www.roche-applied-science.com

Assembly and Finishing of Large/Complex Genomes

James Knight

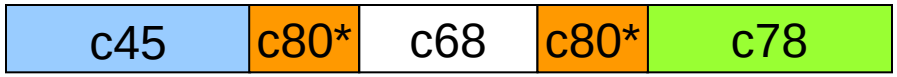
5/28/2009

What we've been doing/dealing-with for the last year

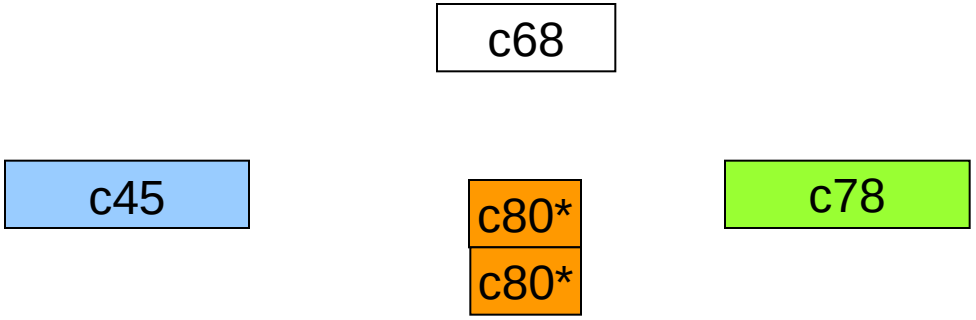
- Scale up
  - Titanium - 5x increase
  - Genome - 15x increase
  - Complexity - ?x increase
  - Crossing the 2 billion and 4 billion number boundaries
- Applications
  - De novo assembly improvements
  - Support for finishing
  - Transcriptome de novo assembly
  - Large genome mapping
    - Full suite of local and structural variations
    - Full/targeted resequencing (such as Sequence Capture)
  - Transcriptome mapping
- Software maturity

Newbler and its Contig Graph

For a genome region (with repeat region c80):

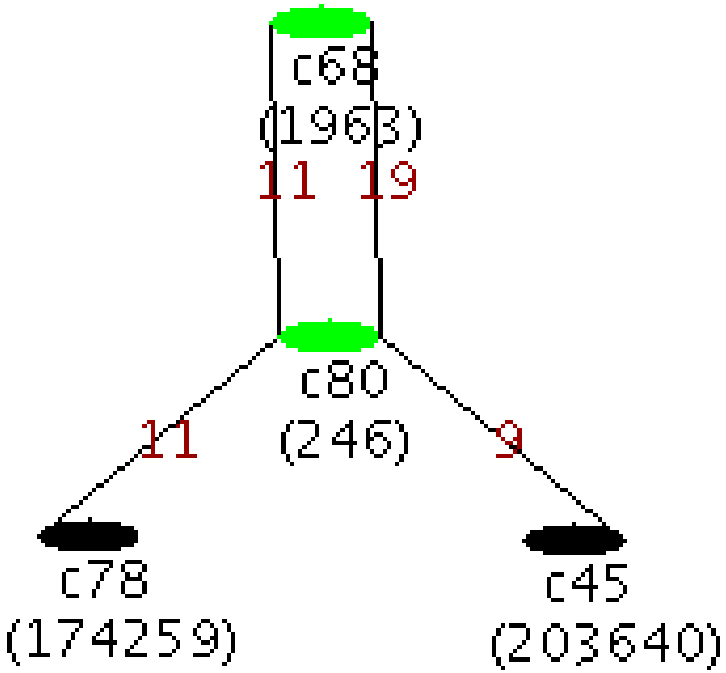


Pairwise overlaps will “collapse” repeat:



Newbler converts the overlaps into a contig graph structure:

- ***Full multiple-alignments for each contig***
- ***Contigs split at repeat boundaries***
- ***Read alignments split across contigs***



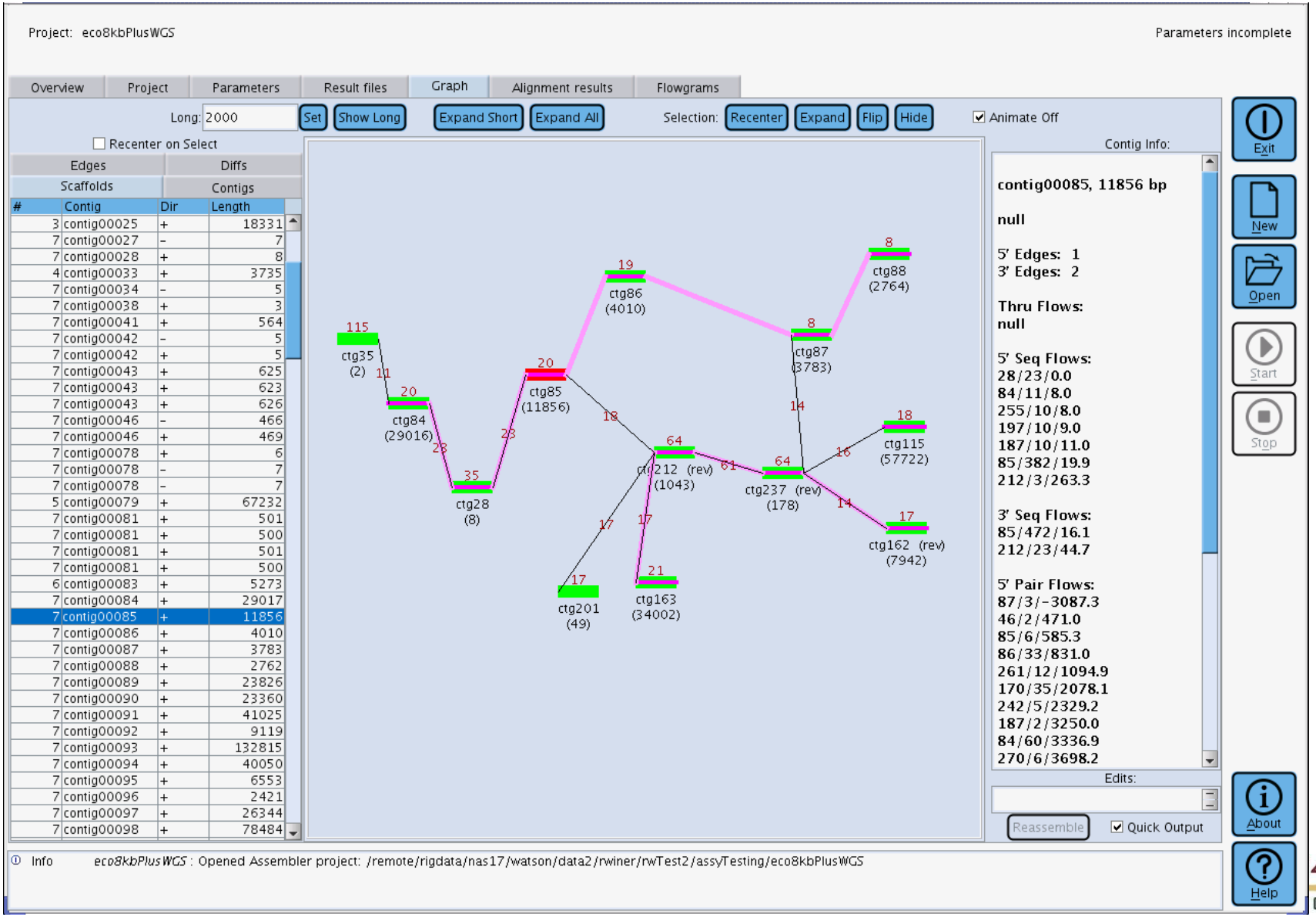
ATATTCACTCAGC*AAC

```
EBEO3TVO2EESBP.243-251.fm81 :gtctgtcagcgcggttatattcact
EBEO3TVO1BKXOR.223-248.fm10 :aacctgactgttcgatatattcactcagc*aac
EBEO3TVO3G19KE.214-251.fm81 gtctgtcagcgacggttatattcactcagc*aac
EBEO3TVO1BQED5.211-250.fm10 :aacctgactgttcgatatattcactcagc*aac
EBEO3TVO1A6PUU.205-245.fm10 :aacctgactgttcgatatattcactcagc*aac
EBEO3TVO1CFVIT.191-243.fm81 :gtctgtcagcgcggttatattcactcagc*aac
EBEO3TVO3G1EA5.182-240.fm81 :gtctgtcagcgcggttatattcactcagc*a*c
EBEO3TVO3GRKYV.182-243.fm10 :aacctgactgttcgatatattcactcagc*aac
EBEO3TVO2EKDO4.172-253.fm10 :aacctgactgttcgatatattcactcagc*aac
EBEO3TVO3GEUK2.171-256.fm10 :aacctgactgttcgatatattcactcagc*aac
EBEO3TVO2DKL4A.152-260.fm10 :aacctgactgttcgatatattcactcagc*aac
EBEO3TVO3FZOKY.150-256.fm10 :aacctgactgttcgatatattcactcagc*aac
EBEO3TVO2D8WE7.105-252.fm81 :gtctgtcagcgcggttatattcactcagc*aac
EBEO3TVO3F6PJ1.99-256.fm10  :aacctgactgttcgatatattcactcagc*aac
EBEO3TVO1BHMAW.85-257.fm10  :aacctgactgttcgatatattcactcagc*aac
EBEO3TVO2EC59P.73-257.fm10  :aacctgactgttcgatatattcactcagc*aac
EBEO3TVO2D3TTU.69-255.fm10  :aacctgactgttcgatatattcactcagc*aac
EBEO3TVO1CBIIE.38-254.fm81  :gtctgtcagcgcggttatattcactcagc*aac
EBEO3TVO2DDQUR.38-250.fm81  :gtctgtcagcgcggttatattcactcagc*aac
DOZMYEEO1EAGN3_left.29-58.fm:aacctgactgttcgatatattcactcagc*aac
EBEO3TVO2DT17H.27-251.fm10  :aacctgactgttcgatatattcactcagc*aac
DOZMYEEO1ELRZ4_right.18-21.f:aacctgactgttcgatatat
DOZMYEEO1BS9O3_right.17-21.f aacctgactgttcgatatatt
EBEO3TVO1BCTBK.3-250.fm10                          atatattcactcagc*aac
```

Finishing Support

- "Each read should appear in only one contig"
  - New -rip option for output generation
  - Concept: Associate each read with the largest contig it is aligned in
    - From largest contig to smallest contig, take reads crossing contig boundaries and rip them from other contigs
- Contigging through repeat regions
  - Darren's BAC pool went from 391 contigs (372 contigs) to 233 gaps*
- Read alignment location reporting
- "Edit directives" to change the contig graph
- Integrating graph viewer and gsAssembler GUI

Finishing Support



Smaller Genomes

- Titanium Improvements
  - More reads, longer reads
  - 3kb, 8kb and 20kb paired-end reads
- 1 Run Examples
  - 4 pad 8kb paired-end reads, 4 bacterial genomes
    - *S. pneumo, E. coli, T. thermophilus, C. jejuni*
    - 1 scaffold per genome/plasmid
  - 4 pad, shotgun, 3kb, 8kb, 20kb
    - *S. cerevisiae*
    - 1 scaffold per chromosome/mitochondrion
  - 1 full shotgun run (sequenced at a core lab)
    - ~30MB fungal genome
    - 300 contigs

Larger Genomes

- Titanium Improvements
  - More reads, longer reads
  - 3kb, 8kb and 20kb paired-end reads
- Sequencing and assembly of larger genomes achieved using only 454 reads
  - Arabidopsis & Drosophila: Newbler
  - Cucumber: Newbler & Celera Assembler
  - Oil Palm: Synamatix
- Current Newbler status
  - Two 1GB genome assemblies completed
  - Memory footprint recently stabilized at 3 bytes per read base
    - Incremental assembly available for lower memory machines
  - Multi-threading a bug fix away from efficient parallelization
  - Handles inbred/haploid datasets well, suffers with non-inbred, polyploid data (plus certain microsatellite repeats)

“Euchromatic” Coverage

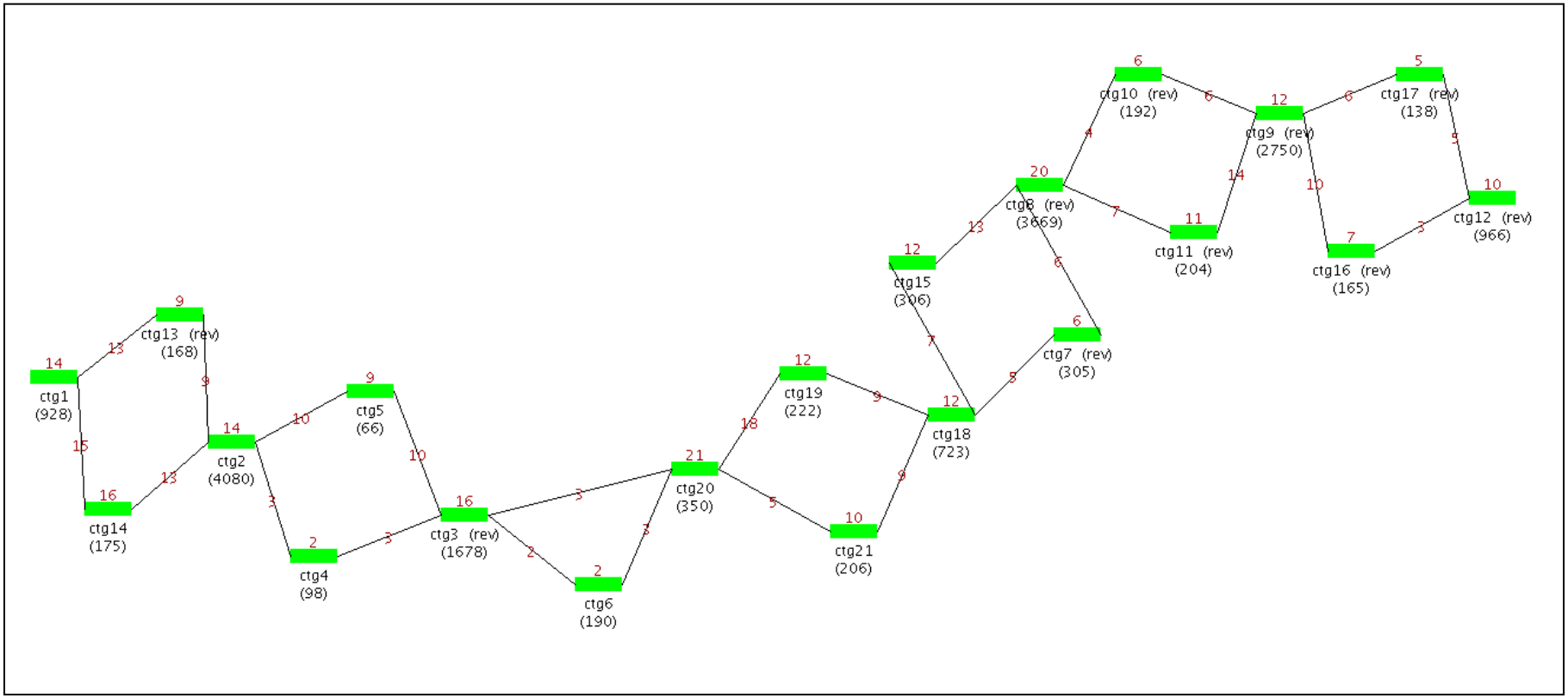
- Arabidopsis & Drosophila assembly size equaled Sanger assemblies
  - Arabidopsis: 114MB of 157MB genome
  - Drosophila: 118MB of 175MB genome
- Oil Palm assembly covers “gene space” and more
  - 1.7 GB genome, over 60% repetitive
  - Combination BAC-pool and shotgun dataset
  - Staged assembler developed by Synamatix
  - Over 99% of transcriptome reads aligned into assembly
  - Customer “very happy” with assembly

“Euchromatic” Coverage

- Cucumber assemblies appear to cover non-repetitive regions
  - Dataset [courtesy of RTHI…“The Hark”]
    - 25x shotgun reads
    - Recommended 3x 3kb and 2x 20kb paired-end reads
    - ~12GB total dataset
  - Celera Assembly and Newbler assembled 200MB of estimated 367MB genome
    - See poster #100 (J. Miller, etal) for Celera Assembly details
      - 87kb N50 contig size, .81MB N50 scaffold size
    - Newbler details
      - 37kb N50 contig size, 1.0MB N50 scaffold size
      - 55% of reads aligned into assembly
      - 21GB of memory, ~18 hours for one-shot assembly
      - Using recommended shotgun (15x) increased contigs from ~13,000 to ~16,000, but retained 1.0MB N50 scaffold size
    - 99% of transcriptome reads aligned into assembly

Inbred or Not Inbred

- Inbred and haploid samples result in longer N50 contig lengths in Newbler assemblies
- Newbler breaking at local, lower identity regions in hopes of detangling
  - Divided regions below were 80-95% identical to each other
- New option being added to adjust assembly algorithms



Improving the Assembly

- Large genome assembly editing
  - Many scaffold gaps found to be covered by individual reads
    - Read ends aligned in both contigs surrounding gaps
  - Higher-throughput editing mode to close those gaps
    - Quick implementation
    - 4-5 hour editing session
  - Arabidopsis editing closed ~1100 gaps
    - # of contigs decreased from 10229 to 9183
    - N50 contig size increased from 29kb to 37kb
    - N50 scaffold size increased from 4.1MB to 4.7MB
  - Will become graph viewer visualization/editing mode

Improving the Assembly

- Long overlaps
  - Seed and extend using long seeds (150-250 bases)
    - “Seeds” resilient to sequencing errors
    - Becomes additional alignment phase (long, then regular, then short)
  - Just implemented first-pass, unoptimized version
    - Arabidopsis
      - 10229 to 9120 contigs
      - 109MB to 113MB contigs
      - 114MB to 119MB scaffolds
      - 4.1MB to 5.6MB N50 scaffold size
    - Cucumber
      - Estimated additional 15MB of contig sequence

Next Steps

- Get the software done
  - Parallelization speed
  - Long overlap phase
  - All of the details of all of the applications
- Get a 144GB machine
  - Quote in the $20-30k range
  - Finally be able to say “Yes, we have done a 3GB genome”
- Update/optimize the assembler for non-inbred, polyploid genomes and repetitive genome regions
- Complete the finishing support work and the graph viewer
- Later
  - Metagenomics
  - Genome annotation software

Acknowledgements

Mapping/Assembly Software Group

Dan Fasulo (just joined, from Celera & Siemens)
Miroslav Kukricar
Roger Winer
Neal Kindlon
Carl Meacham
Phil Dagosto

454 Life Sciences/ Roche

Jason Affourtit
Michael Braverman
Brian Desany
Lei Du
Michael Egholm
Brian Godwin
Tim Harkins
Chinnappa Kodira
Marcel Margulies
Bernard Puc

Users/Collaborators

Arabidopsis: The Salk Institute

Drosophila: Baylor College of Medicine

Cucumber: RTHI

Pretty much most of you…

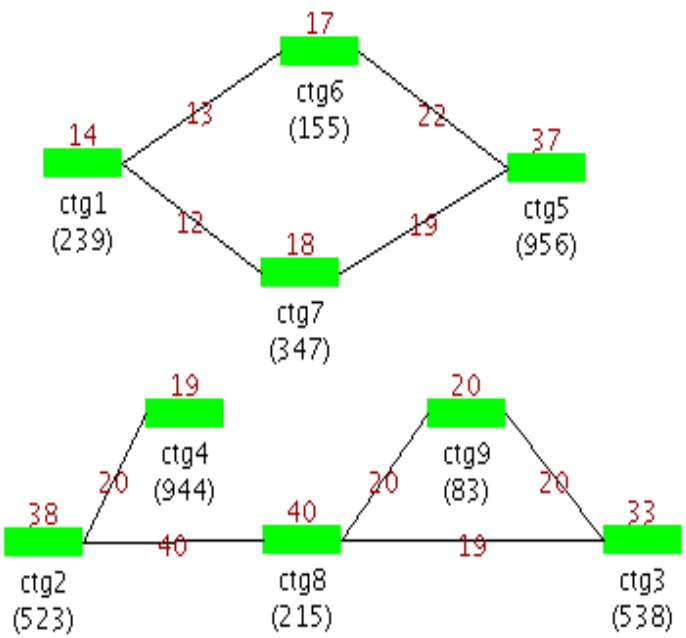
Roche

Transcriptome *de novo* assembly

Example: Dimethyladenosine transferase [Arabidopsis thaliana (Mouse-ear cress)]

5 transcripts: TA40203 1350 bp, TA42933 1181 bp, TA42934 1344 bp, CNS09ZE6 1416 bp, CNS0ADDJ 1459 bp.

- Simulated data
- Standard *de novo* assembly: 9 contigs
- Transcriptome *de novo* assembly: 5 isoforms



Bundle 1:

iso00000:	**ctg00001**	**.**	**ctg00006**	**ctg00005**
iso00001:	**ctg00001**	**ctg00007**	**.**	**ctg00005**

Bundle 2:

iso00002:	**ctg00002**	**ctg00004**	**.**	**.**	**.**
iso00003:	**ctg00002**	**.**	**ctg00008**	**ctg00009**	**ctg00003**
iso00004:	**ctg00002**	**.**	**ctg00008**	**.**	**ctg00003**