**Mitochondrial genome insertion into D5 genome**

(To-do-task in red)

**Genomic and gene sequence analysis**

Genes Gorai.001G159900 - Gorai.001G168200 (68,300 bp) were annotated as putative Mitochondrial DNA-like sequences in the nucleus (NUMTs). How? When? Is it real? Are the conclusions reached about gene expression being maintained and even up-regulated real? See Paterson et al., 2012. A putative nuclear mitochondrial DNA (NUMT) sequence block24 has an intriguing relationship with fibre improvement. A G. raimondii chromosome 1 region includes many genes closely resembling mito-chondrial homologues (Ks , 0.22; Supplementary Table 4.7a). NUMT genes experienced a coordinated change in expression associated with G. barbadense domestication. The 105 (0.2%) genes upregulated in 10 DPA fibre of wild (versus elite) tetraploid G. barbadense (Supplemen-tary Table 5.3) include 30 (37%; P , 0.001) of the 81 NUMT genes, including 8 NADH dehydrogenase and 4 cytochrome-c-related genes. All are within the QTL hotspot Dt01 that affects fibre fineness, length, and uniformity22, suggesting a fibre-specific change in electron transfer in G. barbadense domestication.

Other evidence from synteny analysis? Is this same sequence in the other Gossypium de novo assemblies? If so, which ones? How about in Gossypioides or Kokia?

**MNase-seq analysis**

MNase-seq read mapping and iSeg analysis revealed a hypersensitive region of 1MB long in Ch01 (23-24MB, or minimal range of 23,162,848 to 23,811,737), which overlaps the putative mitochondrial insertion region. This single region is behind the unexpectedly high number of MSFs detected in Chr01.



According to Hank, this kind of hypersensitivity due to mitochondrial insertion in maize is very common. Recently inserted organelle genome sequences lacks the proper GC content or other sequence chemi-physical property to array nucleosomes, thereby remaining as open regions. This can be tested by checking the GC content and the likelihood for nucleosome organization (using prediction program, e.g. [this one](https://genie.weizmann.ac.il/software/nucleo_prediction.html) recommended by Daniel).

Why 1 MB? The insertion is not that big, right?

This observation alone **cannot** tell whether the mt insertion is real or is an assembly error, because the extracted and sequenced MNase-seq DNA (supposed to be nuclear) may still contain mt DNA, as organelle contamination cannot be completely avoided from the nuclei extraction protocol used. If this is the case, for some reason (not clear...but definitely not random because seen in my two reps ), mt DNA fragments are enriched in the light digestion sample, which also gave this hypersensitive peak on the mis-assembled chr01 region.

My favored explanation involves something that pins this contamination to Chromosome 1, even if it is a smaller NUMT, to which everything because stitched during assembly. I cannot think of hypothesis that would localize this to Chromosome 1 without it having some basis in reality.

**Differential gene expression analysis**

Two sets of DE analyses should help:

1. Mt vt nuc. MJ concluded that genes in this region were even upregulated relative to other regions. Huh? Check previous results.
2. D5 vs others. Using D5 reference and polycat, I can compare the the expression of these putative NUMT genes in A2, D5 and polyploids. If expression are generally high in D5, the insertion is likely to be real (mt contamination can be ruled out because all RNA-seq exaction should have similar level of contamination). If no difference between species, there are two explanations: a. mt insertion in D5 is still true, with nuclear copies not expression; b. mt insertion in D5 is an artifact. Use the eflen polycat dataset to check

**FISH**

Ideally, FISH experiment using the 1MB region as probe should tell whether the mt insertion is real. Can we ask Rick Masonbrink to do this?

**Bionano mapping and PacBio long read sequencing**

Datasets are available from Josh. Is this region present in long reads, or in there a consistent boundary flanking this region ?

From Josh email 12/11/17:

Here's the putative mito insertion:



I'm not convinced it is a real insertion. There is the piece Map #328 (blue coverage plot in middle) that aligns with the genome sequence (green bar at top), but that is the only part of 328 that matches to the genome. The rest of #328 does not have a match anywhere in the genome. Map #328 has the same amount of coverage as other Bionano contigs that surround it. It could be a couple things:

I'm guessing the mito region is actually two regions assembled on either side of an actual piece of the nuclear genome. Perhaps the piece of the genome is actually an old mito insertion that assembled with pieces of the real mito sequence.



a view showing the unmatched nick marks.



The flanking contigs in the first image were 2095 and 64. When I look at them in our PacBio assembly, both of the contigs have matches to ends of two different pacbio contigs. So no real answers there.