Improved Error Correction of NGS Data

16 de mayo de 2016

El treball realitzat en el marc d'aquesta tesi doctoral se centra en la correcció d'errors en dades provinents de tècniques de seqüenciació massiva (també anomenades de nova generació, Next Generation Sequencing o NGS) utilitzant tècniques de computació intensiva.

A causa de la reducció de costos i l'increment en les prestacions dels seqüenciadors, així com en els avenços en les ciències mèdiques i biològiques, la quantitat de dades disponibles a NGS s'ha incrementat notablement. La utilització de computadors en l'anàlisi d'aquestes mostres es fa imprescindible per poder donar resposta a l'allau d'informació generada per aquestes tècniques. L'ús de NGS transcendeix la investigació amb nombrosos exemples d'ús clínic i agronòmic (com per exemple la detecció de mutacions en tumors cancerosos), per la qual cosa apareixen noves necessitats quant al temps de procés i la fiabilitat dels resultats.

Per a maximitzar la seua aplicabilitat clínica, les tècniques de procés de dades de NGS han d'accelerar-se i produir dades més precises. En este context és en el que les tècniques de comptuación intensiva juguen un paper rellevant. En l'actualitat, és comú disposar de computadors amb diversos nuclis de procés i inclús utilitzar múltiples computadors per mitjà de tècniques de computació paral·lela distribuïda, inclús fora de l'ús científic. Les tendències actuals cap a arquitectures amb un nombre més gran de nuclis (many-core) posen de manifest que és esta una aproximació rellevant.

Aquesta tesi comença amb una anàlisi dels problemes fonamentals del procés de dades en NGS de forma general i adaptat per a la seua comprensió per una àmplia audiència, a través d'una exhaustiva revisió de l'estat de l'art en la correcció de dades de NGS. Esta revisió introduïx gradualment al lector en les tècniques de seqüenciació massiva, presentant problemes i aplicacions reals de les tècniques de NGS, destacant l'impacte d'esta tecnologia en ciència. D'este estudi es conclouen dos idees principals: La necessitat d'analitzar de forma adequada les característiques de les dades de NGS, atenent a l'enorme varietat intrínseca que tenen les diferents tècniques de seqüenciació massiva; i la necessitat de disposar d'una ferramenta versàtil, eficient i precisa per a la correcció d'errors, com a fase prèvia a qualsevol anàlisi genòmica.

En el context de l'anàlisi de dades, la tesi presenta MuffinInfo. La ferramenta MuffinInfo és una aplicació programari implementada per mitjà de HTML5 per a afavorir la seua portabilitat tant a nivell de sistema operatiu com de

dispositiu. MuffinInfo obté informació rellevant de dades crues de NGS per a afavorir l'enteniment de les seues característiques i l'aplicació de tècniques de correcció d'errors, suportant a més l'extensió per mitjà de funcions que implementen estadístics definits per l'usuari. MuffinInfo emmagatzema els resultats del procés en fitxers JSON que faciliten la seua integració en pipelines de procés. A l'usar HTML5, MuffinInfo pot funcionar en gairebé qualsevol entorn maquinari i programari, donada l'ampli suport que té esta tecnologia. La ferramenta està implementada aprofitant múltiples fils d'execució i gestionant de forma concurrent l'accés a disc i la gestió de l'interfície.

La segona conclusió de l'anàlisi de l'estat de l'art ens porta a l'oportunitat d'aplicar de forma extensiva tècniques de computació d'altes prestacions en la correcció d'errors per a desenrotllar una ferramenta que suport múltiples tecnologies (Illumina, Roche 454, Ió Torrent i experimentalment PacBio). La ferramenta proposada (MuffinEC), suporta diferents tipus d'errors (substitucions, delecciones, insercions i valors desconeguts). MuffinEC supera els resultats obtinguts per les ferramentes existents en este àmbit, en els tres tipus de tests utilitzats en la tesi. Oferix una millor taxa de correcció, en un temps molt inferior i utilitzant menys recursos, la qual cosa facilita a més la seua aplicació en mostres més gran en computadors convencionals, on altres ferramentes no poden funcionar per problemes de recursos. MuffinEC utilitza una aproximació basada en etapes multiples. Primer agrupa totes les seqüències utilitzant la mètrica dels k-mers. En segon lloc realitza un refinament dels grups per mitjà de l'alineament amb Smith-Waterman, generant contigs resultat de l'alineament múltiple de les seqüències compatibles en el grup. Estos contigs resulten de la correcció per columnes d'atenent a la freqüència individual de cada base i l'aplicació de diferents fórmules i tècniques que faciliten discriminar errors de variants significatives.

La tesi s'estructura per capítols la base de la qual ha sigut prèviament publicada en revistes indexades en posicions destacades de l'índex del Journal of Citation Reports i en congressos de prestigi.