Data Management and Analysis Infrastructure for the 1000 Genomes Project

Paul Flicek
Vertebrate Genomics
European Bioinformatics Institute
Co-chair, Data Flow Group, 1000 Genomes Project
Genome sequencing

- Mammals (including humans) have approximately 6,000,000,000 nucleotides in their genome organised in a relatively small number of chromosomes
  - Humans have 46 (1-22,X,Y with copies from each parent)
- Genome sequencing is done by shredding the DNA molecules and collecting many short sequencing “reads” ranging from 25 to 1000 bp
- The reads must be put together to be useful
  - This is complicated and computationally intensive
DNA sequencing technologies

- ABI 3730: “old” Sanger technology
  - 80kb per run in ~800bp reads
- 454: introduced in 2005
  - 400Mb per run in ~350bp reads
  - Run takes ~1 day
- Illumina/Solexa: introduced in 2006
  - 15 Gb per run in ~75 bp reads
  - Run takes ~6 days
- AB SOLiD: introduced late 2007
  - 15-20 Gb per run in ~35 bp reads
  - Run takes ~6 days
- Pacific Biosciences etc. 2009/10
  - Recent science publication
  - Approximately 1 base per second
  - 100 fold faster data production
How next generation sequencing works - Illumina example

1. Prepare DNA
2. Attach to Slide
3. Create Clusters

Images from Illumina.com
How next generation sequencing works - Illumina example

4. Add labelled bases, primers, and DNA polymerase
5. Capture the fluorescence image from each cluster
6. Repeat steps 4 and 5 as many times as possible
A small part of one lane of one flowcell
Evolution of Large-scale Genome Sequencing

- 2000: Human genome working drafts
  - All data freely released
  - Project took about 10 years and cost about $3 billion

- 2008: Major genome centers can sequence the same number of base pairs as were produced for the HGP every 4 days

- 2009: Every 8 hours ($25,000)
2007: The Personal Genome Era Begins

The complete genome of an individual by massively parallel DNA sequencing

David A. Wheeler1,8, Maithreyan Sriharan3, Michael Egholm4, Yufeng Shen1,4, Lei Chen1, Amy McGuire3, Wen Hea, Yi-Ju Chen1, Vinod Makhijani1, G. Thomas Roth2, Xavier Gumés3, Karrie Tartaro4,†, Fahim Niazi1, Cynthia L. Turcotte2, Gerard P. Irykzy2, James R. Luspina3, Craig Chinaluti, Xing-zhi Song1, Yue Liu1, Ye Yuan1, Lynne Nazareth1, Xiang Qin1, Donna M. Muzny3, Marcel Margulies3, George M. Weinstock1,4, Richard A. Gibbs1,4 & Jonathan M. Rothberg†

LUMC brengt als eerste DNA-volgorde van vrouw

Genetici van het Leids Universitair Medisch Centrum (LUMC) hebben de volledige DNA-volgorde opgelost van een vrouw. Het is de eerste vrouw ter wereld en de eerste Europese van wie de DNA-volgorde bekend geworden is. Dat maakt de ondertekenaars masacrangerende bekend tijdens een speciaal ingelijste persconferentie op Bessensap, de jaarlijkse landelijke bijeenkomst voor wetenschap en pers.

De DNA-volgorde en de nabije analyse zullen, op enkele privacy-gevoelige gegevens na, binnenkort openbaar worden gemaakt. De resultaten geven meer inzicht in de genetische verscheidenheid van de mens.

DNA van geneticus Marjolein Kriek

Het DNA is afkomstig van dr. Marjolein Kriek, klinisch geneticus i.o. aan het LUMC. “Als vakgenoot is zij beter dan wie ook in staat om mogelijke voor- en nadelen van openbaar maken goed of slecht te plaatsen”, zegt prof. dr. Gert-Jan van Ommen, leider van het Leids team en trekker van het ‘Center for Medical Systems Biology’ een van de centra van het Nederlands Genoom initiatief.

Gert-Jan van Ommen: “Bovendien geeft het sequentie van een vrouw meer inzicht in het X-chromosoom. Omdat het X-chromosoom het helft van de populaatie, de mensen, het werk alleen moet doen, is daar in de evolutie van de mens harder op geselecteerd. Het X-chromosoom is daardoor minder variabel. Daarnaast werd het tijd om, na het wereldwijd specifiek van vier man, de geëuthen die verhalen wat in evenwicht te brengen.” En glimlachend: “En na Watson vonden we het wel mooi om Kriek te doen.”

The diploid genome sequence of an Asian individual

Jun Wang1,2, Wei Wang1,2, Batang Li1,2, Yingyu Li1,2, Geng Fan1,2, Laura Goodman1, Wei Fan1, Junping Zhang1, Jun Li1, Jiebin Zhang1, Yiran Gu1, Brian Fan1, Hong Lin1, Yue Lu1, Xiaodong Fang1, Huiying Li1, Zhenlin Xu1, Dong Li1, Yaping Zhao1, Yue he1, Zhenjun Yang3, Houcheng Zhang1, Hua Holmanna, Michael Inagawa, John Poul1, Xin Qi1, Jing Zhao1, Junjie Du1, Yan Zhou1, Junjie Qin1, LiLia Ma1, Guoqing Liu1, Zhenhui Yang4, Guojie Zhang1, Bo Yang1, Chang Ya1, Fang Li1, Wenjie Li1, Shaochun Li6, Dandan Li1, Pongnarm N1, Jianjun Li1, Qihua Guo1, Shengming Zhi1, Dongran Liu1, Zhikai Li1, Ning Li1, Qiongwen Guo1, Jiangsu Zhang1, Ya Ye1, Lin Fang1, Qie Han1, Qun Chen1, Lu Li1, Kong Zhi1, A. van der Maas, C. van der Wagt, Zhidong Yang1, Zhiqiang Yan1, Xiaofei Yang1, Xihao Feng1, Koning Krikken, Gao Guo-Bin, Wang1,6,9, Renkun Nolten1, Richard Durbin1, Lars Boland1,11, Xiaoming Zhang1, Songgang Li1,2,11, Haorong Yang1,2,11 & Jan Wang1,2

The Diploid Genome Sequence of an Individual Human

Samuel Levy2, Granger Sutton1, Pauline C. Ng2, Lennart F. S. Lander1, Aaron L. Halpern3, Brian P. Wabena1, Nelson Axelrad4, Jiagui Huang1, Ewen F. Kirkness5, Gernady Denov6, Yuyin Lin7, Jeffrey R. MacDonald8, Andy Wing Chun Pang9, Mary Shage2, Timothy B. Stockwell9, Alexa Tsamouli1, Vivek Bang6, Vikas Bansal2, Saul A. Kravitz1, Dana A. Busam1, Karen Y. Beeson1, Tina C. McIntosh1, Karin A. Remington1, Josep F. Abril1, John Gill1, Jon Berman1, Yu-Hui Rogers1, Marvin E. Frazer1, Stephen W. Scherer1, Robert L. Strousberg1, J. Craig Venter1

Craig Venter

(Photograph: Business Week)
1000 Genomes Project: Primary goals

• Overall Goal: Create a deep catalogue of human variation to provide a better baseline to underpin human genetics

• Discover shared variation (shared = not private to individual) and characterise by allele frequency
  • Aim for effectively all (not just a lot of) common variation
    • For example: any variant down to 1% minor allele frequency in a population in the accessible genome has a 95% chance of being identified
    • The pilot project and simulations will help to determine the precision of this statement
  • Structural variants as well as SNPs
    • Accessible because the project will use paired-end sequencing reads
  • Deeper discovery in gene regions, down to 0.5% to 0.1% MAF
Whole Genome Association Studies

- Genomic regions associated with common diseases
- First major successes in 2007
  - Wellcome Trust Case Control Consortium
    - 17,000 individuals (14,000 cases; 3,000 controls)
    - Diabetes (type I and type II), coronary heart disease, hypertension, bipolar disorder, rheumatoid arthritis, Crohn’s disease
  - Affymetrix 500,000 SNP chip
- There are now hundreds of regions genome wide with more being published every month
- These studies do not identify the causative variants
  - This requires extensive sequencing in many individuals of the associated regions
1000 Genomes Project Details

• Three pilots during 2008
  • Pilot 1: 3x60 samples at 2x (6Gb) per person:
    • European CEU, African YRI, East Asian CHB/JPT
  • Pilot 2: CEU and YRI trios at 20-50x
  • Pilot 3: 1000 genes in 1000 people
  • Multiple platforms/protocols
• Pilots show high quality data collected at scale, and that variants can be called reliably
• Main project design
  • 1,200-1,500 people sequenced each to 4 x coverage
  • Data collection completed within a year
  • Quarterly data releases
• Initial analysis expect to conclude in 2010
1000 Genomes Project Update

- Sequencing started in April 2008
- Over 150 terabytes of data deposited so far
- Approximately 6,000,000,000,000 bp have already been submitted
- Raw and process data is freely available now
  - ftp://ftp.1000genomes.ebi.ac.uk
- Size of GenBank/EMBL Archive/DDBJ at start of project: 235,135,312,328 nucleotides
  - During September and October 2008 the 1000 Genomes project produced the equivalent of EMBL/GenBank every week
- Within the first five months of the project, it had produced more DNA sequence data than all previous DNA sequencing project combined
- Total project data will be approximately 500 terabytes
Putting data scale into perspective

- The 1000 Genomes projects is largest ever data generation project in biology and uses less than 10% of the world-wide sequencing capacity.
- *We know we are not alone and lots of other fields have the same problem.*
- Hardware technology is important, but is not where we are stressed.
- Our single most important problem is the democratization of sequence analysis.
- Biology has become an informatics- and data-heavy science, but we lack a culture that supports pervasive computational analysis.
- Our weak links are computational infrastructure and the training and expertise of bench scientists.

-Sean Eddy

World-wide sequencing data production is now less than an order of magnitude behind CERN

- The Large Hadron Collider produces only 15 petabytes per year from a single point source
- The LHC grid is 140 computer centres in 33 countries centered at CERN (Tier 0)
- Sequencing is producing data in hundreds of centers in dozens of countries and has two Tier 0 sites (EBI & NCBI)
- Sequencing is now used for many experiments of various sizes (“the microscope of modern biology”) each requiring different analysis techniques
- Capacity continues to grow with few if any plans for computational infrastructure
Analysis overview

• Sequence images are processed at the sequencing centers and discarded (about 98% of the data is discarded at this step)
• Results are base calls, “quality scores” and intensities which are submitted for further analysis

• Initial analysis
  • Quality score recalibration
  • Alignment to reference genome sequence

• Secondary analysis
  • Nucleotide variation discovery
    • Trio and population based algorithms
  • Structural variation discovery
    • Read depth and paired-end analysis algorithms
  • Genotype assignments

• Advanced analysis
  • Genome assembly
Data flow and management

- Data production at nine sequencing centers in 4 countries
  - USA data initial collection at NCBI
    - Broad Institute, Washington University, Baylor College of Medicine, 454 Life Sciences, ABI
  - Other data initial collection at EBI
    - Sanger Institute, Illumina (UK); Max Planck Institute for Molecular Genetics (Germany); BGI (China)
- Submitted data exchanged between EBI and NCBI
  - Initial analysis at Sanger Institute
  - Results of initial analysis provided worldwide for secondary analysis
  - Secondary analysis results collected and provided worldwide
- Wellcome Trust Genome Campus, Hinxton, UK
  - European Bioinformatics Institute
  - Wellcome Trust Sanger Institute
- Between the two we have nearly 10 PB of storage and approximately 7500 compute nodes
The need for standard data formats

- Richard Durbin
Computational bottlenecks

- Image processing is done on dedicated hardware at the sequencing machines
- Initial analysis is relatively easy to distribute, but IO heavy
  - Each node stores a copy of the reference genome in memory, reads are distributed for alignment and alignment files merged at the end
    - Initial alignment algorithms used a hash representation
    - New algorithms are based on Burrows-Wheeler transformations
  - Once the alignment files are constructed, analysis of small regions of the genome can be distributed even over many individuals
- Advanced analysis is difficult
  - Assembly algorithms currently require up to 20 terabytes of memory
- Data transfer is a significant problem
Data Transfer Infrastructure

• FTP does not work well for terabytes of data
• “Old fashioned” solutions
  • Copy the data onto a hard drive and mail the hard drive around the world
  • (Significant personnel costs)
• Infrastructure solutions
  • Create/buy dedicated lines for point to point transfer or direct connection to faster points on the backbone
  • Expensive to do collaborative analysis, but will probably be part of the solution
• Advanced technology solutions
  • Asperasoft
  • Uses udp to transfer files to avoid tcp
  • Can quickly saturate connections
Data moving through the Hinxton router

First Data Push

We now regularly exceed 90% of the campus bandwidth for 1000 Genomes data transfers
Data distribution back-up plan

http://www.simbaint.com/
1000 Genomes Browser

- Data visualisation system for most useful data extracts and analysis results
- Based on Ensembl and potentially including the Resembl plugin developed by Illumina
- A separate installation managed and updated at the EBI and available within the 1000genomes.org domain
- SNPs, GLF and coverage data for all individuals
- “Full data” for the trios and other high coverage individuals using Resembl if available
- Built on current version of Ensembl web code (with project specific “skinning”)
  - Expected update to new Ensembl interface in 2009
Data availability through AWS

- Alignment and sequence files moving into Amazon cloud
  - Including tools for sequence extraction
- Ensembl data is already there
- 1000 Genomes Browser may be run from AWS
Going Forward

- Projects are already underway that are 10-50x the size of the 1000 Genomes project
  - Significant focus on cancer genome sequencing

- Data security is a major challenge for these projects (and one not faced by the 1000 Genomes project)
  - Genome sequence is identifiable
  - Research participants sign consent agreements and have the right to withdraw from the study

- Authorisation, access and data use must be addressed such that ethical, legal, and social issues (ELSI) are properly addressed in such a way that scientific discovery is still possible
Final Thoughts

• Annotating the variation catalogue created by the 1000 Genomes project will be one of the major future challenges in human genomics and potentially the most valuable

• The bioinformatics requirements for large scale biology are growing faster than computational capacity

• New and smarter solutions needed
Acknowledgements

- Vertebrate Genomics Team
  - Zam Iqbal, Laura Clarke (1000 Genomes DCC)
  - Fiona Cunningham, Yuan Chen, Will McLaren (Ensembl Variation)
  - Eugene Kulesha, Stephen Keenan (1000 Genomes browser)

- NCBI DCC Team: Steve Sherry, Hoda Khouri, Martin Shumway, Justin Paschall
- EMBL Sequence Archive / European Read Archive Team
- 1000 Genomes and WTCCC Projects
- The Protein and Nucleotide Database Group

- Funding: Wellcome Trust, EMBL, UK MRC, European Union