# Algorithms for structural variant detection

Lin, J.

## Nederlandse samenvatting

Structurele varianten (SV's) vormen eigenlijk de verborgen architectuur van het menselijk genoom, en zijn van cruciaal belang voor ons om ziektes en evolutie te begrijpen. De ontwikkeling van sequencing-technologie en algoritmen maakt de detectie van SV's mogelijk, maar complexe SV's worden soms verkeerd geïnterpreteerd of zelfs over het hoofd gezien. Het ontdekken en karakteriseren van complexe events is een vakgebied dat meerdere disciplines omvat, waaronder domeinkennis en gespecialiseerde algoritmen.

In dit proefschrift introduceren we nieuwe algoritmen om complexe events te detecteren en te valideren, en om de reproduceerbaarheid van huidige SV-detectie pipelines voor klinische toepassingen en onderzoek te beoordelen.

Hoofdstuk 1 begint met de introductie van DNA, verschillende soorten SV's en sequencing-technologie. Vervolgens worden fundamentele technieken en algoritmen uit de informatica die verband houden met het proefschrift kort beschreven. Pattern mining, grafen en deep learning worden toegepast om verschillende soorten complexe SV's (CSV's) te detecteren en te karakteriseren. CSV's bevatten meestal meerdere breakpoints en worden vaak over het hoofd gezien of verkeerd geïnterpreteerd door traditionele strategieën die zijn ontwikkeld voor eenvoudige SV-detectie. Het belangrijkste is dat CSV's grotendeels onderbelicht zijn, waardoor ze zelfs moeilijk te detecteren zijn op basis van bestaande kennis. Rekening houdend met de sequencing-kosten en detectienauwkeurigheid voor verschillende scenario's, hebben we eerst algoritmen ontwikkeld voor zowel short-read als long-read sequencing-technologie zonder patroonovereenkomst ten opzicht van een database met bekende SV-structuren. Momenteel is short-read sequencing aanzienlijk goedkoper en wordt het op grote schaal toegepast in klinische diagnostiek en cohortstudies. Om CSV's via short-read sequencing te detecteren, zijn we van mening dat SV's de verbindingen van aangrenzende segmenten veranderen door middel van een alternatieve verbinding, afgeleid van abnormaal uitgelijnde reads met gepaarde uiteinden.

Daarom stellen we in Hoofdstuk 2 een aanpak (Mako) voor die gebruik maakt van een frequente maximale subgraaf, gebaseerd op abnormale alignments, om zowel SV's als CSV's te detecteren. Deze graaf wordt signal-graaf genoemd, waarbij knopen posities van verbonden genoom-segmenten vertegenwoordigen en takken alternatieve en referentieverbindingen tussen genoomsegmenten aangeven. Vervolgens hebben we een gelineariseerde database met prefix-indexschema toegepast om efficiënt frequente maximale subgrafen in de signal-graaf op te sporen, waaruit SV's en CSV's werden afgeleid uit gedetecteerde subgrafen. In vergelijking met andere benaderingen is een

graaf in staat om complexe genoom-segmentverbindingen weer te geven die afkomstig zijn van CSV's. Bovendien zijn gedetecteerde CSV-subgrafen interpreteerbaar, waardoor het mogelijk wordt om verschillende soorten CSV's beter te begrijpen en te vergelijken. Echter, beperkt door de leeslengte van short-read sequencing, kunnen twee eenvoudige SV's van verschillende haplotypes worden gedetecteerd als een enkele CSV-gebeurtenis. Aan de andere kant zou een korte leeslengte (read length) ook problematisch zijn voor toewijzing in gebieden met potentiële CSV's, waar breakpoints die bij een CSV horen, mogelijk door "callers" zouden kunnen worden gemist. Met de vooruitgang in long-read sequencing, is de kans groter dat een enkele read een hele CSV-gebeurtenis omvat in vergelijking met short-read sequencing. Dit vereenvoudigt de bevestiging van CSV's aanzienlijk door het verschil en de gelijkenis van de read en de corresponderende sequentie op het referentiegenoom te onderzoeken. Als gevolg hiervan is een toenemend aantal CSV's ontdekt door middel van intensieve breakpoint-analyse en visuele bevestiging. Dit is echter alleen van toepassing op kleine hoeveelheden steekproeven, die niet voldoen aan de steeds toenemende vraag naar het bestuderen van CSV's op populatieschaal.

In Hoofdstuk 3 hebben we gebruik gemaakt van menselijke intelligentie voor het identificeren van CSV's op basis van visualisatie, en hebben we een raamwerk voor herkenning van meerdere objecten (SVision) ontwikkeld om zowel SV's als CSV's te detecteren zonder voorafgaande kennis van SV-structuren. We stellen eerst een sequentie-naar-beeld coderingsschema voor, dat niet alleen de verschillen en overeenkomsten van twee sequenties beschrijft, maar ook de context van de achtergrondsequentie verwijdert. Deze coderingsstrategie stelt ons in staat om CSV's efficiënt en effectief te detecteren, zelfs in complexe genoom-gebieden. Verder is de CSV-representatie of -interpretatie een ander uitdagend probleem dat de definitie en studie van CSV's belemmert. Geïnspireerd door de graafstructuur die in Hoofdstuk 2 wordt gebruikt, hebben we ook een graaf gebruikt om CSV's die zijn gedetecteerd uit long-read gegevens weer te geven en te vergelijken, van waaruit we verschillende typen CSV's kunnen classificeren door graafisomorfismen te benutten. Maar anders dan een knoop in de signal-graaf voorgesteld in Hoofdstuk 2, vertegenwoordigt een knoop van de CSV-graaf in Hoofdstuk 3 een gematchte deelsequentie tussen twee sequenties. Deze functie maakt het mogelijk om CSV's te genotyperen op basis van alignment van de graaf. Bovendien biedt dit een nieuw idee voor het detecteren van SV's uit een SV-graaf in plaats van het detecteren van vertekening in de referentie. We verwachten dat deze op grafen gebaseerde SV-detectiebenadering zal helpen om somatische SV's en SV's van tumorsubklonen te detecteren.

Nadat we twee SV-detectiealgoritmen voor trending sequencing-technologie hebben ontwikkeld, willen we vervolgens de mogelijkheden van het gebruik van long-read sequencing in verschillende toepassingen verder onderzoeken. Over het algemeen zijn we van mening dat de meest betrouwbare SV's die zijn gedetecteerd via reproduceerbare analyse-pipelines van cruciaal belang zijn voor long-read toepassingen in klinische of onderzoeksomgevingen. Daarom hebben we in Hoofdstuk 4 eerst een high-throughput SV-validatieaanpak (SpotSV) ontwikkeld om de meest betrouwbare SV's te identificeren. Anders dan SV-detectie, richt SV-validatie zich op het uitsluiten van false negatives en corrigeert onnauwkeurige SV-karakteriseringen, zoals type en breakpoints. Het idee van deze validatieaanpak is ook geinspireerd op de manier waarop menselijke experts SV's visueel karakteriseren. We hebben eerst een eenvoudige lokale herschikkingsmethode toegepast om verschillende segmenten tussen twee sequenties te lokaliseren. Vervolgens hebben we een eenvoudige tweedimensionale berekening gebruikt om de betrouwbaarheid van een gedetecteerde SV te meten.

Daarnaast hebben we in Hoofdstuk 5 de reproduceerbaarheid van bestaande pipelines voor het detecteren van kiembaan SV's en somatische SV's beoordeeld. Dit hoofdstuk onderzoekt systematisch het verschil tussen op assembly gebaseerde en op alignment gebaseerde SV-detectie, waarbij de belangrijkste factoren voor tegenstrijdige ontdekkingen werden benadrukt. We verwachten dat deze evaluatie niet-experts zal helpen om het verschil in methoden te begrijpen en hen zo in staat zal stellen om de juiste analysepipelines in hun eigen toepassingen te selecteren.

Tot slot, in Hoofdstuk 6, beschrijven we toekomstige onderzoeksrichtingen met betrekking tot de nauwkeurige detectie van SV's voor zowel onderzoeks- als klinische instellingen. We zijn er met name van overtuigd dat de combinatie van BioTech en InfoTech, ook wel BT-IT genoemd, een revolutie teweeg zal brengen in de toekomstige gezondheidszorg.