Variant filtering and prioritization using the Ion AmpliSeq exome trio workflow in Ion Reporter Software

Key findings

- Ion Reporter™ Software enables simple push-button informatics, in as few as eight clicks, to allow any laboratory to perform pedigree analysis using the Ion AmpliSeq™ exome trio workflow
- A simplified, stepwise approach enables rapid assessment of important variants, with associated rich annotation and content, from trio exome sequencing data (Figure 1)
- Comparison to a leading genome center bioinformatics pipeline reveals a high degree of sensitivity and precision for the detection of rare *de novo* single nucleotide variants from trio exome data using the Ion Ampliseq exome trio workflow in Ion Reporter Software

Genetic disease research and the implications of recent human evolution

There is increasing evidence from large sequencing studies that the recent explosive growth in human populations is generating an excess of rare genetic variation that could contribute to disease risk [1,2,3]. In a large-scale whole exome sequencing (WES) study of 6,515 people, Fu et al. revealed that 81% of the 709,816 variants found in European Americans and 58% of the 643,128 variants observed in African Americans arose in the past 5,000 years [4]. Nearly all the variants were estimated to be either neutral or deleterious, and due to recent acquisition, were expected to be either rare (variants with minor allele frequency (MAF) of less than 0.5%) or private (variants restricted to a particular population, pedigree, or individual). Additionally, population differences were observed with increased deleterious variants in people

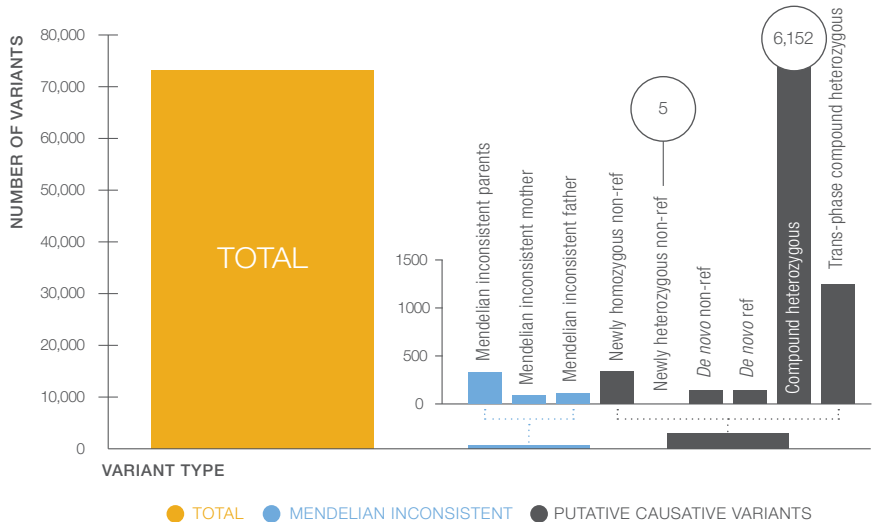

Figure 1. Number of variants identified per Genetic Category Type for NA12878 (female proband) using Ion Reporter Software analysis of pedigree trio data. Sequencing results are from the Ion Proton™ System using the Ion AmpliSeq Exome Kit.

ThermoFisher SCIENTIFIC

of European ancestry due to effective population size differences and weak purifying selection that occurred as a result of a bottleneck during the out-of-Africa dispersion. As a consequence of explosive population growth and evolutionary processes, interpreting the results of sequencing experiments and identifying causative variants from neutral variants is of paramount importance for our understanding of genetic disease.

Variation and genetic disease classification

Genetic disorders can result from variation in the nuclear or mitochondrial genome. Changes that may occur include single nucleotide variants (SNVs), small insertions or deletions (indels), and structural variants in which the anomalies could be a change in the number and structure of chromosomes; this includes copy number variants (CNVs) in which large regions (>1 kilobase (kb)) of the genome can be duplicated (gains) or deleted (losses). These variants can be inherited through the germ line either as preexisting variants or can be generated *de novo* prior to conception during gametogenesis. Post-zygotic *de novo* mutation events may occur during the early stages of embryogenesis and lead to genetic mosaicism in tissue and organs.

Genetic disorders are classified based on the number genes involved. Single-gene disorders are the result of mutations to a single gene and are typically rare (<1/10,000), whereas complex genetic disorders are more frequent (>1/100) but involves multiple genes as well as possible contributing effects from environmental factors. Evidence is emerging, primarily from exome sequencing studies, that *de novo* mutations might explain the heritability of complex genetic disorders [5]. Single-gene disorders can be further subdivided into autosomal and X-linked disorders based on the chromosomal location of the gene. Single-gene disorders predominately observe Mendelian inheritance patterns involving dominant or recessive alleles, although biological phenomena such as genomic imprinting can result in a departure from classical Mendelian inheritance. Additionally, genetic disorders can be the result of mutations to the mitochondrial genome, resulting in a maternal inheritance pattern due to the contribution of mitochondria by the egg cell to the zygote.

Exome sequencing on the Ion Proton System

WES on the Ion Proton System was successfully used to detect heterozygous *de novo* mutations in *STAT1*, a genetic defect common to a group of unrelated subjects with a progressive form of combined immunodeficiency [6].

Similarly, linkage analysis and homozygosity mapping were used to focus exome sequencing results obtained using the Ion Proton System to identify the genetic defect in a family with a congenital disorder of glycosylation [7]. In addition, exome sequencing on the Ion Proton System has been used to orthogonally validate somatic and germ line variants detected by other sequencing platforms [8,9,10].

The power of trio analysis

It is critical that variants of interest can be quickly identified and prioritized while common variants and nondeleterious variants (those predicted to have no effect on protein function or expression) can be reliably filtered from sequencing experiments. Trio analysis is a commonly employed method for identifying causative variants in the affected offspring (proband) of two unaffected parents. Trio analysis classifies the inheritance patterns of variants to identify SNVs and indels that display Mendelian-consistent and Mendelian-inconsistent inheritance patterns. Mendelian-inconsistent

Table 1. Variants identified in NA12878 by Ion Reporter Software analysis of trio exome sequencing results.

Total variants	Ion AmpliSeq exome trio workflow category type SNVs	Newly homozygous SNVs	Trans-compound heterozygote SNVs
73,232	1,466	315	1,227
***De novo* SNVs (MAF ≤0.002 and *P* value ≤0.01)**			
Validated true positives (n = 16) [11]	False negatives	True positives confirmed by additional Sanger sequencing	Putative false positives
15	1	9	11

variants are filtered to identify potential rare *de novo* mutations in the proband that are not present in either parent. Coverage thresholds and other criteria are additionally applied to reduce the number of false positive calls.

Variants in the proband that are consistent with Mendelian inheritance patterns are filtered to identify the acquisition of autosomal recessive alleles—when compared to a reference sequence—including compound heterozygous variants. Using parental genotype information, compound heterozygous variants (in which two heterogeneous recessive alleles in a particular gene are inherited) are further classified as cis-compound heterozygotes (cis-phase affects only one gene copy with inheritance of a wild type homolog as well) and the potentially more informative trans-phase heterozygotes, in which both copies of a gene are affected. Trio analysis can identify newly homozygous variants in the proband through inheritance of a genotype from both of the parents who are heterozygous at that position. The classes of newly homozygous and trans-phase compound heterozygotes can indicate a recessive disorder or condition since the parents can carry the same deleterious mutation but are unaffected because they are heterozygous with a benign allele. In a male proband, Mendelian-consistent maternal inheritance is used to identify X–linked recessive variants of interest.

Experimental design

A WES trio dataset was used to assess the performance of the Ion AmpliSeq exome trio workflow in Ion Reporter Software. A HapMap family trio of European ancestry (CEU samples NA12878 (child), NA12891 (father), and NA12892 (mother)) were sequenced on the Ion Proton System using the Ion AmpliSeq Exome Kit. The trio data for NA12878 is available at: http://ioncommunity.lifetechnologies.com/community/datasets

After multiplexing three samples on an Ion PI™ Chip v2, subsequent primary data analysis resulting in unmapped reads was performed using Torrent Suite™ Software v4.0.2 and then imported into Ion Reporter Software v4.0. The output of the Ion Reporter AmpliSeq exome trio workflow was compared to the results from the same trio that were extensively validated by whole genome sequencing (WGS) at the Wellcome Trust Sanger Institute during the pilot phase of the 1000 Genomes Project [11]. Using CEU trio data, this application note demonstrates a suggested method to prioritize variant discovery using the Ion AmpliSeq exome trio workflow in Ion Reporter Software.

Identify, prioritize, and report the most biologically interesting variants

Ion Reporter Software comprises a suite of bioinformatics tools that streamline and simplify analysis of semiconductor sequencing data. Data generated on the Ion Proton System are automatically uploaded from the Torrent Browser to a local or hosted version of Ion Reporter Software for read mapping, annotation, reporting of common and rare variants, and multi-sample analysis. Biological analysis and interpretation can be enhanced for cancer and genetic disease research through access to reference content from Oncomine™ and Ingenuity™ Variant Analysis™ applications, respectively. Further, data analysis can be automated with predefined workflows and shared with collaborators—all with version control and audit traceability. The identification of candidate *de novo* variants described below can be established in as few as 12 clicks using the predefined Ion AmpliSeq exome trio workflow with application of study-specific filters such as *P* value and minor allele frequency (MAF). Once set up, these study-specific filters can be used in future analyses with no additional setup time.

QC report

The QC report is part of an export package when all identified variants or filtered variants are downloaded from Ion Reporter Software. Reviewing the QC report summary metrics, such as average coverage, can be used to evaluate if the sequencing results are compliant with individual lab-determined standards for quality (Table 1, Supplementary information). Further, variant metrics can be reviewed to assess sequencing quality; these are likely to be sample- and panel-specific. For example, exome sequencing for a sample of western European ancestry would expect a transition/transversion (Ti/Tv) ratio in the range of ~2.5–3.1, a heterozygous/homozygous ratio for SNVs between 1.25–1.7, and a Single Nucleotide Polymorphism Database (dbSNP) concordance of >95%.

Hierarchical variant investigation

There are four categories that benefit from a methodical approach to readily identify and filter for variants of interest from trio sequencing data: (1) *de novo*, (2) trans-phase compound heterozygotes, (3) newly homozygous, and (4) male maternal X-linked. Although NA12878 represents a female proband that is unaffected, this sample can be used to illustrate the power of variant filtering in the AmpliSeq exome trio workflow in Ion Reporter Software. Through selection of the Genetic Category Type when creating a filter chain, a total of 73,232 variants identified in NA12878 were focused by the individual selection of *de novo* (non reference), trans-phase compound heterozygote, and newly homozygous (non reference) category types resulting in a significant reduction to 141, 1,248, and 324 variants, respectively (Figure 1 and Figure 2A). Note that through simple Genetic Category Type filtering in Ion Reporter Software, the ~70,000 variants identified by exome sequencing can be rapidly focused to about ten to thousands of putative causative variants in a proband (Figure 2A and Table 1).

To further focus on variants of significance, prioritization or filtering can be performed using additional Create Filter Chain options such as: (1) *P* value <0.01, (2) variant effect, (3) absence in dbSNP, (4) PhyloP score, and (5) OMIM™ (Online Mendelian Inheritance in Man™) database. Filtering on *P* value helps to remove putative false positives while filtering to eliminate previously identified and likely nondeleterious variants

A.

B.

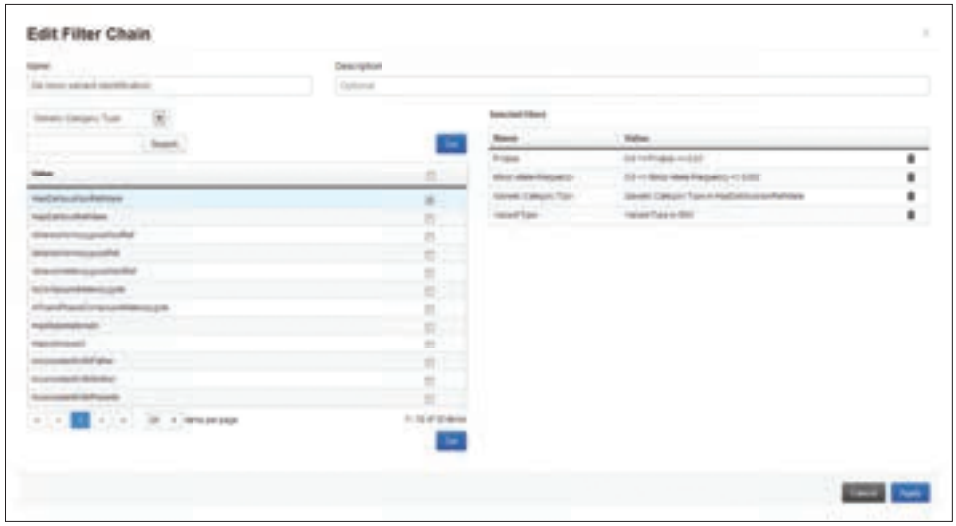

Figure 2. Variant filtering in Ion Reporter Software. (A) Applying the filter chain to the proband variant results rapidly reduces the number of variants for subsequent investigation. In this example, *de novo* (nonreference) SNVs were filtered to remove variants with a minor allele frequency >0.002 and *P* value >0.01. **(B)** Creating a filter chain to filter and prioritize variants can be easily accomplished using the drop-down menu options in the Edit Filter Chain dialog window.

present in dbSNP aides in variant prioritization. Highlighting variants based on variant effect (function) selects for important variants that putatively disrupt gene function such as nonsense, missense, frameshift, and splice junction alterations. Similarly, variant prioritization can be leveraged from the bioinformatics tool PolyPhen, a method that predicts and scores the likely effect of missense or nonsynonymous variants, while annotation sources such as the OMIM database can be used to determine if the variant has associated annotation to a disease.

To investigate *de novo* variants in NA12878, additional filtering criteria were added to the filter chain to identify 35 SNVs for subsequent investigation (Table 1 and Figure 2A). This was accomplished by the creation of a filter chain that removes SNVs with a minor allele frequency (MAF) >0.002 and *P* value >0.01 (Figure 2B). In the case of an affected proband, further prioritization would benefit by selecting those variants that putatively effect function (e.g., nonsense, missense, frameshift, and splice junction disruption), high PhyloP scores, and OMIM annotation. Following application of a filter chain, a focused look at variants of interest is recommended. Reviewing the locus and checking the variant call via Integrative Genomics Viewer (IGV) and evaluation of variant detail such as read depth are possible using Ion Reporter Software (Figure 3).

Stringent filters for the detection of candidate *de novo* variants

Ion Reporter Software analysis of exome sequencing data from the Ion Proton System using the Ion AmpliSeq Exome Kit were compared to the output of a bioinformatics pipeline developed by the Wellcome Trust Sanger Institute [11].The Sanger Institute bioinformatics approach utilized three algorithms to identify *de novo* mutations from WGS data. To facilitate a comparison with the results from Conrad et al. [11], analogous filters were implemented in Ion Reporter Software to filter for common SNVs and problematic genomic regions for variant calling—such as repeats, segmental duplications, and CNVs—to improve detection

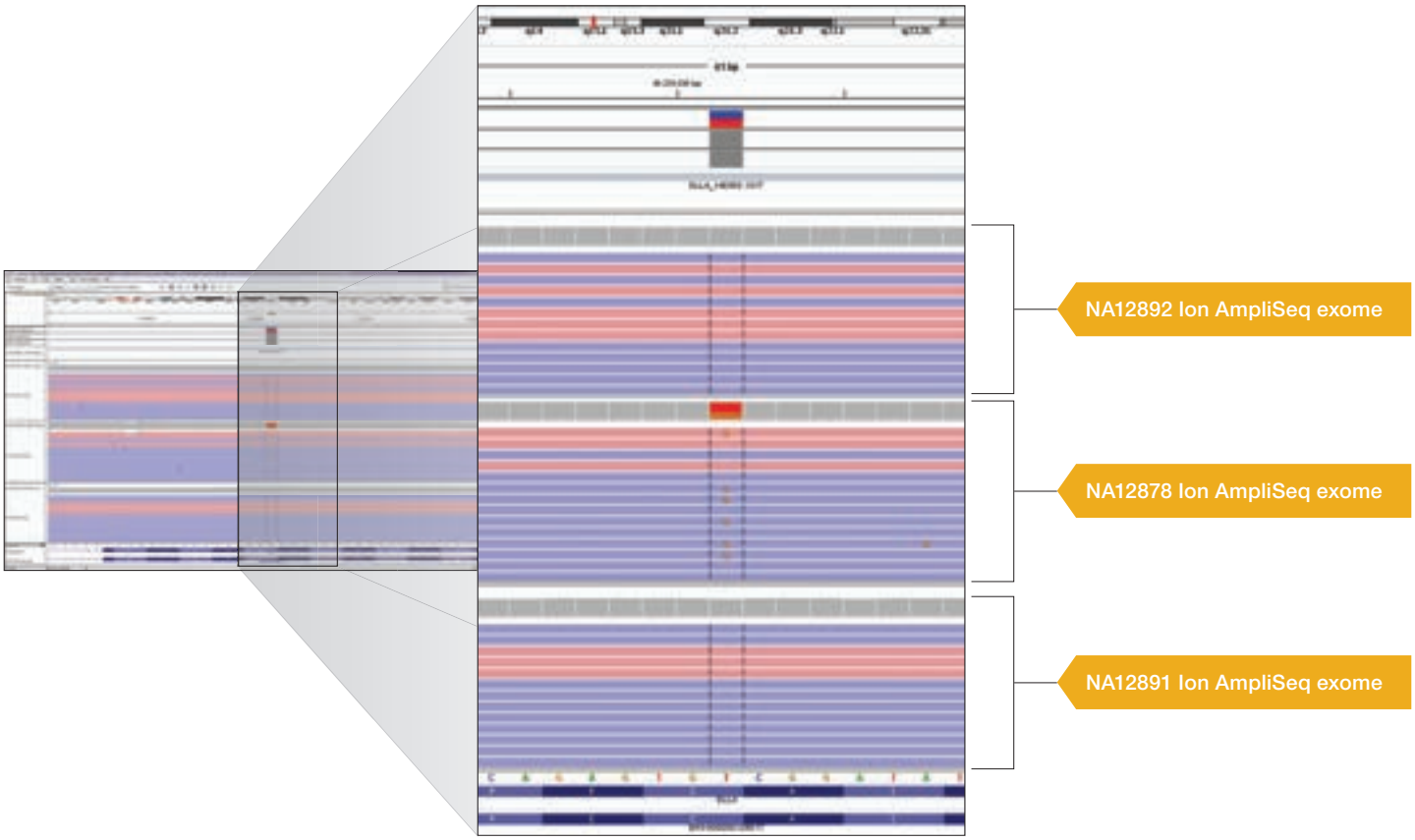

Figure 3. Focused variant investigation of a candidate *de novo* variant. Review of a nonsynonymous SNV (chr15:41229631 T>G, Trp>Cys) in the *DLL4* gene via Integrative Genomics Viewer (IGV) of CEU trio NA12878 (proband), NA12891 (father), and NA12892 (mother).

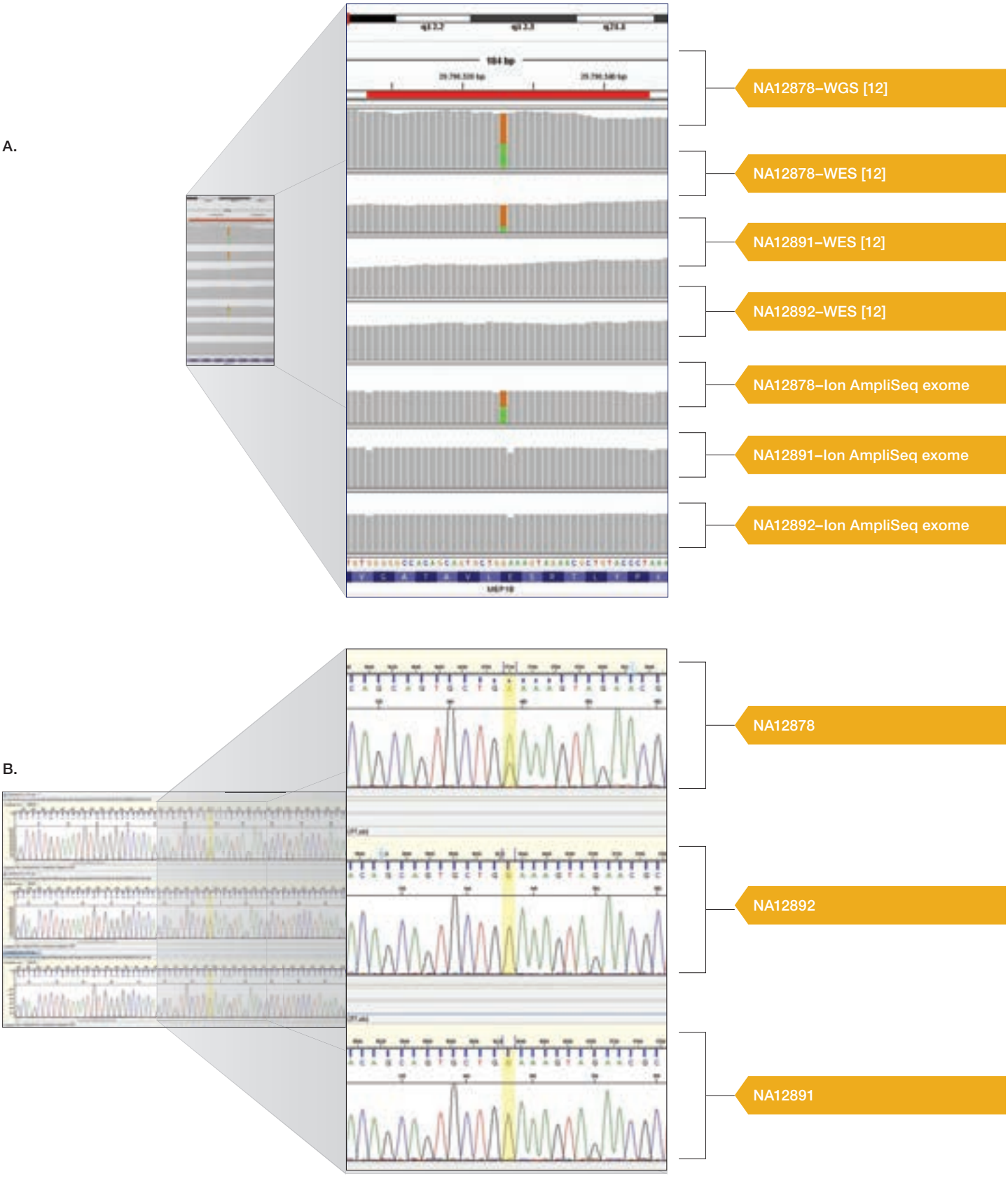

Figure 4. Novel candidate *de novo* SNV in NA12878 detected on the Ion Proton System using the Ion AmpliSeq Exome Kit. (A) Review of a nonsynonymous SNV (chr18:29790526 G>A, Glu>Lys) in the *MEP1B* gene via Integrative Genomics Viewer (IGV) of CEU trio NA12878 (proband), NA12891 (father), and NA12892 (mother) with comparison to independent WGS and WES data [12]. **(B)** Sanger sequencing electropherogram confirms the presence of a *de novo* variant in NA12878.

Table 2. Properties of the 10 additional *de novo* SNVs identified in NA12878 by Ion Reporter Software analysis of trio exome sequencing results.

Position (hg19)	Sanger confirmation	Variant type	Gene	Available annotation
chr1:205210827	Yes	Synonymous CDS C>T SNV	*TMCC2*	rs77588763, submitted by SeattleSeq in dbSNP 131, allele freq = 0% in 1KG
chr3:21462971	Yes	Intronic C>A SNV, 32 bp from CDS	*ZNF385D*	No annotations
chr3:38045851	Yes	Synonymous CDS G>A SNV	*VILL*	rs75449732 submitted by SeattleSeq in dbSNP 131, no allele info
chr3:112251437	Yes	3' UTR C>A SNV	*ATG3*	rs116110015 submitted by 1KG in dbSNP 132, no allele freq info
chr5:120021916	No	Nonsynonymous CDS A>G SNV, Thr>Ala	*PRR16*	rs78018931, submitted by SeattleSeq in dbSNP 131
chr12:56825424	Yes	Intronic A>G SNV, 36 bp from CDS	*TIMELESS*	No annotations
chr14:104209000	Yes	Synonymous CDS C>T SNV	*PPP1R13B*	rs2295142 submitted by U. Tokyo in dbSNP 100, then by SeattleSeq in dbSNP 131, allele freq = 0% in 1KG
chr15:41229631	Yes	Nonsynonymous CDS T>G SNV, Trp>Cys	*DLL4*	No annotations
chr18:29790526	Yes	Nonsynonymous CDS G>A SNV, Glu>Lys	*MEP1B*	rs115830020 submitted by 1KG in dbSNP 132, allele freq = 0%
chrX:63444310	Yes	Nonsynonymous CDS C>T SNV, Val>Ile	*ASB12*	rs200654794 submitted by 1KG in dbSNP 132, allele frequency = 1/1658

of true *de novo* variants (Table 2, Supplementary information). For sample NA12878, a total of 16 *de novo* variants (both germ line and somatic) verified by Conrad et al. were within 58 Mb targeted by Ion AmpliSeq exome enrichment. Using the filter parameters of a MAF ≤0.002 and *P* value ≤0.01 readily eliminated SNVs that were not candidates to be interesting *de novo* variants, resulting in a total of 35 *de novo* SNVs. This includes identification of 15 (94%) of the validated *de novo* SNVs detected by Conrad et al. along with 10 additional interesting candidate novel *de novo* SNVs, and 10 putative false positive *de novo* SNVs (Tables 1 and 2 and Figure 2).

The majority of the 10 false positives could be eliminated through locus review using IGV. For example, two false positive SNVs could be attributed to local alignment issues due to small-scale repeated sequence (e.g., alanine/proline-rich) and one SNV was within 2 bp of an indel. Further, the false negative *de novo* SNV (chr7:10016012 C>T in the *AGFG2* gene) validated by Conrad et al. was present in sequence reads for NA12878 when reviewing the region using IGV but was not called due to default strand bias parameters.

Loci of the remaining 10 candidate *de novo* variants were inspected in NA12878 using WGS and WES data from Ramu et al. [12]. With all 10 of the *de novo* SNVs also present in these datasets, none of these candidates were among the initial 3,236 potential *de novo* mutations identified from WGS data by the Wellcome Trust Sanger Institute's bioinformatics approach detailed in Conrad et al. The 10 candidate *de novo* variants were chosen for Sanger sequencing with 9 (90%) of candidate variants confirmed by an orthogonal sequencing method (Figure 4 and Table 2). Unusually, the remaining candidate *de novo* SNV was supported by WGS and WES data from Ramu et al. (Figure 1A, Supplementary information). The presence of a *de novo* nonsynonymous A>G (Thr>Ala) SNV in the *PRR16* gene was detected at this position, but this variant was not detected by Sanger sequencing (Figure 1B, Supplementary information); however WGS and both Ion AmpliSeq exome and Ramu et al. WES results for this locus indicate a triallelic variant with similar allele frequencies, suggesting that a complex mutational event

occurred at this locus (possibly during cell line creation and culturing used to generate the DNA) that is below the detection limit of Sanger sequencing (Figure 1A, Supplementary information). The inability to Sanger-confirm this *de novo* SNV resulted in 11 putative false positives (Table 1). Thus, the application of simple filtering criteria data using the Ion AmpliSeq exome trio workflow in Ion Reporter Software resulted in 96% sensitivity and a precision of 69% for the detection of rare *de novo* SNVs in sample NA12878 from Ion AmpliSeq Exome pedigree data.

Conclusions

Recent human history has resulted in the acquisition of the majority of genetic variation in the human genome, with most variation occurring at a low frequency. Additionally, the low-frequency variation observed in human populations is enriched in nonsynonymous changes, and as a consequence, could be potentially disease-causing. However, the functional elucidation of rare variants can be difficult due to population-specific effects, so the filtering and prioritization approaches used in Ion Reporter Software are critical for effective variant interpretation. Simple stepwise informatics available using the Ion AmpliSeq exome trio workflow in Ion Reporter Software enables any laboratory to perform trio analysis to rapidly identify important variants from exome sequencing data. Notably, for understanding complex genetic disorders, the detection of rare *de novo* mutations from pedigree exome data using the Ion AmpliSeq exome trio workflow in Ion Reporter Software compares favorably to bioinformatics approaches used by a leading genome center.

References

1. Keinan A, Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336:740–743.
2. Nelson MR, Wegmann D, Ehm MG et al. (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 337:100–104.
3. Tennessen JA, Bigham AW, O'Connor TD et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337:64–69.
4. Fu W, O'Connor TD, Jun G et al. (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493:216–220.
5. Veltman JA, Brunner HG (2012) *De novo* mutations in human genetic disease. *Nat Rev Genet* 13:565–575.
6. Sharfe N, Nahum A, Newell A et al. (2013) Fatal combined immunodeficiency associated with heterozygous mutation in STAT1. *J Allergy Clin Immunol* Epub ahead of print.
7. Monies DM, Al-Hindi HN, Al-Muhaizea MA et al. (2014) Clinical and pathological heterogeneity of a congenital disorder of glycosylation manifesting as a myasthenic/myopathic syndrome. *Neuromuscular Disord* Epub ahead of print.
8. Chong IY, Cunningham D, Barber LJ et al. (2013) The genomic landscape of oesophagogastric junctional adenocarcinoma. *J Pathol* Epub ahead of print.
9. Lupski JR, Gonzaga-Jauregui C, Yang Y et al. (2013) Exome sequencing resolves apparent incidental findings and reveals further complexity of SH3TC2 variant alleles causing Charcot-Marie-Tooth neuropathy. *Genome Med* 5:57.
10. Boland JF, Chung CC, Roberson D et al. (2013). The new sequencer on the block: comparison of Life Technology's Proton sequencer to an Illumina HiSeq for whole-exome sequencing. *Hum Genet* 132:1153–1163.
11. Conrad DF, Keebler JEM, DePristo MA et al. (2011) Variation in genome-wide mutation rates within and between human families. *Nat Genet* 43:712–714.
12. Ramu A, Noordam MJ, Schwartz RS et al. (2013). DeNovoGear: *de novo* indel and point mutation discovery and phasing. *Nat Methods* 10:985–987.

Find out more at **thermofisher.com/iontorrent**

ThermoFisher SCIENTIFIC

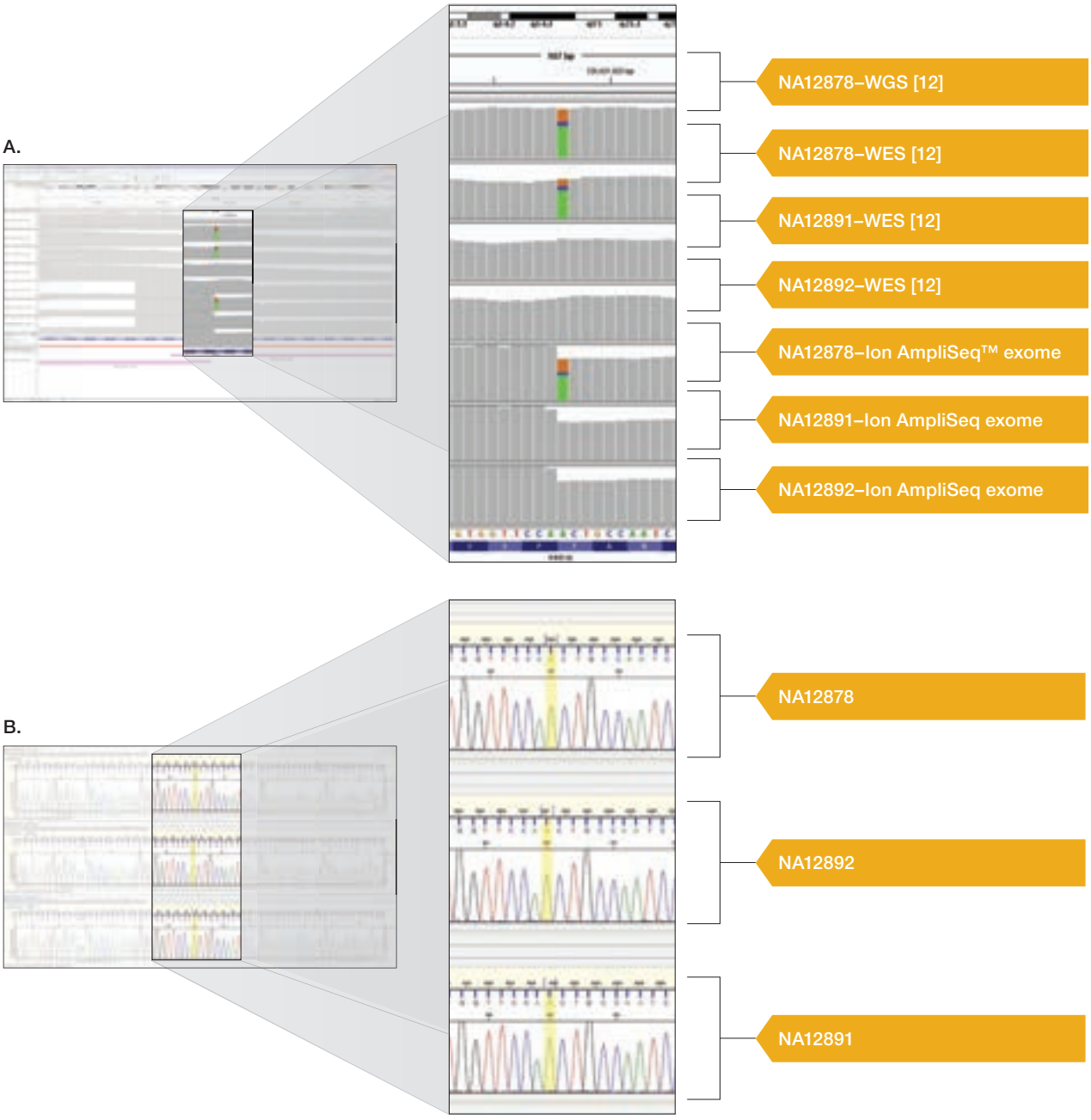

Figure 1. Focused variant investigation of a putative *de novo* variant in NA12878. (A) Review of a nonsynonymous SNV (chr5:120021916 A>G SNV, Thr>Ala) in the *PRR16* gene via Integrative Genomics Viewer (IGV) of CEU trio NA12878 (proband), NA12891 (father), and NA12892 (mother) with comparison to independent WGS and WES data [12]. **(B)** Sanger sequencing electropherogram does not confirm the presence of this *de novo* variant in NA12878.

Table 1. Ion Reporter Software QC metrics for CEU trio exome sequencing results from the Ion Proton System using the Ion AmpliSeq Exome Kit.

Metric	NA12878, daughter (proband)	NA12892, mother	NA12891, father
Total number of reads	29,182,512	30,242,667	29,956,931
Total number of bases (Mbp)	5,308.6354	5,525.1876	5,478.3422
Total number of bases (AQ20) (Mbp)	4,244.3844	4,458.9920	4415.6495,
Mean coverage depth (fold)	89.7383	93.4046	92.6057
Coverage within target region	98.6725	98.6982	98.8382
Mean read length (AQ20)	159.7818	161.9907	161.6471
Mean read length (AQ30)	120.7127	122.7163	122.7377
Number of homozygous SNVs	16,588	16,708	16,681
Number of homozygous INDELs	988	1,055	977
Number of heterozygous SNVs	28,433	28,220	28,251
Number of homozygous MNVs	4	6	7
Number of heterozygous MNVs	42	30	35
Number of heterozygous INDELs	1,728	1,622	1,739
Ti/Tv Ratio (SNPs)	2.523	2.542	2.501
dbSNP concordance	0.972	0.973	0.972
Heterozygotes/homozygotes	1.716	1.679	1.698
Indels/total	0.057	0.056	0.057
Indels/kb	0.047	0.046	0.047
SNPs/kb	0.780	0.778	0.778

Table 2. Details of the filters and excluded base pair (bp) sizes used for the identification of candidate *de novo* variants.

Filter	Wellcome Trust Sanger Institute [11]	Ion Reporter Software
Simple repeats	UCSC hg18	UCSC hg19
	53,240,703 bp	71,333,825 bp
Segmental duplications	UCSC hg18	UCSC hg19
	142,256,390 bp	159,499,260 bp
CNV	GSVC 42M Probe	DGV 1kG hg19
	157,719,857 bp	161,353,041 bp
dbSNP SNVs	dbSNP129 (excluding NA12878 private)	dbSNP138 (minor allele frequency >0.01)
	14,944,456 bp	12,032,300 bp
Read depth	0x coverage ≥1 member	Implemented in the default Ion AmpliSeq exome trio workflow
	191,658,532 bp	
Broad Institute GATK multiple realignment	Sites in which a heterozygous call becomes homozygous	Not performed
	1,278,665 bp	
Short indel calls	Pindel small indel call +/-100 bp	1kG Trio small indels +/-100 bp
	17,065,637 bp	166,665,073 bp
Total excluded base pairs	468,051,448 bp	488,682,600 bp

Find out more at **thermofisher.com/iontorrent**

ThermoFisher SCIENTIFIC