Learning cell identities and (post-)transcriptional regulation using single-cell data

Michielsen, L.C.M.

SAMENVATTING

Weefsels in het menselijk lichaam, in het bijzonder de hersenen, zijn heterogeen en bestaan uit veel verschillende celtypen. Celtypen kunnen worden gedefinieerd door de genen die tot expressie komen in een cel, wat gecontroleerd wordt door unieke celtypespecifieke (post-) transcriptionele mechanismen. Ziekten kunnen deze controlemechanismen verstoren en dus een verschillend effect hebben op celtypes. Begrijpen welk celtype wordt beïnvloed door een ziekte is daarom cruciaal bij het ontwikkelen van nieuwe medicijnen. Single-nucleotide-polymorfismen (SNPs) in het DNA kunnen geassocieerd worden met ziekten, maar ongeveer 95% van de SNPs maakt deel uit van het niet-coderende DNA. Meestal is het onbekend of een SNP de oorzaak van een ziekte is en welk gen en celtype wordt beïnvloed. Het bestuderen van genexpressie op celniveau zou zulke verstoorde mechanismen kunnen onthullen.

De vooruitgang in single-cell RNA sequencing heeft ons begrip van heterogene weefsels sterk verbeterd en geleid tot de ontdekking van veel nieuwe celtypes. Deze nieuwe technologie brengt echter ook computationele uitdagingen met zich mee. Bij het vergelijken van datasets van verschillende cohorten (bijvoorbeeld van veel verschillende individuen) is het belangrijk om cellen consistent te annoteren. Om deze consistentie te garanderen, is het essentieel om cellen te annoteren met behulp van classificatiemethoden in plaats van de huidige cluster-methoden, die subjectief en tijdrovend zijn. Om deze overgang te vergemakkelijken, hebben we in dit proefschrift classificatiemethoden voor celtypen gebenchmarkt en computationele methoden ontwikkeld om automatisch referentie-atlassen te bouwen met behulp van meer-dere reeds gelabelde single-cell datasets. We laten zien hoe dergelijke referentie-atlassen kunnen worden ingezet om automatisch nieuwe (ongelabelde) single-cell datasets te annote-ren en hoe ze continu kunnen worden bijgewerkt met behulp van nieuwe single-cell datasets.

Met de meer consistente annotatie van celtypen in single-cell data, gaan we terug naar de relatie tussen mutaties en hun effect op genexpressie. Hiertoe bestuderen we sequentie-naar-expressie modellen die een verandering in expressie kunnen voorspellen wanneer een mutatie wordt waargenomen. Aangezien genexpressie celtypespecifiek is, introduceren we sequentie-naar-expressiemodellen getraind op single-cell data om celtypespecifieke voorspellingen te doen. We gebruiken deze modellen om aan te tonen dat bepaalde mutaties inderdaad genexpressie veranderen, wat ons begrip van transcriptionele regulatie vergroot.

Naast verschillen in genexpressie tussen celtypen, kunnen celtypen ook verschillende isovormen van een gen tot expressie brengen (d.w.z. verschillende combinaties van exonen in een mRNA-molecuul). Ook dit kan worden veranderd door mutaties in het DNA. Single-cell long-read sequencing maakt het mogelijk om expressie van isovormen in celtypes te meten. We gebruiken deze data en stellen een nieuwe aanpak voor waarbij we onze sequentie-naar-expressiemodellen gebruiken om celtypespecifiek isovormgebruik te voorspellen. Dit opent een nieuwe weg voor het bekijken van celtypespecifieke veranderingen.

Alles bij elkaar introduceren we een scala aan computationele methoden om single-cell RNA sequencing data te gebruiken om ons begrip van cellulaire heterogeniteit te verbeteren.