

From: [John West](#)
To: [Richard Chen](#)
Subject: Accuracy next steps, spreadsheet to organize the discussion
Date: Thursday, June 7, 2012 6:24:18 PM
Attachments: [Error mech Detect n Poss solutions JW 6Jun2012.xls](#)

Hi Rich,

I put this together earlier this week to begin organizing my thoughts on how we move towards accuracy solutions, somewhat in parallel with ongoing analysis. I would be happy to go through it when you have time.

John

MECHANISM OF ERROR GENERATION	CAN WE DETECT & COUNT EACH ERROR FROM THIS MECHANISM ? HOW ? ARE WE SET UP TO LOOK FOR THIS ?	HOW WE COULD ELIMINATE / REDUCE THIS ERROR TYPE		
		BIOINFORMATICALY		LAB METHODS
		ALGORITHMS	REFERENCE	

Mis-alignment due to:

Repeats / degeneracy	MIE due to false positive het SNP's, clustering in known degenerate regions Reproducibly abnormal coverage in known degenerate regions Ratio of reads at het loci far from 50/50 Tri-allelic in a single genome (fairly easy to see) Triple-haplotype (more work to make QC tool) Systematically different result from paired-end reads of different insert lengths	Chain-link SNP alleles by shared raw reads and PE-reads across a degen region Special focus on junction sequences Variable alignment stringency across the genome, tailored to known differences in degeneracy, likely as a 2nd stage re-alignment	Solve degenerate regions in "gold standard" genome by "spare no expense" Develop MAF data in degenerate regions to guide later alignment of others	Longer reads ILMN 2x250 Ion Proton Oxford Nanopore Moleculo Longer PE inserts Multiple PE insert lengths Custom pulldut for long insert paired ends or other assays
SNP/InDel cluster	Reproducible coverage drop around SNP clusters ? Ratio of reads at het loci far from 50/50 InDel MIE due to failure to detect InDel due to cluster Apparent failure of LD - Alleles at nearby loci don't "travel together" as expected	New version of GATK is better ? Use BreakSeq approach for SNP/InDel clusters Check for allele combinations which are unlikely from known LD	Do SNP clusters near 50% MAF split 49/51 in non-real combinations in our ethnically specific major allele refs ? Add InDels to our ref where major allele	Custom SNP array or pulldut incorporating the allele sets known to be in LD

Low raw sequence quality:

Raw Q-score depression downstream from specific sequence motifs	Directional Q-score average dips, particularly in standard locations	Motif-specific pattern recognition including length, not just substitutions		If GC or hairpin driven, use nucleotide analog kits
Clusters systematically too faint in certain sequence contexts	Low raw Q-scores near the start of a read, e.g. bases 2-20. We don't currently look for this, but could. We could confirm with access to fluorescence intensity data - an option on the system but not usually on	Look specifically for these & if they are reproducible. Should we exclude them ? What motifs cause them ?		Lower cluster density, longer imaging exposure times ? Different cluster growth, e.g. # cycles
Phasing worse than normal in certain sequence contexts	We could confirm with access to fluorescence intensity data - an option on the system but not usually on	Motif-specific algorithms		Motif-based sequencing primers
Pile-up of low raw Q-scores by random near-registration of the ends of reads, which are inherently low-Q	We don't currently look for this, but could May be a good refinement of the directional Q-score metric	Look for this and flag it (at least)		Some shorter reads, which are highly accurate all the way to the ends

Low coverage:

Leading to a lack of calls, missed heterozygosity, or missed SV The low coverage being due to:	Look for directional coverage reproducible dropouts	Look for where this is unusual, i.e. not seen in test genomes sequenced the same way		
Extremes of %GC	Need to find best scale and lateral offset for GC correlation with coverage	Need to understand mechanism of GC bias to know how to fix it		Nucleotide analogs Extreme GC pulldut(s)
Hairpin	JW spreadsheet-based model can find these but is too slow for genome scale. Also, not all large hairpins lead to low coverage - needs work	Refine model to understand correlation between hairpin properties and error occurrence, so we can think about how to fix the problem	Genome-wide database of hairpins, their properties & their impact	Nucleotide analogs Pullduts or PCR sets which have just one side of the hairpin
Poisson	The part of coverage variability which is not reproducible run to run on same library			Probably worst problem where combined with other sources low coverage, so fix those
Exome pulldut imperfection	Coverage vs position will show this	Maybe room to add SV detection to exomes, esp via BreakSeq		Supplementary pulldut

Failure to detect small InDel's, due to:

Algorithm weakness	InDel MIE's may identify some of these ? Can BAM crawler identify InDels in alignment of raw reads which were not called in the variant lists (possible false negatives) ?	New GATK may improve this quite a bit If BAM crawler can do this (see left) then local re-alignment may correct miss InDels should be in LD with nearby SNP's so we could look for those and use that	Ethnically-specific major InDel allele reference, on Hugo's plan.	Longer illumina reads should help: Minimal false positive but a much bigger % of the read will be usable
Difficult seq contexts Seq with InDel maps elsewhere InDel creates degeneracy InDel creates hairpin InDel near reference gap InDel close to, but not in, STR InDel close to, but not in, SV	Special case not currently handled Special case not currently handled Special case not currently handled Special case not currently handled Special case not currently handled	Could create dedicated algorithm if priority Could create dedicated algorithm if priority Could create dedicated algorithm if priority Could create dedicated algorithm if priority Could create dedicated algorithm if priority		See other hairpin options, above
Many reads too close to either end at location of InDel to be able to detect	Similar to issue of pile-up of low raw Q scores by random near-registration of ends of reads, but either end is a problem here.	Extend the InDel detection algorithms for loci with known high-allele-frequency InDels, to support detection even near either end of a read. Different from detection of novel InDels.		

STR errors due to:

Biochemistry fails	Shows up as a drop in coverage and/or a reproducible drop in raw Q-scores downstream	Failure is generally for very long STR's which are generally heterozygous. This may be detectable as an allelic read depth imbalance at nearby het SNP's		Detectable by electrophoresis Standard technology we could incorporate Long STR's can also be strong hairpins, so hairpin-reducing nucleotide analogs could help
Seq OK but algorithm fails, esp if het Worse (?) if both alleles don't match reference	Would this show up as a low alignment Q-score ? Would we notice a pile-up of those ? Should we add this to our BAM crawler ? STR MIE's, if we look for these	Since all STR loci should already be known we could write a dedicated algorithm for this. IckSTR may already be this. Mark has suggested trying it.	Current reference almost certainly does not have the most common # STR units at each position, let alone by ethnicity. This should be easy to change in the reference, though it requires mapping coordinate systems.	

Homopolymer errors due to:

Biochemistry fails	Shows up as a drop in coverage and/or a reproducible drop in raw Q-scores downstream	A simple dedicated algorithm could prioritize getting the sequence right downstream of the homopolymer, vs the length of the homopolymer, by working with reads which begin near the end of the		Dedicated sample prep which intentionally and reproducibly spikes different bases into these regions, to break them up and make them more sequenceable.
--------------------	--	---	--	---

		homopolymer, or end just into it. Almost all homopolymer loci in the genome should be known now, so we can use that info	Could use a targeted meganuclease. What priority / value ?
Seq OK but algorithm fails, esp if het Worse (?) if both alleles don't match reference	Would this show up as a low alignment Q-score ? Would we notice a pile-up of those ? Should we add this to our BAM crawler ? Homopolymer MIE's, if we look for these	Does lobSTR potentially help here ?	Current reference likely does not have the most common # homopolymer base units in all of these regions, let alone by ethnicity

Translocation errors:

False positive due to repeat seq	Translocation MIE's, if we look for them	Paper circulated fairly recently (by Gemma ?) showed success dealing with this by filtering out repeat regions & using long mate pairs	Supplement sequencing with long mate pairs
Missed reads which span junction	Translocation detected by paired-end algorithm but not split-read algorithm	HugeSeq has two algorithms which look for junction sequences & could filter out potential translocations for which it did not find them	
Robertsonian translocations missed Not detectable with next gen seq.			Supplement next gen seq with karyotyping

Larger (>50bp) deletion errors:

Zygosity not reported, or wrong	MIE in large deletion detection	Map out regions of systematically low coverage (in the absence of a deletion) Potentially useful in distinguishing deletions from normal coverage variation	Longer reads should be proportionately better at detecting junction sequences (with & w/o deletion, in case of het del). If we can put more weight on junction-based algorithms, we will be less subject to false positives from coverage variation.
Del detected but not junction seq so length & position inexact	HugeSeq's SV algorithm requires detection by at least two algorithms, but doesn't require junction seq detection Should it ?	Junction sequences are most likely to be missed where coverage is systematically low and a deletion is heterozygous. Also where deletion removes one unit of a multiple unit tandem repeat or STR because junction seq can be same with or without deletion. Potentially solvable with a dedicated algorithm if prioritized	Potential for pull-out in junction sequence regions, to increase ability to detect these known junctions
False positive from read-depth algorithm due to coverage variation caused by other mechanisms	MIE in large deletion detection, and HugeSeq requires two algorithms to detect SV's confidently	See above	

Larger insertion errors:

Haven't investigated enough to categorize	MIE's and reproducibility with multiple PE insert lengths	Haven't investigated enough to propose approaches	Can old standard genomes and 1,000 Genomes' SV data help produce a better ref sequence re insertions ?	Longer insert read-pairs might be quite helpful.
---	---	---	--	--

CNV errors:

Copy number wrong due to mis-alignment esp in repeats	CNV MIE's, if we look for these	Algorithms to improve alignment in degenerate & compression regions might also help with CNV's - another variation on the same type of problem	Gold standard genomes, esp using molecules 10-100kbp, might help solve some of this and create a framework on which to organize the population statistics
Total copy # OK, but heterozygosity wrong e.g. 4x as 2+2 vs 3+1	CNV MIE's, if we look for these		
Junction sequences of CNV's not detected	HugeSeq's SV algorithm requires detection by at least two algorithms, but doesn't require junction seq detection Should it ?		
Complex CNV's: differing, overlapping blocks Also re-arrangements, deletions, etc	MIE's and reproducibility with multiple PE insert lengths	More complex cases could likely be addressed with more sophisticated and dedicated algorithms. Development of these would benefit from interaction with an effort to build gold standard genomes using maximal technology.	

Reference sequence errors:

Incorporates rare SNP alleles, leading to mis-alignment and missed variant calls	Reference is not 1,000 Genomes major allele Done by Rick Dewey for PLoS paper	Update with newer 1,000 Genomes data ? Cross-check against NHLBI exome major alleles ? Call all interpretable loci regardless of variance vs reference sequence	
Incorporates rare InDel alleles, leading to mis-alignment	Hugo plans to look for InDels with >50% Alt Allele Frequency in 1,000 Genomes result he calculated	Addressing this is on Hugo's project plan (right ? When ?)	
Incorporates rare SV alleles, leading to SV inverse (e.g. Insert for Del)	Hugo plans to look for SV's with >50% Alt Allele Frequency in 1,000 Genomes result he calculated	Addressing this is on Hugo's project plan (right ? When ?)	
Compressions (large duplications not recognized in the reference)	All quartet family members het where inheritance state would not allow that Split of reads at het loci is far from 50/50 esp in very high coverage genome Blocks of elevated coverage seen identically in other ethnicities Loci tri-allelic in a single individual Three haplotypes in one region of a single genome	Make an effort to validate more compressions from Rick Dewey's list: Top 36 was not a limit - We just ran out of time manually validating from the original "top 50". Alignment algorithm proposed based on population data, supplemented by gold standard genomes	Reads of very long molecules would help. Not just mate pairs, but possibly very long insert clones, Molecule approach or Shendure approach or CGI "long fragment read" approach.
Gaps in the reference	These are well documented in the reference itself.	Look for reads which align at edge of gap and extend into the "unknown" to build our own extensions to narrow the gaps. Others have done some of this, which we might leverage. If we wanted to make significant progress here, it could be very hard and we would certainly want to understand what kept the original human genome project from closing these gaps. They certainly tried.	