Toward highly accurate and fast variant and de novo mutation identification from high-throughput sequencing data by joint Bayesian family calling

Francisco M. De La Vega, Mehul Rathod, Richard Littin, Len Trigg, John G. Cleary ✉
Real Time Genomics Inc., San Bruno, United States

Motivation and Objectives

Whole-genome sequencing (WGS) has become a fundamental tool in human disease research and is being adopted in clinical settings at an unprecedented rate. Whole-genome and exome sequencing has been successful in the elucidation of highly penetrant genes in early childhood diseases and its making inroads in complex trait studies entailing thousands of samples. As WGS becomes faster and moves into the clinic, e.g. into neonatal ICUs (Saunders *et al.*, 2012) and in prenatal screening (Talkowski *et al.*, 2012), there is an unmet need for both speed and accuracy in the analysis workflow. Due to its shotgun nature, mis-mapping of short reads in complex genomic regions and high sequencing error rates, calling variants from human high-throughput sequencing (HTS) data still results in substantial false positives and false negatives (Ajay *et al.*, 2011). The problem is magnified when looking for *de novo* mutations in affected offspring of families, as this enriches for sequencing artifacts (Veltman and Brunner, 2012). This is problematic since *de novo* mutations are thought to be responsible for about half of all early neurodevelopmental childhood disorders (Veltman and Brunner, 2012) and likely a similar fraction of neonatal/prenatal cases (Saunders *et al.*, 2012; Talkowski *et al.*, 2012).

Methods

In order to alleviate these problems, we developed a joint Bayesian calling framework which calls variants simultaneously across a pedigree leveraging shared haplotypes in its members and incorporating a Mendelian segregation model, to produce accurate variant and *de novo* mutation calls from HTS data. We present how our Bayesian framework escapes combinatorial explosion (as compared to more simplistic approaches), is highly scalable to large pedigrees, can deal with low coverage and missing data, and can call *de novo* mutations if desired (Conrad *et al.*, 2011). Coupled with our fast alignment method, a family of three 40X whole genomes can collectively be analyzed from reads to variant calls in ~30 hours on a single commodity server, and is amenable to large-scale parallelization for further speed improvements. To validate our method, we analyzed WGS data from a 3-generation CEPH family of 17 members produced by Illumina Inc. as part of their "Platinum Genomes" resource(). Each genome was sequenced with the HiSeq® 2500 system to 40X average depth using 2x100bp libraries of ~350bp insert size. We aligned reads and performed calls in 3 nuclear family subsets and the entire pedigree for comparison.

Results and Discussion

We focus our analysis on NA12878, a female in the second generation, for which extensive orthogonal validation data exists including fosmid-end Sanger sequence data (Kidd *et al.*, 2008), Complete Genomics WGS data, OMNI SNP-array genotype data (Consortium *et al.*, 2013) and experimentally validated germline and cell-line somatic *de novo* mutation data (Conrad *et al.*, 2011). As compared to naïve singleton calling, our family caller produced more high quality SNV/indel/MNP calls and eliminates low quality calls, as judged by commonly used quality metrics such as Ti/Tv, Het/Hom ratios, and dbSNP/OMNI array concordance. All this with a low 2.5% FP rate as assessed by variants called at monomorphic sites in the OMNI array (Consortium *et al.*, 2013); cf. Table 1, below.

Table 1. Summary statistics and quality metrics comparing singleton and family calling.

Quality metrics	SNVs	Indels/ MNP	Ti/Tv	Het/Hom	% dbSNP (r129)	OMNI TP	OMNI FP
Singleton calls	3573672	775857	2.05	1.66	89.5	98%	2.4%
Family calls	3469745	665964	2.1	1.59	89.2	98%	2.5%
Pedigree analysis	**Mendelian errors**	**de novo candidates**	**de novo segregants**	**de novo germline**	**Germline sensitivity (%)**	**de novo somatic**	**Somatic sensitivity (%)**
Singleton calls	101204	16902	14341	47	96%	878	92%
Family calls	8672	2667	295	47	96%	872	92%

As compared with the Conrad *et al* (Conrad *et al.*, 2011) validated *de novo* mutations set, we observed 96% and 92% sensitivity in detecting reported germline and *de novo* mutations, respectively (note that the cell line batch may be different and thus have different somatic mutations). While high sensitivity can be achieved by simply reporting variants that pass less stringent accuracy thresholds (and in so doing increasing substantially the number of variants that violate Mendelian segregation), our family calling achieves high sensitivity, delivering a 10X reduction in Mendelian errors from 101,204 to 8,672 (cf. Table 1). A further 10X reduction in Mendelian violations can be achieved without using the *de novo* priors, which would be appropriate when assuming inherited disease. Through the analysis of variant segregation to the third generation, we confirmed 99% of the Conrad *et al*(Conrad *et al.*, 2011) germline mutations (somatic variants do not segregate, as expected) and observed about ~250 new *de novo* mutation candidates, which is close to expectation (about 100 from previous studies (Conrad *et al.*, 2011)). Importantly, the high de novo sensitivity of 96% was achieved while reducing the number of candidate *de novo* mutations by greater than 6-fold, from 16,902 candidates to 2,667 *de novo* candidates, without using empirical filters (this is ongoing work we will report at the conference). Our results suggest that joint family calling produces more accurate calls than singleton calling and allows for the assessment of *de novo* mutation candidates with much less noise. We illustrate the impact of an improved call set in the downstream interpretation analysis of a simulated cased from the literature, and a real case from a cardio-pulmonary syndrome. We believe the analytical advances we present are crucial for the clinical adoption of genome and exome sequence data in family disease studies and beyond.

References

Ajay,S.S. *et al.* (2011) Accurate and comprehensive sequencing of personal genomes. *Genome Res.* **21**, 1498–1505.

Conrad,D.F. *et al.* (2011) Variation in genome-wide mutation rates within and between human families. *Nat. Genet.*, **43**, 712–714.

Consortium,T.1.G.P. *et al.* (2013) An integrated map of genetic variation from 1,092 human genomes. *Nature* **490**, 56–65.

Kidd,J.M. *et al.* (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64.

Platinum Genomes Platinum Genomes Illumina, Inc.

Saunders,C.J. *et al.* (2012) Rapid Whole-Genome Sequencing for Genetic Disease Diagnosis in Neonatal Intensive Care Units. *Sci Transl Med* **4**, 154ra135–154ra135.

Talkowski,M.E. *et al.* (2012) Clinical Diagnosis by Whole-Genome Sequencing of a Prenatal Sample. *N Engl J Med* **367**, 2226–2232.

Veltman,J.A. and Brunner,H.G. (2012) De novo mutations in human genetic disease. *Nat. Rev. Genet.* **13**, 565-575