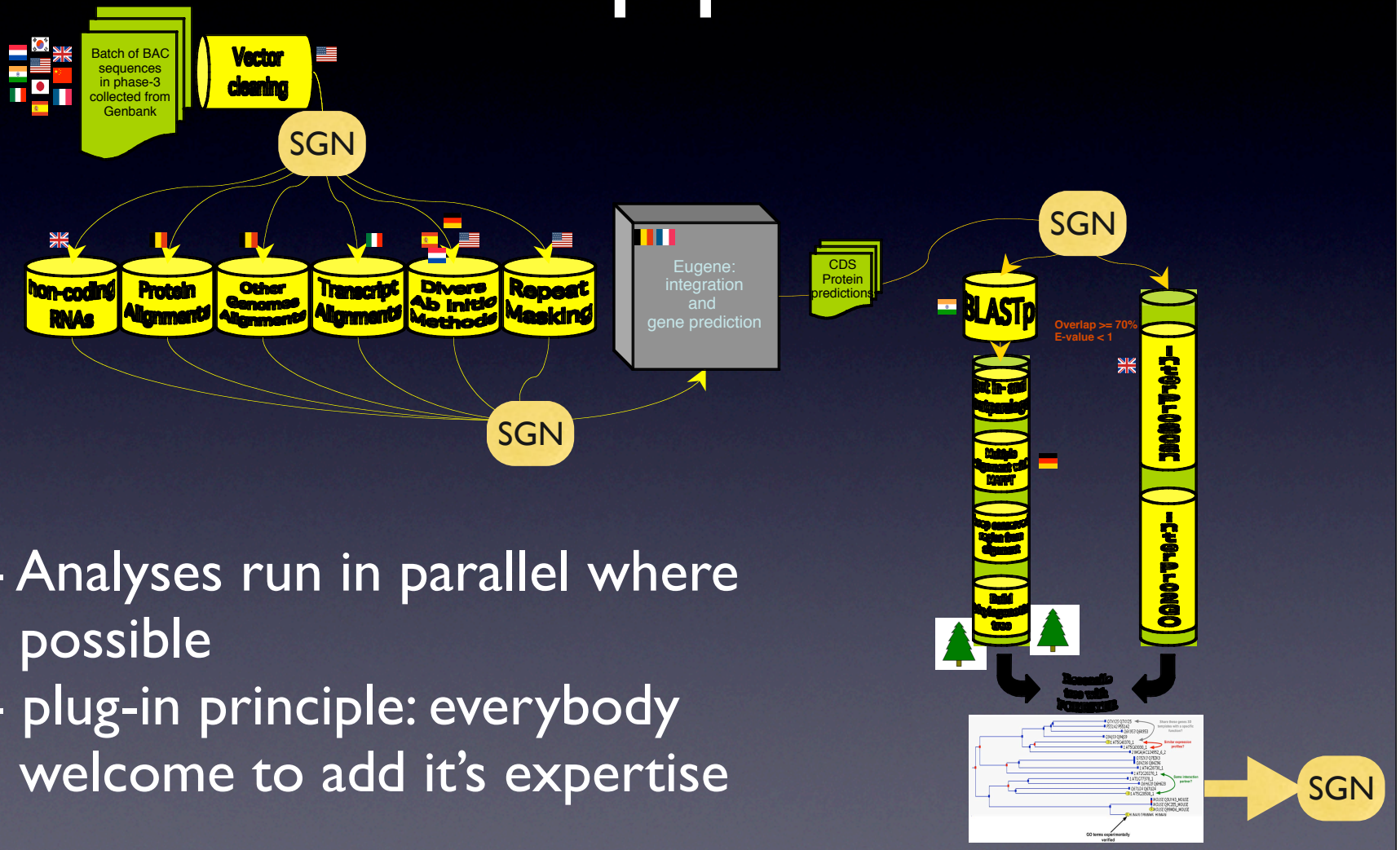
iTAG

Stephane Rombauts

iTAG, why?

- aim
  - high quality
  - homogeneous
  - state of the art tools
  - include all SOL members, divide the work

iTAG pipeline



- Analyses run in parallel where possible
- plug-in principle: everybody welcome to add it's expertise

iTAG members:

Jeffrey Fawcett, Stephane Rombauts, Yves Van de Peer, VIB Department of Plant Systems Biology, University Gent, Gent, Belgium

Thomas Schiex, Céline Noirot, INRA - UR 875, France

Jérôme Gouzy, UMR 441-2594 INRA-CNRS, France

Yu Wang, Remy Bruggmann, Klaus Mayer, MIPS/IBI Inst. for Bioinformatics, GSF Research Center for Environment and Health, Neuherberg, Germany

Anika Joecker, Heiko Schoof, Max Planck Institute for Plant Breeding Research,Cologne, Germany

Saloni Mathur1, Debasis Chattopadhyay2, Nagendra K. Singh3 and Akhilesh K. Tyagi1, Indian Initiative on Tomato Genome Sequencing, 1Inter-disciplinary Centre for Plant Genomics, Department of Plant Molecular Biology, University of Delhi, South Campus, New Delhi-110021, India;2National Institute for Plant Genome Research, New Delhi-110067, India; 3National Research Centre on Plant Biotechnology, Indian Agricultural Research Institute, New Delhi-110012, India

Maria Luisa Chiusano,Alessandra Traini, Nunzio D'Agostino, Department of Soil, Plant, Environmental and Animal Production Sciences, University of Naples "Federico II", Portici (NA), Italy

Erwin D. Datema, Mark W.E.J. Fiers, Roeland C.H.J. van Ham, Plant Research International, Wageningen University and Research Centre, Wageningen, The Netherlands.

Francisco Camara, Roderic Guigo, Genome Bioinformatics Laboratory GRIB -- IMIM/UPF/CRG, Barcelona, Spain

Daniel Buchan, Sarah Butcher, James Abbot , Division of molecular biosciences, Imperial College, London, UK

Gerard Bishop, Division of Biology, Imperial College London, UK

Robert Buels, Lukas Mueller, Dept of Plant Breeding and Genetics, Emerson Hall, Cornell University, Ithaca NY 14853, US

itag@sgn.cornell.edu

pipeline v000

- on phase-3 BACs only
- exchange format is GFF3,
- Eugene adapted to take also GFF3 as input
- data repository at SGN
- all the steps towards v001 are in place
  - only thing to do still is to train eugene to a fully Tomato Eugene

training set

- procedure on 229 BACs:
  - with a not fully trained eugene, make a prediction using all available extrinsic data
  - generated gene structures were pre-selected by script from Daniel (428 genes selected)
  - results are on the itag-wiki
  - manual final QC on selected models

QC of structures

- QC involves alignments with ESTs, proteins from SWISS-PROT, and NR
  - 172 BACs with 428 genes to check
  - BACs from chr01-04 done (±)
    = 54 BACs with 183 genes that passed QC
    (3.4 genes/BAC)
  - not all selected genes are actually good, while not-selected genes could be added after minor correction, so QC is really needed.
  - main problem are split and truncated genes

future plans

- finish QC for training set
- train eugene
- run pipeline v000
- switch to v001
- prepare websites to distribute data
- put annotation in GenBANK, update of the BACs (contact resp. BAC-owners!!)

people involved in Gent

- Jeffrey Fawcett
- Stephane Rombauts
- Yves Van de Peer