üzerindeki genlerin sayısını temsil eder.
3) $k$ kümesinde yüksek ekspresyona sahip genlerin küme içi ekspresyon değerlerini içeren $D$ matrisinden hangi genlerin $k$ kümesinde maksimum ortalamaya sahip olduğu bulunur ve bu genler ile $D$ matrisinden $F \in R^{k \times l}$, $C$ matrisinden de E $\in R^{k \times l}$ matrisi elde edilir. Burada $l$, $k$ kümesinde hem yüksek ekspresyona hem de maksimum ortalama ekspresyon değerine sahip genlerin sayısını ifade eder.
4) Kalan tüm genler için $E$ ve $F$ matrisini kullanarak dağılım metriği hesaplanır. Dağılım metriği Eşitlik 2.2'deki gibi ifade edilir ve herhangi bir $g$ geni için hesaplanırken aşağıdaki değerler kullanılır.

$$disp = \frac{(a - b) * (c^e)}{d^e} \qquad \text{Eşitlik 2.2}$$

**a**: $g$ geninin $k$ kümesindeki ortalama ekspresyon değeri ($F$ matrisinden elde edilir.)

**b**: $g$ geninin $k$ kümesi harici kalan kümelerdeki ortalama ekspresyon değerlerinin maksimum değeri ($F$ matrisinden elde edilir.)

**c**: $g$ geninin $k$ kümesindeki ekspresyon oranı

**d**: $g$ geninin $k$ kümesi haricinde kalan kümelerdeki ekspresyon oranlarının ortalaması ($E$ matrisinden elde edilir.)

**e**: Ekspresyon oranının önemini ayarlayan sabit (Default = 10)

5) Hesaplanan dağılım metrikleri sıralanır ve en yüksek metrik değerine sahip olan $t$ adet (Default = 10) gen $k$ kümesi için işaretçi seçiminde kullanılmak üzere aday işaretçi gen olarak saklanır.
6) Eğer $t$ gen elde edilemezse 1. adıma dönülür ve ekspresyon eşiği kısıtlaması %1 oranında düşürülür. Tüm işlemler tekrarlanır ve istenilen sayıda gen elde edilip edilmediği kontrol edilir. Bu işlem $t$ adet gen elde edilene kadar 10 kez tekrarlanır.

7) Eğer 6. adımdaki işlem sonucunda yine $t$ adet gen elde edilemezse bu kez ortalamanın genlerin filtrelenen kümede maksimum olmasına dair olan kriter kaldırılır ve ilk 5 işlem tekrarlanır.

Aday işaretçi genler seçildikten sonra kümeye özgü aday işaretçi gen setinden işaretçi genleri seçmek için Silhouette veya Calinski-Harabasz indeksi ile birleştirilmiş normal veya dinamik olmak üzere iki yaklaşım kullanılır. Bu yaklaşımlar için gerekli olan Silhouette indeksi aşağıda belirtilen şekilde hesaplanır;

Bir $o_i$ nesnesinin Silhouette değerini hesaplamak için öncelikle $o_i \in C_k$ ve $C_k$ kümesindeki tüm $o_i$ nesneleri arasındaki küme içi ortalama uzaklığı temsil eden $a(i)$ değeri Eşitlik 2.3'teki gibi hesaplanır. Eşitlik 2.3'teki $n_k$, $k$ kümesindeki nesnelerin sayısını, $d(o_i, o_{i'})$ ise $o_i$ ile $o_i$ nesneleri arasındaki mesafeyi temsil eder.

$$a(i) = \frac{1}{n_{k^{-1}}} \sum_{\substack{i' \in I_k \\ i' \neq i}} d(o_i, o_{i'}) \qquad \text{Eşitlik 2.3}$$

$o_i$ ile diğer tüm kümelerdeki nesneler arasındaki ortalama uzaklıkları temsil eden $\delta(o_i, C_{k'})$ ise Eşitlik 2.4'teki gibi hesaplanır.

$$\delta(o_i, C_{k'}) = \frac{1}{n_{k'}} \sum_{i' \in I_{k'}} d(o_i, o_{i'}) \qquad \text{Eşitlik 2.4}$$

Bu ortalama mesafelerin en küçüğü $b(i)$ olsun,

$$b(i) = \min_{k' \neq k} \delta(o_i, C_{k'}) \qquad \text{Eşitlik 2.5}$$

Her $o_i$ nesnesi için $s(i)$ ile temsil edilen silhouette değeri Eşitlik 2.6'daki gibi hesaplanır.

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i),\ b(i)\right)} \qquad \text{Eşitlik 2.6}$$

Belirli bir $C_k$ kümesi için silhouette değerlerinin ortalaması, ortalama silhouette değeri olarak adlandırılır ve Eşitlik 2.7'deki gibi hesaplanır.

$$\sigma_k = \frac{1}{n_k} \sum_{i \in I_k} s(i) \qquad \text{Eşitlik 2.7}$$

Son olarak verilerin global silhouette indeksi, Eşitlik 2.8'deki gibi tüm kümelerin ortalama silhouette değerlerinin ortalaması alınarak hesaplanır [86].

$$S = \frac{1}{K}\sum_{k=1}^{K}\sigma_k \qquad \text{Eşitlik 2.8}$$

$10^5$'ten fazla sayıda hücre içeren veri setleri için kullanılan Calinski-Harabasz indeksi ise aşağıda belirtilen şekilde hesaplanır;

Küme etiketleri bilinen ve $k$ adet küme içeren bir veri setinin Calinski-Harabasz indeksinin hesaplanması için öncelikle $W_k$ ve $B_k$ matrislerinin hesaplanması gerekir. $W_k$ ve $B_k$ matrisleri aşağıdaki Eşitlik 2.9 ve 2.10'daki gibi hesaplanır.

$$B_k = \sum_{q=1}^{k} n_q(\mu_q - \mu_E)(\mu_q - \mu_E)^T \qquad \text{Eşitlik 2.9}$$

Buradaki $B_k$ kümeler arası toplam varyansı $\mu_q$, $q$. kümenin merkezini, $\mu_E$, genel merkezi ifade eder.

$$W_k = \sum_{q=1}^{k}\sum_{x\in\mu_q}(x-\mu_q)(x-\mu_q)^T \qquad \text{Eşitlik 2.10}$$

$W_k$ ise küme içi varyansı temsil eden matristir. Elde edilen $W_k$ ve $B_k$ matrislerinin diyagonal (ilk köşegen) toplamları ve gözlem sayısını temsil eden $N_E$ kullanılarak Calinski-Harabasz indeksi Eşitlik 2.11'deki gibi hesaplanır [87].

$$s = \frac{tr(B_k)}{tr(W_k)} \times \frac{n_E - k}{k-1} \qquad \text{Eşitlik 2.11}$$

Eşitlik 2.11 ve Eşitlik 2.7'de verilen indekslerden herhangi biri hesaplanırken her $k$ kümesi için küme içerisindeki elemanlar aynı kümede, ancak küme dışındaki elemanların hepsi bir kümede toplanacak şekilde yeni küme etiketleri oluşturulur. İndeksler bu etiketler üzerinden hesaplanır.

Bu etiketlere bağlı olarak normal programlama yaklaşımında veri büyüklüğüne göre seçilen indeks her $k$ kümesi için, bu küme için belirlenen tüm aday genler için ayrı ayrı hesaplanır.

Hesaplanan indeks değerlerinden maksimum $t$ adet gen $k$ kümesi için işaretçi gen olarak seçilir.

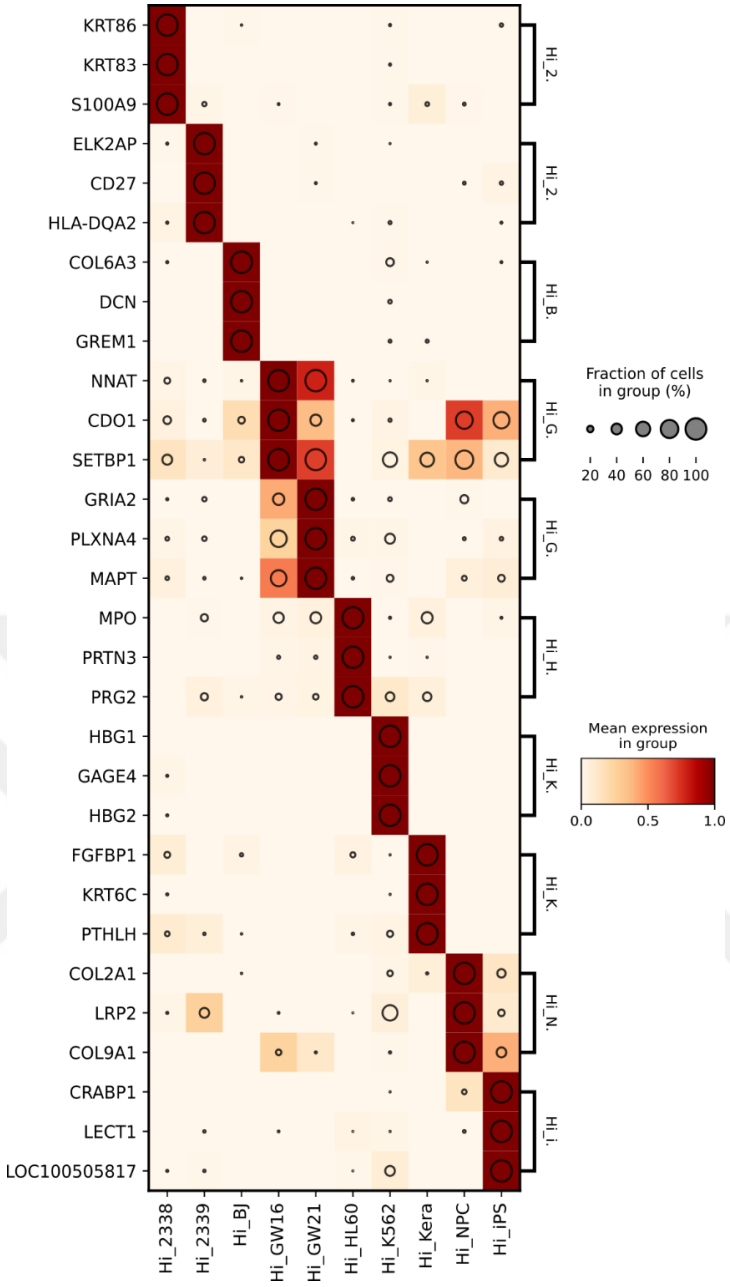
Dinamik programlama yaklaşımında ise yine aynı şekilde her küme için oluşturulan etiketler kullanılarak tüm aday genler için indeks değerleri hesaplanır. Daha sonra en yüksek indeks değerine sahip gen $t$ adet gen içerecek işaretçi kombinasyonu için 1. işaretçi gen olarak nitelendirilir. 2. gen için kalan genler 1. gen ile kombine edilir, her kombinasyon için indeks hesaplanır ve en yüksek indeks değerine sahip kombinasyondaki 2. gen 2. işaretçi gen olarak nitelendirilir. Bu işlemler $t$ adet optimum gen seçilene kadar tekrarlanır.

## 2.5 Geliştirilen Yazılım Paketi Dahilindeki Görselleştirme Yöntemleri

İşaretçi genler seçildikten sonra seçilen genler scMAGS paketi dahilinde 4 farklı şekilde görselleştirilebilir. Bunlar seçilen genler kullanılarak oluşturulan Dotplot, Heatmap, t-SNE ve k-NN ile sınıflama sonuçlarını değerlendirmek için kullanılan Karmaşıklık Matrisinden oluşur. Bu çizimler işaretçi genlerin görsel olarak analiz edilmesi ve kullanılabilirliğinin değerlendirilmesi açısından önem arz etmektedir.

### 2.5.1 Dotplot

Dotplot görselleri scanpy paketinden yararlanılarak oluşturulmuştur [77]. Bu görsel, seçilen işaretçi genlerin tüm kümelerdeki küme içi ortalama ekspresyon seviyelerini renklerle, küme içi ekspresyon oranlarını ise halkalarla temsil etmektedir. Her bir satır bir gene, her bir sütün bir kümeye, yani bir hücre tipine karşılık gelmektedir. Oluşturulabilmesi için öncelikle işaretçi genlerin seçilmiş olması gerekir. İşaretçi genler seçildikten sonra ilk adım olarak $A \in R^{kxt}$ ve $B \in R^{kxt}$ matrislerinin oluşturulması gerekir. Burada $t$ seçilen işaretçi gen sayısına $k$ ise küme sayısına karşılık gelir. $A$ matrisi seçilen işaretçilerin küme içi ekspresyon değerlerinin ortalamalarını, $B$ matrisi ise küme içi ekspresyon oranlarını içerir. Hesaplanan $A$ ve $B$ matrisleri tüm genlerin kendi içerisinde değerlendirilebilmesi için sütun yani gen bazında 0-1 arasında ölçeklenir. Bu matrislerden $A$ görselin renklerini, $B$ ise halkaların büyüklüklerini ayarlamak için kullanılır. Şekil 2.3'te Pollen veri seti için çizilmiş Dotplot görselinin örneği görülmektedir. Görselin sağ ve alt kenarlarında hücre isimleri görülmektedir. Alt kenardaki etiketler bulunduğu sütunun verilen etiketteki hücre tipine karşılık geldiğini göstermektedir. Sağ kenardaki braket (köşeli parantez) şeklindeki etiketler ise kapsadığı bölgedeki genlerin etiketteki verilen hücre tipi için seçildiğini gösterir. Sol kenardaki etiketler ise gen isimlerine karşılık gelir.

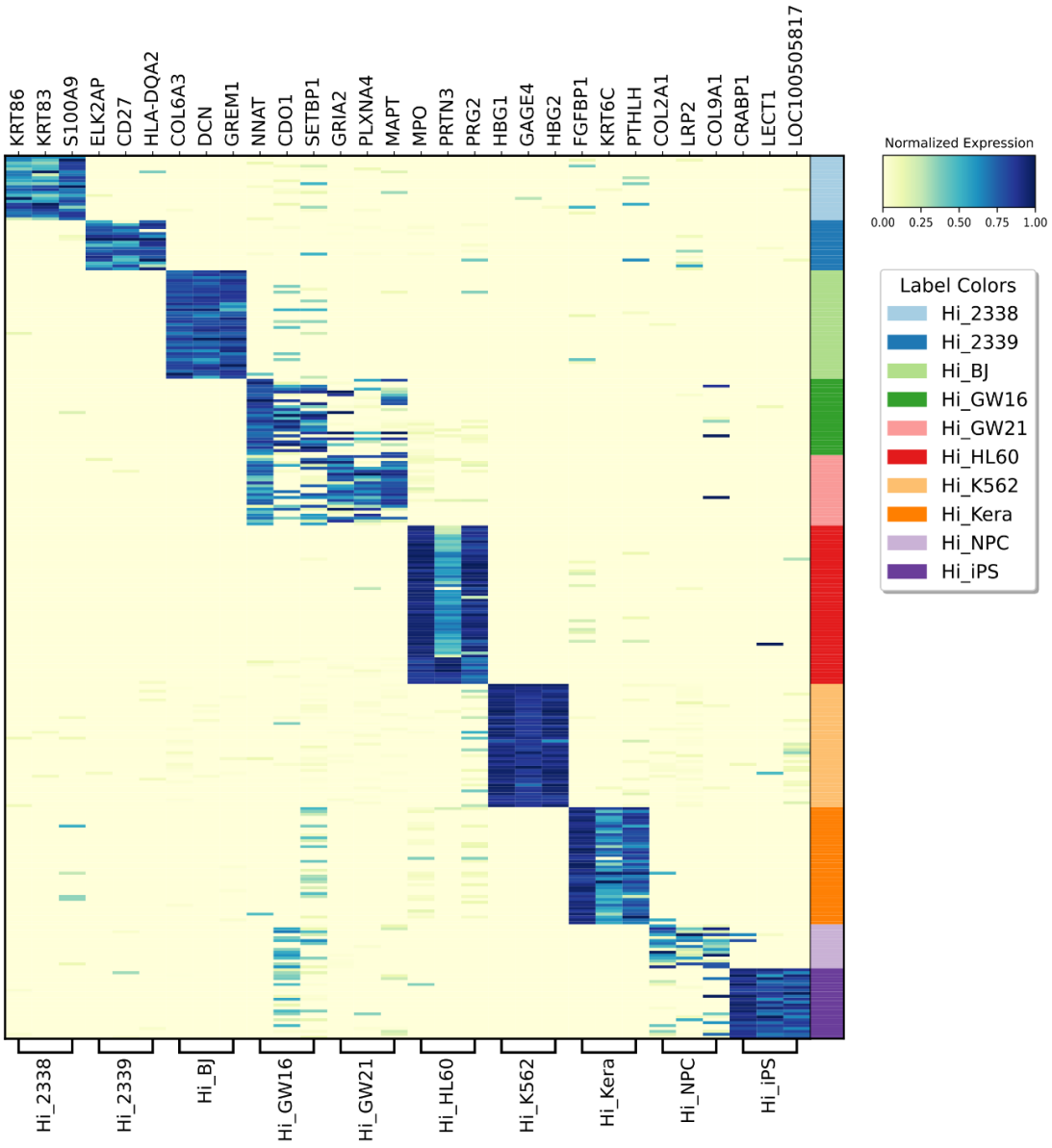



**Şekil 2.3:** Pollen veri seti için seçilen işaretçi genlerin Dotplot grafiği

### 2.5.2 Heatmap

Dotplot grafiği seçilen genler hakkında önemli bilgiler sunar, ancak genlerin ayrıntılı ekspresyon profillerini görmek için bir heatmap tasarlanmıştır. Bu heatmap'te alt kenara braketler eklemiştir ve her braket bir hücre tipini kapsamaktadır. Braketin kapsadığı alan dahilindeki genler, braketin etiketindeki hücre tipi için seçilen işaretçilere karşılık gelir. Şeklin sağ kenarına hücre tiplerini ayırt eden bir renk sütunu eklenmiştir. Ayrıca şeklin üst kenarına da yine gen isimlerine karşılık gelen etiketler eklenir. Bu heatmap'in oluşturulabilmesi için öncelikle seçilen işaretçi genlerin oluşturduğu veri matrisine, aykırı

ve yüksek ekspresyon değerlerinin etkisini gidermek için $log(1+x)$ dönüşümü uygulanır. $log(1+x)$ dönüştürülmüş matris renklendirme için kullanılır. Şekil 2.4'te Pollen veri seti için seçilen işaretçilerden oluşturulmuş heatmap görülmektedir.




**Şekil 2.4:** Pollen veri seti için seçilen işaretçi genlerin Heatmap grafiği

### 2.5.3 t-SNE

t-SNE 2008 yılında Geoffrey Hinton ve Laurens van der Maaten tarafından önerilmiş ve biyoinformatik analizleri için vazgeçilmez bir araç olmuştur [88]. scMAGS paketi içerisine, seçilen işaretçi kombinasyonunun hücre tiplerini nasıl ayırdığını 2 boyutta görebilmek için eklenmiştir. Şekil 2.5'te Tasic veri seti için seçilen işaretçi genlerin t-SNE grafiği görülmektedir. t-SNE $log(1+x)$normalize edilmiş seçilen işaretçi genlerin oluşturduğu

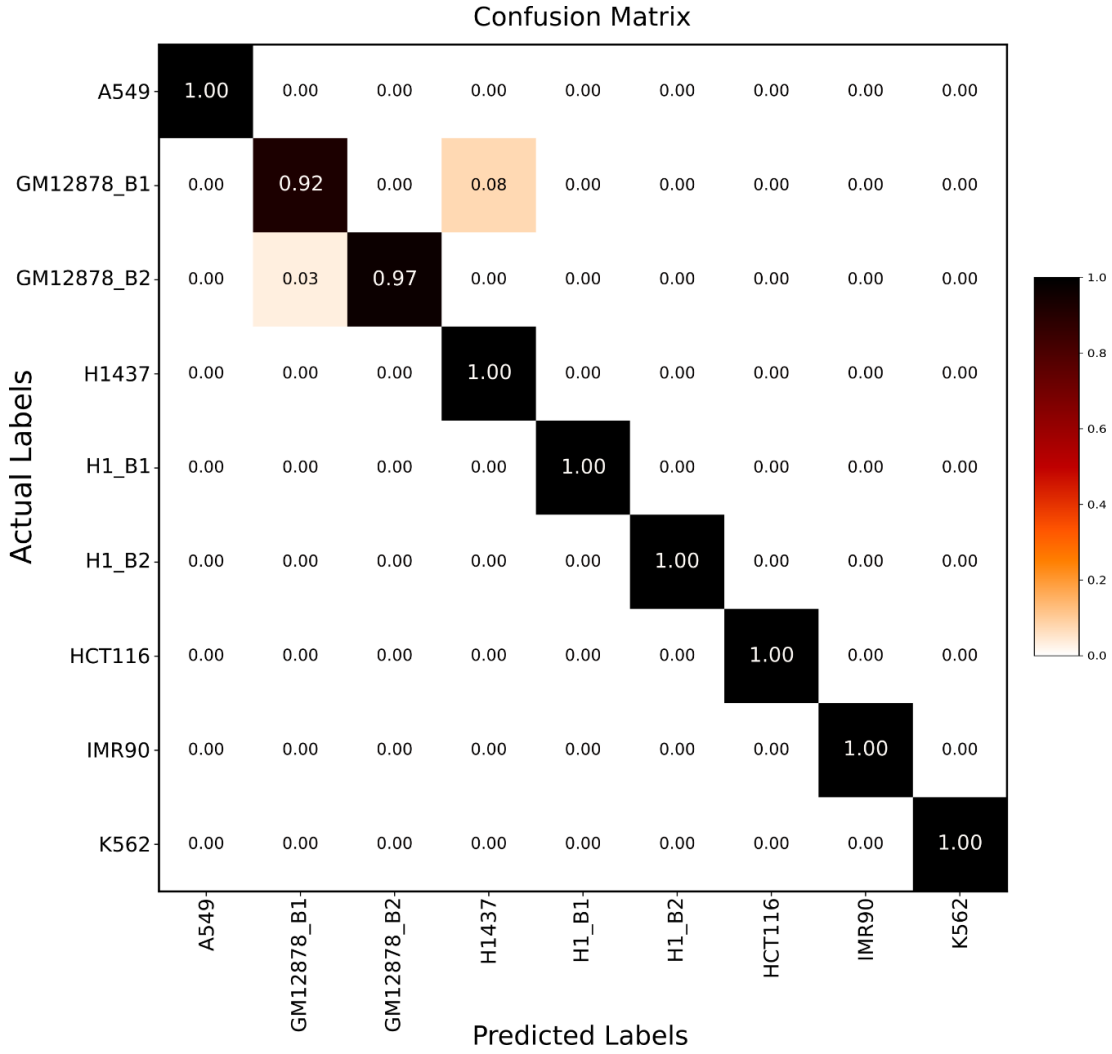
veri matrisi üzerinden hesaplanır. Hesaplanan t-SNE boyutlarından ilk ikisi ile dağılım grafiği çizilir ve kümelere göre renklendirilir.


**Şekil 2.5:** Tasic veri seti için seçilen işaretçi genlerin t-SNE grafiği

### 2.5.4 Confusion matrix (Karmaşıklık Matrisi)

scMAGS paketi içerisinde seçilen işaretçilerin k-NN sınıflandırması sonuçlarını değerlendirmek için sonuçlar karmaşıklık matrisi (Confusion Matrix) ile görselleştirilir. Sınıflandırma algoritmalarının performansını artırmaya yönelik gen seçmek yanlış gen seçimine yol açabilir. Ancak seçilen işaretçilerin değerlendirilmesi için k-NN sınıflandırma kullanılabilir. Bu nedenle seçilen işaretçileri k-NN ile sınıflandıran ve karmaşıklık matrisi hesaplayıp görselleştiren bir fonksiyon pakete dahil edilmiştir. Bu fonksiyon içerisinde

işaretçilerden oluşan veri seti %30 test %70 eğitim verisi olarak ayrıştırılır ve ayrıştırılan veriye k-NN algoritması uygulanıp sınıflama yapılır. Elde edilen sonuçlar ise karmaşıklık matrisi kullanılarak görselleştirilir. Şekil 2.6'da Li veri seti için seçilen işaretçilerle yapılan sınıflamanın sonuçlarını içeren karmaşıklık matrisi görülmektedir.




**Şekil 2.6:** Li veri setinin k-NN sonuçlarını içeren karmaşıklık matrisi

# 3. ALGORİTMALARIN PERFORMANSLARININ DEĞERLENDİRİLMESİ

## 3.1 Amaç

Bu bölümün amacı tez kapsamında geliştirilen algoritmanın doğruluğunun ve verimliliğinin çeşitli yöntemler kullanılarak değerlendirilmesi ve literatürde önerilen diğer yöntemlerle karşılaştırılmasıdır. Bu amaç doğrultusunda Bölüm 2.2'de belirtilen kriterlerin sağlanıp sağlanmadığının kontrol edilmesi için halka açık 15 adet scRNA-seq ve 2 adet sekanslama bazlı uzamsal transkriptomik veri seti kullanılarak literatürde önerilen yöntemler ile tez kapsamında geliştirilen algoritma karşılaştırılmıştır. Algoritmalar hesaplama süreleri, RAM kullanımı miktarları, seçtikleri genlerin ekspresyon karakterleri ve seçtikleri genlerin sınıflama performansı ile ilişkisi olmak üzere birkaç kritere bağlı olarak değerlendirilmiştir. Bu bölümde gerçekleştirilen tüm hesaplamalar 2 Adet Intel Xeon® Silver 4210R 2.40GHz 20 Çekirdekli işlemciye ve 128 Gb RAM'e sahip Dell Precision T7820 iş istasyonu üzerinde gerçekleştirilmiştir.

## 3.2 Kullanılan Veri Setleri

Çizelge 3.1'de kullanılan veri setleri, veri setlerinde bulunan hücre, gen ve küme sayıları görülmektedir. Veri setleri hem küçük hem de 2 milyona kadar hücre içeren büyük veri setlerini içermektedir. Yapılan değerlendirmeler yüksek sayıda gen içermeleri ve literatürde de yöntemlerin bu şekilde değerlendirilmesi sebebiyle scRNA-seq verileri üzerinden değerlendirilmiştir. Veri setlerinden 10 tanesi NCBI-GEO (Gene Expression Omnibus) halka açık genomik veri deposundan, 2 tanesi ArrayExpress (EMBL-EBI) halka açık işlevsel genomik veri havuzundan, 1 tanesi NCBI-SRP veri deposundan 1 tanesi Sanger enstitüsüne bağlı oluşturulmuş çalışmaya özgü web sitesinden, 1 tanesi 10X Genomics web sitesinden olmak üzere tamamı halka açık veri tabanlarından temin edilmiştir. Ayrıca scRNA-seq veri setlerine ek olarak sekanslama bazlı uzamsal transkriptomik veri setleri de değerlendirmelere dahil edilmiştir.

### 3.2.1 SRP041736 (Pollen)

Pollen ve arkadaşları sığ (düşük kapsamlı) sekanslamanın derinliklerini araştırmak amacıyla, mikroakışkanlar kullanarak 11 popülasyondan 301 tek hücre yakalamış ve transkriptomlarını scRNA-seq ile analiz etmişlerdir [89]. Düşük kapsamlı sekanslamanın hücre tipleri arasında ayrım yapıp yapmayacağını incelemek amacıyla sağlam farklılıklar göstermesi beklenen

kaynaklardan gelen hücreleri karşılaştırmışlardır. Bunlar pluripotent, cilt, kan ve sinir hücreleridir Sonuç olarak hücre başına en az 50.000 okuma gerektiğini ve hücreler arasında değişen bol miktarda gen kombinasyonunun hücrelerin sınıflandırmasına izin verdiğini keşfetmişlerdir. Buna ek olarak düşük kapsamlı scRNA-seq'nın yakından ilişkili hücre tiplerini heterojen popülasyondan ayırt etmekte yeterli olup olmadığını araştırmak için, nörojenez aşamaları sırasında gelişen insan korteksinden türetilen tek hücreleri de analiz etmişlerdir.

**Çizelge 3.1:** Kullanılan veri setleri ve erişim kodları

| Data-Set | Number Of Cells | Number Of Genes | Number Of Cell Types | Accession Code |
|---|---|---|---|---|
| **Biase** | 56 | 25737 | 4 | GSE57249 |
| **Yan** | 90 | 20124 | 7 | GSE36552 |
| **Zeisel** | 3005 | 19972 | 9 | GSE60361 |
| **Li** | 561 | 55186 | 9 | GSE81861 |
| **Tasic** | 1679 | 24150 | 17 | GSE71585 |
| **Xin** | 1600 | 39851 | 8 | GSE81628 |
| **Darmanis** | 466 | 22088 | 9 | GSE67835 |
| **Baron** | 1937 | 20125 | 14 | GSE84133 |
| **Treutlein** | 80 | 23271 | 5 | GSE52583 |
| **Kolodziejczyk** | 704 | 38653 | 3 | E-MTAB-2600 |
| **Goolam** | 124 | 41480 | 4 | E-MTAB-3321 |
| **Pollen** | 301 | 23730 | 10 | SRP041736 |
| **Kleshchevnikov** | 40532 | 31053 | 9 | Sanger |
| **Bhaduri** | 1.3M | 27998 | 20 | 10X-Genomics |
| **Cao** | 2M | 26183 | 38 | GSE119945 |

### 3.2.2 GSE52583 (Treutlein)

Treutlein ve arkadaşları gelişen akciğerdeki çeşitli hücre tiplerinin, hiyerarşilerini tanımlamak ve transkripsiyonel durumlarını ölçmek için alveolar farklılaşmayı kapsayan 4 farklı aşamadaki 198 hücre üzerinde mikroakışkan scRNA-seq kullanmışlardır [90]. İlk aşamada gelişmekte olan bronşio-alveolar epitelinin hücresel bileşimini çözmek amacıyla gelişen fare akciğer epitelinin 80 ayrı canlı hücresinin transkriptomlarının sakkülasyonunun sonunda sekanslamışlardır. Bunun sonucunda bilinen işaretçi genleri kullanılarak; Clara-*Scgb1a1* ile, Ciliated-*Foxj1* ile, AT1-*Pdpn Ager* ile ve AT2-*Sftpc Sftpb*, genleri ile karakterize edilmiş ve bu hücre tipleri için önceden hiç belirtilmemiş veya önceden bilinen birçok işaretçi (marker) genler bildirilmiştir. Her hücre için en yüksek *p* değerine sahip birkaçı şu şekilde sıralanabilir; Ciliated: *1110017D15Rik-1700007G11Rik*, Clara: *Chad-Cyp2f2*, AT1: *Clic5-Akap5*, AT2: *Etv5-Lamp3*. İşaretçi seçimi yapılırken veri setinin 80 adet hücre içeren kısmı kullanılmıştır.

### 3.2.3 GSE67835 (Darmanis)

Darmanis ve arkadaşları yetişkin ve fetal insan beyninin hücresel karmaşıklığını tam bir transkriptom düzeyinde yakalamak için 466 hücre üzerinde scRNA-seq kullanmışlardır [91]. Bu 466 hücre beyindeki tüm ana hücre tiplerine göre sınıflandırılabilmiştir. Ayrıca sınıflandırılan toplulukların, klasik internöron belirteçleri kullanılarak tipik olarak gözlemlenen internöron alt tiplerinin kategorizasyonunu koruduğu gösterilmiştir. Kategorizasyonun korunduğunu göstermek amacıyla, fare internöronları sınıflandırmak için geleneksel olarak kullanılan genlerin; yani *GAD1, VIP CALB2 CCK, RELN, PVALB ve SST* genlerinin ekspresyonlarını araştırmışlardır. Elde edilen verileri analiz ederek; yetişkin beyninin kortikal nöronlarındaki çeşitliliği sorgulamış doğum öncesi ve sonrası nöronlar arasındaki gen ekspresyon progillerinin karşılaştırmalı bir analizini yapmış, yetişkin nöronlarında *MHCI* geninin kesin ekspresyonunu doğrudan gözlemlemiş ve fare insan arasındaki ekspresyon paternlerindeki farklılıkları belgelemişlerdir.

### 3.2.4 GSE84133 (Baron)

Baron ve arkadaşları, dört insan ve iki fare türünden 12.000'den fazla bireysel pankreas hücresinin transkriptomlarını analiz etmek için damlacık tabanlı (droplet-based) scRNA-seq kullanmışlardır [92]. Çalışmanın amacı pankreas hücre tipleri arasındaki heterojeniteyi tanımlamak, hücre tipleri içindeki alt popülasyonları keşfetmek ve hücre tiplerine özgü yeni işaretçi (marker) genleri tespit etmektir. Örneğin endokrin hücrelerde bilinen transkripsiyon

faktörlerini doğrulamışlardır; beta hücreleri için: *MAFA, NKX6-1,* alfa hücreleri için *IRX1, IRX2,* delta ve kanal hücreleri için *HHEX*, alfa ve beta hücreleri için *MAFB,* alfa, gamma ve epsilon hücreleri için *ARX*. Ancak bunlara ek olarak yeni endokrin spesifik transkripsiyon faktörlerini de keşfetmişlerdir; örneğin *POU3F1, SIX3* ve *OLIG1* delta hücrelerine özgü önceden bildirilmemiş bir faktördür. Çalışmada sağlıklı ve diyabetik donörler arasında gen ekspresyon profilleri karşılaştırılmış, diyabetik ve sağlılı donörler arasında farklı şekilde ifade edilen toplu (bulk-RNA-seq) verilerde tanımlanan çok sayıda genin muhtemelen yalnızca hücre tipi oranı farklılıkları nedeniyle değişken olduğunu, yani bu iki grup arasında diğer genlerin farklı ekspresyonunun maskelendiğini ortaya koymuşlardır. Hücre tipi oranı ayarlamalarından sonra alfa ve beta hücreleri arasında farklı şekilde eksprese edilen genler tespit etmişlerdir.

### 3.2.5 GSE81608 (Xin)

Xin ve arkadaşları diyabetik olmayan ve tip 2 diyabet organ donörlerinden 1492 insan pankreas alfa, beta, gamma ve PP (pancreatic polypeptide) hücrelerinin transkriptomlarını belirlemek için scRNA-seq kullanmışlardır [93]. Tip 2 diyabette ekspresyonu bozulmuş 245 genin yanı sıra hücre tipine özgü genleri ve yolları belirlemişlerdir. Fare ve insan alfa beta hücrelerindeki ekspresyon profillerini karşılaştırarak türe özgü ekspresyonu ortaya çıkarmışlardır. Örneğin fare alfa hücrelerinde düşük ekspresyona sahip ancak insan alfa hücrelerinde yüksek ekspresyona sahip genler ALDH1A1, PEMT, FXYD5 olarak listelenmiştir. Ayrıca hücre tipine özgü genleri de tanımlamışlardır; bu genler alfa hücreleri için: *GCG, DPP4, FAP, PLCE1, LOXL4, IRX2, TMEM236, IGFBP2, COTL1, SPOCK3* ve *ARDC4*, beta hücreleri için: *INS, ADCYAP1, IAPP, RGS16, DLK1, MEG3, INS-IGF2* ve *MAFA*, gama hücreleri için: *SST, BCHE, HHEX, RPLJP19* ve PP hücreleri için *PPY*' dir.

### 3.2.6 GSE71585 (Tasic)

Tasic ve arkadaşları yetişkin erkek farelerde birincil görsel korteksten 1600'den fazla hücreyi karakterize etmek, sınıflandırmak ve transkriptomik analizlerini gerçekleştirmek için FACS (Fluorescence-activated Cell Sorting) izolasyonu tabanlı scRNA-seq kullanmışlardır [94]. Sekanslama ile önceden bilinen ve yeni keşfedilmiş birçok işaretçi gen tanımlamışlardır. Bunlardan bazıları sadece tek bir hücre tipinde eksprese edilen benzersiz (unique markers) belirteçler, bazıları ise kombinatoryal (combinatorial markers) belirteçlerdir. Analiz sonucunda hücreleri 49 ayrı çekirdek kümeye ayırmış ve ana hücre sınıfları için bilinen işaretçilere dayanarak 23 GABAergic nöronal küme

($Snap25^{+}, Slc17a7^{-}, Gad1^{+}$), 19 glutametergic küme ($Snap25^{+}, Slc17a7^{+}, Gad1^{-}$), 7 nöronal olmayan küme ($Snap25^{-}, Slc17a7^{-}, Gad1^{-}$) tanımlamışlardır. Glutamaterjik hücreler için L2/3, L4, L5a, L5b, L6a ve L6b olmak üzere altı ana transkriptomik bazlı sınıfı tanımlamışlardır. Çalışmanın ana odağı nöronal hücre tipleri olsa da yedi nöronal olmayan hücre tipi bulmuşlardır bunlar: oligodendrositler, öncü hücreleri (precursor cells OPCs), iki tip oligodendrosit, astrositler, mikroglia, endotel hücreleri ve düz kas hücreleridir.

### 3.2.7 GSE81861 (Li)

Li ve arkadaşları 11 primer kolorektal tümörden ve uyumlu normal mukozadan scRNA-seq kullanarak, 561 tek hücrenin kolorektal tümörlerde ve bunların mikro ortamlarındaki transkripsiyonel heterojenitesini analiz etmişlerdir [95]. Tümör ve normal hücreler arası diferansiyel ekspresyon analizleri için tümör örneklerinden elde edilen en büyük 5 kümeye yani epitel, fibroblast, B hücreleri, T hücreleri ve miyeloid hücrelerine öncelik verip bu hücre tiplerinin normal mukozadaki ekspresyonlarını karşılaştırmışlardır. Karşılaştırma 129 adet, diferansiyel olarak eksprese edilmiş gen ile sonuçlanmıştır ve bu genlerin çoğu toplu (bulk) analizlerde tespit edilememiştir.

### 3.2.8 E-MTAB-3321 (Goolam)

Goolam ve arkadaşları amacı embriyonik ve ekstra embriyonik hücre kaderlerini belirlemek olan implantasyon öncesi gelişimin, ne zaman ve nasıl başlatıldığını incelemek için farelerde implantasyon öncesi ardışık aşamalardaki 124 hücrenin transkriptomlarını, Smart-seq2 scRNA-seq protokolü ile karakterize etmişlerdir [96]. Gelişim ilerledikçe farklı şekilde eksprese edilen genleri tanımlamak için 2 hücreli, 4 hücreli, 8 hücreli, 16 hücreli, ve 32 hücreli evresindeki hücreleri izole etmişlerdir. Bu evrelerden 2, 4, ve 8 hücreli evrelerinde sırasıyla 659, 1339, 813 oldukça değişken ekspresyona sahip gen ortaya çıkarmışlardır. Bu genlerden *Sox21* ve *Oct4* genlerinin 4 hücreli aşamasında, oldukça heterojen ekspresyon profili sergilediğini göstermiş ve *Sox21* seviyelerinin tükenmesinin, *Cdx2*'nin ve ekstra embriyonik kaderin düzenlenmesine yol açtığını belirlemişlerdir.

### 3.2.9 GSE36552 (Yan)

Yan ve arkadaşları farklı evrelerdeki insan preimplantasyon embriyolarından ve insan embriyonik kök hücrelerinden (hESC) 124 hücrenin transkriptomlarının analizi için scRNA-seq kullanmışlardır [97]. Embriyoları preimplantasyon gelişiminin yedi önemli aşamasındayken yani; oosit, zigot, 2 hücreli, 4 hücreli, 8 hücreli, morula ve geç blastosist

aşamalarındayken dizilemişlerdir. Gen ekspresyonundaki en büyük değişiklikleri 4-8 hücreli geçiş aşamasında olduğunu bildirmişlerdir. İşaretçi seçimi için bu veri setinin sadece embriyonik hücrelerden oluşan 90 hücrelik kısmı kullanılmıştır.

### 3.2.10 E-MTAB-2600 (Kolodziejczyk)

Kolodziejczyk ve arkadaşları üç farklı koşulda kültürlenen 704 adet fare embriyonik kök hücresinin (mESC) transkriptom analizi için scRNA-seq kullanmışlardır [98]. Analiz sonucunda alt popülasyon yapılarını karşılaştırmış ve gen ekspresyon seviyelerinde hücreden hücreye varyasyonları, serum/LIF, 2i/LIF, ve alternatif temel a2i/LIF olmak üzere 3 farklı durumda kültürlenen mES hücreleri üzerinde incelemişlerdir. Farklı kültür ortamlarında yetiştirilen mESC'lerin transkriptomik olarak farklı hücre popülasyonları oluşturduğunu, 2i ve a2i'de kültürlenen hücrelerin birbirine çok benzediğini bildirmişlerdir.

### 3.2.11 GSE57249 (Biase)

Biase ve arkadaşları eşleşen kardeş blastomerlerin scRNA-seq analizi ile 9 adet 1 hücreli zigot, 10 adet 2 hücreli, 5 adet 4 hücreli olmak üzere 49 fare blastomer hücresinin arasındaki farklılıkları bildirmişlerdir [99]. Eşdeğerlik hipotezi'nin 2 ve 4 hücreli embriyolardaki bireysel blastomerlerin homojen olduğunu öne sürmesinin aksine; 2 hücreli aşamada 138, 4 hücreli aşamada 205 genin bimodal ekspresyon paterni gösterdiğini ve embriyolar arası varyasyon çıkarıldıktan sonra embriyo içi bimodaliteye sahip 2 hücreli için 12 ve 4 hücreli için 13 adet gen bulmuşlardır. Analizler sonucunda 2 ve 4 hücreli embriyolarında blastomerler arasındaki farklılıkları ortaya çıkarmış ve bu farklılıkların genellikle embriyolar arası farklılıklardan daha büyük olduğunu göstermişlerdir.

### 3.2.12 GSE119945 MOCA (Cao)

Cao ve arkadaşları fare organogenezinin transkripsiyonel dinamiklerini araştırmak amacıyla, tek bir deneyde gebeliğin 9,5 ila 13,5 günleri arasında evrelenen 61 embriyodan türetilmiş yaklaşık 2 milyon hücrenin transkiptomlarını, scRNA-seq kullanarak ortaya çıkarmışlardır. 38 adet ana kümeden bahsetmiş ve spesifik işaretçi genleri belirlemişlerdir [84]. Örneğin epitelyal hücreler için *Epcam* ve *Trp36,* hepatositler için *Afp* ve *Alb,* melanositler için *Tyr* ve *Trpm1*, lens hücreleri için *Cryba2*, duyusal nöronlar için Mpz geni belirtilen hücreler için spesifik ekspresyona sahip genler olarak belirlenmiştir. Ana hücre tipleri arasında genlerin %68'inin yani 17.789 genin diferansiyel ekspresyona sahip olduğunu ve bu genler arasından hücre başı ortalama 75 gen olmak üzere toplamda 2863, hücre tipine özgü işaretçi gen

tanımladıklarını ve bu tanımlanan genlerin çoğunun daha önceden bu hücre tipleri için belirteç olarak tanımlanmadığını bildirmişlerdir. Örneğin notokord hücrelerinde *Ntn1* ve *Shh* için en yüksek ifadeyi tespit etmiş ve daha önce notokord hücreleri için tanımlanmayan *Tox2, Stxbp6, Schip1* ve *Frmd4b* genlerini bu hücreler için işaretçi olarak tanımlamışlardır. En büyüğü 144.648 hücre ve en küçüğü 1000 hücre içeren 38 ana hücre tipi için ortalama hücre sayısı 1869 olan 655 adet alt küme belirlemiş bazı alt kümeleri eledikten sonra kalan 571 tanesi için ekspresyon farklılığına bağlı alt küme işaretçilerini belirlemişlerdir. Son olarak bu atlası diğer hücre atlaslardan ayırt etmek için "Fare Organogenez Hücre Atlası" (Mouse Organogenesis Cell Atlas MOCA) olarak adlandırmışlardır.

### 3.2.13 GSE60361 (Zeisel)

Zeisel ve arkadaşları fare somatosensoriyel korteks ve hipokampal CA1 bölgesindeki 3005 hücreyi sınıflandırmak için scRNA-seq kullanmışlardır [100]. S1 piramidal, CA1 piramidal, internöron, oligodendrosit, astrosit, mikroglia, endotel, mural ve ependimal olmak üzere 9 ana sınıfın her birinde ikili kümeleme (biclustering) ile 47 adet moleküler olarak farklı hücre alt sınıf tanımlamışlardır. Ana hücre tiplerinden S1 piramidal hücreler için *Tbr1,* oligodendrositler için *Hapln2*, mural hücreler için *Acta2,* endotel hücreler için *Lyc61* genlerini ve hiç keşfedilmemiş olarak; S1 piramidal hücreler için *Gm11549*, internöronlar için *Pnoc* ve hipokampal piramidal hücreler için *Spink8* geni ile hücre tiplerini işaretlemişlerdir. Tanımlanan 47 alt hücre sınıfı içinde birçok işaretçi tanımlamış ve beyin hücre tiplerinin transkripsiyonel çeşitliliğini ortaya çıkarmışlardır.

### 3.2.14 Kleshchevnikov

Kleshchevnikov ve arkadaşları Cell2location ismiyle sundukları algoritmalarını değerlendirmek için çeşitli nöronal ve glial hücre tiplerini içeren bir doku olan fare beynini snRNA-seq kullanarak incelemişlerdir [101]. 40532 hücrenin snRNA-seq verilerinin referans hücre tipi imzalarını tanımlamak amacıyla öncelikle Louvain kümeleme ve ardından geleneksel scRNA-seq iş akışını uygulamış ve 59 hücre kümesi tanımlamışlardır. snRNA-seq analizinde kullanılan veri seti tüm profilli beyin alanlarından veri içermekte olduğu için yüksek kaliteli bir hücre tipi referansı sağlamaktadır.

### 3.2.15 Bhaduri (10X, 1.3M)

Bhaduri ve arkadaşları E18.5 fare beynindeki transkripsiyonel heterojeniteyi incelemek için 1.3 milyon hücrenin transkriptomlarını scRNA-seq ile analiz etmişlerdir [102]. Tek hücreli

RNA sekanslamanın uygulanabilirliğinin artması nedeniyle birçok bilim insanı insan vücudunun kapsamlı bir hücre atlasını oluşturmaya çalışmaktadır, ancak insan vücudu mevcut teknolojilerin analiz edebileceğinden daha fazla hücre içermektedir. Bhaduri ve arkadaşlarının bu çalışmadaki asıl amacı etkili bir örnekleme stratejisi geliştirerek atlas çalışmalarının maliyetini düşürmek ve daha verimli hale getirmektir. Bu amaçla fare beyninden alınan 1.3 milyon hücre ile popülasyon yapısının altında yatan heterojeniteyi ortaya çıkarmak için gerekli olan örneklem arasındaki ilişkiyi keşfetmeye çalışmışlardır. Sonuç olarak ilk araştırmalar için daha az sayıda hücrenin yeterli olabileceğini bildirmişlerdir.

### 3.2.16 Sekanslama bazlı uzamsal transkriptomik veri setleri

Önerilen yöntem in situ transkriptomik metotlar için tasarlanmış olsa da sekanslama bazlı uzamsal transkriptomik veri setlerinde de çalışabilmektedir. Zaten sekanslama bazlı uzamsal transkriptomik yöntemleri scRNA-seq verisine ek olarak konumları içermektedir. Dolayısıyla scRNA-seq verisinden konumlar haricinde yapısal olarak pek bir farkı yoktur. Ek olarak bu veri setleri için oluşturulan görseller, yöntemin ve sonuçların daha iyi biçimde anlaşılmasına olanak sağlamaktadır. Bu sebepten dolayı birçok farklı doku kesitini içeren 2 farklı sekanslama bazlı uzamsal transkriptomik veri seti değerlendirmelere dahil edilmiştir.

Jaffe ve arkadaşları, insan beyninin dorsolateral prefrontal korteksinde 10X visium platformu kullanarak, birçok nöropsikiyatrik bozuklukla ilişkilendirilen bölgenin laminer topografisini tanımlamaya çalışmışlardır. 3 farklı nörotipik yetişkin donörden ölüm sonrasında DLPFC (dorsolateral prefrontal cortex) doku bölümlerinin uzamsal gen ekspresyon profillerini incelemişlerdir. Literatürde daha önceden farklı teknolojiler kullanarak çeşitli çalışmalarda belirlenen, laminar-spesifik işaretçi genlerin doğruluğunu değerlendirmiş ve doğrulamışlardır [103]. Dahil edilen diğer veri seti ise 10X Genomics web sitesinden temin edilmiştir. Bu veri setinde ise FFPE fare beyin dokusundan elde edilen 5 $\mu m$'lik doku kesitleri Visium gen ekspresyonu slaytları vasıtasıyla sekanslanmıştır [104].

## 3.3 RAM Kullanım Miktarları ve Hesaplama Süreleri

RAM kullanım miktarı ve hesaplama sürelerinin değerlendirilebilmesi için tüm algoritmalar varsayılan parametreleri ile 15 adet veri seti üzerinde, her hücre tipi için 3 işaretçi gen seçecek şekilde çalıştırılmıştır. SMaSH ve scGeneFit algoritmaları seyrek matrislerle çalışmaya yazılımsal olarak uygun değildir. Bu nedenle seyrek matrisler bu iki algoritma

için yoğun (dense) matrise çevrilmiştir. Ancak 1.3 milyon ve 2 milyon hücre içeren veri setlerini çevirmek yüksek miktarda RAM (+ 400 Gb) ihtiyacı nedeniyle mümkün olmamıştır.

**Çizelge 3.2:** Tüm algoritmaların hesaplama süreleri

| **Calculation Time (sec)** | | | | |
|---|---|---|---|---|
| **Method** / **Data-Set** | scMAGS | SMaSH | scGeneFit | COSG |
| **Biase** | 4.035 | - | 1192.183 | 0.202 |
| **Yan** | 4.110 | - | 773.103 | 0.250 |
| **Zeisel** | 6.132 | 76.998 | 745.508 | 0.762 |
| **Li** | 5.143 | - | - | 0.513 |
| **Tasic** | 6.368 | - | - | 0.610 |
| **Xin** | 5.784 | 65.968 | - | 0.815 |
| **Darmanis** | 4.360 | - | 9056.897 | 0.228 |
| **Baron** | 5.411 | - | 1018.471 | 0.527 |
| **Treutlein** | 3.986 | - | 2535.693 | 0.227 |
| **Kolodziejczyk** | 4.360 | 27.013 | 3332.770 | 0.385 |
| **Goolam** | 4.147 | - | 7325.235 | 0.368 |
| **Pollen** | 4.410 | - | - | 0.446 |
| **Kleshchevnikov** | 159.897 | 912.117 | 4047.900 | 4.055 |
| **Bhaduri** | 287.940 | - | - | 278.964 |
| **Cao** | 551.196 | - | - | 58.355 |

Çizelge 3.2'de kıyaslanan algoritmaların çalışma süreleri, saniye cinsinden görülmektedir. SMaSH test-train ayrımı yapması ve bunun sonucu olarak küçük kümeler içeren veri setlerinde yetersiz test kümesi boyutları sebebiyle hata oluşturmuş ve birçok veri setinde çalıştırılamamıştır. scGeneFit ise yüksek sayıda gen içeren veri setlerinde, hücre sayısından bağımsız olarak yüksek RAM miktarlarına ihtiyaç duymuş ve çalıştırılamamıştır. Çalıştığı veri setlerinde ise en yavaş hesaplamayı gerçekleştirmiştir. Sonuçlardan görüleceği üzere en hızlı hesaplamaları COSG ve scMAGS gerçekleştirmiştir. scMAGS, COSG'ye göre daha

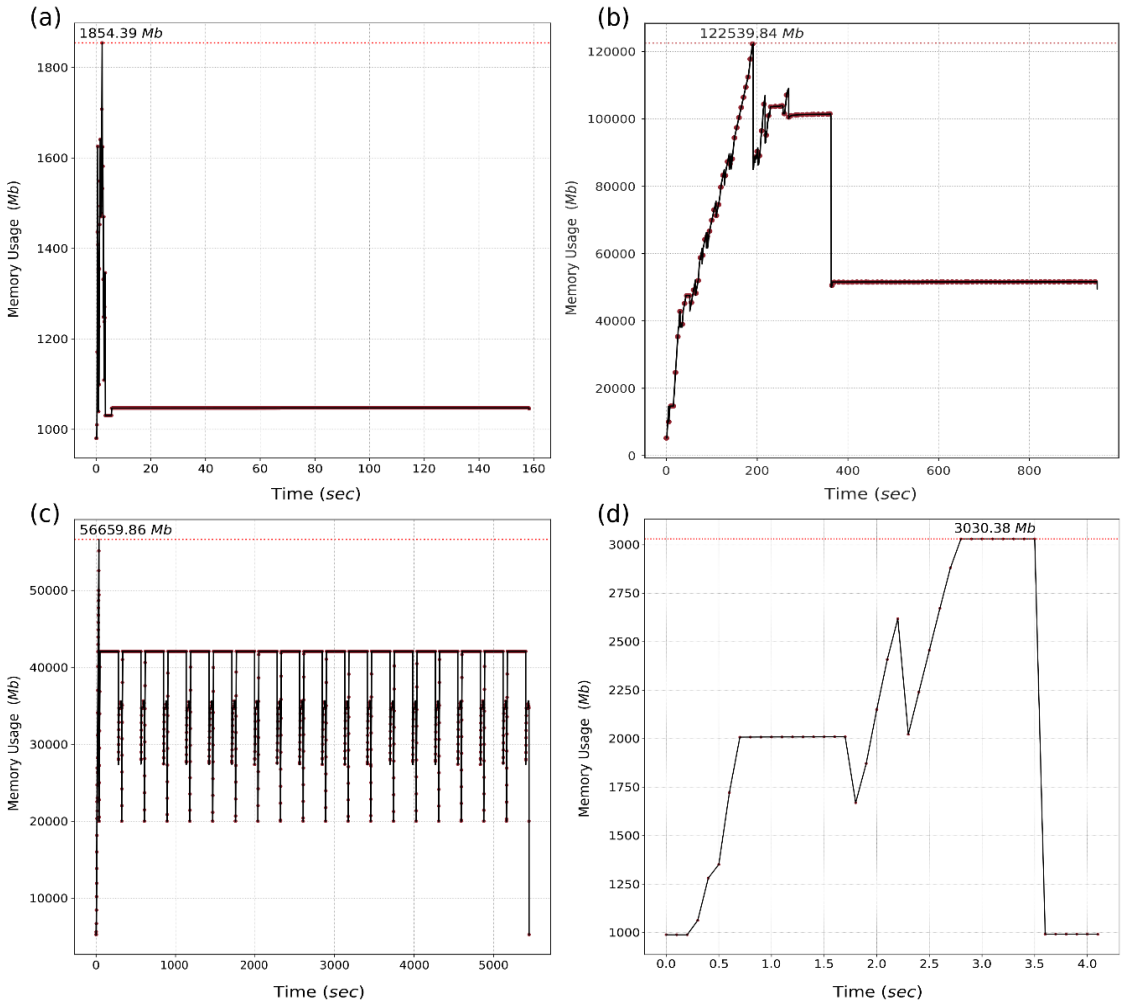
yavaş hesaplamalar gerçekleştirse de bağımsız olarak değerlendirildiğinde gayet hızlı bir biçimde hesaplamaları bitirmiştir. Örneğin 1.3M hücre içeren Bhaduri veri setinde COSG ile neredeyse aynı sürede hesaplamaları bitirmiş 2M hücre içeren veri setinde ise 9.5 dakikada hesaplamaları tamamlamıştır.

**Çizelge 3.3** Tüm algoritmaların RAM kullanım miktarları (Mb)

| **Peak Memory (RAM) Usage (Mb)** | | | | |
|---|---|---|---|---|
| **Method** / **Data-Set** | scMAGS | SMaSH | scGeneFit | COSG |
| **Biase** | 201 | - | 35342 | 315 |
| **Yan** | 203 | - | 21940 | 326 |
| **Zeisel** | 1267 | 6930 | 22299 | 2407 |
| **Li** | 704 | - | - | 1331 |
| **Tasic** | 885 | - | - | 1704 |
| **Xin** | 1431 | 7006 | - | 2508 |
| **Darmanis** | 353 | - | 26237 | 602 |
| **Baron** | 794 | - | 18945 | 1416 |
| **Treutlein** | 216 | - | 29099 | 329 |
| **Kolodziejczyk** | 739 | 3402 | 80312 | 1155 |
| **Goolam** | 271 | - | 92316 | 457 |
| **Pollen** | 294 | - | - | 557 |
| **Kleshchevnikov** | 1854 | 122539 | 56460 | 3030 |
| **Bhaduri** | 46949 | - | - | 112249 |
| **Cao** | 17282 | - | - | 50870 |

Çizelge 3.3'te algoritmaların RAM kullanım miktarları megabayt cinsinden görülmektedir. RAM kullanımını ölçmek için gerçekleştirilen hesaplamalarda, algoritmalar yine varsayılan parametrelerle çalıştırılmıştır. RAM kullanımını raporlamak için bir Python paketi olan "*memory_reporter*" tercih edilmiş ve örnekleme süresi 0.1 saniye olarak belirlenmiştir. scMAGS diğer algoritmalara kıyasla RAM kullanımında en başarılı sonuçları göstermiştir. Bazı durumlarda hesaplama süresi ve RAM kullanımı arasında bir ödünleşim (trade-off) bulunur ve bu durumlarda scMAGS daha düşük RAM kullanacak şekilde hesaplamalarını

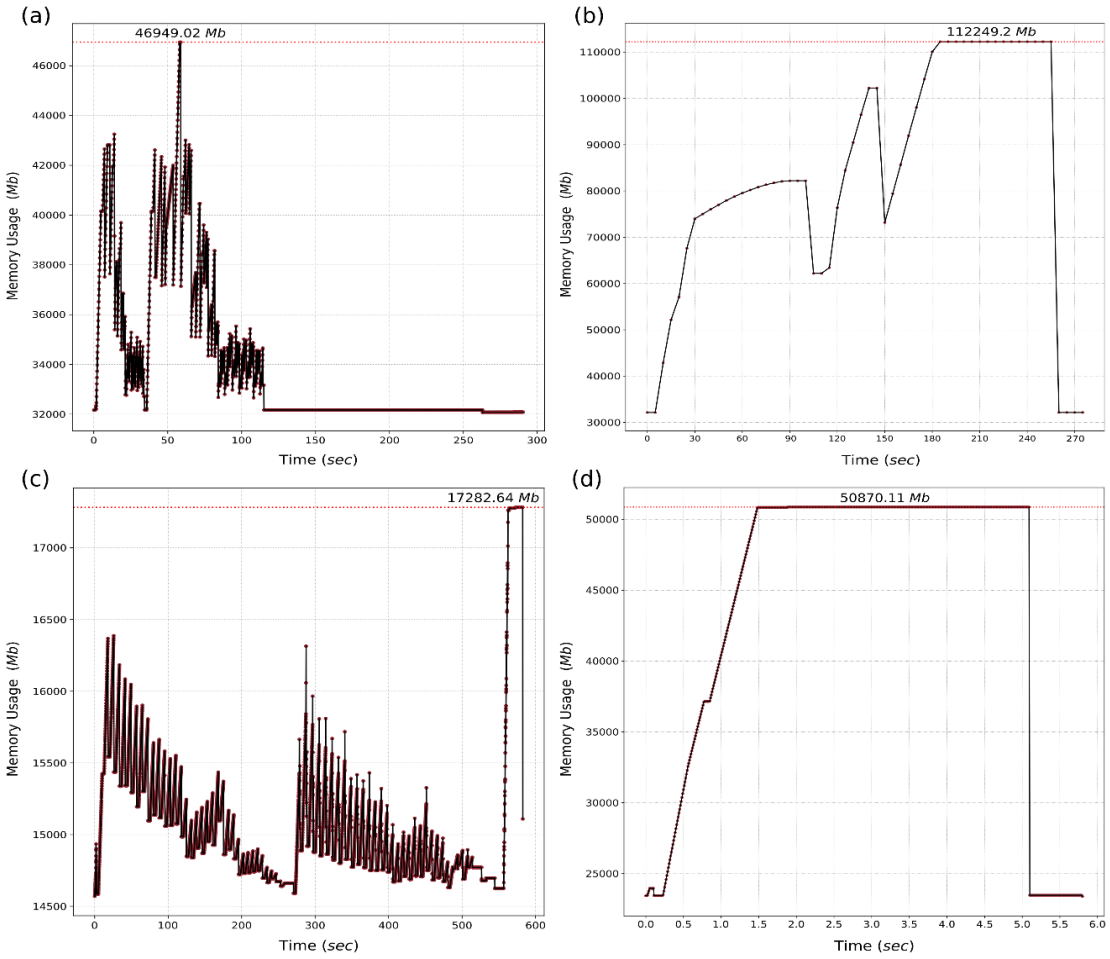
gerçekleştirmektedir, bu nedenle hesaplama sürelerinde COSG'ye göre daha yavaş çalışsa da daha düşük miktarda RAM kullanmaktadır çünkü bu şekilde programlanmıştır. RAM kullanımında scGeneFit en verimsiz sonuçları göstermiş, SMaSH ise scGeneFit'ten sonra en kötü sonuçları elde etmiştir. COSG tüm veri setlerinde scMAGS'tan daha yüksek miktarda RAM kullanmış, büyük ve seyrekliği (sıfır oranları) düşük olan veri setlerinde ise çok yüksek miktarda RAM kullanmıştır.




**Şekil 3.1:** Kleshchevnikov veri seti için algoritmaların RAM kullanımı raporları. (a) scMAGS, (b) SMaSH, (c) scGeneFit, (d) COSG

Şekil 3.1'de diğer algoritmaların tamamının çalışabildiği en büyük veri seti olan Kleshchevnikov veri seti için RAM kullanımı raporları görülmektedir. scGeneFit ve SMaSH seyrek matrislerle çalışamadığından sayım matrisleri yoğun matrise çevrilmiştir. Bu nedenle RAM kullanım eğrileri daha yüksek noktalardan başlamaktadır. SMaSH bu veri setinde yaklaşık 123 Gb, scGeneFit ise yaklaşık 57 Gb RAM kullanmıştır. Bu durum iki

algoritmanın da RAM kullanımı açısından verimsiz olduğunu göstermektedir. scMAGS en düşük miktarda, COSG ise scMAGS'ın neredeyse iki katı RAM kullanarak hesaplamalarını tamamlamıştır.

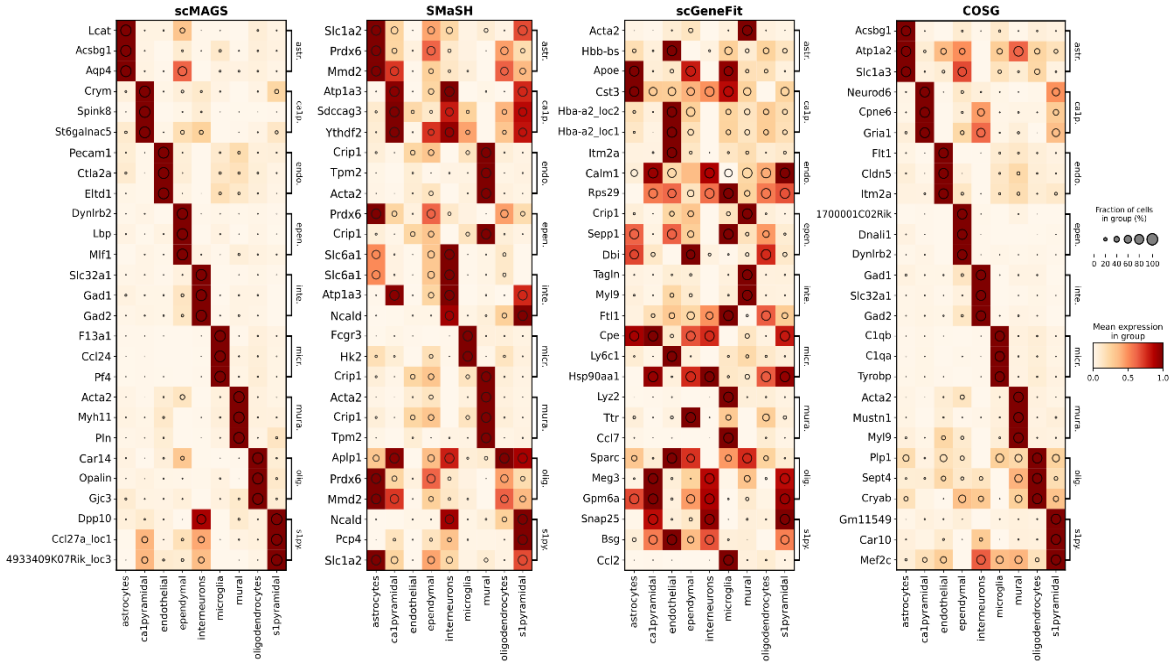



**Şekil 3.2:** Bhaduri (1.3M) ve Cao (2M) veri setinde scMAGS ve COSG'nin RAM kullanımı raporları. (a) scMAGS (10X), (b) COSG (10X), (c) scMAGS (Cao), (d) COSG (Cao)

Şekil 3.2'de scMAGS 1.3M hücre içeren 10X veri setinde yaklaşık 47 Gb, 2M hücre içeren Cao veri setinde ise yaklaşık 18 Gb tepe RAM kullanımı değerlerine ulaşmıştır. Buna karşın COSG 10X veri setinde 113 Gb Cao veri setinde ise yaklaşık 51 Gb RAM kullanmıştır. Bu durum scMAGS'ın çok büyük veri setlerinde bile düşük miktarlarda RAM kullandığını, COSG'nin ise büyük veri setlerinde RAM kullanımı açısından scMAGS'a göre verimsiz olduğunu göstermektedir.

### 3.4 Seçilen İşaretçi Genlerin Ekspresyon Profilleri

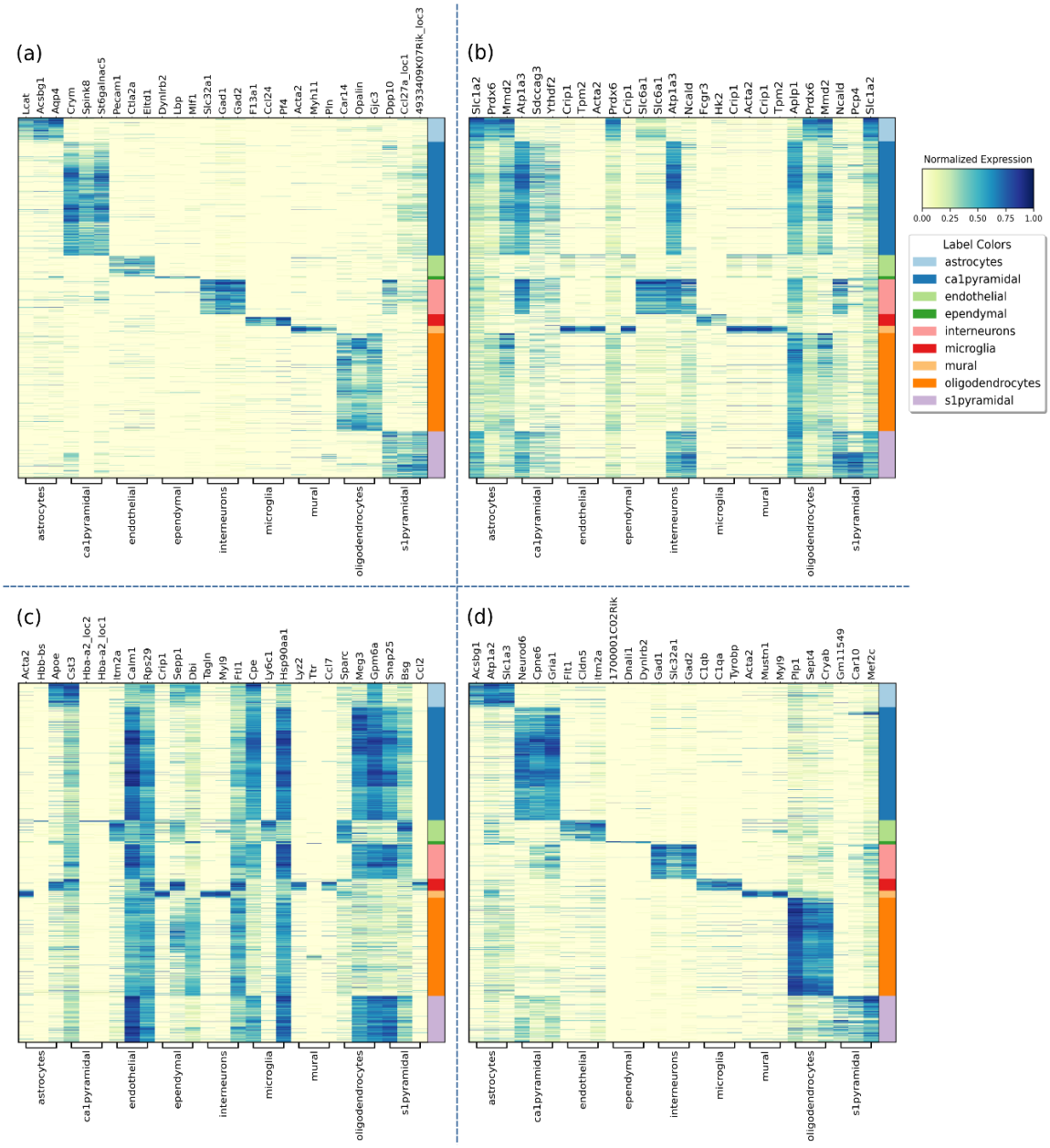
Seçilen genlerin Bölüm 2.2'de belirlenen hücre tipine özgü işaretçiler de aranan kriterlere sahip olup olmadığının değerlendirilebilmesi ve ekspresyon profillerinin incelenebilmesi için Dotplot, Heatmap, t-SNE figürleri oluşturuldu ve işaretçilerin ekspresyon profilleri değerlendirilmiştir. Ek olarak seçilen genlerin sınıflama performansının değerlendirilebilmesi için k-NN ile sınıflandırma yapılmış ve sonuçlar karmaşıklık matrisi ile görselleştirilmiştir.




**Şekil 3.3:** Tüm yöntemlerin Zeisel veri seti için seçtiği işaretçi genlerin Dotplot grafikleri

Şekil 3.3' deki Dotplot, Zeisel veri seti için tüm yöntemlerin seçtiği işaretçilerin ekspresyon profillerini özetlemektedir. scMAGS'ın seçtiği işaretçilerin seçildikleri hücre tiplerinde yüksek ekspresyona ve ekspresyon ortalamasına sahip oldukları ve Bölüm 2.2'de belirtilen kriterleri sağladıkları görülmektedir. SMaSH'in seçtiği işaretçilerin büyük bir kısmı belirtilen kriterleri sağlamamakta ve buna ek olarak bazı genler (*Slc1a2, Prdx6, Crip1, Mmd2, Atp1a3, Acta2, Slc6a1*) birden çok hücre tipi için işaretçi olarak seçilmiştir. SMaSH'in seçtiği genler diferansiyel eksprese edilmiş ve bu sebeple sınıflandırma algoritmasının performansını artırmış olabilir. Ancak bu durum uzamsal transkriptomik deneyleri için doğru işaretçiler olacağı anlamına gelmemektedir. scGeneFit'in seçtiği genlerde yine SMaSH'teki gibi diferansiyel ekspresyona sahip olabilir ve oluşturdukları uzayda hücre tiplerini ayırt ediyor olabilir ancak uzamsal transkriptomik açısından anlamsızdır. Örneğin astrositler için seçilen *Acta2* veya *Hbb-bs* seçildikleri hücreler

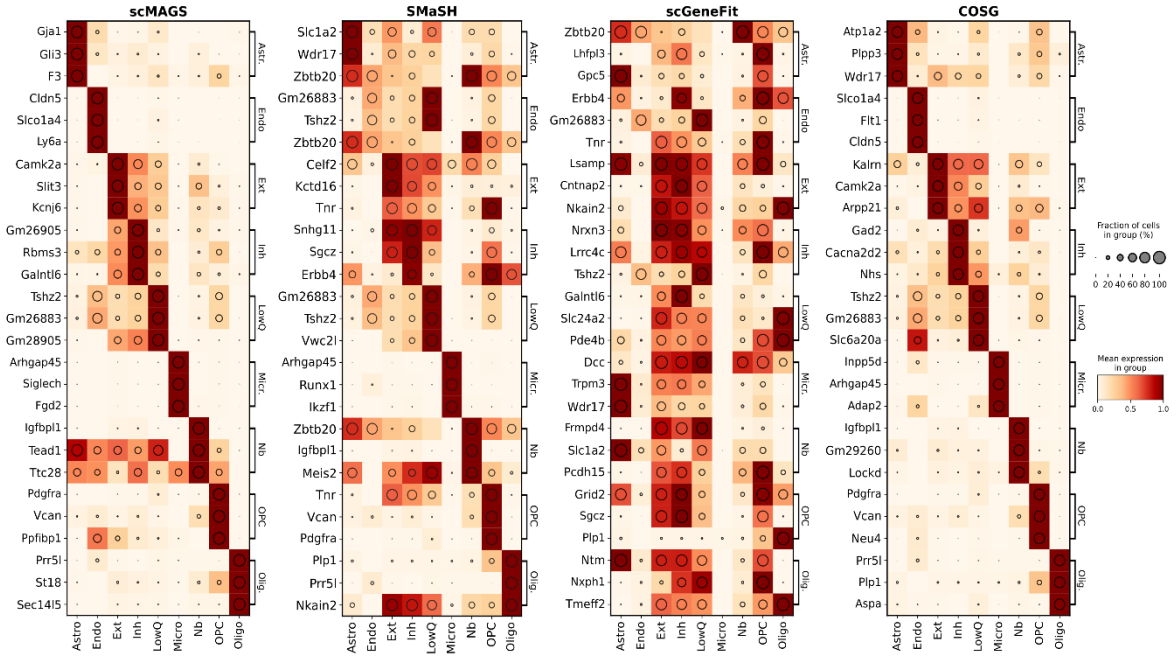
içerisinde yüksek ekspresyona sahip olmaları gerekirken *Acta2* mural hücrelerinde *Hbb-bs* endotelyal hücrelerde yüksek ekspresyona sahiptir. COSG ise scMAGS'a en yakın ve belirtilen kriterlerin birçoğuna uyan sonuçları vermektedir.




**Şekil 3.4:** Tüm yöntemlerin Zeisel veri seti için seçtiği işaretçi genlerin Heatmap grafikleri (a) scMAGS, (b) SMaSH, (c) scGeneFit, (d) COSG

Şekil 3.4 (a)'da görüldüğü üzere scMAGS'ın seçtiği işaretçiler sadece seçildikleri hücre tiplerinde koyu renklere yani yüksek ekspresyon değerlerine sahiptir. Seçilen işaretçilerin her birinin farklı renkleri içeren problara bağlanacağı düşünülürse, işaretçilerin hücre tiplerini birbirinden ayırt edebileceği açıkça görülmektedir. Şekil 3.4 (b)'de SMaSH'in

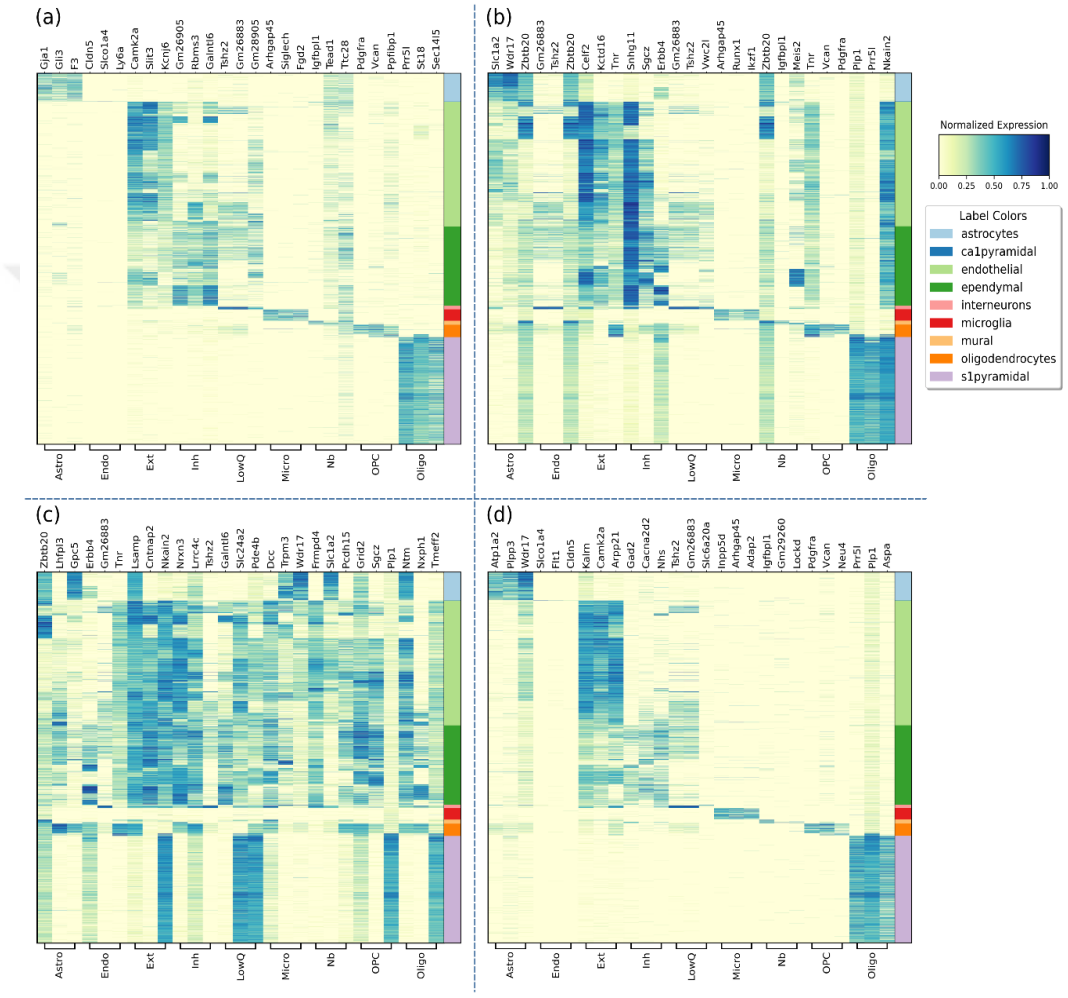
seçtiği genlerin ise hücre tiplerini birbirinden ayırt edemeyeceği açık biçimde görülmektedir. Örneğin oligodendrositler için seçilen *Aplp1*, oligodendrositler haricinde de birçok hücre tipinde eksprese edilmiştir veya mikroglia hücreleri için seçilen *Crip1*, sadece mural hücrelerinde eksprese edilmiş ve mikroglia hücrelerinde ise neredeyse hiç ekspresyona sahip değildir. Bu ve bunun gibi birçok işaretçi istenilen ekspresyon profilinin tam aksine sahiptir. Ancak SMaSH, *Aplp1* gibi genleri uzayda seçildikleri hücre tiplerini diğer hücre tiplerinin tamamından farklı bir konuma yerleştirdiği ve buna bağlı olarak sınıflama performansını iyileştirdiği için işaretçi olarak nitelendirmektedir. scGeneFit'in seçtiği genlerin de neredeyse tamamı ya hücre tiplerinin tamamında eksprese edilmiş ya da seçildiği hücre tipinde ekspresyona sahip değildir, dolayısıyla anlamsız genlerdir. COSG'nin seçtiği genler ise seçildikleri hücrelerde yüksek ekspresyona sahiptir ancak astrositler ve oligodendrositeler için seçilen işaretçiler diğer hücre tiplerinde de ekspresyona sahiplerdir.




**Şekil 3.5:** Tüm yöntemlerin Kleshchevnikov veri seti için seçtiği işaretçi genlerin Dotplot grafikleri

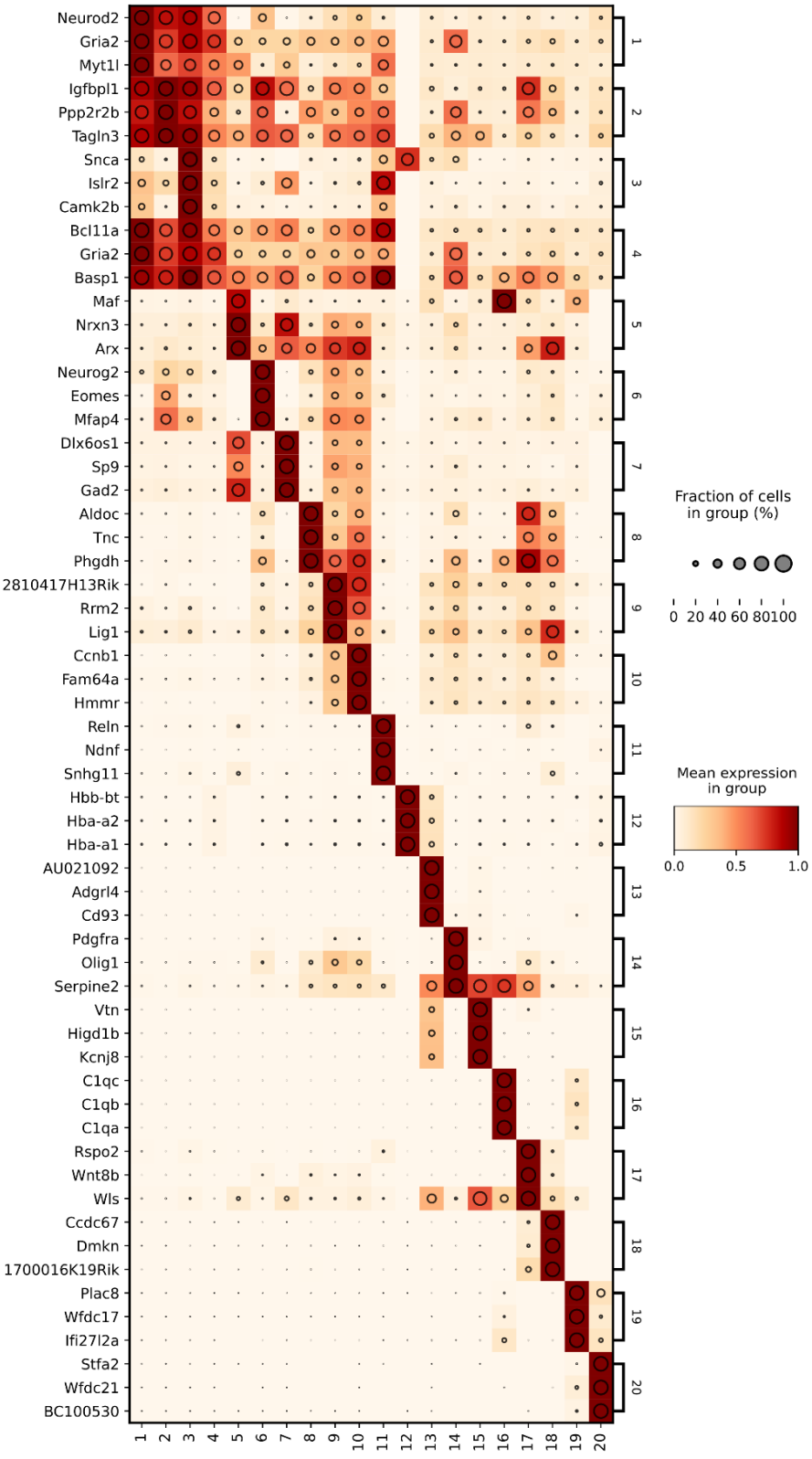
Şekil 3.5 diğer yöntemlerin hepsinin ortak olarak çalışabildiği en büyük veri seti olan Kleshchevnikov veri seti için seçilen işaretçilerin ekspresyon profillerini özetlemektedir. scGeneFit bu veri setinde de yine uzamsal transkriptomik açısından anlamsız genleri işaretçi olarak seçmiştir, seçtiği genler seçildikleri hücre tiplerinde istenilen ekspresyon karakterine sahip değildir. SMaSH'in sonuçlarında ise Zeisel veri setinde görülen problemler devam etmektedir. Örneğin astrositler için seçilen *Zbtb20*, endositler için seçilen *Gm26883* ve

*Tszh2*, Ext hücreleri için seçilen *Tnr* genleri sınıflama performansını iyileştiren ancak istenilen ekspresyon karakterine sahip olmayan genlerdir. Ayrıca *Zbtb20* astrositler ve endositler için, *Tnr* Ext ve OPC için, *Gm26883* Endosit ve LowQ için, *Tszh2* ise Endosit ve LowQ için işaretçi olarak seçilmiştir. Ancak bir genin birden çok hücre tipi için işaretçi olarak kullanılması hücre tiplerinin ayrılamamasına sebep olacaktır. scMAGS ve COSG'nin seçtiği işaretçiler ise şekilden de görüleceği üzere belirlenen kriterleri sağlamaktadır.




**Şekil 3.6:** Tüm yöntemlerin Kleshchevnikov veri seti için seçtiği işaretçi genlerin Heatmap grafikleri (a) scMAGS, (b) SMaSH, (c) scGeneFit, (d) COSG

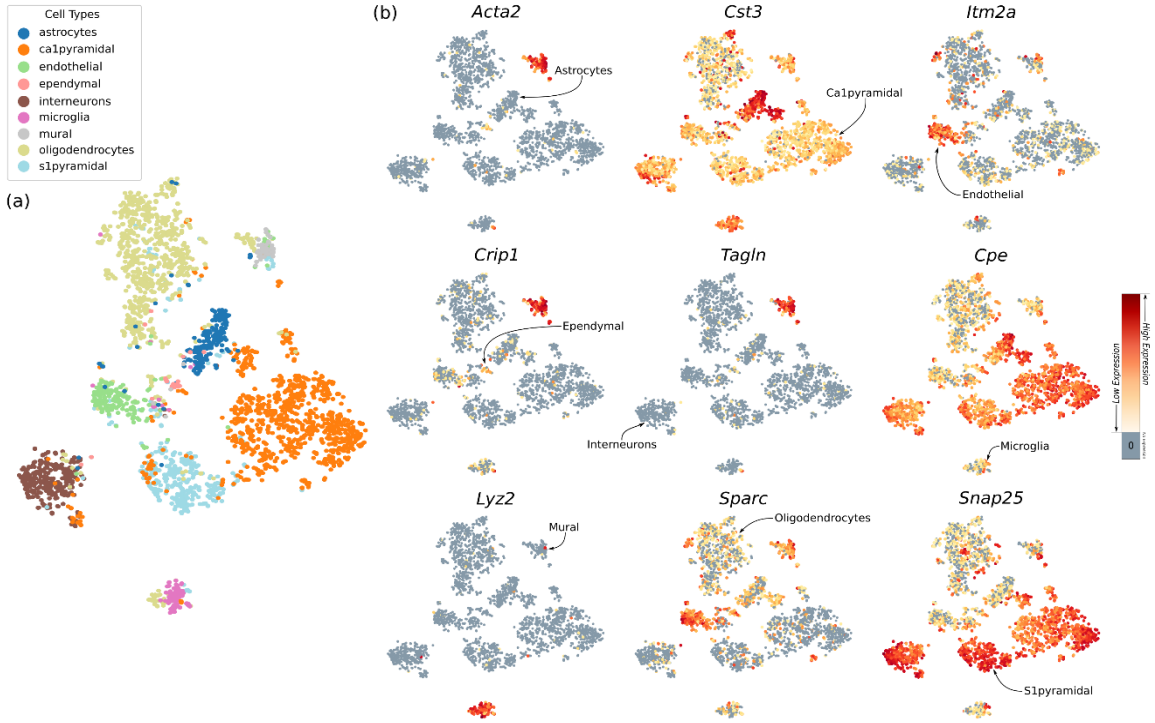
Şekil 3.6 Kleshchevnikov için seçilen işaretçileri daha ayrıntılı bir biçimde görselleştirmektedir. SMaSH'in ve özellikle scGeneFit'in seçtiği işaretçilerin hücreleri ayırt etmediği burada daha açık bir biçimde görülmektedir. scMAGS ve COSG ise birbirlerine yakın ve karşılaştırılabilir sonuçlar göstermişlerdir.




**Şekil 3.7:** 10X (1.3M) veri seti için scMAGS'ın seçtiği işaretçilerin Dotplot grafiği

Şekil 3.7 En büyük veri setlerinden olan 10X veri seti için scMAGS'ın seçtiği işaretçileri göstermektedir. Şekilden görüleceği üzere seçilen işaretçiler 4. küme haricinde seçildikleri

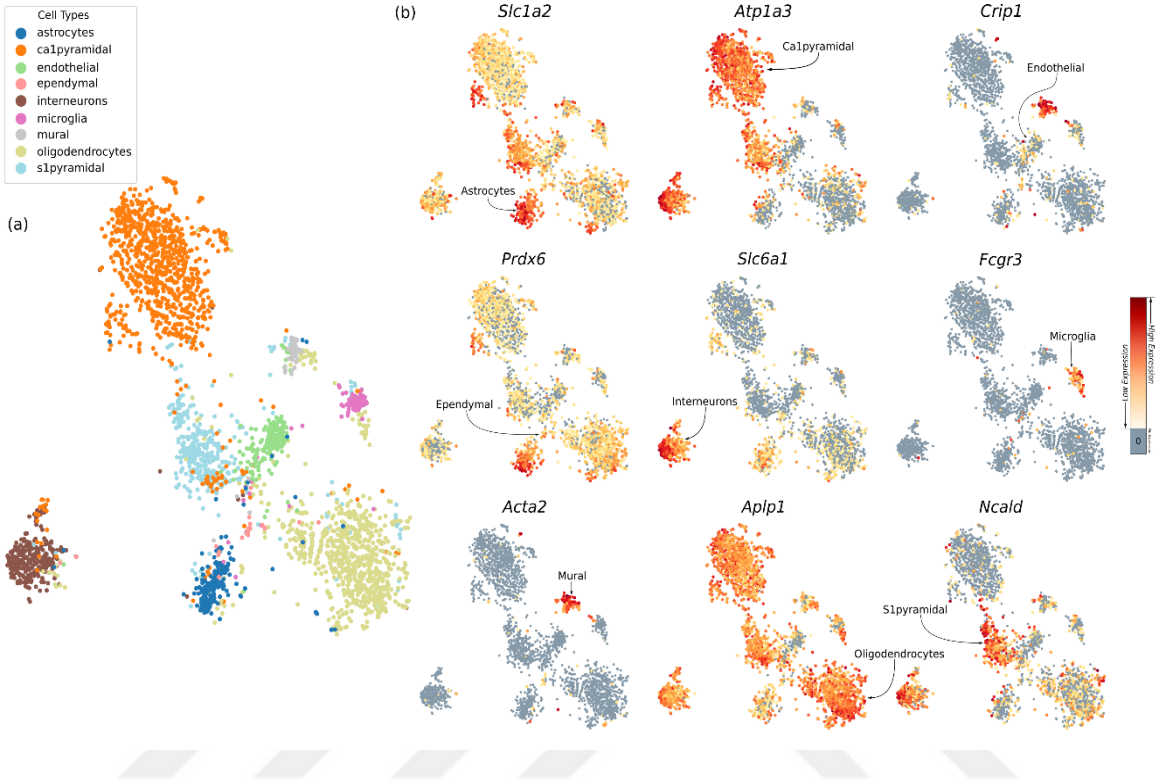
küme de en yüksek ekspresyon ortalamalarına ve eksprese edilme oranlarına sahiptirler. 4. küme için ise belirlenen kriterlere uygun bir gen bulunamadığından 4. küme haricinde en az eksprese edilen genler işaretçi olarak seçilmiştir. Bu sonuçlar scMAGS'ın büyük veri setlerinde de stabil olarak çalıştığını, yani ölçeklenebilir olduğunu göstermektedir. 10X veri setine ek olarak Çizelge 3.1'de bahsedilen diğer veri setleri için scMAGS ve COSG tarafından seçilen işaretçi genlerin ekspresyon profilleri Şekil A.2, Şekil A.3, Şekil A.4, Şekil A.5 ve Şekil A.6'da Dotplotlar vasıtasıyla karşılaştırılmalı biçimde görselleştirilmiştir. Şekillerden scMAGS'ın başarılı sonuçlar gösterdiği bariz bir şekilde görülmektedir.




**Şekil 3.8:** scGeneFit'in Zeisel veri seti için seçtiği işaretçilerin t-SNE grafikleri

Şekil 3.8, scGeneFit'in Zeisel veri seti için seçtiği işaretçilerin, t-SNE ile 2 boyutta nasıl gözüktüğünün anlaşılabilmesi için çizilmiştir. Şekil 3.8 (a)'da gömülü uzayın ilk 2 boyutu dağılım (scatter) grafiği ile görselleştirilmiş, daha sonra hücre tipleri için seçilen işaretçilerin hangi bölgelerde eksprese edildiğinin görülebilmesi için aynı boyutlar kullanılarak, Şekil 3.8 (b)'de scGeneFit'in seçtiği en iyi işaretçiler ekspresyon olmayan bölgeler için sadece gri, ekspresyon olan bölgeler için turuncu-kırmızı renk haritası (colormap) kullanılarak görselleştirilmiştir. Buradan yola çıkarak işaretçilerin her biri Şekil 3.8 (a)'da bulunan kümeleri ayırt eden renklerden, sadece seçildikleri hücre tipinin olduğu bölgede ekspresyona sahip olmalıdır. scGeneFit'in seçtiği işaretçiler ise seçildikleri hücre tiplerinde eksprese

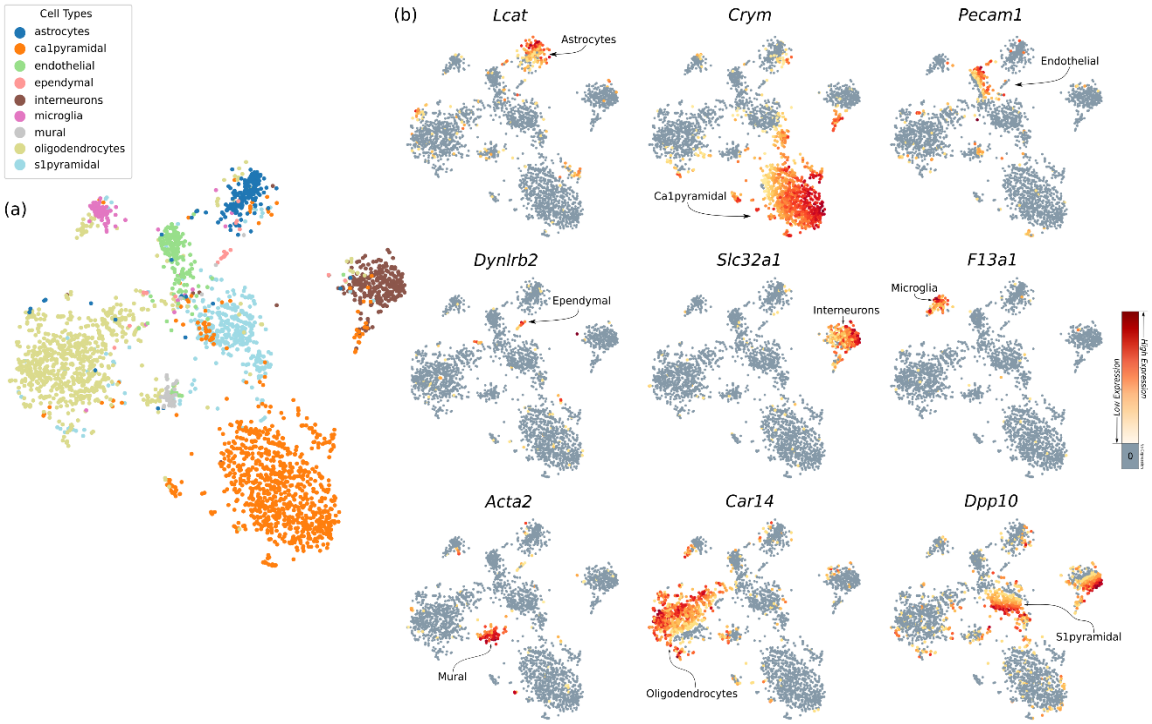
edilmekle beraber bazen tüm hücre tiplerinde (*Cpe, Cst3*) eksprese edilmiş, bazen de sadece seçildikleri hücre tipinden farklı, bir hücre tipinde eksprese edilmişlerdir (*Acta2, Tagln, Lyz2*).



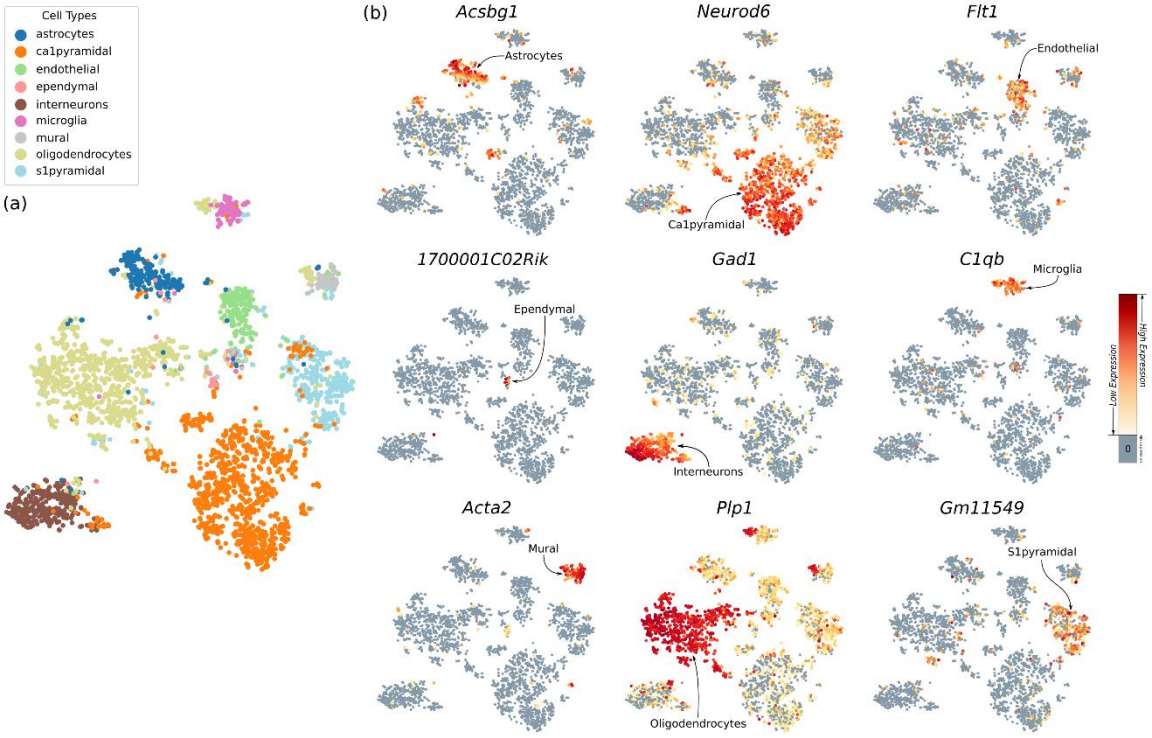

**Şekil 3.9:** SMaSH'ın Zeisel veri seti için seçtiği işaretçilerin t-SNE grafikleri

Şekil 3.9'da SMaSH'in seçtiği işaretçilerin t-SNE grafiğinde, hücre tipleri işaretçilerin sınıflama performansını artırmaya yönelik seçilmesi sebebiyle iyi biçimde ayrılmıştır. Ancak seçilen işaretçiler bu sebepten dolayı istenilen ekspresyon profiline sahip değildir ve aynı anda birçok hücre tipinde eksprese edilmiştir. Örneğin *Aplp1* neredeyse hücre tiplerinin tamamında yüksek ekspresyona sahiptir ve ayırt edici bir gen olmadığı açıkça görülmektedir. Bu durum *Prdx6* ve *Slc1a2* genleri içinde geçerlidir. Ek olarak Ncald geni S1piramidal hücrelerinde ve oligodendrositlerin birçok kısmında ekspresyona sahiptir.

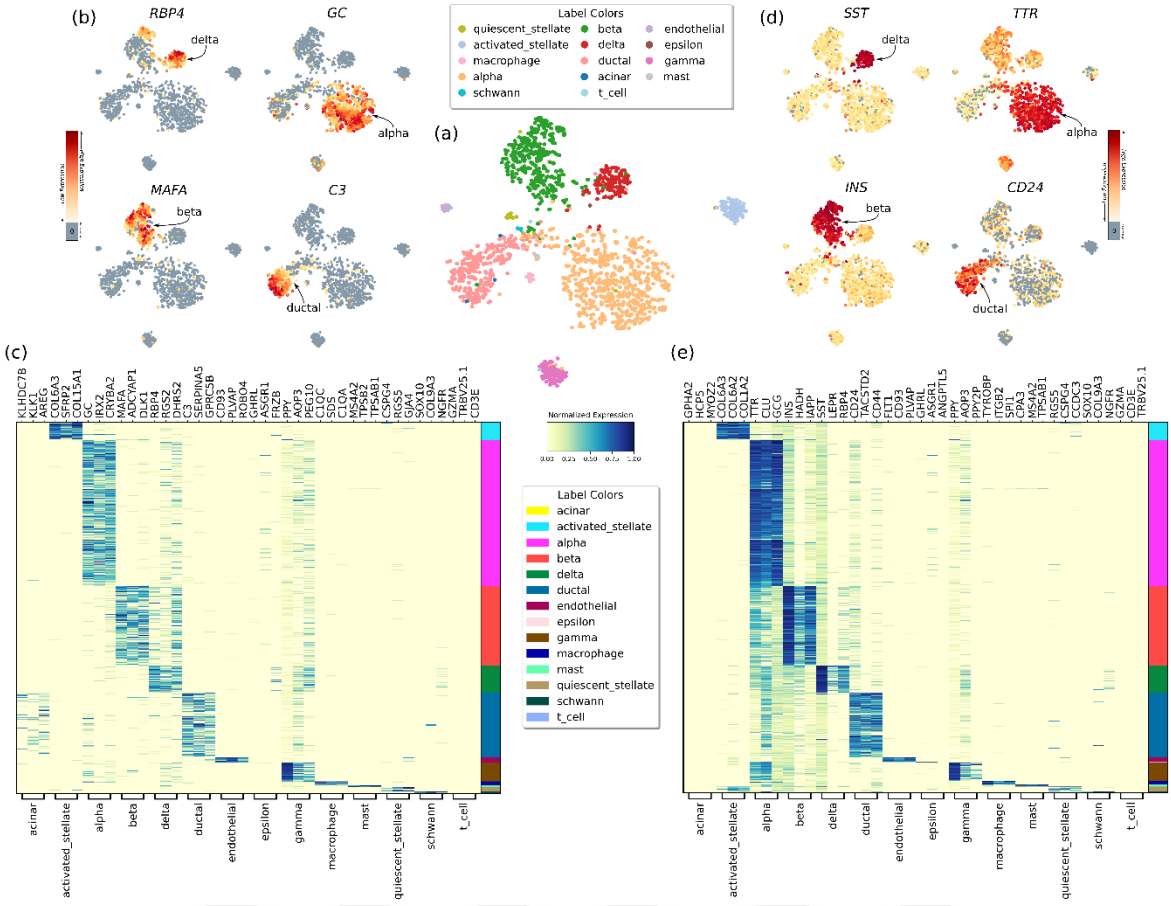
Şekil 3.10'da ise scMAGS'ın seçtiği işaretçiler, sadece seçildikleri hücre tipinde eksprese edilmiş ve diğer hücre tiplerinde eksprese edilmemişlerdir. Bu durum sonucunda seçilen işaretçilerin farklı hücre tiplerini birbirinden ayırt edebilecekleri açıkça görülmektedir. Şekil 3.11'de görülen COSG'nin seçtiği işaretçiler genel olarak sadece seçildikleri hücre tiplerinde ekspresyona sahiptir ancak *Plp1* geni diğer hücre tiplerinin neredeyse tamamında ekspresyona sahiptir.




**Şekil 3.10:** scMAGS'ın Zeisel veri seti için seçtiği işaretçilerin t-SNE grafikleri



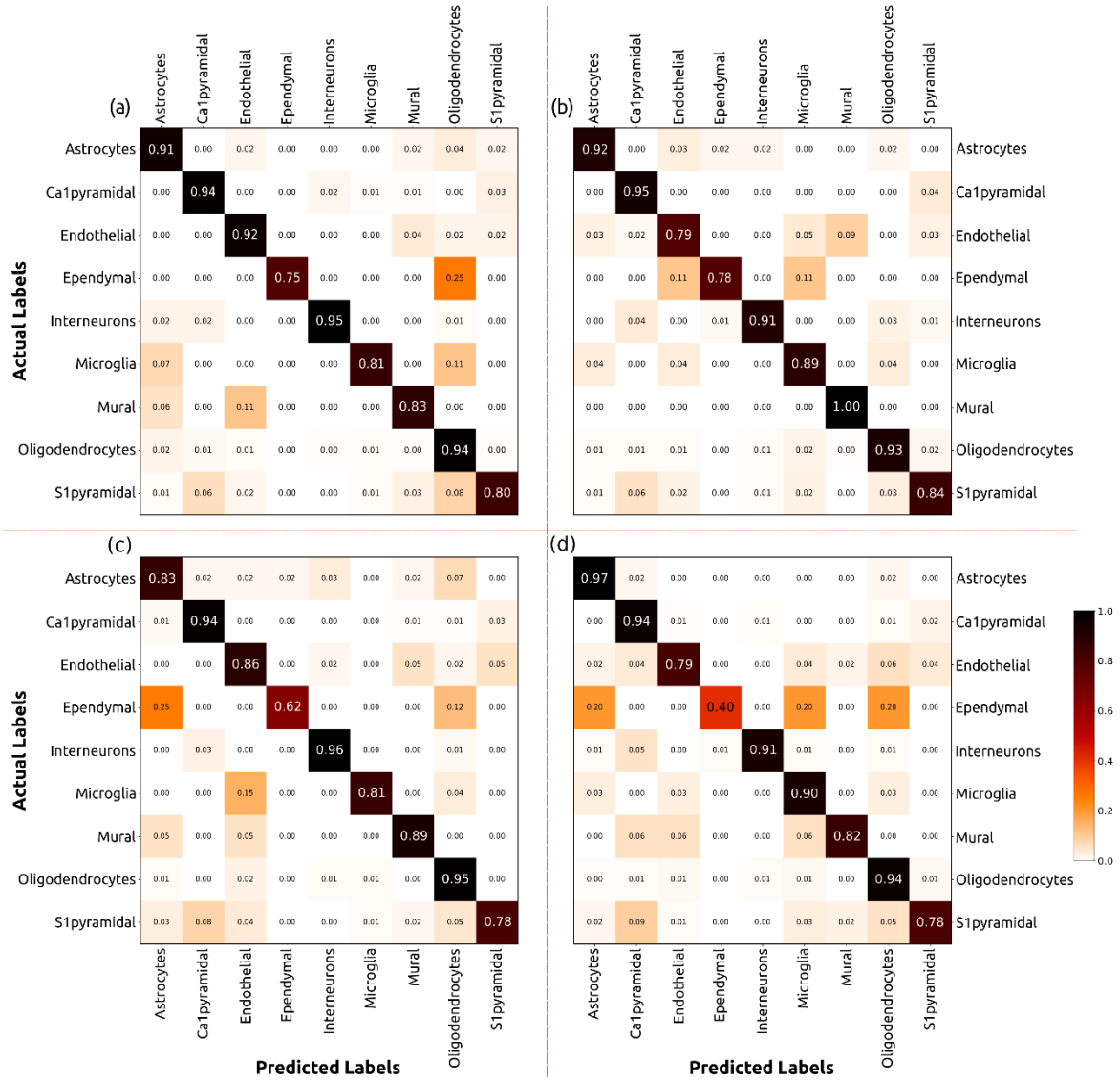

**Şekil 3.11:** COSG'nin Zeisel veri seti için seçtiği işaretçilerin t-SNE grafikleri




**Şekil 3.12:** Baron Human 2 veri seti için scMAGS ve COSG tarafından seçilen işaretçiler

scGeneFit ve SMaSH'in seçtiği işaretçilerin belirtilen kriterleri sağlamadıkları değerlendirmelerde ek şekillerde görülmektedir. Yapılan değerlendirmelere göre COSG scMAGS ile karşılaştırabilir sonuçlar göstermekle beraber bazı veri setlerinde seçtiği işaretçiler hedef hücre tipleri haricinde de eksprese edilmiştir. Bu durum Şekil 3.12 ve ek şekillerde görülmektedir. Şekil 3.12 (d)'de COSG'nin seçtiği işaretçiler nereyse hedef hücrelerin haricindeki hücrelerin tamamında eksprese edilmiştir. Örneğin alpha ve beta hücreleri için seçilen işaretçiler neredeyse diğer tüm hücre tiplerinin tamamında eksprese edilmiştir. Ancak bu durum scMAGS için geçerli değildir. scMAGS hedef hücre tipleri haricindeki hücrelerde ekspresyonu minimum olan genleri bulmayı amaçlar. Bu durum uzamsal transkrptomik deneylerinde hücre tiplerinin ayrılamamasına sebep olabilir.

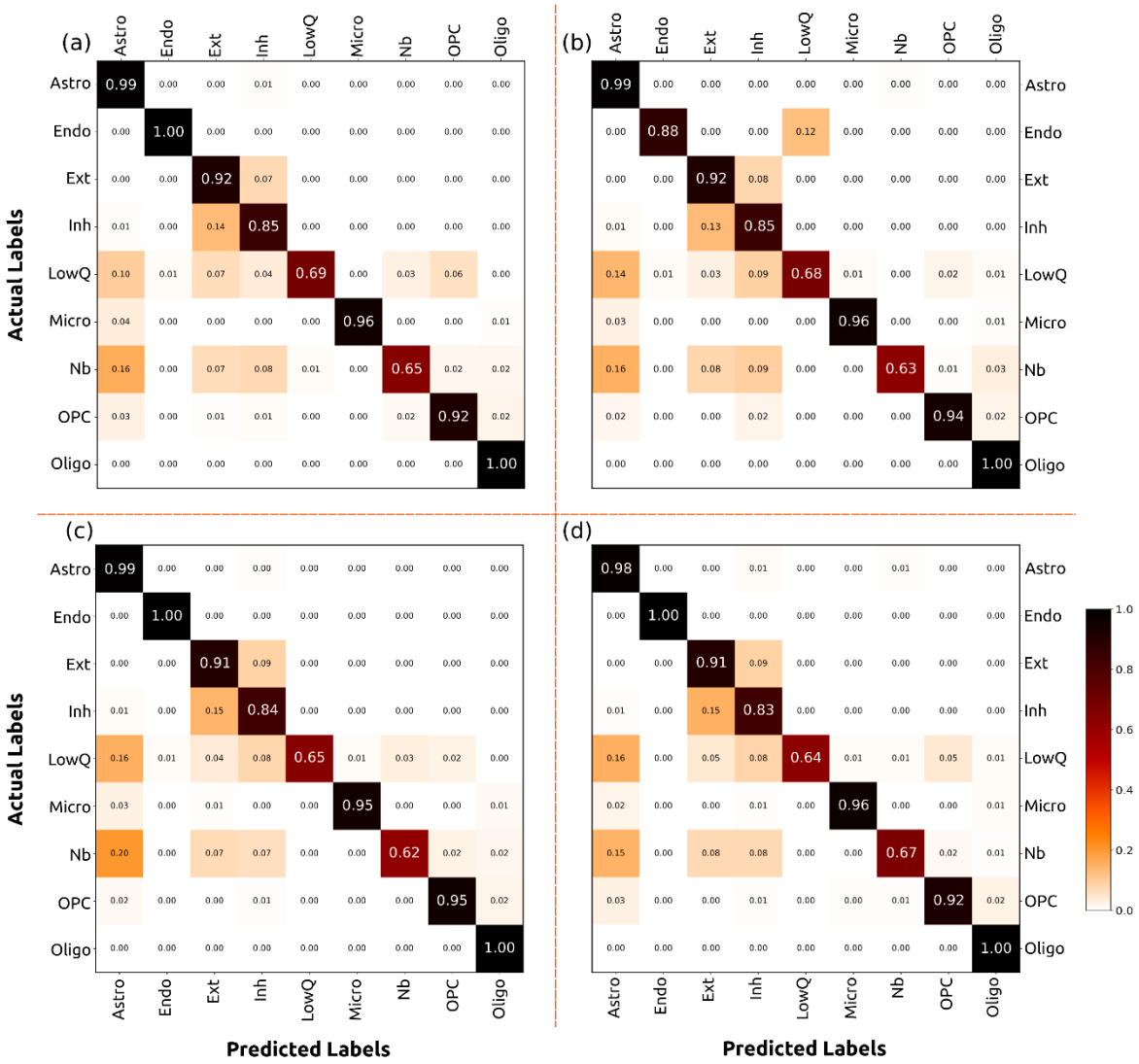
Bu duruma başka bir örnek olarak Şekil A.1 gösterilebilir. COSG bu veri setinde de bazı hücre tipleri için kendi hücre tipi haricinde ekspresyona sahip işaretçileri seçmiştir. Örneğin beta hücreleri için seçilen *Ins1* ve *Ins2* tüm hücre tiplerinde yüksek ekspresyona sahiptir.




**Şekil 3.13:** Zeisel veri seti için seçilen işaretçilerle gerçekleştirilen k-NN sınıflandırmasının Karmaşıklık Matrisleri. (a) scMAGS, (b) SMaSH, (c) scGeneFit, (d) COSG

Şekil 3.13 seçilen işaretçi genlerle gerçekleştirilen k-NN sınıflandırması sonuçlarının karmaşıklık matrislerini görselleştirmektedir. k-NN sınıflandırması için sadece seçilen işaretçiler kullanılarak veri setlerindeki hücrelerin %30'u test %70'i eğitim için kullanılmak üzere ayrılmıştır. Eğer seçilen işaretçiler doğruysa sınıflama sonuçlarının iyi çıkması beklenir. Ancak sınıflama sonuçlarının iyi olması seçilen işaretçilerin doğru olduğu anlamına da gelmez. Burada amaç tüm algoritmaların sınıflandırma sonuçlarını karşılaştırıp aslında doğru genler seçildiğinde de sınıflama performansının yüksek olacağını göstermektir. Bu kısımdan önce yapılan değerlendirmelerde SMaSH'in seçtiği işaretçilerin kriterlere uymadığı gösterilmiştir. Ancak bu durum Bölüm 2.3'te açıklandığı gibi

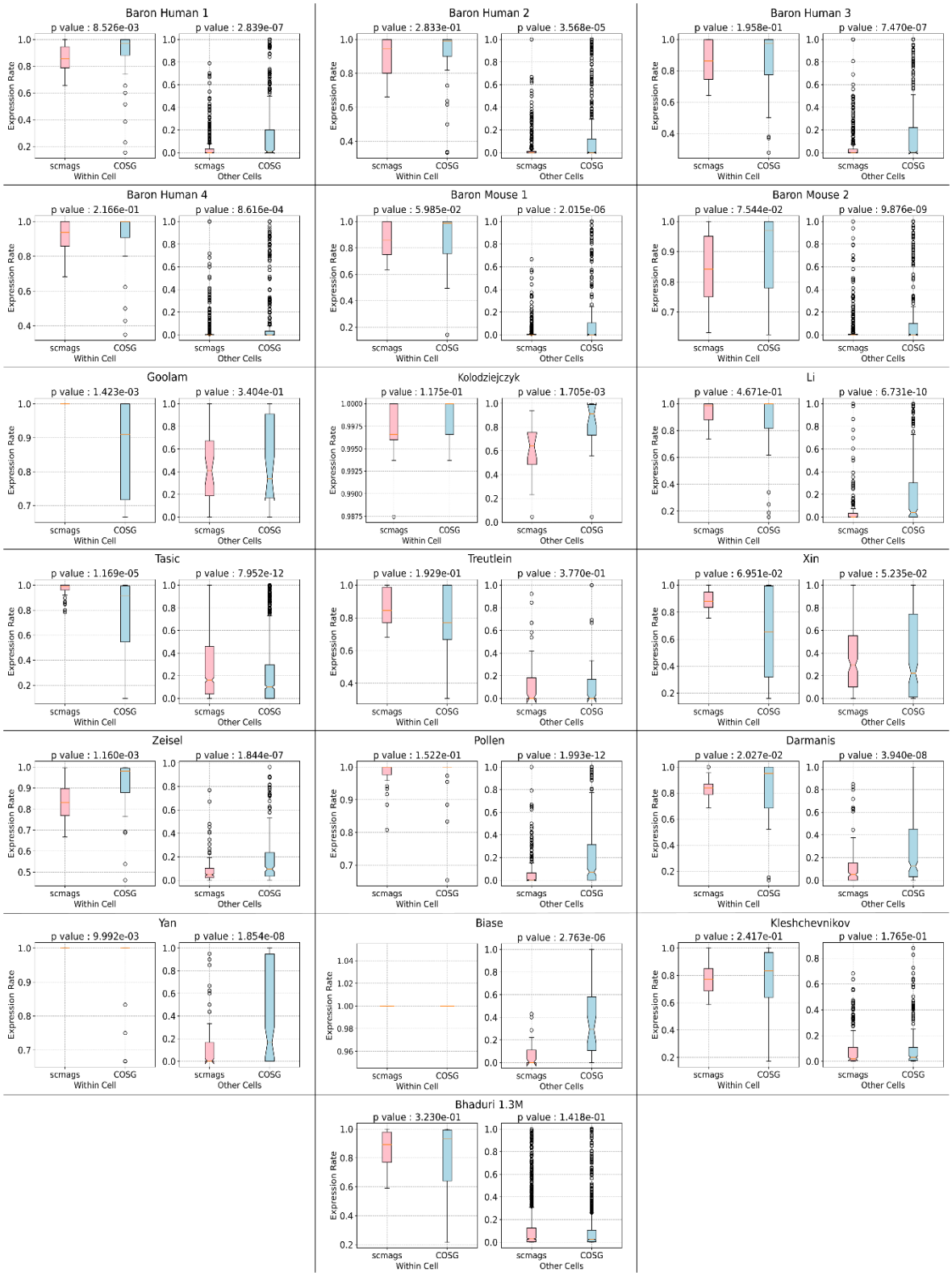
işaretçilerin sınıflama performansını düşüreceği anlamına gelmez. scMAGS ve SMaSH Şekil 3.13'ten görüleceği üzere birbirlerine en yakın ve en iyi sonuçları göstermektedir. scMAGS sınıflama sonuçlarını iyileştirmeye çalışmasa da doğru genleri seçtiği için sınıflama sonuçları SMaSH ile aynıdır. SMaSH ise işaretçileri sınflama sonuçlarına göre seçtiği için beklendiği gibi sınıflama da iyi sonuçlar göstermektedir. Ayrıca COSG ve scGeneFit' te genel olarak iyi sınıflandırma sonuçları göstermiştir.




**Şekil 3.14:** Kleshchevnikov veri seti için seçilen işaretçilerle gerçekleştirilen k-NN sınıflandırmasının Karmaşıklık Matrisleri. (a) scMAGS, (b) SMaSH, (c) scGeneFit, (d) COSG

Şekil 3.13'te görülen durum Şekil 3.14'te de devam etmektedir. Tüm algoritmalar k-NN sınıflandırması için başarılı ve birbirine yakın sonuçlar göstermiştir. Ancak daha önce de

bahsedildiği gibi sınıflandırma sonuçları yöntemin doğru çalıştığını doğrulamaz ama doğru çalışıyorsa bunu teyit etmek için kullanılabilir.

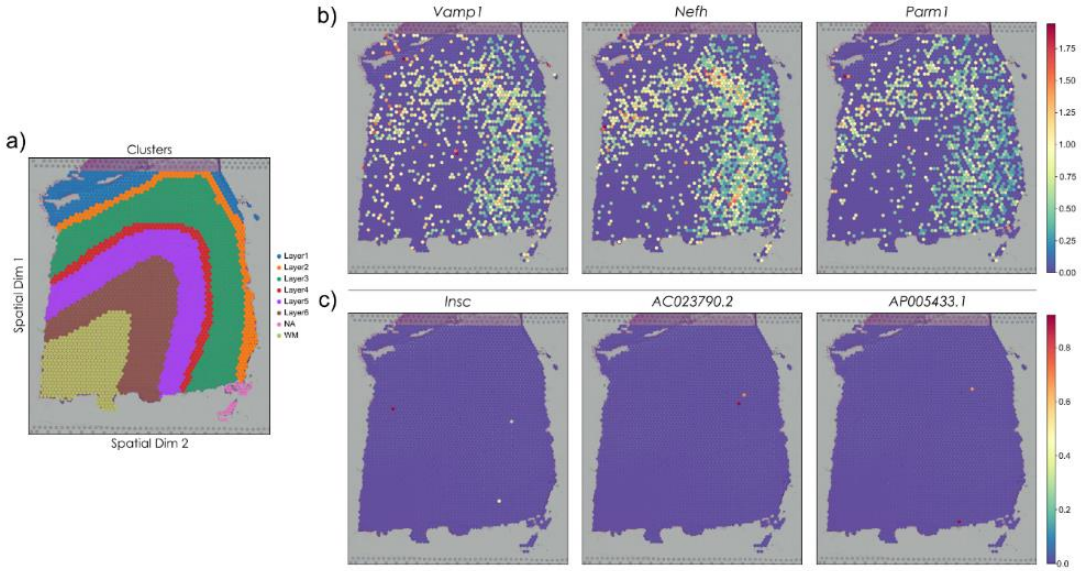



**Şekil 3.15:** Tüm veri setleri için scMAGS ve COSG'nin seçtiği işaretçi genlerin küme içi ve küme dışı ekspresyon oranlarının boxplot grafikleri

Şekil 3.15'te tüm veri setleri için scMAGS ve COSG'nin seçtiği işaretçi genlerin tüm kümelerdeki küme içi ve küme dışı ekspresyon oranlarının genel özeti görülmektedir. Burada her küme için seçilen genlerin küme içi ekspresyonları soldaki kutu grafiklerinde (boxplot), küme dışı ekspresyonları ise sağdaki kutu grafiklerin de görülmektedir. Bu grafikte olması gereken durum küme içi ekspresyon oranları maksimum düzeyde iken küme dışı ekspresyonların minimum düzeyde olmasıdır. Baron Human 3, Baron Mouse 1, Goolam, Li, Tasic, Treutlein, Xin, Darmanis, Kleshchevnikov ve Bhaduri veri setleri için seçilen genlerin küme içi ekspresyon oranlarında scMAGS'ın COSG'ye göre hem daha yüksek eksprese edilen genleri seçtiği hem de kutu grafiklerinin varyansının çeyrekler arası aralıklarının daha düşük olduğu açıkça görülmektedir. Küme içi ekspresyonları için belirtilen durum küme dışı ekspresyonlarında da geçerli olduğu gözükmektedir. Görüldüğü üzere scMAGS COSG'ye göre küme dışında daha az eksprese edilen genleri seçmiştir.

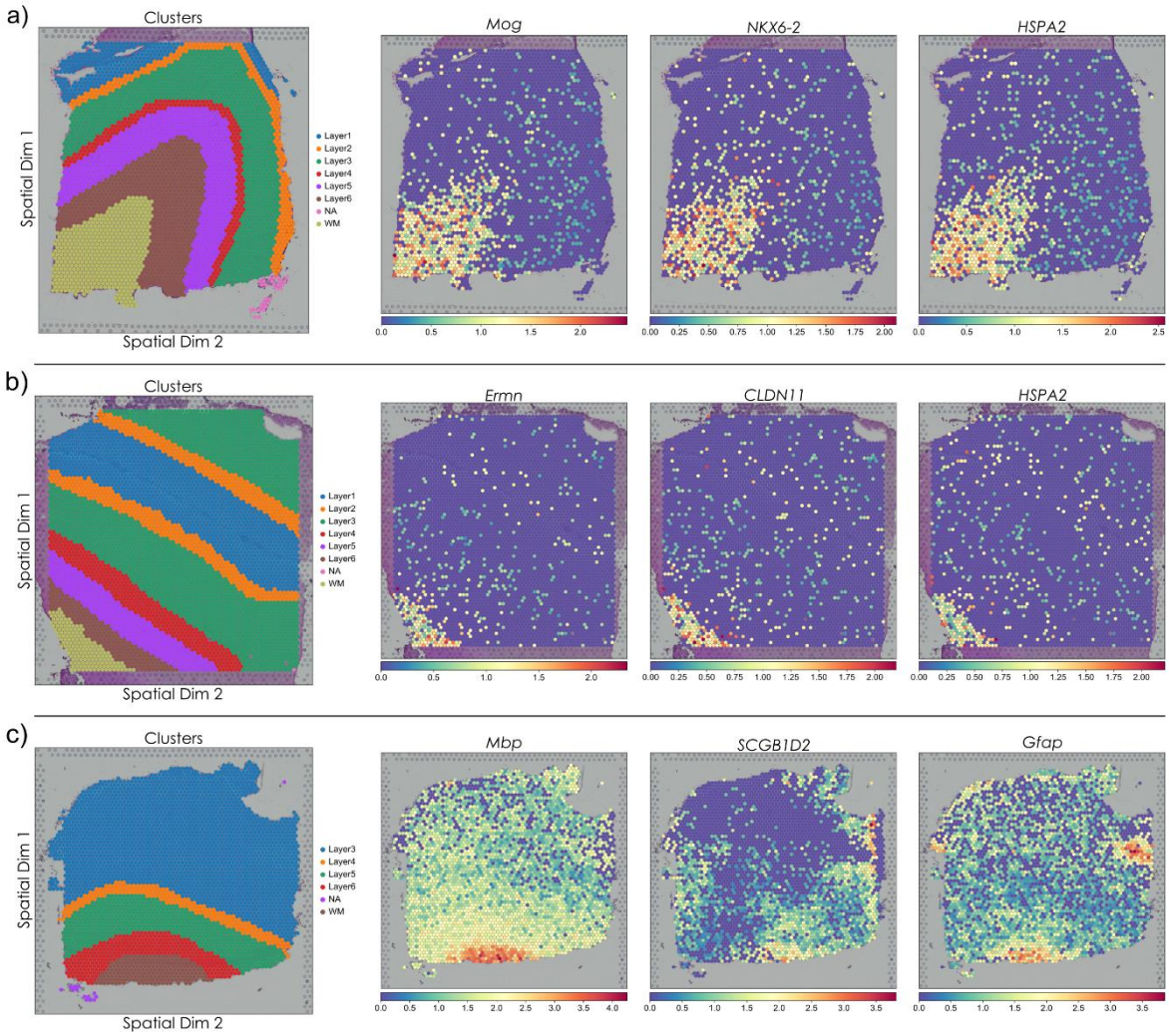
## 3.5 scMAGS'ın Sekanslama Bazlı Uzamsal Transkriptomiks Veri Setlerinde Değerlendirilmesi

Tez kapsamında önerilen yöntemin asıl amacı in situ transkriptomik yöntemlerin ihtiyaç duyduğu işaretçi gen seçimini gerçekleştirmektir. Ancak önerilen yöntem, sekanslama bazlı uzamsal transkriptomik verilerinde veya scRNA-seq veri setlerinde de hücreye özgü işaretçi gen seçimini gerçekleştirebilmektedir. Bu nedenle yöntem iki adet 10X Visium tabanlı sekanslama bazlı uzamsal transkriptomik veri seti üzerinde çalıştırılmıştır.



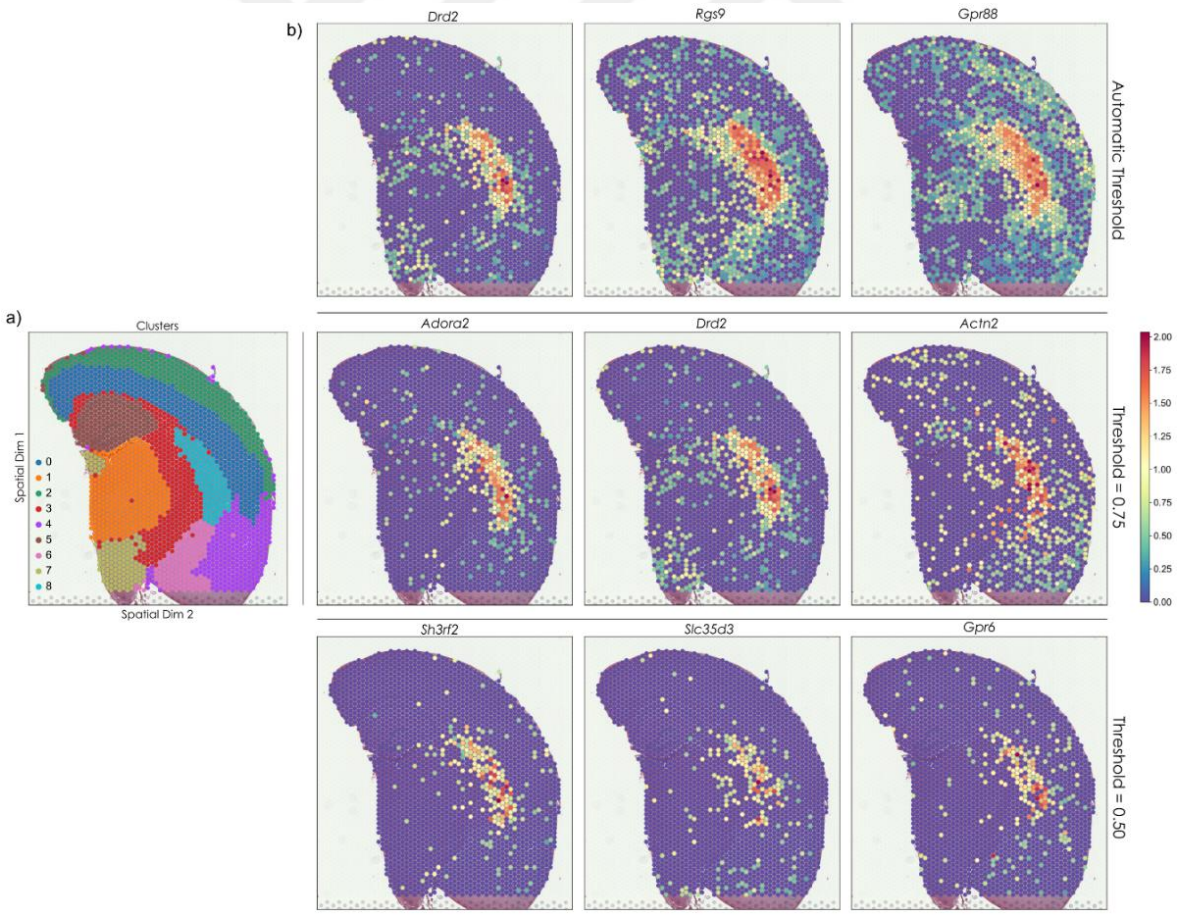

**Şekil 3.16:** DLFPC veri setinin 151673 no'lu doku kesitinin Layer 4 kümesi için scMAGS ve COSG'nin seçtiği işaretçiler (a) Hücre tipleri (b) scMAGS (c) COSG

Şekil 3.16'da (c)'de görüldüğü üzere COSG'nin seçtiği genler Layer 4 dahil olmak üzere hiçbir bölgede eksprese edilmemiştir. scMAGS için küme içi ekspresyon eşiği 0.5 olarak ayarlanmış ve seçtiği genler Layer 4 bölgesinin sahip olduğu şekil ile benzerlik sergilemektedir. Şekil 3.16 (a)'da görüldüğü üzere Layer 4 bölgesi çok küçük bir alanı kapsamaktadır. Dolayısıyla sadece o bölgeyi temsil eden bir gen bulunmayabilir. Ancak böyle bir durum olsa bile COSG'nin seçtiği genlerin ekspresyona sahip olmaması algoritmanın büyük bir kısıtlamaya sahip olduğuna işaret etmektedir. COSG'nin aksine scMAGS keskin hatlarla bölgeyi işaret eden bir gen olmasa da Layer 4 bölgesinin şekline benzer ve bulunduğu bölge de ve mümkün olduğunca yakın kısımlarında eksprese edilen genleri seçmiştir. Yani seçtiği genlerin ekspresyon bölgeleri ile Layer 4 bölgesi benzer bir paterne sahiptir. Zaten seçilen genlerin en yoğun ekspresyona sahip olduğu bölgeler incelenecek olursa Layer 4 katmanına benzer olduğu görülmektedir.




**Şekil 3.17:** scMAGS'ın DLPFC veri setinin (a) 151673, (b) 151509, (c) 151670 no'lu doku kesitlerindeki WM kümesi için seçtiği işaretçiler

Şekil 3.17 scMAGS'ın DLPFC veri setlerinde default parametrelerle WM kümesi için seçtiği genleri görselleştirmektedir. Şekil 3.17 (a) ve (b)' de görüldüğü üzere eğer istenilen kümeyi yani bölgeyi temsil edebilen ve diğer kümelerden ayırt eden bir gen mevcutsa, bu işaretçiler tüm genler arasından seçilebilmektedir. Ancak tüm genler bu şekilde belirlenen kümeyi temsil etmeyebilir. Yani sadece o bölgede ekspresyona sahip olmayabilir. Böyle durumlar da Bölüm 2.4.3'te belirtilen filtreleme adımındaki kriterler devreye girmektedir. Bu durum Şekil 3.17 (c)'de görülebilir. Öncelik belirlenen kümenin tamamında eksprese edilen genlere verilmiştir ve aynı zamanda WM kümesi haricindeki bölgelerde ekspresyon seviyesinin düşük olduğu dolayısıyla bölgeyi ayırt ettiği görülmektedir. 151670 için seçilen 3. gen olan *Gfap* ise küme içinde 1. gen olan *Mbp*'ye göre görece daha düşük ekspresyon oranına sahiptir. Ancak *Gfap* diğer küme dışında daha düşük ekspresyona sahiptir dolayısıyla diğer kriteri sağlamaktadır. Bu şekil ile anlatılmak istenen durum şudur ki; her kümeyi keskin bir şekilde temsil eden bir işaretçi gen bulunmayabilir. Bu şartlar altında, Bölüm 2.4.3'te belirtilen durumlar arasında ödün vermek gerekmektedir.




**Şekil 3.18:** FFPE doku kesitlerinde 8 no'lu küme için değişen eşik değerlerine bağlı olarak scMAGS tarafından seçilen genler.

Şekil 3.18'de görülen grafik önerilen yöntemin küme içi ekspresyon eşiği oranının etkisini göstermek için oluşturulmuştur. Şekil 3.18 (a)'da görülen farklı renk kümeleri hücre tiplerini işaret etmektedir. Dolayısıyla herhangi bir küme için seçilen işaretçi gen ideal olarak sadece o bölgede eksprese edilmelidir. Şekil 3.18 (b)'de 8. küme için seçilen işaretçi genler görülmektedir. İlk satırda bulunan 3 şekilde default parametrelerle seçilen *Drd2*, *Rgs9* ve *Gpr88* genleri görülmektedir. Algoritma otomatik olarak belirlediği eşik üzerinden genleri filtrelemiş ve kalan genler üzerinden gen seçimini gerçekleştirmiştir. *Drd2* görüldüğü üzere en yoğun olarak 8. kümede eksprese edilmiş ve diğer hücrelerde çok düşük ekspresyona sahiptir ve 1. sırada, dolayısıyla en iyi işaretçi gen olarak seçilmiştir. *Rgs9* ve *Gpr88* genleri görüldüğü üzere küme dışında daha çok ekspresyona sahiptir. Bunun sebebi algoritmanın küme içi ve küme dışı ekspresyon değerleri üzerinde seçim yapmak üzere programlanmış ve küme içi ekspresyona öncelik veriyor olmasıdır. 2. satırda seçilen genler küme içi ekspresyon eşiği 0.75 olarak belirlenerek seçilmiş ve görüldüğü üzere algoritma filtreleme adımında daha farklı genleri filtrelediği için seçim aşamasında daha farklı genler seçilmiştir. 3. satırda ise eşik 0.5'e düşürülmüş ve daha farklı sonuçlar görülmüştür. Buradaki amaç kullanıcının seçtiği genleri görselleştirerek seçebileceğini göstermektir. Önerilen yöntem otomatik olarak bir eşik belirlemektedir ancak aynı zamanda kullanıcı bu parametreyi değiştirebilmektedir. Eşik belirlenirken mümkün olduğunca küme içindeki tüm hücrelerde ekspresyon olması amaçlanmaktadır (yani yüksek eşik değerleri seçilmektedir) ve bu durum Bölüm 2.4.2'de açıklamıştır. Dolayısıyla kullanıcı bu parametreyi değiştirerek amacına uygun genleri görsel analizlerle kontrol ederek elde edebilmektedir. Şekil A.7'de, Şekil 3.17'de Layer 4 katmanı için işaretçi seçimi gerçekleştirilmiş olan DLFPC veri setinin diğer kümeler için scMAGS ve COSG tarafından yapılan işaretçi gen seçimleri incelenebilir. Burada scMAGS tarafından yapılan seçimler küme içi ekspresyon eşiği değiştirilerek yapılmış ve değişen eşik değerlerine göre sonuçlar incelenmiştir. Önerilen yazılım paketini kullanacak olan araştırmacı da paket içerisinde sunulan görselleştirme fonksiyonları vasıtasıyla otomatik eşikleme ile seçilen genleri değerlendirip gerekirse; eşik değerlerini, aday işaretçi gen sayısını veya ekspresyon oranının önemini ayarlayan sabiti değiştirerek amacına uygun genleri elde edebilir.

### 3.6 Sonuçlar, Tartışma ve Öneriler

Yapılan değerlendirmeler sonucunda önerilen yöntemin in situ transkriptomik deneyleri için gereken kriterleri sağladığı ve diğer yöntemlere nazaran daha verimli bir biçimde çalıştığı açıkça görülmektedir. Yöntem deneylerin gerektirdiği kriterler dikkate alınarak ve aynı

zamanda giderek büyüyen veriler sebebiyle ortaya çıkan yüksek RAM ihtiyaçları ve hesaplama gücü düşünülerek geliştirilmiştir ve elde edilen sonuçlarda bunu açıkça ortaya koymaktadır. Bununla birlikte scMAGS ancak belirtilen kriterle uygun genler mevcutsa veri seti içerisinde mevcutsa bu genleri seçebilmektedir. Çünkü transkriptomik deneyleri sadece anlık ölçümler vermekte ve deney sonuçlarını etkileyen birçok faktör bulunmaktadır. Dolayısıyla aynı bölgelerden elde edilen veri setlerinde bile farklı genlerin seçilmesi muhtemeldir. Bu sebeplerden ötürü tüm veri setlerinde kriterlere tamamen uyan genlerin seçilmiş olması beklenmemelidir. Ayrıca scMAGS Bölüm 2.4.2'de belirtildiği üzere belli kriterlerin önem sırasına göre seçim yapmaktadır ve küme içi ekspresyon eşiğini otomatik olarak belirlemektedir. Dolayısıyla eşikler kümeler içerisindeki durumlara göre değişkenlik gösterebilir. Bu durumlar göz önünde bulundurularak, kullanıcı seçtiği genleri yazılım paketi dahilindeki görselleştirme fonksiyonlarını kullanarak görselleştirmeli, eğer beklenen kriterler sağlanamadıysa algoritma içerisinde bulunan parametreler değiştirilerek sonuçların değişimi gözlenmelidir.

Diğer bir açıdan scMAGS küme içinde düzgün dağılımla eksprese edilmiş genleri seçmeyi amaçlamaktadır. Ancak ileriye yönelik çalışmalarda küme içi altkümelerin tanımlanmasına ve analiz edilmesine olanak sağlayabilen veya kümeler arası belli paternler sergileyen, gradyana sahip genleri tespit edebilen algoritmalar geliştirilebilir. Kümeler arasında gradyana sahip genlere ek olarak, gelişimsel yörüngeleri analiz etmemize olanak sağlayabilecek algoritmaların geliştirilmesi de birçok probleme ışık tutabilir.

## KAYNAKÇA


1. **Alberts, Bruce Bray, Dennis Hopkin, Karen Johnson, Alexander Lewis, Julian Raff, Martin Roberts, Keith Walter, P.** (2014). *Essential Cell Biology* (4th ed.). New York: Garland Science.

2. **Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., … Canaider, S.** (2013). An estimation of the number of cells in the human body. *Annals of Human Biology*, *40*(6), 463–471. https://doi.org/10.3109/03014460.2013.807878

3. **Smith, G., Rev., J. G.-F. L.,** & 1996, undefined. (n.d.). The admission of DNA evidence in state and federal courts. *HeinOnline*. Retrieved from https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/flr65&section=105

4. **WATSON, J. D., & CRICK, F. H.** (1953). THE STRUCTURE OF DNA. *Cold Spring Harbor Symposia on Quantitative Biology*, *18*, 123–131. https://doi.org/10.1101/SQB.1953.018.01.020

5. **Strobel, E. J., Yu, A. M., & Lucks, J. B.** (2018). High-throughput determination of RNA structures. *Nature Reviews Genetics 2018 19:10*, *19*(10), 615–634. https://doi.org/10.1038/s41576-018-0034-x

6. 'The Double Helix' Review - Twists in the Tale of the Great DNA Discovery - The New York Times. (n.d.). Retrieved October 12, 2022, from https://www.nytimes.com/2012/11/13/science/the-double-helix-review-twists-in-the-tale-of-the-great-dna-discovery.html

7. **Goodwin, S., McPherson, J. D., & McCombie, W. R.** (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics 2016 17:6*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

8. **Ball, B., Langille, M., & Geddes-Mcalister, J.** (2020). Fun(Gi)omics: Advanced and diverse technologies to explore emerging fungal pathogens and define mechanisms of antifungal resistance. *mBio*, *11*(5), 1–18. https://doi.org/10.1128/MBIO.01020-20/ASSET/95D2479F-DE43-4897-96AA-A2036AC5FF32/ASSETS/GRAPHIC/MBIO.01020-20-F0002.JPEG

9. **Hawkins, R. D., Hon, G. C., & Ren, B.** (2010). Next-generation genomics: an integrative approach. *Nature Reviews Genetics 2010 11:7*, *11*(7), 476–486. https://doi.org/10.1038/nrg2795

10. **Wang, Z., Gerstein, M., & Snyder, M.** (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics 2008 10:1*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

11. **Byron, S. A., van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D., & Craig, D. W.** (2016). Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nature Reviews Genetics 2016 17:5*, *17*(5), 257–271. https://doi.org/10.1038/nrg.2016.10

12. **Hwang, B., Lee, J. H., & Bang, D.** (2018, August 1). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental and Molecular Medicine*. Nature Publishing Group. https://doi.org/10.1038/s12276-018-0071-8

13. **Huang, S.** (2009). Non-genetic heterogeneity of cells in development: more than just noise. *Development, 136*(23), 3853–3862. https://doi.org/10.1242/DEV.035139

14. Single cell RNA-seq: An introductory overview and tools for getting started - 10x Genomics. (n.d.). Retrieved October 13, 2022, from https://www.10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started

15. **Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., … Surani, M. A.** (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods, 6*(5), 377–382. https://doi.org/10.1038/nmeth.1315

16. **Potter, S. S.** (2018). Single-cell RNA sequencing for the study of development, physiology and disease. *Nature Reviews Nephrology 2018 14:8, 14*(8), 479–492. https://doi.org/10.1038/s41581-018-0021-7

17. **Tang, X., Huang, Y., Lei, J., Luo, H., & Zhu, X.** (2019). The single-cell sequencing: New developments and medical applications. *Cell and Bioscience*, *9*(1), 1–9. https://doi.org/10.1186/S13578-019-0314-Y/FIGURES/2

18. **Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T**. (2017, August 18). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine*. BioMed Central Ltd. https://doi.org/10.1186/s13073-017-0467-4

19. **Stuart, T., & Satija, R.** (2019). Integrative single-cell analysis. *Nature Reviews Genetics 2019 20:5, 20*(5), 257–272. https://doi.org/10.1038/s41576-019-0093-7

20. **Saliba, A. E., Westermann, A. J., Gorski, S. A., & Vogel, J.** (2014, August 18). Single-cell RNA-seq: Advances and future challenges. *Nucleic Acids Research*. Oxford University Press. https://doi.org/10.1093/nar/gku555

21. **Hashimshony, T., Wagner, F., Sher, N., & Yanai, I.** (2012). CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports*, *2*(3), 666–673. https://doi.org/10.1016/j.celrep.2012.08.003

22. **Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., … McCarroll, S. A.** (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, *161*(5), 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002

23. **Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., … Amit, I.** (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, *343*(6172), 776–779. https://doi.org/10.1126/science.1247651

24. **Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., & Sandberg, R.** (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods, 10*(11), 1096–1100. https://doi.org/10.1038/nmeth.2639

25. **Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., … Kirschner, M. W.** (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, *161*(5), 1187–1201. https://doi.org/10.1016/j.cell.2015.04.044

26. **Sasagawa, Y., Nikaido, I., Hayashi, T., Danno, H., Uno, K. D., Imai, T., & Ueda, H. R**. (2013). Quartz-Seq: A highly reproducible and sensitive single-cell RNA sequencing method, reveals nongenetic gene-expression heterogeneity. *Genome Biology*, *14*(4), 1–17. https://doi.org/10.1186/gb-2013-14-4-r31

27. **Islam, S., Zeisel, A., Joost, S., la Manno, G., Zajac, P., Kasper, M., … Linnarsson, S.** (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, *11*(2), 163–166. https://doi.org/10.1038/nmeth.2772

28. **Kuret, T., Sodin-Šemrl, S., Leskošek, B., & Ferk, P.** (2022). Single Cell RNA Sequencing in Autoimmune Inflammatory Rheumatic Diseases: Current Applications, Challenges and a Step Toward Precision Medicine. *Frontiers in Medicine*, *8*, 3067. https://doi.org/10.3389/FMED.2021.822804/BIBTEX

29. **Lafzi, A., Moutinho, C., Picelli, S., & Heyn, H.** (2018). Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nature Protocols 2018 13:12*, *13*(12), 2742–2757. https://doi.org/10.1038/s41596-018-0073-y

30. **Shapiro, E., Biezuner, T., & Linnarsson, S.** (2013, September 30). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*. Nature Publishing Group. https://doi.org/10.1038/nrg3542

31. **Wang, Y., & Navin, N. E.** (2015, May 21). Advances and Applications of Single-Cell Sequencing Technologies. *Molecular Cell*. Cell Press. https://doi.org/10.1016/j.molcel.2015.05.005

32. **Hu, P., Zhang, W., Xin, H., & Deng, G.** (2016). Single cell isolation and analysis. *Frontiers in Cell and Developmental Biology*, *4*(OCT), 116. https://doi.org/10.3389/FCELL.2016.00116/BIBTEX

33. **Whitesides, G. M.** (2006). The origins and the future of microfluidics. *Nature 2006 442:7101*, *442*(7101), 368–373. https://doi.org/10.1038/nature05058

34. **Zeb, Q., Wang, C., Shafiq, S., & Liu, L.** (2019). An overview of single-cell isolation techniques. In *Single-Cell Omics: Volume 1: Technological Advances and Applications* (pp. 101–135). Elsevier. https://doi.org/10.1016/B978-0-12-814919-5.00006-3

35. **Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., & Teichmann, S. A.** (2015, May 21). The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell*. Cell Press. https://doi.org/10.1016/j.molcel.2015.04.005

36. **Sena, J. A., Galotto, G., Devitt, N. P., Connick, M. C., Jacobi, J. L., Umale, P. E., … Bell, C. J.** (2018). Unique Molecular Identifiers reveal a novel sequencing artefact with implications for RNA-Seq based gene expression analysis. *Scientific Reports 2018 8:1*, *8*(1), 1–13. https://doi.org/10.1038/s41598-018-31064-7

37. **Eberwine, J., Sul, J. Y., Bartfai, T., & Kim, J.** (2014, December 30). The promise of single-cell sequencing. *Nature Methods*. Nature Publishing Group. https://doi.org/10.1038/nmeth.2769

38. **Luecken, M. D., & Theis, F. J.** (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, *15*(6), e8746. https://doi.org/10.15252/msb.20188746

39. **Stegle, O., Teichmann, S. A., & Marioni, J. C.** (2015, March 26). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*. Nature Publishing Group. https://doi.org/10.1038/nrg3833

40. **Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., & Chen, J.** (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology*, *21*(1), 1–32. https://doi.org/10.1186/s13059-019-1850-9

41. **Kharchenko, P. v.** (2021). The triumphs and limitations of computational methods for scRNA-seq. *Nature Methods*, 1–10. https://doi.org/10.1038/s41592-021-01171-x

42. **Lee, J. H.** (2017). Quantitative approaches for investigating the spatial context of gene expression. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, *9*(2), e1369. https://doi.org/10.1002/WSBM.1369

43. **Williams, C. G., Lee, H. J., Asatsuma, T., Vento-Tormo, R., & Haque, A.** (2022). An introduction to spatial transcriptomics for biomedical research. *Genome Medicine*, *14*(1), 1–18. https://doi.org/10.1186/S13073-022-01075-1/FIGURES/3

44. **Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., & Zhuang, X.** (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, *348*(6233). https://doi.org/10.1126/SCIENCE.AAA6090/SUPPL_FILE/CHEN-SM.PDF

45. **Crosetto, N., Bienko, M., & van Oudenaarden, A.** (2014). Spatially resolved transcriptomics and beyond. *Nature Reviews Genetics 2014 16:1*, *16*(1), 57–66. https://doi.org/10.1038/nrg3832

46. **Tian, L., Chen, F., & Macosko, E. Z.** (2022). The expanding vistas of spatial transcriptomics. *Nature Biotechnology 2022*, 1–10. https://doi.org/10.1038/s41587-022-01448-2

47. **Marx, V.** (2021). Method of the Year: spatially resolved transcriptomics. *Nature Methods 2021 18:1*, *18*(1), 9–14. https://doi.org/10.1038/s41592-020-01033-y

48. **Borm, L. E., Mossi Albiach, A., Mannens, C. C. A., Janusauskas, J., Özgün, C., Fernández-García, D., … Linnarsson, S.** (2022). Scalable in situ single-cell profiling by electrophoretic capture of mRNA using EEL FISH. *Nature Biotechnology 2022*, 1–10. https://doi.org/10.1038/s41587-022-01455-3

49. **Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., … Frisén, J.** (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, *353*(6294), 78–82. https://doi.org/10.1126/SCIENCE.AAF2403/SUPPL_FILE/AAF2403_STAHL_SM.PDF

50. **Larsson, L., Frisén, J., & Lundeberg, J.** (2021). Spatially resolved transcriptomics adds a new dimension to genomics. *Nature Methods 2021 18:1*, *18*(1), 15–18. https://doi.org/10.1038/s41592-020-01038-7

51. **Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., … Macosko, E. Z.** (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, *363*(6434), 1463–1467. https://doi.org/10.1126/SCIENCE.AAW1219/SUPPL_FILE/AAW1219S1.MOV

52. Spatial Gene Expression for Fresh Frozen - Official 10x Genomics Support. (n.d.). Retrieved October 19, 2022, from https://www.10xgenomics.com/support/spatial-gene-expression-fresh-frozen

53. **Zhuang, X.** (2021). Spatially resolved single-cell genomics and transcriptomics by imaging. *Nature Methods 2020 18:1*, *18*(1), 18–22. https://doi.org/10.1038/s41592-020-01037-8

54. **Volpi, E. v., & Bridger, J. M.** (2018). FISH glossary: an overview of the fluorescence in situ hybridization technique. *https://doi.org/10.2144/000112811*, *45*(4), 385–409. https://doi.org/10.2144/000112811

55. **Amann, R., & Fuchs, B. M.** (2008). Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques. *Nature Reviews Microbiology 2008 6:5*, *6*(5), 339–348. https://doi.org/10.1038/nrmicro1888

56. **Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A., & Tyagi, S.** (2008). Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods 2008 5:10*, *5*(10), 877–879. https://doi.org/10.1038/nmeth.1253

57. **Gall, J. G.** (1968). Differential synthesis of the genes for ribosomal RNA during amphibian oögenesis. *Proceedings of the National Academy of Sciences of the United States of America*, *60*(2), 553–560. https://doi.org/10.1073/PNAS.60.2.553/ASSET/6F28E228-C20B-4A56-B33F-AA45646DE160/ASSETS/PNAS.60.2.553.FP.PNG

58. **Levsky, J. M., & Singer, R. H.** (2003). Fluorescence in situ hybridization: past, present and future. *Journal of Cell Science*, *116*(14), 2833–2838. https://doi.org/10.1242/JCS.00633

59. **Lubeck, E., & Cai, L.** (2012). Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nature Methods 2012 9:7*, *9*(7), 743–748. https://doi.org/10.1038/nmeth.2069

60. **Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M., & Cai, L.** (2014). Single-cell in situ RNA profiling by sequential hybridization. *Nature Methods 2014 11:4*, *11*(4), 360–361. https://doi.org/10.1038/nmeth.2892

61. **Shah, S., Lubeck, E., Zhou, W., & Cai, L.** (2017). seqFISH Accurately Detects Transcripts in Single Cells and Reveals Robust Spatial Organization in the Hippocampus. *Neuron*, *94*(4), 752-758.e1. https://doi.org/10.1016/J.NEURON.2017.05.008

62. **Shah, S., Lubeck, E., Zhou, W., & Cai, L.** (2016). In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. *Neuron*, *92*(2), 342–357. https://doi.org/10.1016/J.NEURON.2016.10.001

63. **Lein, E., Borm, L. E., & Linnarsson, S.** (2017). The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science*, *358*(6359), 64–69. https://doi.org/10.1126/SCIENCE.AAN6827

64. **Shah, S., Takei, Y., Zhou, W., Lubeck, E., Yun, J., Eng, C. H. L., … Cai, L.** (2018). Dynamics and Spatial Genomics of the Nascent Transcriptome by Intron seqFISH. *Cell*, *174*(2), 363-376.e16. https://doi.org/10.1016/J.CELL.2018.05.035

65. **Polonsky, M., Round, K., Seewaldt, V., & Cai, L.** (2020). Abstract B117: Analysis of luminal breast cancer tumor microenvironment using seqFISH. *Cancer Epidemiology, Biomarkers & Prevention*, *29*(6_Supplement_1), B117–B117. https://doi.org/10.1158/1538-7755.DISP18-B117

66. **Codeluppi, S., Borm, L. E., Zeisel, A., la Manno, G., van Lunteren, J. A., Svensson, C. I., & Linnarsson, S.** (2018). Spatial organization of the somatosensory cortex revealed by osmFISH. *Nature Methods 2018 15:11*, *15*(11), 932–935. https://doi.org/10.1038/s41592-018-0175-z

67. **Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., & Zhuang, X.** (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*. https://doi.org/10.1126/SCIENCE.AAA6090

68. **Codeluppi, S., Borm, L. E., Zeisel, A., la Manno, G., van Lunteren, J. A., Svensson, C. I., & Linnarsson, S.** (2018). Spatial organization of the somatosensory cortex revealed by

osmFISH. *Nature Methods 2018 15:11*, *15*(11), 932–935. https://doi.org/10.1038/s41592-018-0175-z

69. **Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wählby, C., & Nilsson, M.** (2013). In situ sequencing for RNA analysis in preserved tissue and cells. *Nature Methods 2013 10:9*, *10*(9), 857–860. https://doi.org/10.1038/nmeth.2563

70. **Eng, C. H. L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., … Cai, L.** (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature 2019 568:7751*, *568*(7751), 235–239. https://doi.org/10.1038/s41586-019-1049-y

71. **Wang, X., Allen, W. E., Wright, M. A., Sylwestrak, E. L., Samusik, N., Vesuna, S., … Deisseroth, K.** (2018). Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, *361*(6400). https://doi.org/10.1126/SCIENCE.AAT5691/SUPPL_FILE/AAT5691_WANG_SM_TABLE-S2.XLSX

72. **Alon, S., Goodwin, D. R., Sinha, A., Wassie, A. T., Chen, F., Daugharthy, E. R., … Boyden, E. S.** (2021). Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. *Science*, *371*(6528). https://doi.org/10.1126/SCIENCE.AAX2656/SUPPL_FILE/AAX2656_TABLESS1-S6ANDS9-S14.XLSX

73. **Biancalani, T., Scalia, G., Buffoni, L., Avasthi, R., Lu, Z., Sanger, A., … Regev, A.** (2021). Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nature Methods 2021 18:11*, *18*(11), 1352–1362. https://doi.org/10.1038/s41592-021-01264-7

74. **Zhuxia, L., Guangdun, P., Zhuxia, L., & Guangdun, P.** (2021). Spatial transcriptomics: new dimension of understanding biological complexity. *Biophysics Reports, Uncorrected proof*, *7*(0), 1–17. https://doi.org/10.52601/BPR.2021.210037

75. **Tsanov, N., Samacoits, A., Chouaib, R., Traboulsi, A. M., Gostan, T., Weber, C., … Mueller, F.** (2016). smiFISH and FISH-quant – a flexible single RNA detection approach with super-resolution capability. *Nucleic Acids Research*, *44*(22), e165–e165. https://doi.org/10.1093/NAR/GKW784

76. **Nelson, M. E., Riva, S. G., & Cvejic, A**. (2021). SMaSH: A scalable, general marker gene identification framework for single-cell RNA sequencing and Spatial Transcriptomics. *bioRxiv*, 2021.04.08.438978. https://doi.org/10.1101/2021.04.08.438978

77. **Wolf, F. A., Angerer, P., & Theis, F. J.** (2018). SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, *19*(1), 1–5. https://doi.org/10.1186/S13059-017-1382-0/FIGURES/1

78. **Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., … Raychaudhuri, S.** (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods 2019 16:12*, *16*(12), 1289–1296. https://doi.org/10.1038/s41592-019-0619-0

79. **Dumitrascu, B., Villar, S., Mixon, D. G., & Engelhardt, B. E.** (2021). Optimal marker gene selection for cell type discrimination in single cell analyses. *Nature Communications 2021 12:1*, *12*(1), 1–8. https://doi.org/10.1038/s41467-021-21453-4

80. **Dai, M., Pei, X., & Wang, X. J.** (2022). Accurate and fast cell marker gene identification with COSG. *Briefings in Bioinformatics*, *23*(2). https://doi.org/10.1093/BIB/BBAB579

81. **Balki, I., Amirabadi, A., Levman, J., Martel, A. L., Emersic, Z., Meden, B., … Tyrrell, P. N.** (2019). Sample-Size Determination Methodologies for Machine Learning in Medical Imaging Research: A Systematic Review. *Canadian Association of Radiologists Journal*, *70*(4), 344–353. https://doi.org/10.1016/j.carj.2019.06.002

82. **Liu, S., & Johnson, V. E.** (2016). lanternpharma. *Biostatistics*, *17*(2), 249–263. https://doi.org/10.1093/BIOSTATISTICS

83. **Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., … Bielas, J. H.** (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*. https://doi.org/10.1038/ncomms14049

84. **Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., … Shendure, J.** (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, *566*(7745), 496–502. https://doi.org/10.1038/s41586-019-0969-x

85. **Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I.** (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, *46*(1), 243–256. https://doi.org/10.1016/J.PATCOG.2012.07.021

86. **Rousseeuw, P. J.** (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*(C), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

87. **Caliñski, T., & Harabasz, J.** (1974). A Dendrite Method Foe Cluster Analysis. *Communications in Statistics*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

88. **van der Maaten, L., & Hinton, G.** (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

89. **Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. A., Lui, J. H., … West, J. A. A.** (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature Biotechnology*, *32*(10), 1053–1058. https://doi.org/10.1038/nbt.2967

90. **Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., … Quake, S. R.** (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, *509*(7500), 371–375. https://doi.org/10.1038/nature13173

91. **Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., … Quake, S. R.** (2015). A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(23), 7285–7290. https://doi.org/10.1073/pnas.1507125112

92. **Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., … Yanai, I.** (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Systems*, *3*(4), 346-360.e4. https://doi.org/10.1016/j.cels.2016.08.011

93. **Xin, Y., Kim, J., Okamoto, H., Ni, M., Wei, Y., Adler, C., … Gromada, J.** (2016). RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. *Cell Metabolism*, *24*(4), 608–615. https://doi.org/10.1016/J.CMET.2016.08.018

94. **Tasic, B., Menon, V., Nguyen, T. N., Kim, T. K., Jarsky, T., Yao, Z., … Zeng, H.** (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience*, *19*(2), 335–346. https://doi.org/10.1038/nn.4216

95. **Li, H., Courtois, E. T., Sengupta, D., Tan, Y., Chen, K. H., Goh, J. J. L., … Prabhakar, S.** (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature Genetics*, *49*(5), 708–718. https://doi.org/10.1038/ng.3818

96. **Goolam, M., Scialdone, A., Graham, S. J. L., MacAulay, I. C., Jedrusik, A., Hupalowska, A., … Zernicka-Goetz, M.** (2016). Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. *Cell*, *165*(1), 61–74. https://doi.org/10.1016/J.CELL.2016.01.047

97. **Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., … Tang, F.** (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature Structural and Molecular Biology*, *20*(9), 1131–1139. https://doi.org/10.1038/nsmb.2660

98. **Kolodziejczyk, A. A., Kim, J. K., Tsang, J. C. H., Ilicic, T., Henriksson, J., Natarajan, K. N., … Teichmann, S. A.** (2015). Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. *Cell Stem Cell*, *17*(4), 471–485. https://doi.org/10.1016/j.stem.2015.09.011

99. **Biase, F. H., Cao, X., & Zhong, S.** (2014). Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. *Genome Research*, *24*(11), 1787–1796. https://doi.org/10.1101/gr.177725.114

100. **Zeisel, A., M̂oz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., Manno, G. la, Juréus, A., … Linnarsson, S.** (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, *347*(6226), 1138–1142. https://doi.org/10.1126/SCIENCE.AAA1934/SUPPL_FILE/ZEISEL-SM.PDF

101. **Kleshchevnikov, V., Shmatko, A., Dann, E., Aivazidis, A., King, H. W., Li, T., … Bayraktar, O. A.** (2020). Comprehensive mapping of tissue cell architecture via integrated single cell and spatial transcriptomics. *bioRxiv*, 2020.11.15.378125. https://doi.org/10.1101/2020.11.15.378125

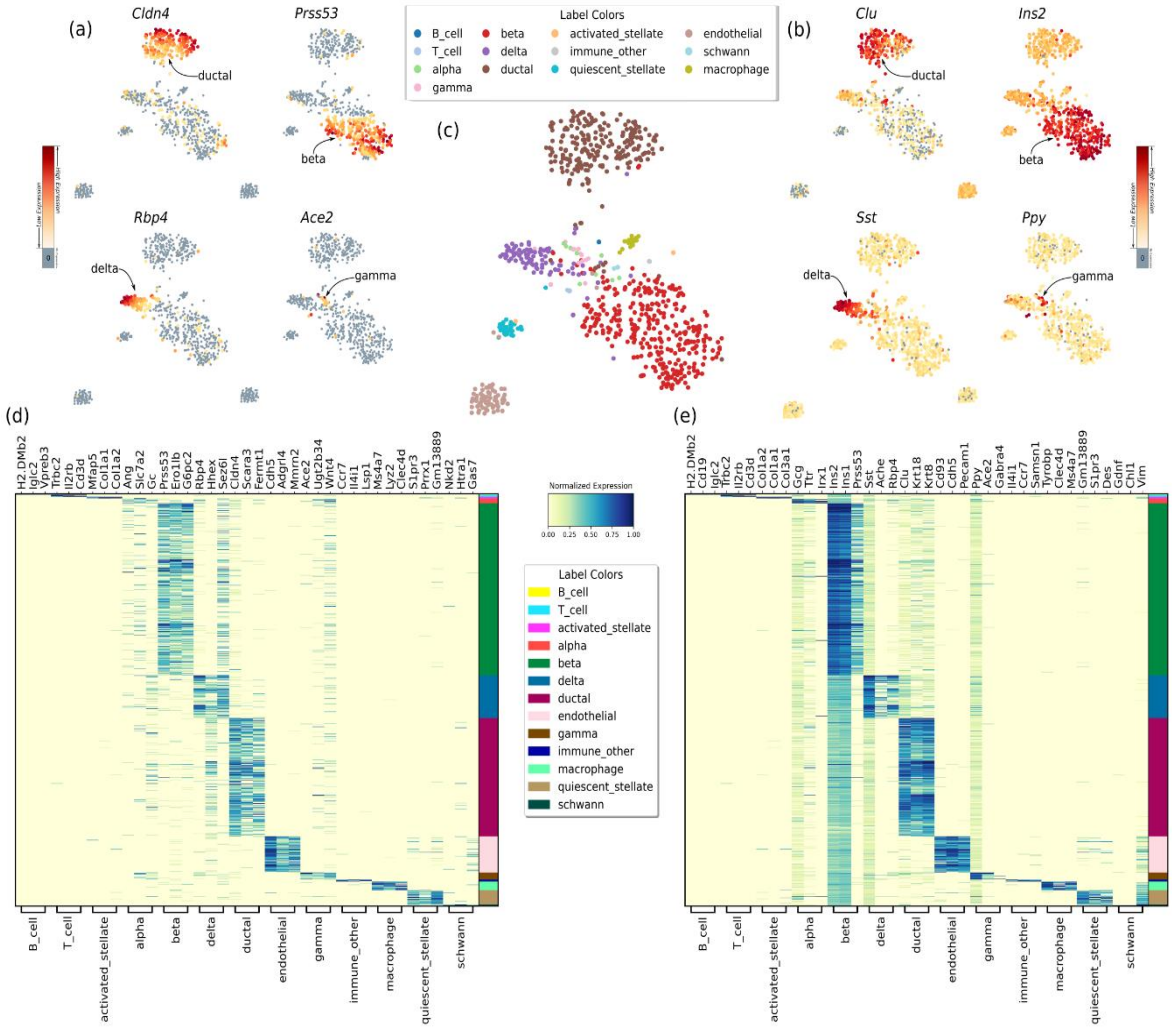
102. **Bhaduri, A., Nowakowski, T. J., Pollen, A. A., & Kriegstein, A. R.** (2018). Identification of cell types in a mouse brain single-cell atlas using low sampling coverage. *BMC Biology*, *16*(1), 1–10. https://doi.org/10.1186/S12915-018-0580-X/FIGURES/4

103. **Maynard, K. R., Collado-Torres, L., Weber, L. M., Uytingco, C., Barry, B. K., Williams, S. R., … Jaffe, A. E.** (2021). Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience 2021 24:3*, *24*(3), 425–436. https://doi.org/10.1038/s41593-020-00787-0

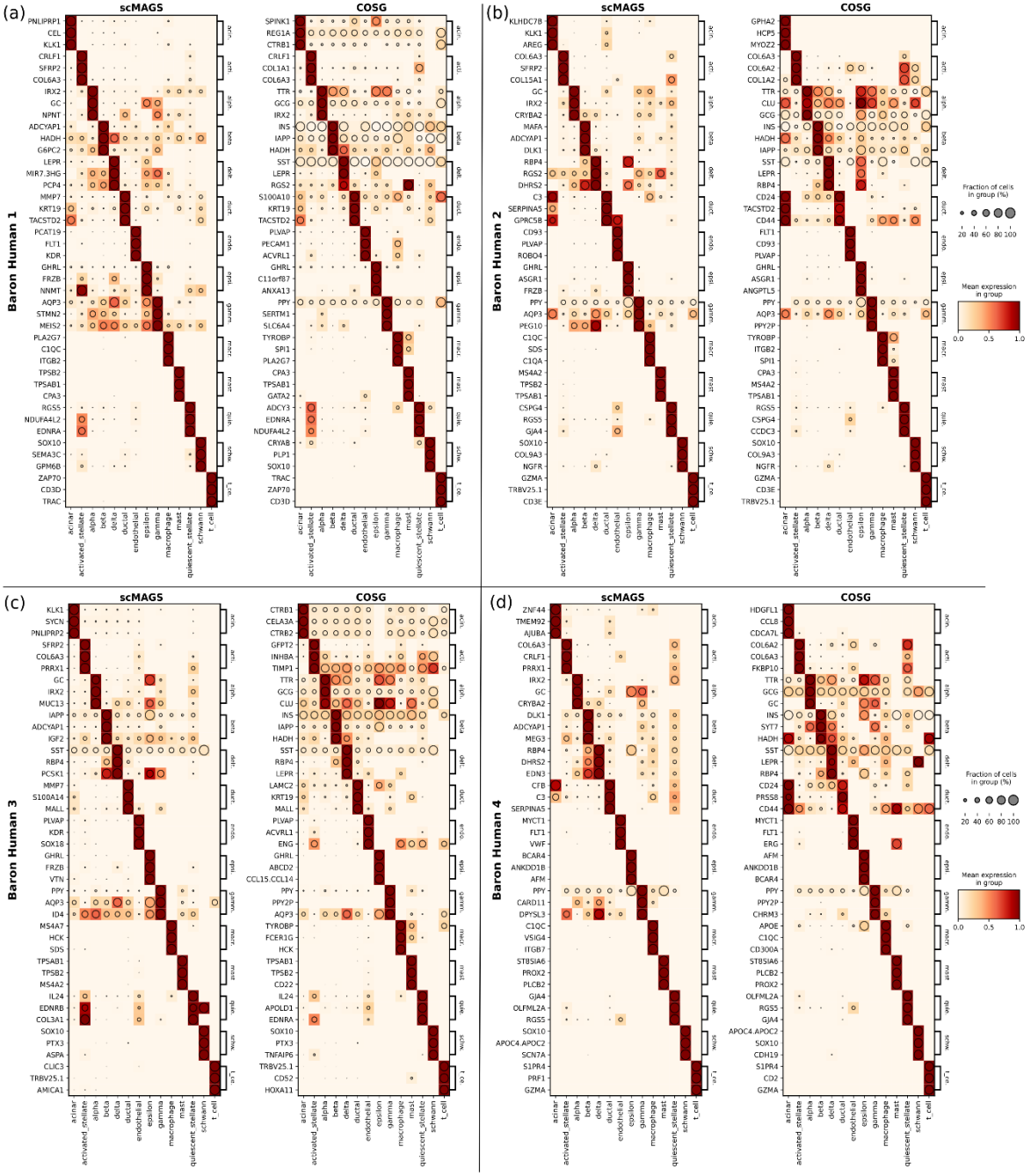
104. Adult Mouse Brain (FFPE) - 10x Genomics. (n.d.). Retrieved October 25, 2022, from https://www.10xgenomics.com/resources/datasets/adult-mouse-brain-ffpe-1-standard-1-3-0

# EKLER

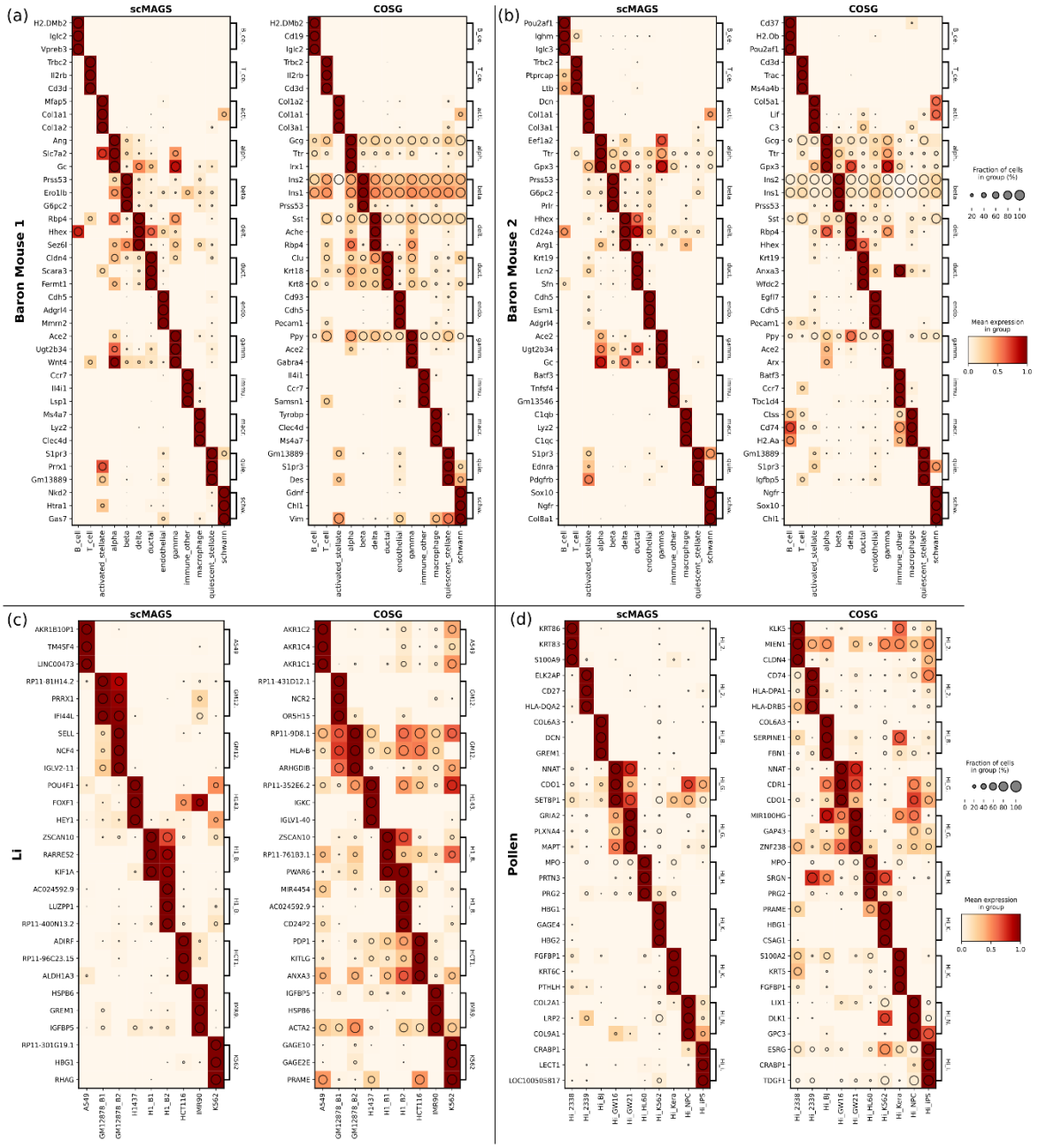
## EK-A



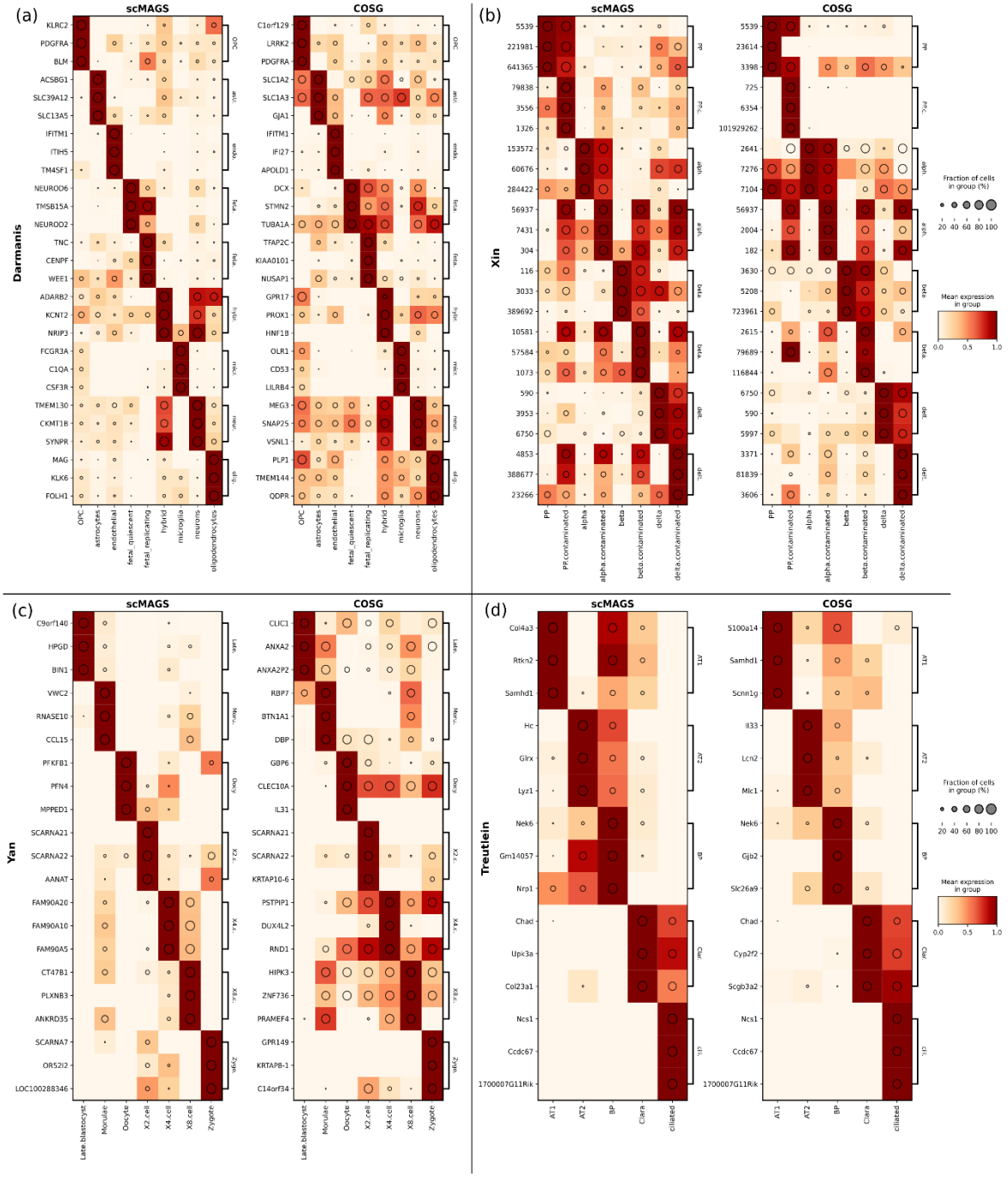

**Şekil A.1:** Baron Mouse 2 veri seti için scMAGS ve COSG tarafından seçilen işaretçiler



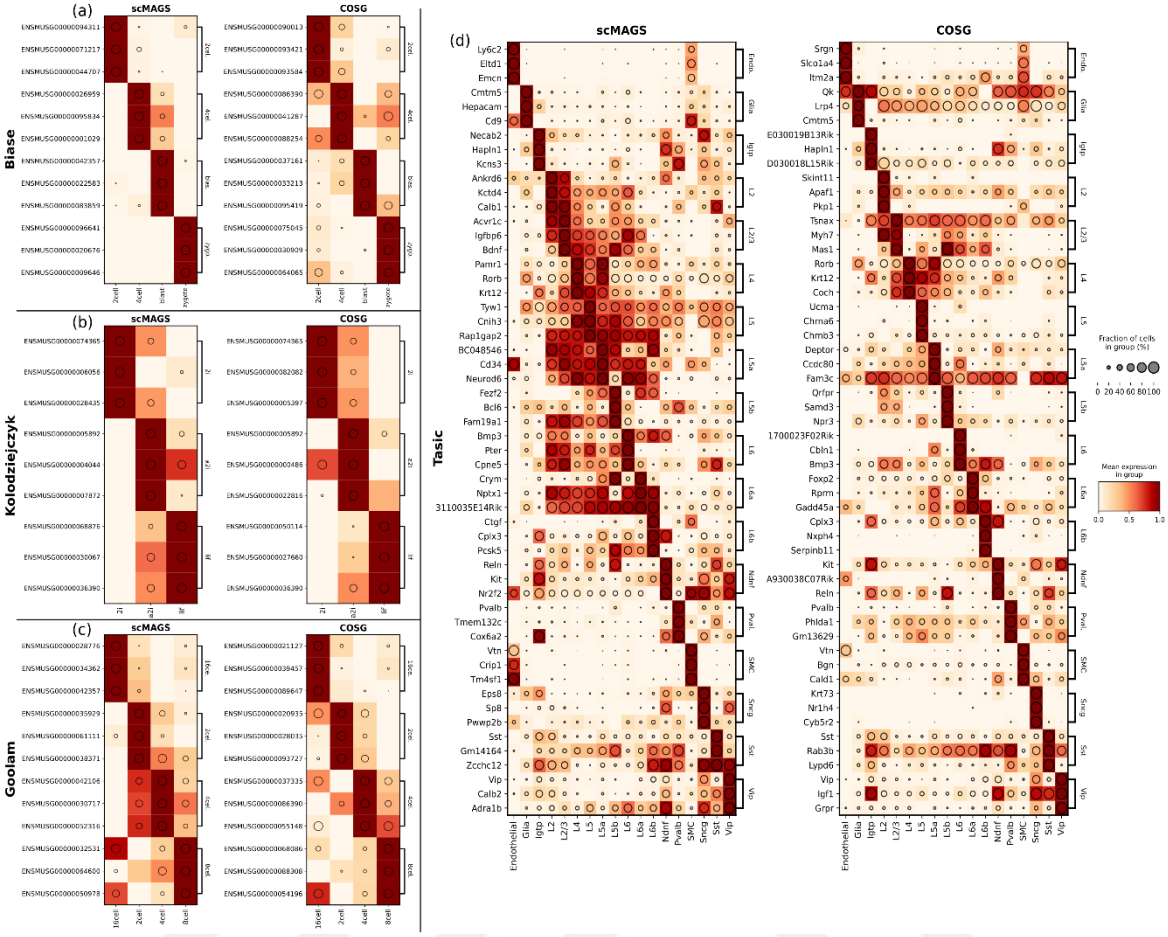

**Şekil A.2:** Baron Human 1-2-3-4 veri setleri için scMAGS ve COSG tarafından seçilen işaretçilerin Dotplot grafikleri



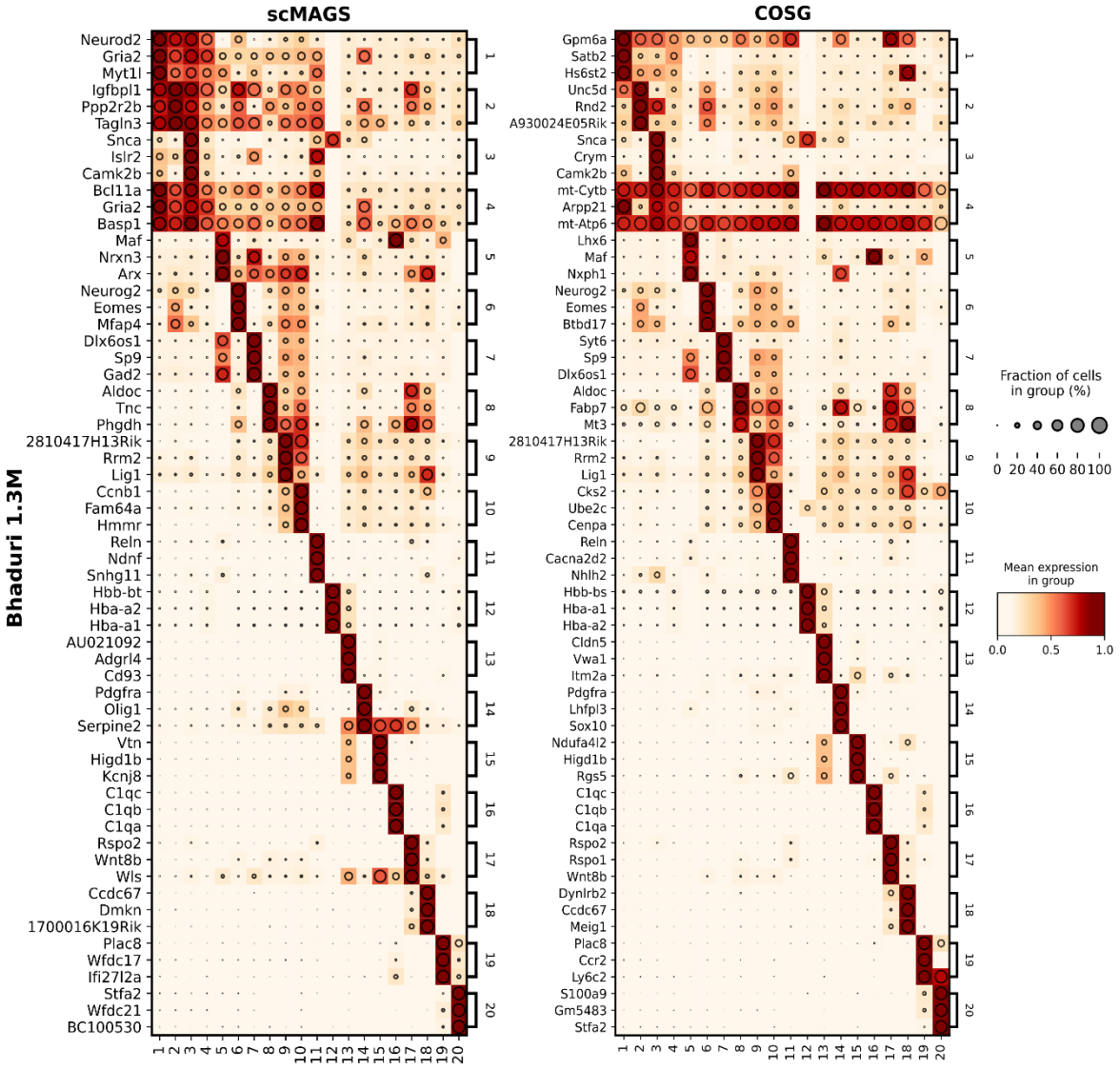

**Şekil A.3:** Baron Mouse 1-2, Li ve Pollen veri setleri için scMAGS ve COSG tarafından seçilen işaretçilerin Dotplot grafikleri



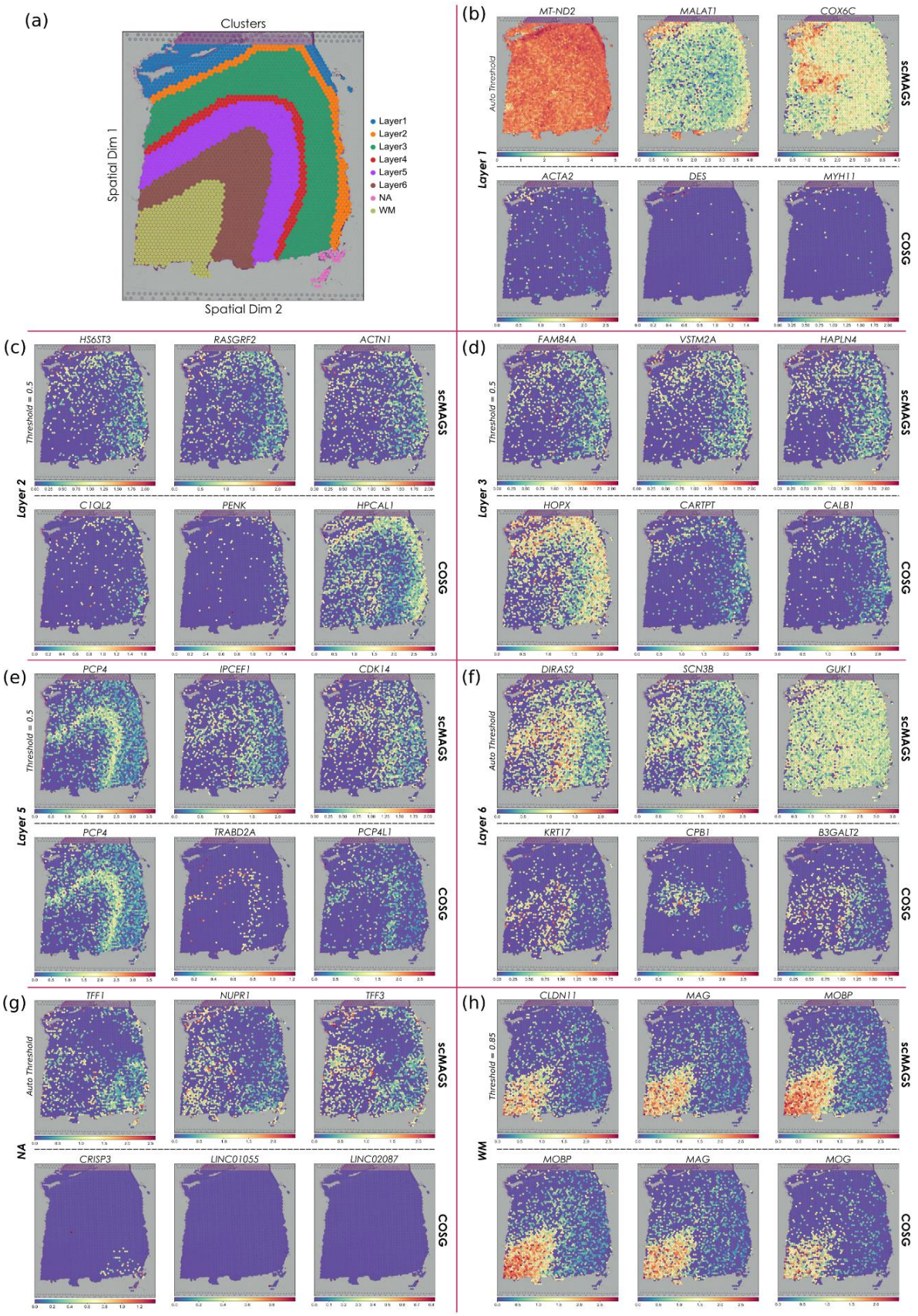

**Şekil A.4:** Darmanis, Xin, Yan ve Treutlein veri setleri için scMAGS ve COSG tarafından seçilen işaretçilerin Dotplot grafikleri




**Şekil A.5:** Biase, Kolodziejczyk, Goolam ve Tasic veri setleri için scMAGS ve COSG tarafından seçilen işaretçilerin Dotplot grafikleri




**Şekil A.6:** Bhaduri veri seti için scMAGS ve COSG tarafından seçilen işaretçilerin Dotplot grafikleri




**Şekil A.7** DLFPC veri setinin 151673 no'lu doku kesitin de Layer 1, 2, 3, 5, 6, NA ve WM kümeleri için scMAGS ve COSG tarafından seçilen işaretçi genler

## ÖZGEÇMİŞ

**Ad-Soyad: Yusuf BARAN**

**ÖĞRENİM DURUMU:**

- **Lisans: 2020, İnönü Üniversitesi, Mühendislik Fakültesi, Biyomedikal Mühendisliği**

**YÜKSEK LİSANS VEYA DOKTORA TEZİNDEN TÜRETİLEN ÇALIŞMALAR**

- **Baran, Y., & Doğan, B. (2022).** scMAGS: Marker gene selection from scRNA-seq data for spatial transcriptomics studies. *bioRxiv,* 2022.03.22.485261. https://doi.org/10.1101/2022.03.22.485261