

**From:** [John West](#)  
**To:** [Richard Chen](#); [Scott Kirk](#); [Mark Pratt](#); [Christian Haudenschild](#)  
**Subject:** Slide for accuracy requirements walkthrough meeting 3-5pm today  
**Date:** Tuesday, October 9, 2012 1:58:57 PM  
**Attachments:** [Accuracy mtg 9Oct2012 JW.pptx](#)

---

All,

I put together one slide which summarizes my understanding of where we are headed on accuracy and how that should tie in to the product specifications we have for upcoming versions. Perhaps we can start our 3-5pm meeting by discussing this framework ?

Thanks,

John

# JW Accuracy Overview

## CHARACTERIZING ERRORS

# by type, mechanism, medical application

### **Methods to identify errors:**

A / B repeatability or vs Gold Std

MIE (need SV, not just SNP)

Single genome methods:

Tri-allelic loci

Odd allele ratios at het loci

Low / no coverage (WGS & exomes)

Coverage variability sample to sample

Particularly for SV detection in exomes

### **Define regions linked to mechanisms**

Download from UCSC browser (mostly done)

Run alignment degeneracy alg on MiSeq :

Read lengths

Molecule lengths

Coverage statistics

To define low or variable coverage regions

### **Pareto analysis**

# errors by type, error mechanism

Focus this by medical application

### **Re-Do this with modern data**

v3 seq chemistry, gel & PCR-free WGS

v3 seq chem exomes

CEPH1463, Venter, Diversity panel

Data purchased, downloaded, seqc'd

## FIXING ERRORS

### **Bioinformatic approaches:**

Better reference (NEMAR, SV version, hg20)

Improve SV algorithms (many parts)

Upstream-only base call around homopolymers

LobSTR

Use of array data (Genotyping and SV)

Better alignment algs (3<sup>rd</sup> party & ours)

### **Lab & bioinformatic approaches**

Pullout & integration with regular data

Exome supplement

Long read / molecule (SD's, compressions)

High GC

Optimize operation for accuracy vs \$

Novel Assays (Pulsed field, phasing, etc)

### **Content development**

Data downloaded, purchased, seqc'd in-house

e.g. for BreakSeq, norms, GT ever seen ?

MAF for SV/CNV/InDel, ethnic controls

## FLAGGING ERRORS

### **Descriptors, not just Q-Scores:**

Why we think the Q is low

### **More accurate Q-Scores**

Difficult sell due to competitive noise

## PRODUCT SPEC / FEATURE

### **Reduced error rates**

Focused on medical content

Quantitative Personalis vs Standard

By method to ID errors

by variant type & mechanism:  
SNP

NEMAR hg20 vs Public hg19

Coverage filled in

Homopolymer region caller

Alignment degen w long reads

Optimize for accuracy vs \$

STR's, InDel's

SV's

Qualitative too:

e.g. SV zygosity & in exomes

e.g. STR analyzed at all

e.g. Reporting tri-allelic loci

### **More content covered**

Finished genes in exomes

Medical content outside exons

Whole genome isn't:

Degenerate alignment (e.g. HLA)

Low coverage

### **Flags which guide the type of validation / follow-up**