The Pan-Cancer QPQ project

Jan Korbel
EMBL Heidelberg,
For the PCAWG steering committee

EMBL
40 YEARS | 1974–2014

25 Sept. 2015

https://dcc.icgc.org/pcawg

Costs of human genome sequencing

2003

2015

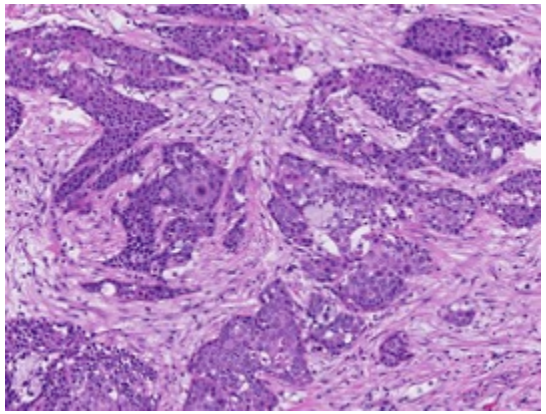
The International Cancer Genome Consortium (ICGC)

Objective: Characterize patterns of mutation in >50 types of cancer

1. Sequence patient's normal genome

…GATTATTGCAGGTAT…

2. Sequence patient's tumor genome



…GATTATTCCAGGTAT…

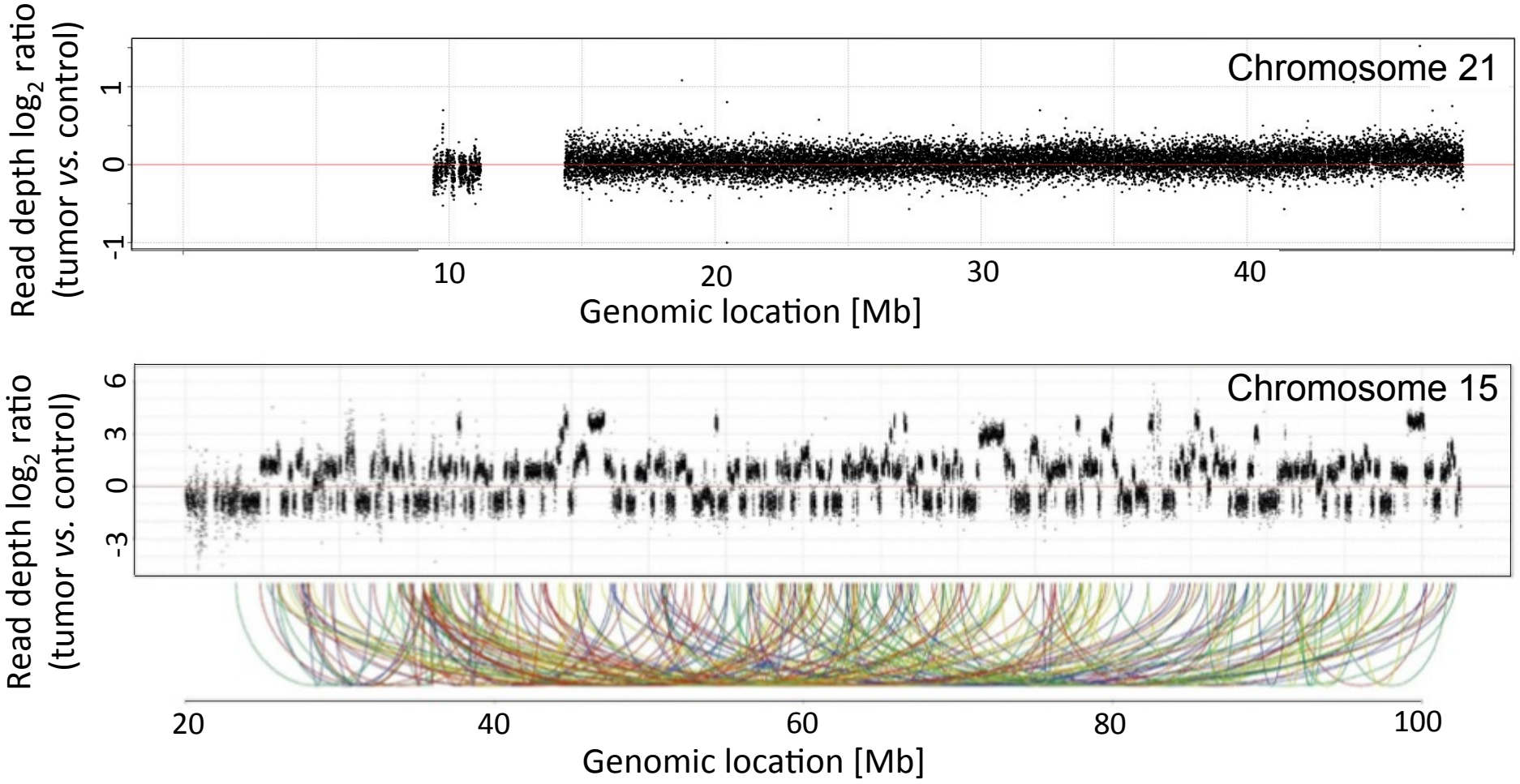
3. Identify cancer-associated (somatic) DNA alterations

…GATTATTCCAGGTAT…

4. Obtain basic insights into human disease biology through data analysis.

5. If feasible translate knowledge into diagnostic & treatment approaches.

Sequencing genomes of the childhood brain tumor medulloblastoma revealed catastrophic alterations of individual chromosomes



- Rearrangements cannot be readily reconciled (statistically) with a stepwise process.
- *Chromothripsis*: *chromo* for chromosome, *thripsis*, shattering into pieces.
- Chromothripsis linked with mutations of the gene encoding the p53 tumor suppressor.

Rausch *et al. Cell* 2012; Korbel & Campbell, *Cell* 2013; Mardin et al. *MSB in press*

Future of cancer genomics in Europe (and worldwide)

Commoditization of genome sequencing is changing the way we do science!

- Genome sequencing is becoming a regular molecular biology “tool”.
- Millions of cancer genomes will likely be sequenced within 5-10 yrs.
- This commoditization offers new opportunities in research (*e.g.* to link rare genetic variants to clinical responses in cancer, and to answer basic research questions using integrative analyses).

Current status of cancer genomics

- Basic research: systematic cancer genome analyses across centres worldwide, within the ICGC and other research consortia.
- Clinics: patient genome sequencing in clinical studies & uptake into clinical practice to assess treatment options.
- Humans to become best genotyped & phenotyped organism in biology.

Dissemination of cancer genomics leads to increased data fragmentation:

- Data submitted to different repositories using distinct data formats.
- Lack of harmonization of analysis methodologies makes data essentially incomparable.
- Repositories lack suitable computing resources for downstream analyses (e.g. mutation detection).
- Data security and privacy rules differing between countries.

Pan-Cancer Analysis of Whole Genomes (PCAWG)

Deeply sequenced cancer & normal genomes from >2,600 cancer patients

- Harmonization of the world's cancer genomic data (including International Cancer Genome Consortium and Cancer Genome Atlas project data; nearly 1 PB), to enable joint integrative analyses.
- "Big data" analytics framework: based on cancer genomes, transcriptomes, epigenomes (DNA methylation), and clinical data.

PCAWG Steering Committee: Gad Getz (USA), Jan Korbel (EMBL), Lincoln Stein (Canada), Josh Stuart (USA), Peter Campbell (UK) [chair]

PCAWG's genome analysis framework:

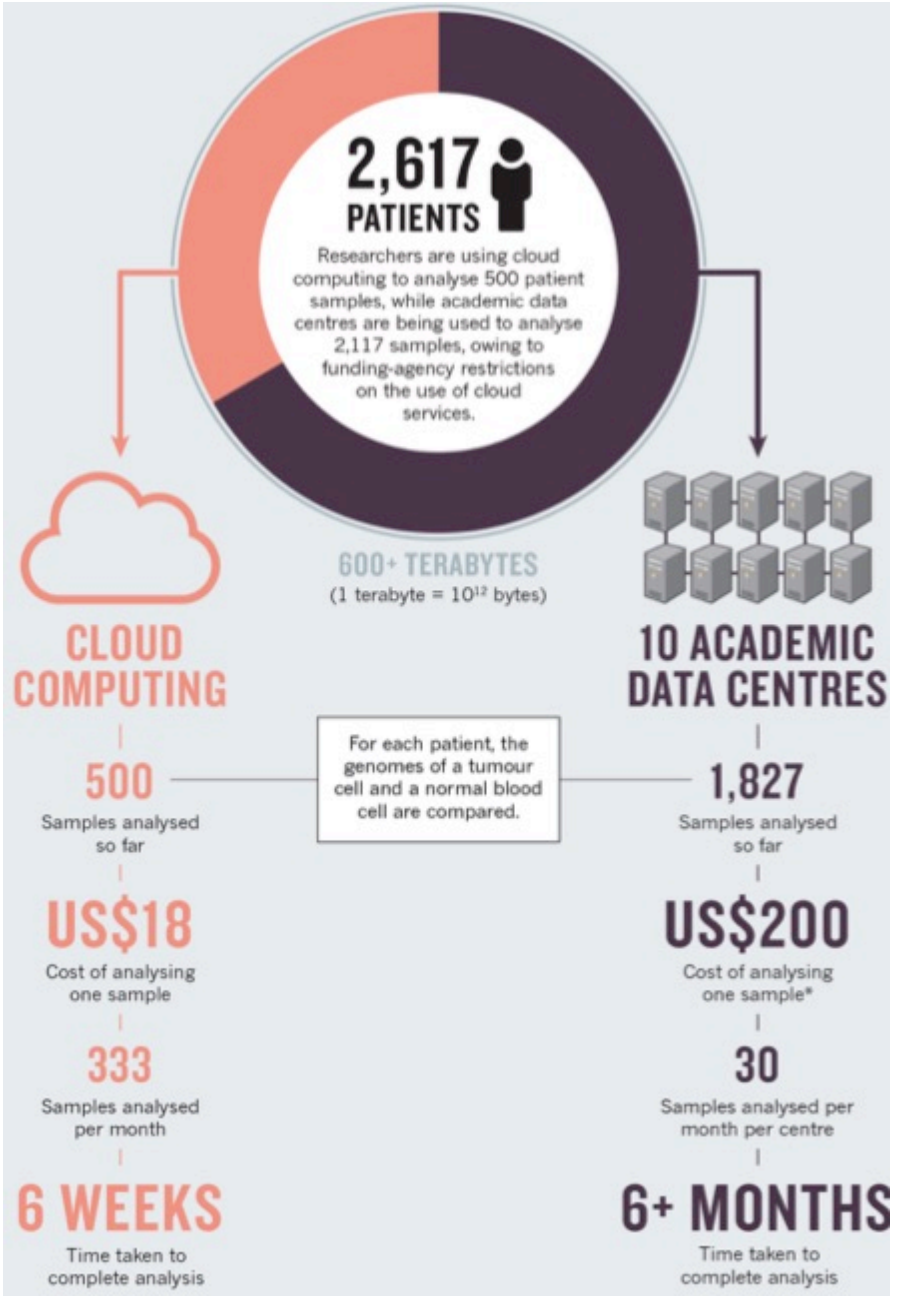
Three standardized somatic analysis pipelines: Broad (USA), Sanger (UK), Heidelberg (EMBL/DKFZ).

Additional germline genome pipeline: Annai Systems data center.

Status: tumor-normal pairs aligned; somatic variant calls for 2,600 patients ('August freeze'), germline calls for 2,100 patients

Cloud computing & IT partners: Seven Bridges Genomics, Intel, Amazon Web Services, SAP, Fujitsu

https://dcc.icgc.org/pcawg



Stein LD, Knoppers BM, Campbell P, Getz G & Korbel JO, *Nature* 2015

Objectives of PCAWG

Facilitate comparative analyses among diverse tumor types by use of standardized analysis pipelines, and covering a range of research themes (see below).

Publish a marker paper (likely in *Nature)*, together with a set of companion papers (likely to be published in *Nature* & *Nature Genetics*) reflecting PCAWG working groups.

Research activities organized by steering committee, which advises a series of working groups comprising >700 scientists (including many leaders in the field of cancer in Europe and Northern America), and broadly covering the following themes:

Analysis of mutations in regulatory regions and non-coding RNAs
Integration of the transcriptome and genome
Integration of the epigenome and genome
Consequences of somatic mutations on pathway and network activity
Patterns of genomic structural variations
Mutation signatures and processes
The germline cancer genome
Inferring driver mutations and identifying cancer genes and pathways
Translating cancer genomes to the clinic
Evolution and heterogeneity
Portals, visualization and software infrastructure
Molecular subtypes and classification
Mitochondrial genomes
Pathogens
Novel somatic mutation calling methods

Proposal: future of cancer genomics

Pan-Cancer Analysis of Whole Genomes "QPQ" ("quid pro quo")* - in our view necessary follow-up of PCAWG to ensure sustainability and creation of a virtual marketplace for aggregation & analysis of genomes & associated data, which can act as a commons of cancer genomic data facilitating biomedical science.

- Interactive repository for cancer genomes, epigenetic data, clinical data.

Incentives for data generators:

- In exchange for depositing data, will get (prioritized) access to point-and-click menu of bioinformatics tools to run on data (forming incentives: high quality data analysis as "virtual payment").

Envisioned users:

- Life scientists/data scientists as well as clinicians with less expertise in cancer genomes analyses (allowing for IaaS and SaaS models).

Envisioned scenario: sets of federated clouds in different countries or regions (*e.g.*, in several European countries).

***The PCAWG Steering Committee: Peter Campbell (UK), Gad Getz (USA), Jan Korbel (EMBL), Lincoln Stein (Canada), Josh Stuart (USA).**