RNA-seq Quality Control in BIOS freeze 2

OPENING UP THE BBMRI GENOMICS INFRASTRUCTURE IN THE NETHERLANDS

AMSTERDAM, 21ST OF SEPTEMBER, 2016

ANNIQUE CLARINGBOULD

RNA-seq data in BIOS
Freeze 1

- Freeze 1: 2,116 unrelated samples
- Several current papers use freeze 1 data

Hypothesis-free identification of modulators of genetic risk factors

Daria V. Zhernakova[1*], Patrick Deelen[1,2*], Martijn Vermaat[3*], Maarten van Iterson[4*], Michiel van Galen[3], Wibowo Arindrarto[5], Peter van 't Hof[5], Hailiang Mei[5], Freerk van Dijk[1,2], Harm-Jan Westra[6,7,8], Marc Jan Bonder[1], Jeroen van Rooij[9], Marijn Verkerk[9], P. Mila Jhamai[9], Matthijs Moed[4], Szymon M. Kielbasa[4] , Jan Bot[10], Irene Nooren[10], René Pool[11], Jenny van Dongen[11], Jouke J. Hottenga[11], Coen D.A. Stehouwer[12], Carla J.H. van der Kallen[12], Casper G. Schalkwijk[12], Alexandra Zhernakova[1], Yang Li[1], Ettje F. Tigchelaar[1], Marian Beekman[4], Joris Deelen[4], Diana van Heemst[13], Leonard H. van den Berg[14], Albert Hofman[15], André G. Uitterlinden[9] , Marleen M.J. van Greevenbroek[12], Jan H. Veldink[16], Dorret I. Boomsma[11], Cornelia M. van Duijn[17], Cisca Wijmenga[1], P. Eline Slagboom[4], Morris A. Swertz[1,2], Aaron Isaacs[17,18], Joyce B.J. van Meurs[9], Rick Jansen[19], Bastiaan T. Heijmans[4#], Peter A.C. 't Hoen[3#], Lude Franke[1#]

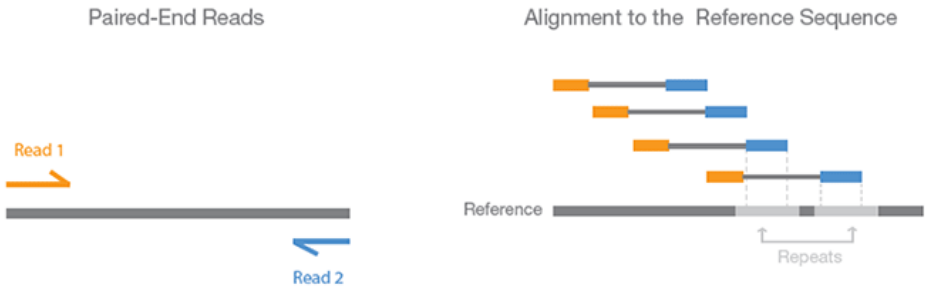
Disease variants alter transcription factor levels and methylation of their binding sites

Marc Jan Bonder[1,*], René Luijk[2,*], Daria V. Zhernakova[1,**], Matthijs Moed[2,**], Patrick Deelen[1,3,**], Martijn Vermaat[4,**], Maarten van Iterson[2], Freerk van Dijk[1,3], Michiel van Galen[3], Jan Bot[5], Roderick C. Slieker[2], P. Mila Jhamai[6], Michael Verbiest[3], H. Eka D. Suchiman[2], Marijn Verkerk[6], Ruud van der Breggen[2], Jeroen van Rooij[6], Nico Lakenberg[2], Wibowo Arindrarto[8], Szymon M. Kielbasa[7], Iris Jonkers[2], Peter van 't Hof[7], Irene Nooren[5], Marian Beekman[2], Joris Deelen[2], Diana van Heemst[9], Alexandra Zhernakova[1], Ettje F. Tigchelaar[1], Morris A. Swertz[1,3], Albert Hofman[10], André G. Uitterlinden[6], René Pool[11], Jenny van Dongen[11], Jouke J. Hottenga[11], Coen D.A. Stehouwer[12], Carla J.H. van der Kallen[12], Casper G. Schalkwijk[12], Leonard H. van den Berg[13], Erik. W van Zwet[8], Hailiang Mei[7], Yang Li[1], Mathieu Lemire[14], Thomas J. Hudson[14,15,16], the BIOS Consortium, P. Eline Slagboom[2], Cisca Wijmenga[1], Jan H. Veldink[13], Marleen M.J. van Greevenbroek[12], Cornelia M. van Duijn[17], Dorret I. Boomsma[11], Aaron Isaacs[17,##], Rick Jansen[18,##], Joyce B.J. van Meurs[6,##], Peter A.C. 't Hoen[4,#], Lude Franke[1,#], Bastiaan T. Heijmans[2,#]

Refined mapping of autoimmune disease associated genetic variants with gene expression suggests an important role for non-coding RNAs

Isis Ricaño-Ponce[a], Daria V. Zhernakova[a], Patrick Deelen[a, b], Oscar Luo[c], Xingwang Li[c], Aaron Isaacs[d], Juha Karjalainen[a], Jennifer Di Tommaso[a], Zuzanna Agnieszka Borek[a], Maria M. Zorro[a], Javier Gutierrez-Achury[a], Andre G. Uitterlinden[d], Albert Hofman[d], Joyce van Meurs[d], BIOS consortium[e],

RNA-seq data in BIOS
Freeze 2



Freeze 2 initial numbers	Number of samples
BIOS + NTR extra	4543
BIOS only	3824

- Samples available from
  CODAM, LLDeep, LLS, NTR, PAN, RS
- Paired-end sequencing Illumina's Hiseq2000
- Intended number of reads: >15M
  - Re-sequenced several samples
- Many factors can influence the quality of the data
- Goal: usable set of unrelated samples with gene expression (and genotype) data of good quality

Picture source: Paired-End Sequencing web page Illumina,
http://www.illumina.com/technology/next-generation-sequencing/paired-end-sequencing_assay.html

Main strategy

1. Match RNA to DNA for each individual based on data
   - Resolve mix ups
   - Merge topped up samples
2. Use quality metrics as cut-offs
3. Sample selection

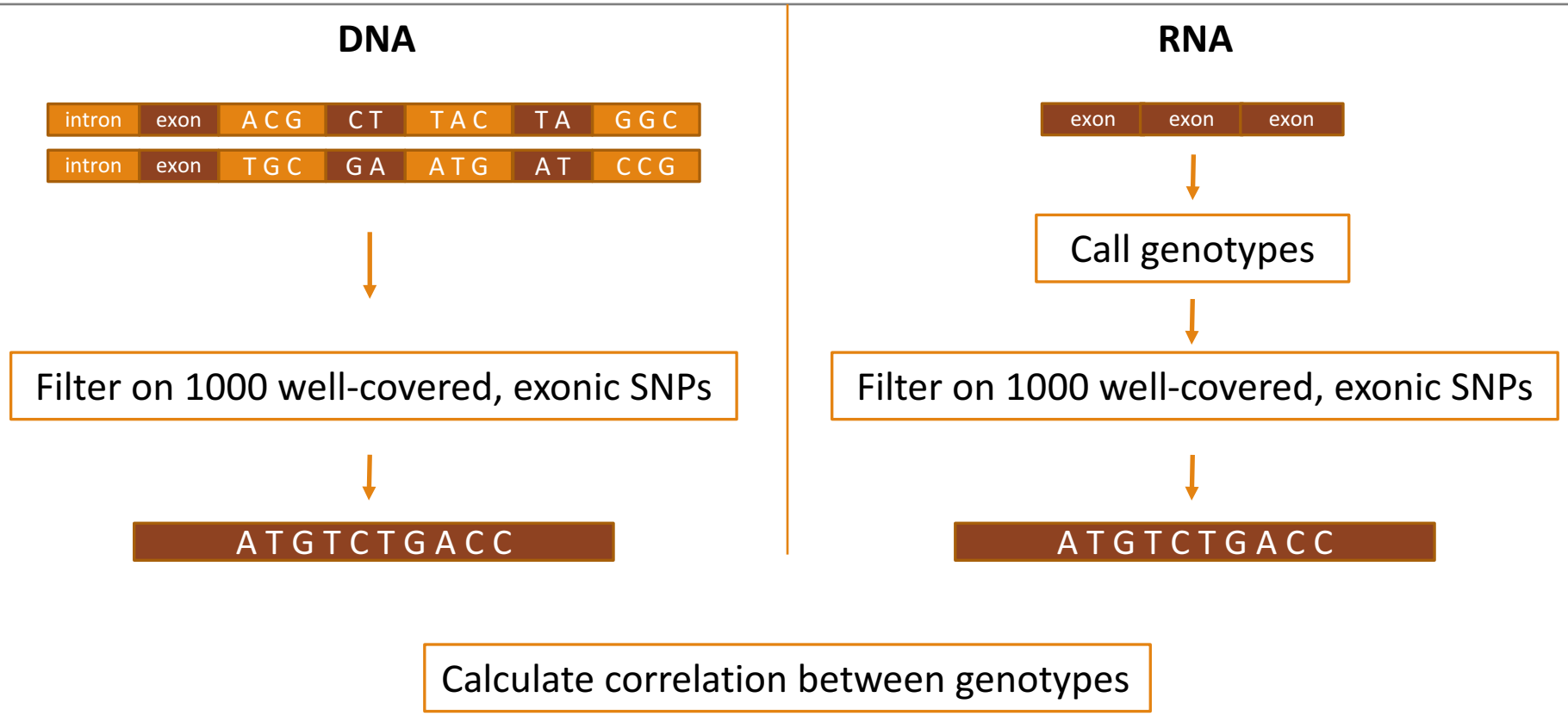
DNA-RNA matching
Method

- Extract genotypes from RNA-seq data
  - Alignment
  - Quantification
  - Genotype calling (Unified Genotyper)

- DNA-RNA matching
  - Check that correct samples are matched

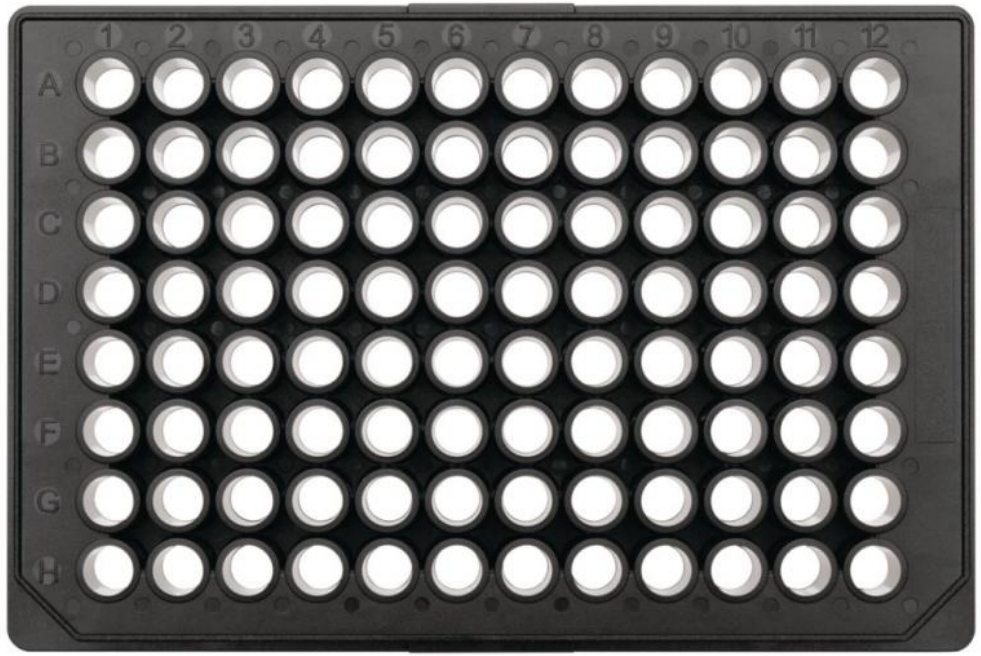
- **Remove** outliers
  - Ethnic
  - Heterozygosity

DNA-RNA matching
Method



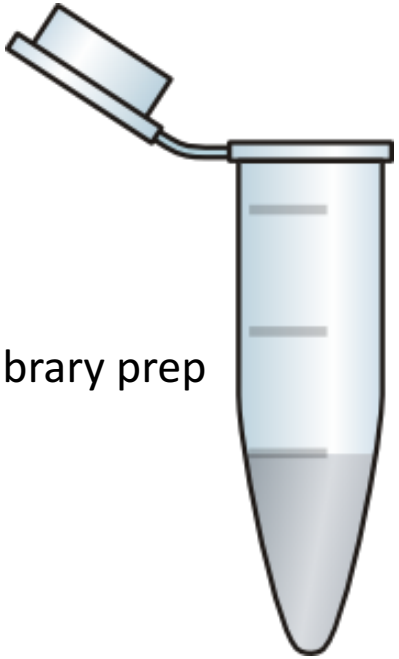
DNA-RNA matching
Resolving twin pairs and mix ups

- 96-well plate turned around
- Sample mix ups
- Identifying monozygotic & dizygotic twins
- **Remove** unresolved 39 mix ups



DNA-RNA matching
Merging RNA-seq runs

- Some low-read samples were re-sequenced
- Merge two RNA samples from one individual



One library prep

First round sequencing
(many reads)

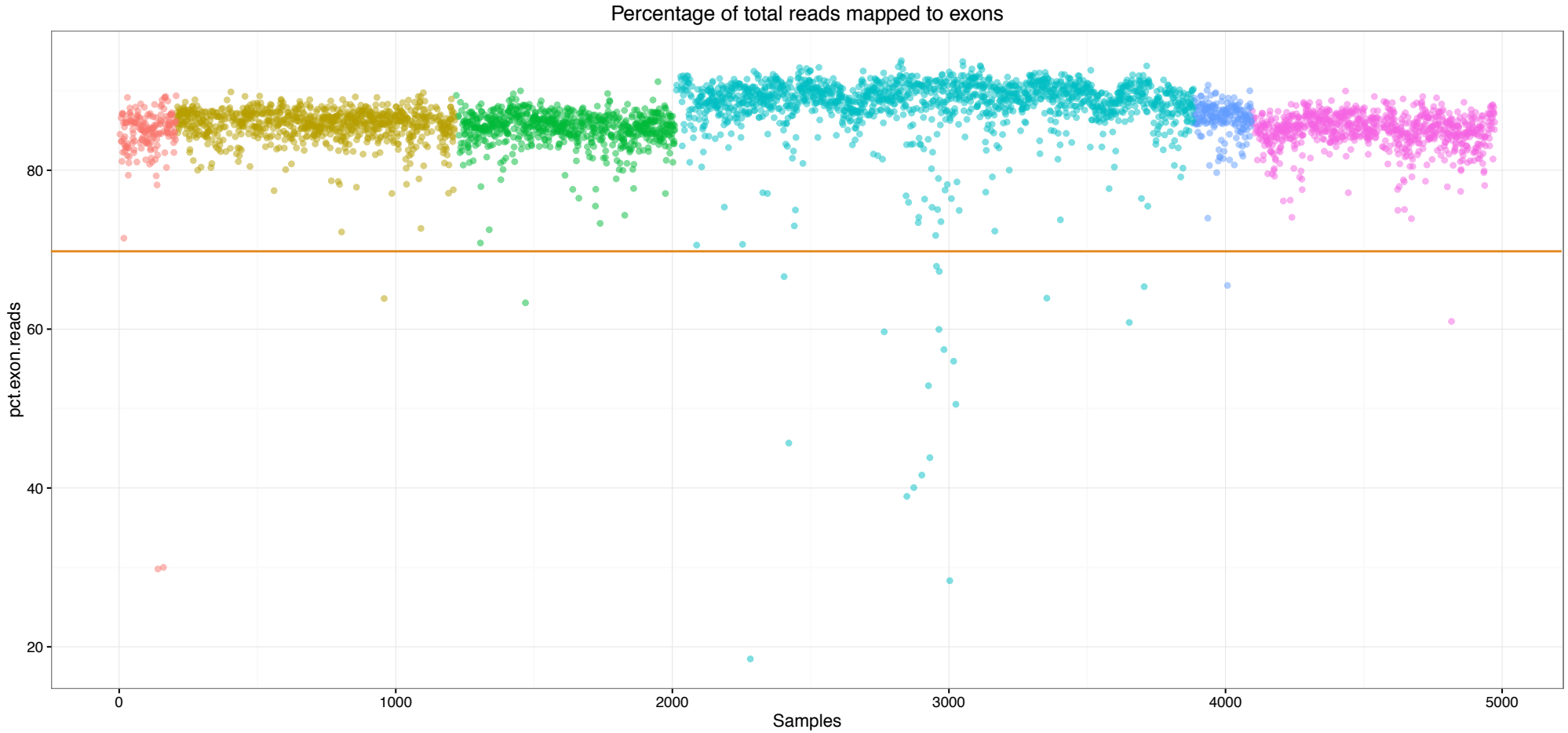
Second round sequencing
(top up reads)

QC metrics
Pipeline output

- Output from QC pipeline includes many variables
- All of these were plotted to check which thresholds are informative
- exon mapped / genome total = percentage of total reads mapping to exons

avg_deletion_length	avg_mapped_length
start_mapping_time	avg_input_length
pct_unique_mapped	end_time
num_unique_mapped	pct_unmapped_mismatch
num_splice_annotated	genome_duplicates
num_splice_noncanonical	exon_duplicates
pct_unmapped_other	exon_mapped
num_splice_total	genome_total
num_splice_atac	genome_mapped
num_splice_gcag	exon_total
num_input	genome_insert_std
rate_insertion_per_base	genome_insert_mean
pct_mapped_multiple	R2_raw_GC_mean
rate_mismatch_per_base	R2_raw_GC_std
start_job_time	R1_raw_GC_mean
pct_unmapped_short	R1_raw_GC_std
mapping_speed	R1_clean_GC_mean
avg_insertion_length	R2_clean_GC_mean
pct_mapped_many	R2_clean_GC_std
rate_deletion_per_base	R1_clean_GC_std
num_splice_gtag	MEDIAN_5PRIME_TO_3PRIME_BIAS
num_mapped_many	MEDIAN_3PRIME_BIAS
num_mapped_multiple	MEDIAN_5PRIME_BIAS

QC metrics
Reads mapping to exons



Remove samples with < 70% exon mapping

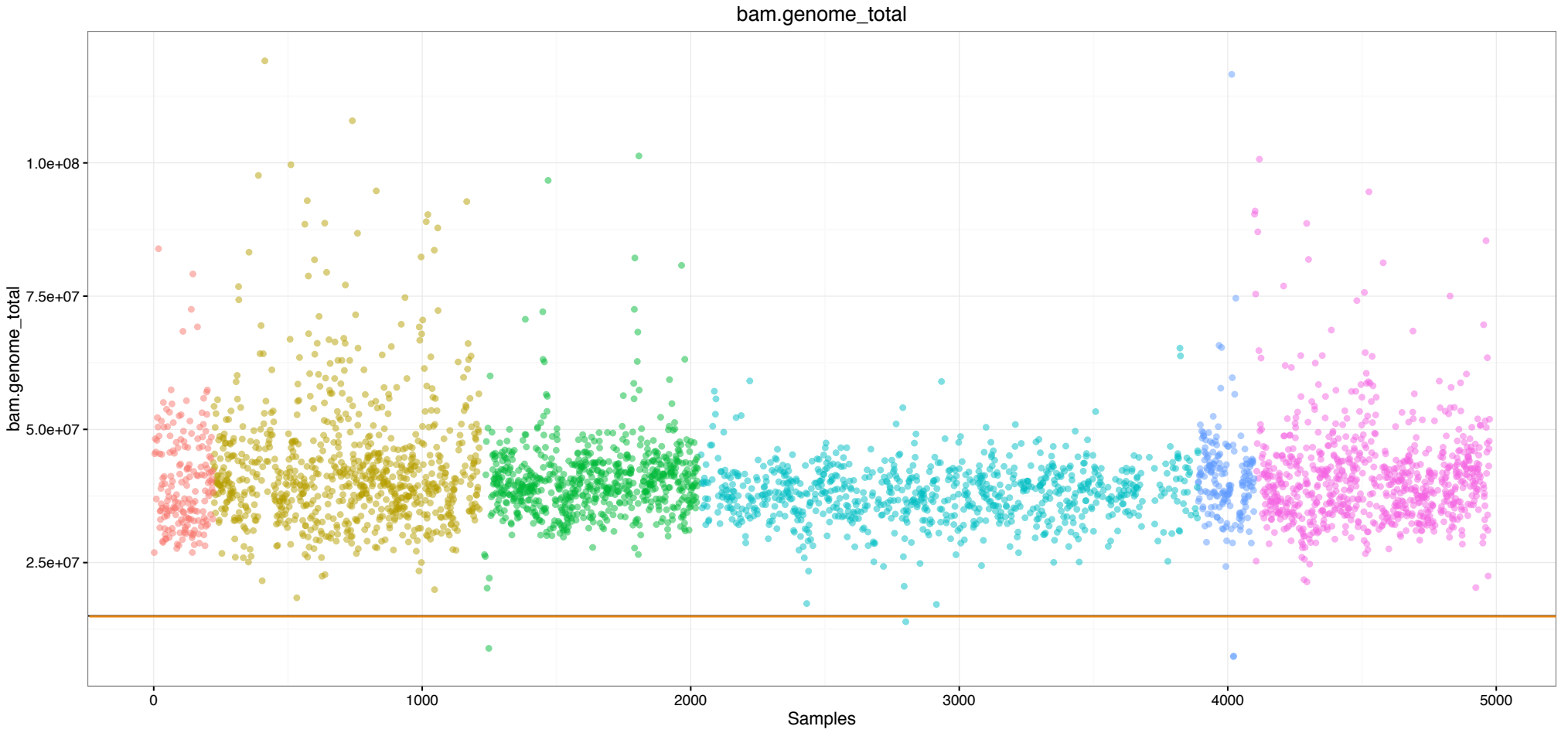
Sample selection
Family members

- **Remove** samples with first degree relationship
- Twins: take the RNAseq sample with higher % mapping to exons
- GoNL: include parents (more samples)

QC metrics
Mixupmapper

- Matching RNA to DNA again with a different method (MixupMapper)
- **Keep** samples with no match MixupMapper, but good match in earlier DNA-RNA mapping
- **Remove** samples with no match between RNA and DNA in both analyses
  - These samples cannot be recognized as a match
  - No match is an indication of RNA-seq quality: used as a QC metric

QC metrics
Total number of reads



Remove samples with total number of reads < 15M

Freeze 2 RNA-seq dataset

Step	Removed	Samples left
RNA sequencing		4543
Outliers	- 96	4447
Unsolved mix ups	- 39	4408
<70% mapped to exon	- 23	4385
No genotypes available	- 109	4276
No DNA – RNA match	- 109	4167
Related samples	- 684	3483
<15M reads	- 4	**3479**

Cohort	Freeze 2 DNA + RNA unrelated samples	Freeze 2 RNA only unrelated samples
CODAM	186	186
PAN	175	175
RS	777	777
LLS	697	697
LLDeep	825	825
NTR	819	928
Total	**3479**	3588

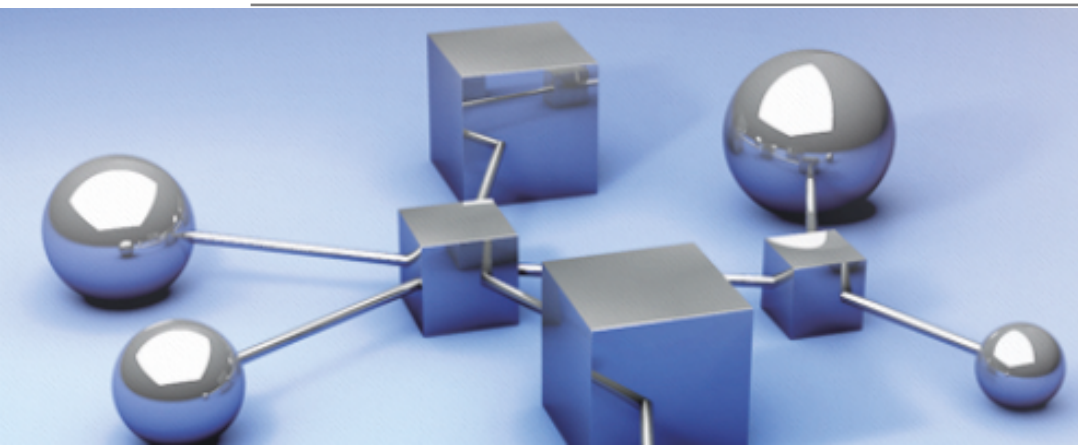
Concluding remarks

- Freeze 2 RNA-seq data is **ready for use**

- Keep in mind
  - Fairly lenient QC parameters
  - Cohort effects → adjust with covariate
  - Unrelated samples

- Quality control is a necessary **foundation for good research**

Acknowledgments



BBMRI.nl
Biobanking and
BioMolecular resources
Research Infrastructure
The Netherlands

Team effort by among others

Maarten van Iterson
Rick Jansen
Joost Verlouw
Niek de Klein
Freerk van Dijk
Leon Mei
Peter-Bram ‘t Hoen
Joyce van Meurs
Bas Heijmans
Morris Swertz
Lude Franke