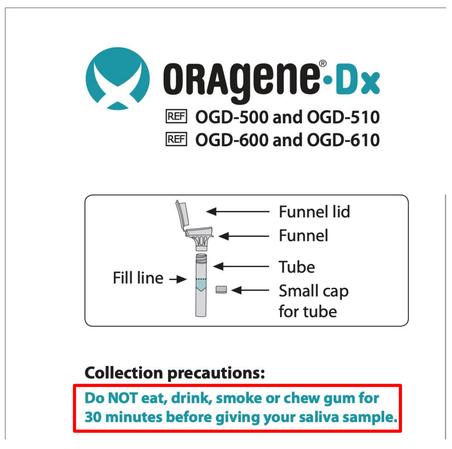
BACON:
Baited Abrogation of CONtamination

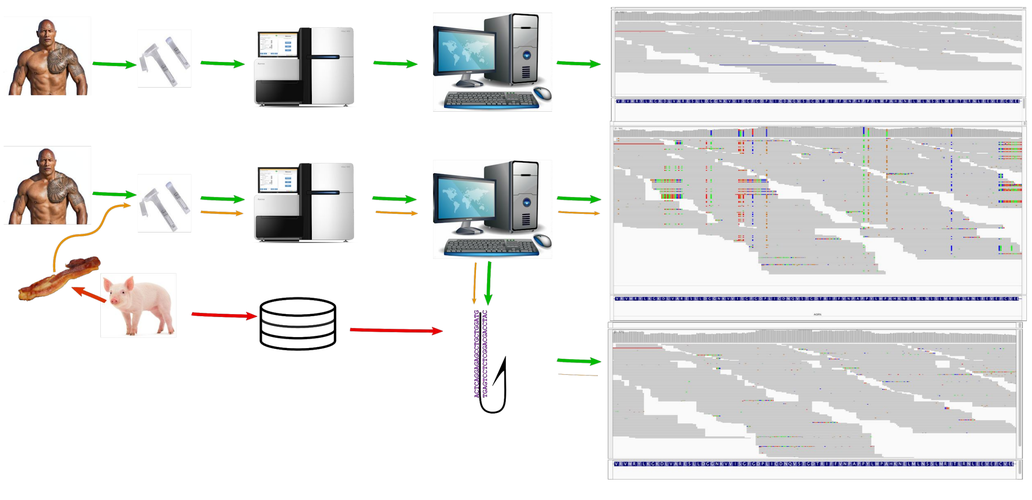
Fernando L. Mendez[1] Ruomu Jiang[1] Simon White[1] William Lee[1]

[1]Helix Opco LLC, San Mateo, CA

Introduction

- **Human DNA** is often collected from **saliva**
- **Saliva** may contain bacterial and **food DNA**
- **Food DNA** may be **confounded as human**
- **Variant calls** may be **polluted by food DNA**
- Polluting variants may look as **rare and pathogenic**





- **Pig sequences would predict for ACMG59 genes changes of HIGH impact[1]:**
  - **20 missense mutations in 8 genes**
  - **64 frameshifts in 12 genes**
  - **4 stop gains in 3 genes**

(including 24 mutations in BRCA2 and 6 mutations in TP53)

Methods

- Identify sequences in a number of animals that are used as food that could map to the human genome
- Filter out animal regions that could affect mapping of human reads
- Assess the impact of the filtering on genotype calls

Results

- Generated experimental DNA mixtures of NA12878 and pig samples and processed them through a standard pipeline.
- Generated genotype calls and found (Figure 1):
  - Number of calls due to contamination increases with animal DNA
  - Contaminating calls are preferentially located in or near exons (not shown)
  - A significant number of wrong calls predicted as having high impact
- Generated supplementary relevant animal sequences as decoys to **bait contaminating read sequences**.
- The addition of the decoys reduced the effect of contamination for all concentrations of pig DNA (Figure 1).

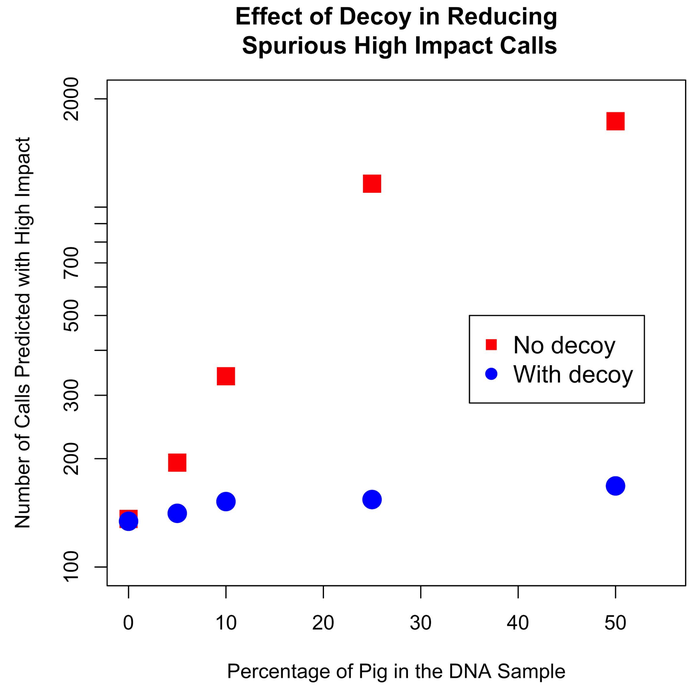


Figure1: The number of spurious calls grows with the percentage of animal contamination, including those predicted as having high impact (logarithmic scale).
The growth of the number spurious calls is controlled by adding sequence decoys to the reference.

Lowdown

- **Sampling saliva** simplifies collection of human DNA,
  but
- **Food contamination** poses a risk of wrong variant calling,
  and
- **Spurious calls** may lead to unnecessary additional testing,
  however
- **Bioinformatic** recognition and **removal of contaminating reads** significantly reduces the number of wrong variant calls and its negative consequences

[1]A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3., Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. Fly (Austin). 2012 Apr-Jun;6(2):80-92. PMID: 22728672

Visit Helix.com